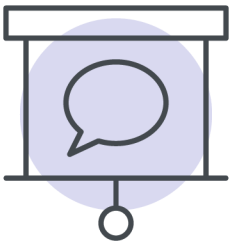
McGill

MGSC 372
FINAL EXAM
STUDY GUIDE



Lecture Notes

MGSC 372 – Notes

Introduction to Estimation

Introduction to Least Squares Estimation (LSE)

Example: Estimate the mean μ of a distribution based on a sample of n values $x_1, x_2, x_3, \dots, x_n$.

Let $\mu =$ the population mean

Then the least squares estimate of μ is the value that minimizes the sum of squares.

$$SS = (x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \cdots + (x_n - \mu)^2 = \sum_{i=1}^{n} (x_i - \mu)^2$$

Intuitively, the quantity SS measures total (squared) deviation from the mean of the population, so its minimum value will be the least squares estimate.

Differentiation: Brief Review

The Power Rule:

$$\frac{d}{dx}(x^n) = nx^{n-1}$$

The Chain Rule:

$$\frac{d}{dx}(u^n) = nu^{n-1}\frac{du}{dx}$$

A special example of the chain rule is:

$$\frac{d}{d\mu}(x - \mu)^2 = 2(x - \mu)\frac{d}{d\mu}(x - \mu) = 2(x - \mu)(-1)$$

Minimizing SS

To find the value of μ that minimizes SS we first calculate the derivative with respect to μ:

$$\frac{d}{d\mu}(SS) = 2(x_1 - \mu)(-1) + 2(x_2 - \mu)(-1) + 2(x_3 - \mu)(-1) + \cdots + 2(x_n - \mu)(-1)$$

$$\frac{d}{d\mu}(SS) = -2[(x_1 - \mu) + (x_2 - \mu) + (x_3 - \mu) + \cdots + (x_n - \mu)]$$

$$\frac{d}{d\mu}(SS) = -2[x_1 + x_2 + x_3 + \cdots x_n - n\mu]$$

Next, set the derivative to equal 0:

$$\frac{d}{d\mu}(SS) = 0 \rightarrow x_1 + x_2 + x_3 + \cdots + x_n - n\mu = 0 \rightarrow \mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Therefore, the least squares estimate of μ is the sample mean:

$$\bar{x} = \frac{\sum_{1}^{n} x_i}{n}$$

Conclusion

For a sample of n observations, the LSE of the population mean is given by the sample mean:

$$\bar{x} = \frac{\sum x}{n}$$

Notation: It is common practice in statistics to denote an estimator of a parameter θ by $\hat{\theta}$.

Thus, since $\bar{x}$ is an estimator of μ, we can write $\hat{\mu} = \frac{\sum x}{n}$.

Example

A sample of seven values is obtained from a population:

$$x_1 = 230, x_2 = 275, x_3 = 317, x_4 = 382, x_5 = 305, x_6 = 315, x_7 = 284$$

Find the LSE of the population mean μ.

Solution

$$\bar{x} = \frac{\sum x_i}{7} = \frac{230 + 275 + 317 + 283 + 305 + 315 + 291}{7} = 288$$

Maximum Likelihood Estimation

Introduction to Maximum Likelihood Estimation

Least squares estimation is only one of several methods used to estimate population parameters.

An alternative approach to least squares estimation is the method of **maximum likelihood estimation**.

Maximum likelihood estimates are often more accurate than least squares estimates.

Definition: Likelihood Function

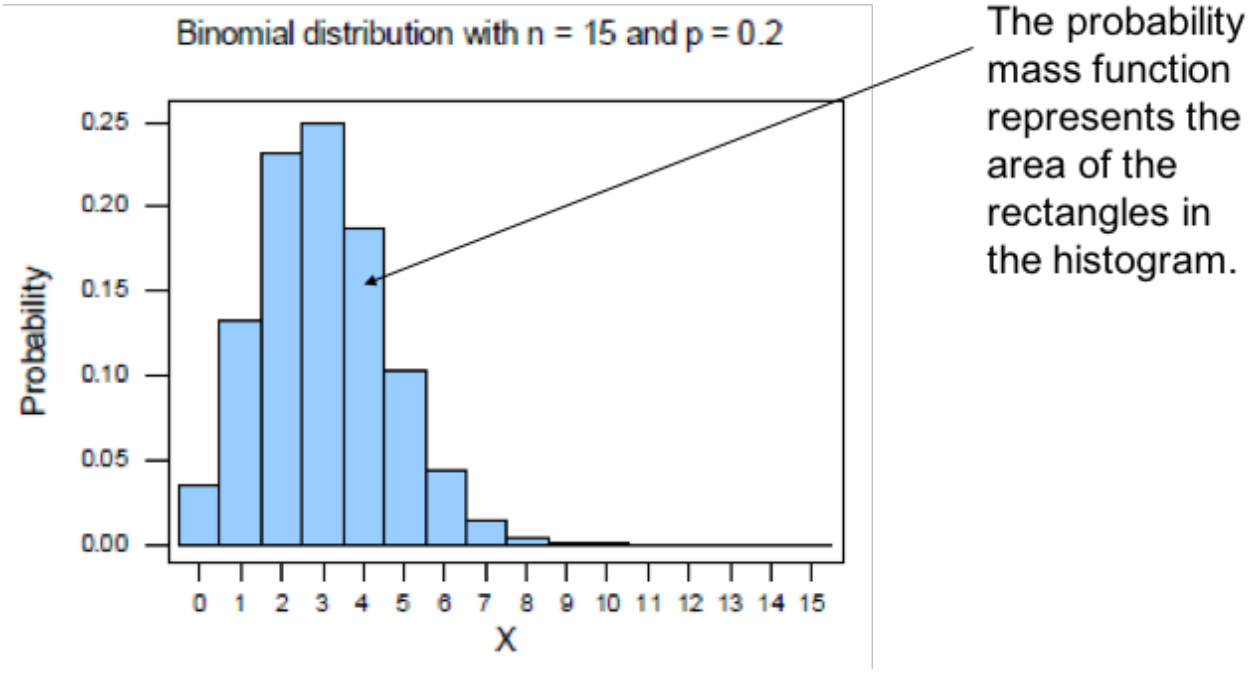
For a discrete random variable:

Let $x_1, \dots, x_n$ be a sample of n values of a discrete random variable X with parameter θ. Then the likelihood function of the sample is given by

$$L(\theta) = P(x_1)P(x_2)P(x_3) \dots P(x_n) = \prod_{i=1}^{n} P(x_i)$$

Where $P(x)$ is the population probability mass function.

Example of Discrete Probability Distribution



Likelihood Function

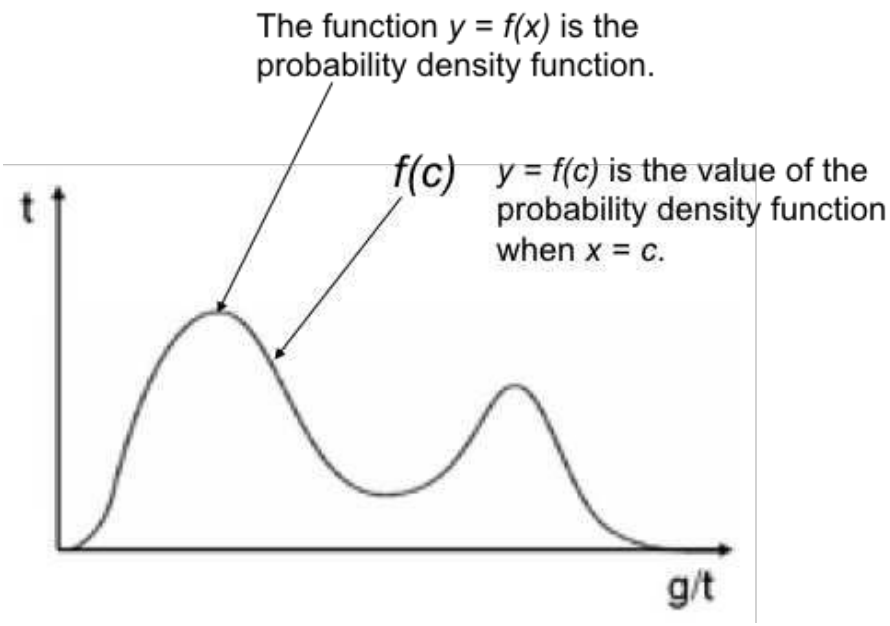
For a continuous random variable:

Let $x_1, \dots, x_n$ be a sample of n values of a continuous random variable X with parameter θ. Then the likelihood function of the sample is given by

$$L(\theta) = f(x_1)f(x_2)f(x_3) \dots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

Where $f(x)$ is the population probability density function.

Example of a Continuous Probability Distribution



Definition: Maximum Likelihood Estimation

Maximum likelihood estimation: The value of μ that maximizes the likelihood function.

Example of MLE

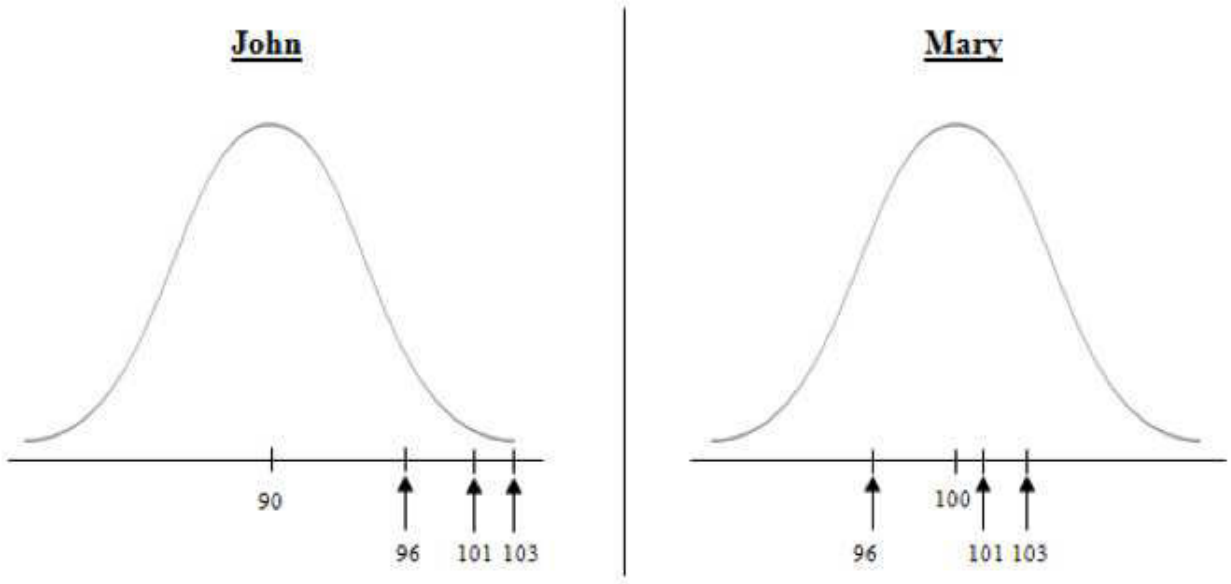
We illustrate the method of maximum likelihood estimation by demonstrating how to estimate the mean of a normal population with known standard deviation of $\sigma = 5$.

Assume that two different experimenters have a preconceived idea of the population mean. John thinks the mean is 90 and Mary thinks it is 100.

To resolve their dispute they select a random sample of three values:

$$x_1 = 96, x_2 = 103, x_3 = 101$$

Graphs of normal curves with means at 90 and 100, respectively

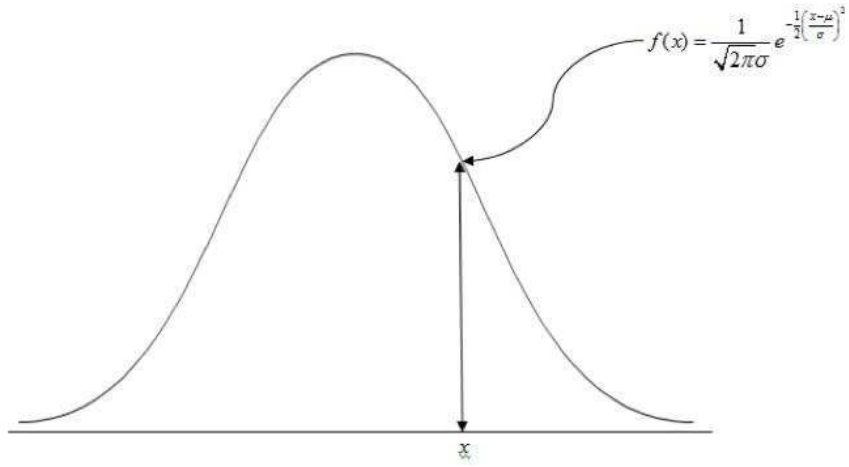


Calculating the value of the likelihood function

We recall that for a normal distribution, the probability density function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

This represents the height of the normal distribution curve above the horizontal axis.



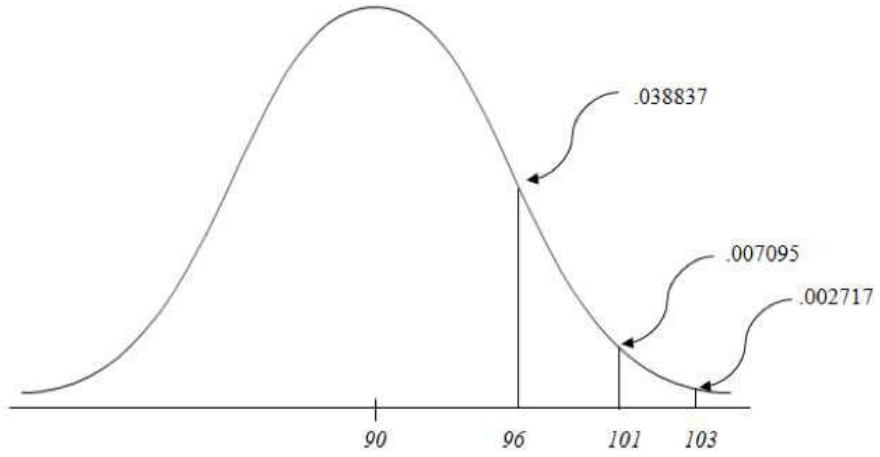
Calculating the likelihood function for John ($\mu = 90$)

$$f(96) = \frac{1}{\sqrt{2\pi(5)}} e^{\frac{-1}{2}\left(\frac{96-90}{5}\right)^2} = 0.038837$$

$$f(101) = \frac{1}{\sqrt{2\pi(5)}} e^{\frac{-1}{2}\left(\frac{101-90}{5}\right)^2} = 0.007095$$

$$f(103) = \frac{1}{\sqrt{2\pi(5)}} e^{\frac{-1}{2}\left(\frac{103-90}{5}\right)^2} = 0.002717$$

The likelihood function is the product of these three heights.



$$L(90) = f(96) * f(101) * f(103) = 7.48549 * 10^{-7}$$

Calculating the likelihood function for Mary ($\mu = 100$)

$$L(100) = f(96) * f(101) * (103) = 3.01986 * 10^{-4}$$

Conclusion

Mary wins.

Since $L(100) = 3.01986 * 10^{-4}$ is greater than $L(90) = 7.48549 * 10^{-7}$ it follows that Mary's guess of 100 has a higher likelihood of being close to the population mean than John's guess of 90.

NOTE: The least squares estimate of the population mean is

$$\bar{x} = \frac{96 + 101 + 103}{3} = \frac{300}{3} = 100$$

So Mary's guess is right on the mark!

MLE may be generated by the following methods:

Tabular/graphical: In this numerical approach, a table of likelihood values is generated (e.g. using Excel) and the maximum value is identified.

Use a nonlinear numerical solver to find the maximum value of the likelihood function (e.g. Excel's Solver).

Analytical/mathematical: This method generally involves differentiating the likelihood function (or its logarithm) with respect to the parameter(s) to be estimated and setting the derivative equal to zero.

The Mathematical Approach

MLE of μ (normal population, σ known) - Mathematical Derivation

Assume that x is normal and σ is known (i.e. $x = N(\mu, \sigma^2)$ where μ is unknown and σ is a known constant).

Consider a sample of three observations x_1, x_2, x_3 then

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

And the likelihood function is as shown:

The Likelihood Function

$$L(\mu) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} * \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2} * \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2}\left(\frac{x_3-\mu}{\sigma}\right)^2}$$

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^3 e^{\frac{-1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 * \frac{-1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2 * \frac{-1}{2}\left(\frac{x_3-\mu}{\sigma}\right)^2}$$

$$\boldsymbol{L(\mu) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^3 e^{\frac{-1}{2\sigma^2}[(x_1-\mu)^2+(x_2-\mu)^2+(x_3-\mu)^2]}}$$

Log Likelihood

We want to find the value of μ that maximizes the likelihood function.

Since the logarithm of the likelihood function attains its maximum value for the same value of μ as the likelihood function itself, we will take the logarithm of the above expression, and use the properties of logarithms shown.

Properties of Logs

$$\ln(AB) = \ln(A) + \ln(B)$$

$$\ln\left(\frac{A}{B}\right) = \ln(A) - \ln(B)$$

$$\ln(A)^c = c\ln(A)$$

$$\ln(e) = 1$$

$$\ln(e)^c = c$$

Simplifying and differentiating the log-likelihood function

$$\ln\big(L(\mu)\big) = \ln\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^3 - \frac{1}{2\sigma^2}[(x_1-\mu)^2 + (x_2-\mu)^2 + (x_3-\mu)^2]$$

Differentiating, we get:

$$\frac{d}{d\mu}[\ln(L(\mu))] = -\frac{1}{2\sigma^2}[2(x_1-\mu)(-1) + 2(x_2-\mu)(-1) + 2(x_3-\mu)(-1)]$$

$$\frac{d}{d\mu}[\ln(L(\mu))] = \frac{1}{\sigma^2}[x_1 + x_2 + x_3 - 3\mu]$$

Solving for the MLE

Setting $\frac{d}{d\mu}[\ln(L(\mu))] = 0$ we get $\frac{1}{\sigma^2}[x_1 + x_2 + x_3 - 3\mu] = 0$

So that

$$\hat{\mu} = \frac{x_1 + x_2 + x_3}{3}$$

It follows that the maximum likelihood estimate of μ is the sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MLE and LSE

We observe that the maximum likelihood estimate of the population mean for a normal distribution with constant variance is the same as the least squares estimate.

In both cases, the estimate is the arithmetic mean of the sample data.

Example

x is normally distributed with $\sigma = 10$. A sample of five values is obtained:

$$x_1 = 230, x_2 = 275, x_3 = 317, x_4 = 285, x_5 = 303$$

Find the maximum likelihood estimate of the population mean μ.

Solution

The MLE of μ is:

$$\frac{230+275+317+285+303}{5}=\frac{1410}{5}=282$$

A discrete example MLE for binomial distribution

Recall that $P(X=x_i)=\binom{n}{x_i}p^{x_i}(1-p)^{n-x_i}$

Likelihood function for three values of x: $\{x_1,x_2,x_3\}$

$$L(p)=\binom{n}{x_1}p^{x_1}(1-p)^{n-x_1}\times\binom{n}{x_2}p^{x_2}(1-p)^{n-x_2}\times\binom{n}{x_3}p^{x_3}(1-p)^{(n-x_3)}$$

$$L(p)=\binom{n}{x_1}\binom{n}{x_2}\binom{n}{x_3}p^{(x_1+x_2+x_3)}\times(1-p)^{(n-x_1)+(n-x_2)+(n-x_3)}$$

$$L(p)=\binom{n}{x_1}\binom{n}{x_2}\binom{n}{x_3}p^{x_1+x_2+x_3}\times(1-p)^{(3n-[x_1+x_2+x_3])}$$

Taking Logarithms

$$\ln L(p)=\ln\binom{n}{x_1}\binom{n}{x_2}\binom{n}{x_3}+\ln p^{x_1+x_2+x_3}+\ln(1-p)^{(3n-[x_1+x_2+x_3])}$$

$$\ln L(p)=\ln\binom{n}{x_1}\binom{n}{x_2}\binom{n}{x_3}+(x_1+x_2+x_3)\ln p+(3n-[x_1+x_2+x_3])\ln(1-p)$$

Differentiating,

$$\frac{d}{dp}\ln L(p)=0+\frac{x_1+x_2+x_3}{p}+\frac{3n-[x_1+x_2+x_3]}{1-p}$$

$$\therefore\frac{d}{dp}\ln L(p)=0\rightarrow\frac{x_1+x_2+x_3}{p}=\frac{3n-[x_1+x_2+x_3]}{1-p}$$

Solving for the MLE of p

So that:

$$(1-p)(x_1+x_2+x_3)=p(3n-[x_1+x_2+x_3])$$

$$(x_1+x_2+x_3)-p(x_1+x_2+x_3)=3pn-p(x_1+x_2+x_3)$$

$$x_1+x_2+x_3=3pn$$

$$\therefore p = \frac{x_1 + x_2 + x_3}{3n} = \frac{1}{n} \times \frac{x_1 + x_2 + x_3}{3} = \frac{1}{n} \times \bar{x}$$

The General Result

Generalizing to the case of k observations we get:

$$p = \frac{x_1 + x_2 + x_3 + \cdots + x_k}{kn} = \frac{\sum_{i=1}^{k} x_i}{kn} = \frac{1}{n} \times \bar{x}$$

Binomial Example #1

A coin is tossed 10 times and the number of heads is recorded. This process is repeated 8 times. The number of heads obtained in each of the 8 repetitions of the experiment is as follows:

$$x_1 = 3, x_2 = 5, x_3 = 8, x_4 = 6, x_5 = 4, x_6 = 7, x_7 = 5, x_8 = 4$$

Does this appear to be a fair coin?

Solution

$$MLE = \frac{\sum x_i}{nk} = \frac{3 + 5 + 8 + 6 + 4 + 7 + 5 + 4}{10 * 8} = \frac{42}{80} = 0.525$$

The probability of heads is close to 50% so the coin appears to be fair.

Binomial Example #2

An assembly line produces a fixed, but unknown, percentage of defective items. Five random samples of 100 items are selected and the number (%) of defective items items in each sample is recorded as follows:

Sample #	1	2	3	4	5
% Defectives	4	7	5	3	4

Solution

The MLE for the percentage of defective items produced by the assembly line is given by:

$$MLE = \frac{\sum x_i}{nk} = \frac{4 + 7 + 5 + 3 + 4}{100 * 5} = \frac{23}{500} = 0.046$$

We conclude that $\hat{p} = 0.046$ so approximately 4.6% of items produced.

Some comments on the method of maximum likelihood estimation

Note 1: For estimating a population mean for normal distributions, the least squares estimate (LSE) and the maximum likelihood estimate (MLE) are identical. Both methods result in the sample mean as the estimator. In general, if the error terms in a model are normally and independently distributed with constant variance, then the MLE and LSE are identical.

Note 2: Least squares estimation is purely a descriptive procedure – it is independent of the probability distribution of the variable. Maximum likelihood estimation explicitly requires knowledge of the probability distribution.

Note 3: MLE is a preferred method of parameter estimation in statistics and is an indispensable tool for many statistical modeling techniques, in particular in non-linear modeling with non-normal data. An example is logistic regression.

Note 4: MLE has many desirable properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator); consistency (estimate improves with increasing sample size); efficiency (lowest possible variance of parameter estimates). In contrast, LSE is not a sufficient estimator since it ignores the probability distribution. Note, however, that LSE estimators are unbiased whereas MLE estimators are biased.

Note 5: In nonlinear, heteroskedastic (unequal variance) models, such as ARCH and GARCH models for volatility, LSE is an inappropriate method of estimation and MLE is required.

The Lognormal Distribution and Application

The Lognormal Distribution

A continuous random variable x follows a **lognormal distribution** if its natural logarithm, $\ln(x)$, follows a normal distribution.

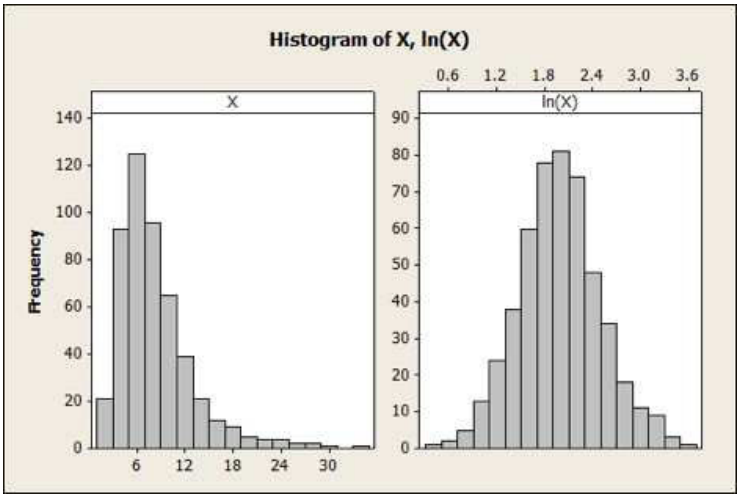
In other words, if $\ln(x)$ is normal, then x is lognormal.

Properties of the lognormal distribution:

- Skewed to the right
- Strictly positive (i.e. bounded below by 0)

The lognormal distribution may be used to model data on asset prices (note that prices are bound below by 0).

Histograms of lognormal variable x and normal distribution $\ln(x)$



The Lognormal Distribution

The lognormal distribution is described by two parameters, its mean and variance, as in the case of a normal distribution.

The **mean** of a lognormal distribution is given by:

$$E(x) = e^{(\mu+\frac{1}{2}\sigma^2)}$$

The **variance** of a lognormal distribution is given by:

$$Var(x) = e^{(2\mu+\sigma^2)}(e^{\sigma^2} - 1)$$

Where μ and σ^2 are the mean and variance of the normal distribution of the $\ln(x)$ variable and $e \cong 2.718$ is the natural base for logarithms.

The **median** of a lognormal distribution x is given by e^{μ}

Recall that the exponential and logarithmic functions are inverse functions so that

$$\ln(x) = y \rightarrow x = e^{y}$$

This relationship is used to switch between the lognormal variable x and the normal variable $y = \ln(x)$.

Lognormal Distribution Example

The material failure mechanism for a specific metal alloy has determined that the tensile strength x of the material, measured as the force per unit of cross section area N/m^2, has a lognormal distribution with parameters $\mu = 5$ and $\sigma = 0.1$.

(a) Computer $E(x)$ and $var(x)$
(b) Compute $P(x > 120)$
(c) Compute $P(110 \leq x \leq 130)$

(d) What is the value of median tensile strength?
(e) If the smallest 5% of strength values were unacceptable, what would the minimum acceptable strength be?

Solutions

(a)

$$E(x) = e^{\mu+\frac{\sigma^2}{2}} = e^{5+0.005} = e^{5.005} = 149.16$$

$$Var(x) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) = e^{10.01}(e^{0.01} - 1) = 223.59$$

(b)

$$P(x > 120) = P(\ln x > \ln 120) = P\left(z > \frac{\ln 120 - 5}{0.1}\right)$$

$$= P\left(z > \frac{4.7875 - 5}{0.1}\right) = P(z > -2.13)$$

$$= P(z < 2.13) = 0.9834$$

(c)

$$P(110 < x < 130) = P(\ln 110 < \ln x < \ln 130)$$

$$= P(4.700 < \ln x < 4.868) = P(-3 < z < -1.32)$$

$$= P(1.32 < z < 3) = P(z < 3) - P(z < 1.32)$$

$$= 0.9987 - 0.9066 = 0.0921$$

(d)

Median $= e^{\mu} = e^{5} = 148.41$

(e)

The value of x, call it L, for which $P(x < L) = 0.05$ is determined as follows:

$$P\left(z < \frac{\ln L - 5}{0.1}\right) = 0.05$$

But we know that $P(z < -1.645) = 0.05$ from the table,

$$\therefore \frac{\ln L - 5}{0.1} = -1.645$$

So that $\ln L = 5 - 1.645(0.1) = 4.8355$

$$\ln L = 4.8355 \rightarrow L = e^{4.8355} = 125.9$$

Example: Relative Asset Prices and the Lognormal Distribution

Consider the **relative price** of an asset between periods 0 and 1, defined as S_1/S_0, which is equal to $1 + R_{0,1}$.

For example, if $S_0 = \$30$ and $S_1 = \$34.5$, then the relative price is $\$34.5/\$30 = 1.15$ or $1 + 0.15$, meaning that the **holding period return** $R_{0,1}$ is 15%.

The **continuously compounded return** $r_{T,T+1}$ associated with a holding period return of $R_{T,T+1}$ is given by the natural log of the relative price.

$$r_{T,T+1} = \ln\left(\frac{S_{T+1}}{S_T}\right) = \ln(1 + R_{T,T+1})$$

For the above example, the continuously compounded return is given by $r_{0,1} = \ln\left(\frac{\$34.5}{\$30}\right) = \ln(1.15) = 0.1397 = 13.97$, which is lower than the holding period return of 15%.

To generalize, note that between periods 0 and T, $r_{0,T} = \ln\left(\frac{S_T}{S_0}\right)$, so we can write:

$$S_T = S_0 e^{r_{0,T}}$$

Recall that $\ln(XY) = \ln(X) + \ln(Y)$ so that:

$$\ln\left(\frac{S_T}{S_0}\right) = \ln\left(\frac{S_T}{S_{T-1}}\right) + \ln\left(\frac{S_{T-1}}{S_{T-2}}\right) + \cdots + \ln\left(\frac{S_1}{S_0}\right)$$

Or equivalently,

$$r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + \cdots + r_{0,1}$$

We see that the mean continuously compounded return between periods 0 and T is the sum of the continuously compounded returns of the interim one-period returns.

If the one-period continuously compounded returns are normally distributed, their sum will also be normal.

Even if they are not, by the CLT, their sum will be approximately normal, provided the number of terms is large enough.

Therefore, we can model the relative stock price as a lognormal variable whose natural log, given by the continuously compounded return is distributed normally.

Application: Option pricing models like Black-Scholes include the volatility of continuously compounded returns on the underlying asset obtained through historical data.

Stock Option Prices

The price of a stock option is a function of the underlying stock's price and time.

The properties of the lognormal distribution of future stock prices are obtained by assuming that prices follow a random process called **geometric Brownian motion**.

From this assumption we can derive the following equation:

$$\ln S_T \cong N\left[\ln S_0 + \left(\mu - \frac{\sigma^2}{2}\right)T, \left(\sigma\sqrt{T}\right)^2\right]$$

This equation shows that $\ln S_T$ is approximately normally distributed with:

Mean $= \ln S_0 + \left(\mu - \frac{\sigma^2}{2}\right)T$

Standard deviation $= \sigma\sqrt{T}$

Stock Prices are Lognormal

We conclude that a stock's price at time T, denoted by S_T, given its price today, is lognormally distributed.

Note

The standard deviation of the logarithm of the stock price is $\sigma\sqrt{T}$. It is proportional to the square root of how far ahead we are looking. In other words, volatility increases with time.

Example 1

Initial stock price $S_0 = \$40$

Expected return $\mu = 16\%$ per annum

Volatility $= 20\%$ per annum

Estimate the stock price in 6 months

$$S_0 = 40$$
$$\mu = 0.16$$
$$\sigma = 0.20$$
$$T = 0.5$$

$$\ln S_T \cong N\left[\ln 40\left(0.16 - \frac{0.2^2}{2}\right) * 0.5, \left(0.2\sqrt{0.5}\right)^2\right]$$

$$\ln S_T \cong N(3.759, 0.02)$$

95% CI for $\ln S_T$

There is a 95% probability that a normal variable has a value that is within 1.96 standard deviations of its mean.

Since $\ln S_T$ is normal with mean 3.759 and standard deviation $\sqrt{0.02} \cong 0.141$ we can write, with 95% confidence:

$$3.759 - 1.96(0.141) \leq \ln S_T \leq 3.759 + 1.96(0.141)$$

$$3.483 \leq \ln S_T \leq 4.035$$

$$e^{3.483} < S_T \leq e^{4.035}$$

$$32.56 \leq S_T \leq 56.54$$

We conclude that there is a 95% probability that the stock price in 6 months time will be between $32.56 and $56.54.

Example 2

Initial stock price $S_0 = \$74$

Expected return $= 6\%$ per annum

Volatility $= 10\%$ per annum

Estimate the stock price in 18 months

$$\ln S_T \cong N\left[\ln S_0 + \left(\mu - \frac{\sigma^2}{2}\right)T, \left(\sigma\sqrt{T}\right)^2\right]$$

$$\ln S_T \cong N[\ln 74 + \left(0.06 - \frac{0.1^2}{2}\right) \times 1.5, (0.1\sqrt{1.5})^2$$

$$\ln S_T \cong N(4.387, 0.1225^2)$$

90% CI for $\ln S_T$

$$4.387 - 1.645(0.1225) \leq \ln S_T \leq 4.387 + 1.645(0.1225)$$

$$4.185 \leq \ln S_T \leq 4.589$$

$$e^{4.185} \leq S_T \leq e^{4.589}$$

$$\$65.69 \leq S_T \leq \$98.40$$

Simple Linear Regression Review

Simple Linear Regression

Simple linear regression analysis is used to analyze the nature of the relationship between two variables.

The dependent (response) variable is designated by Y and the independent (predictor) variable is designated by X. For a given independent variable, there may be many values of the dependent variable.

The decision regarding which variable to designate Y and which variable to designate X must be based upon theory, knowledge of the subject matter, and the objectives of the analysis. The relationship between the two variables is estimated and then used to make predictions for Y.

Scatter Diagram

A scatter diagram is a graph showing the shape and direction of the underlying relationship between the independent variable X and the dependent variable Y.

Observations are plotted in pairs (x, y) with one variable plotted on each axis.

Linear Relationships Between Two Variables Intercept and Slope

The relationship between the two variables is described by a straight line mode in general form:

True regression line:

$Y = \beta_0 + \beta_1 X + \epsilon$ or $E(Y) = \beta_0 + \beta_1 X$

Estimated regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

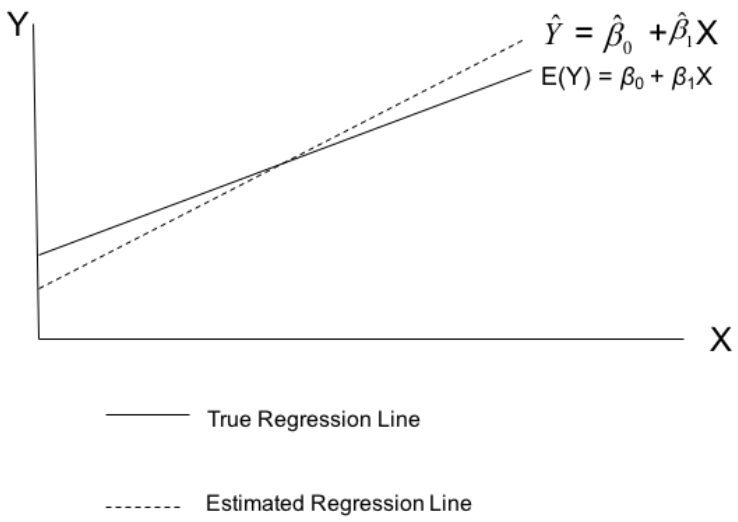
Parameter Estimation

$\hat{\beta}_0$ is a point estimation of β_0

$\hat{\beta}_1$ is a point estimation of β_1

We note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are variables while β_0 and β_1 are constants.

Simple Linear Regression Model



Definitions of Terms

$X =$ the independent (predictor) variable

$Y =$ the dependent (response) variable

$\beta_0 =$ the true Y-intercept

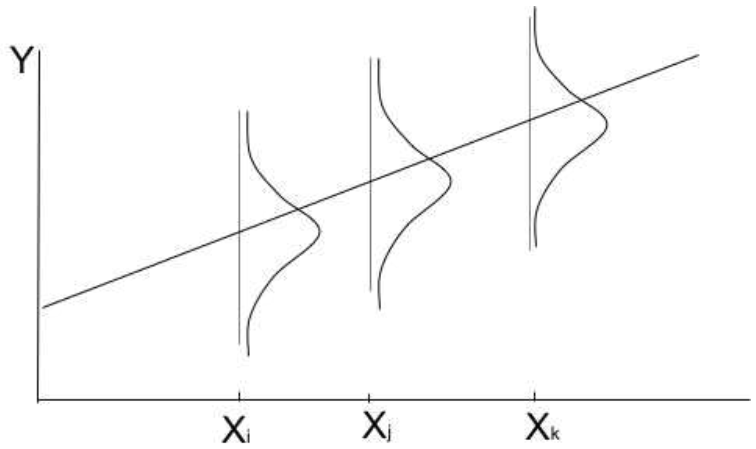
$\beta_1 =$ the true slope

$\epsilon =$ the error term

$\hat{\beta}_0 =$ the estimated Y-intercept

$\hat{\beta}_1 =$ the estimated slope

Simple Linear Regression Model



Assumptions

1. In Simple Linear Regression, we have the assumption of linearity.
2. For each value of X, the Y values are normally distributed.
3. For each value of X, the variance of the Y-values is the same (homoscedasticity).
4. Independence (independent sample of Y are chosen for different values of X i.e. error terms are not correlated).

Intercept and Slope

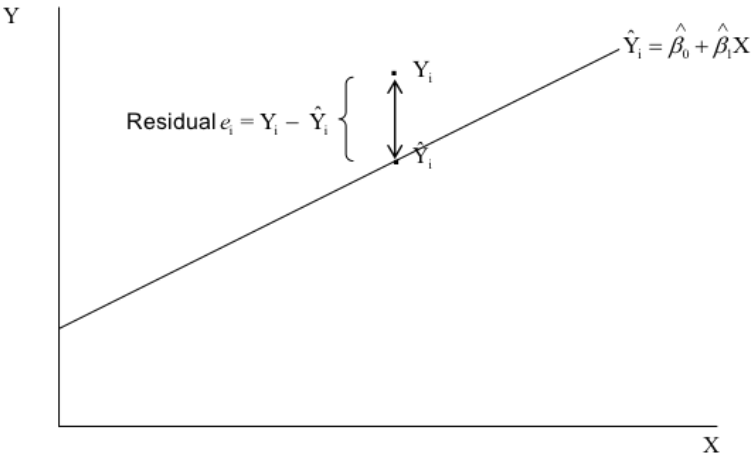
The Y-intercept β_0 is the point on the Y-axis where the true regression line crosses and is the average value of Y when $X = 0$.

The slope of the true regression line β_1 represents the average change in Y when X is increased by one unit.

β_0 and β_1 are called the **parameters** of the regression line.

Residual Error

The difference between an observed value Y and an estimated $\hat{Y}$ is called a residual.



SSE - Sum of Squares Error

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

SSE represents the sum of the squared deviations between the observed Y-values in the data set and the Y-values predicted by the estimated regression line. **This is the amount of variation in Y that is not explained by the regression line.**

Note: The **Least Squares** regression line is the line that has an intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ that will minimize SSE.

Minimizing SSE

To find the line that best fits the points in the scatter diagram, we minimize the quantity:

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X)^2$$

We note that $SSE = f(\hat{\beta}_0, \hat{\beta}_1)$. We obtain the two partial derivatives:

$\frac{\delta(SSE)}{\delta\hat{\beta}_0}$ and $\frac{\delta(SSE)}{\delta\hat{\beta}_1}$

And solve the equations:

$\frac{\delta(SSE)}{\delta\hat{\beta}_0} = 0$ and $\frac{\delta(SSE)}{\delta\hat{\beta}_1} = 0$

Least Squares Estimates of Regression Parameters

$$\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Interpretation of Regression Coefficients

The regression equation is:

$$\hat{Y} = 7.905 + 0.715X$$

$\hat{\beta}_1 = 0.715$ implies that, on average, a one unit increase in X will result in an increase of 0.715 in Y, i.e. if advertising increases by \$1000 then profit will increase, on average, by \$715.

$\hat{\beta}_0 = 7.905$ implies that the average profit in stores with no advertising budget will be \$7905.

Extrapolation

The interpretation of $\hat{\beta}_0 = 7.905$ is not reliable.

The problem is that we are making a claim about a value of X for which we have no experimental evidence.

All of our experimental data is for values of X in the range \$3000 to \$7000. Therefore we cannot make a reliable claim about the relationship between X and Y when $X = 0$.

This is called **extrapolation**.

Point Estimates

The regression equation can be used to predict a value of Y based on a given value of X by substituting the X-value into the regression line.

Example

Estimate the profit of a store which spends $3500 on advertising.

Let $X = 3.5$

Then $\hat{Y} = 7.905 + 0.715(3.5) = 10.4$ (i.e. $10400)

Caution: Do not use the regression line to predict Y with values of the independent variable significantly beyond the range of those represented in the sample. The nature of the relationship outside the range of X-values represented in the sample may not be linear and **extrapolation** may lead to false conclusions.

Partitioning Total Deviation

Total deviation $= Y_i - \bar{Y}$

Unexplained deviation $= Y_i - \hat{Y}_i$

Explained deviation $= \hat{Y}_i - \bar{Y}$

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Total deviation = Unexplained deviation + Explained deviation

Sum of Squares

We can compute sums of squares in regression analysis and construct an analysis of variance (ANOVA) table for the regression.

Partitioning Sums of Squares

It can be shown that:

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

$$SSTO = SSE + SSR$$

Note: In simple linear regression $df_{reg} = 1$. Therefore:

$$MSR = SSR$$

The ratio MSR/MSE follows an F-distribution

Note that:

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2/df_1}{\sum(Y_i - \hat{Y}_i)^2/df_2}$$

This shows that MSR and MSE are both variances and it is well known that the ratio of two variances follows the F-distribution.

Testing the Significance of the Regression Line

A linear relationship exists between the dependent variable and the independent variable only if the slope of the regression line is significantly different from zero.

If $\beta_1 \neq 0$, then a linear relationship exists between X and Y.

Note that the magnitude of $\hat{\beta}_1$ by itself does not tell us the strength of the linear relationship between X and Y. In fact, by changing the units of measurement we can always make $\hat{\beta}_1$ larger or smaller. That is why we used standardized z or t values in tests of hypothesis.

Testing the Slope: F-distribution

$H_0: \beta_1 = 0$ (no significant linear relationship between X and Y)
$H_1: \beta_1 \neq 0$ (significant linear relationship between X and Y)

$$TS: F^* = \frac{MSR}{MSE}$$

$AL: F_\alpha$ where $df_{num} = 1$, $df_{den} = n - 2$

DR: Do not reject H_0 if $F^* \leq F\_\alpha$, reject H_0 *if* $F^* > F_\alpha$

Example: Test of Hypothesis for β_1 (F-test)

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$TS: F^* = 14.79$$

$$AL: F_{0.05} = 5.32 \ (df_{num} = 1, df_{den} = 8)$$

DR: Do not reject H_0 if $F^* \leq 5.32$, reject H_0 if $F^* > 5.32$.

Conclusion: Reject $H_0 \rightarrow \beta_1 \neq 0$

Testing the Slope: t-distribution

$H_0: \beta_1 = 0$ (no significant linear relationship between X and Y)

$H_1: \beta_1 \neq 0$ (significant linear relationship between X and Y)

$$TS: t^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$$

$AL: \pm t_{\frac{\alpha}{2}}$ where $df = n - 2$

DR: Do not reject H_0 if $-t_{\frac{\alpha}{2}} \leq TS \leq +t_{\frac{\alpha}{2}}$, reject otherwise

Example: Test of Hypothesis for β_1 (t-test)

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$TS: \hat{\beta}_1 = 0.715 \rightarrow t^* = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} = \frac{0.715 - 0}{0.1859} = 3.846$$

$$AL: A_1 = -t_{0.025} = -2.306, A_2 = t_{0.025} = 2.306$$

DR: Do not reject H_0 if $-2.306 \leq t^* \leq 2.306$, reject H_0 otherwise.

Conclusion: Reject $H_0 \rightarrow \beta_1 \neq 0$

Relationship between t and F in SLR

$t^* = 3.846$ and $F^* = 14.79$

Since $3.846^2 = 14.79 \rightarrow t^{*2} = F^*$

Also $t_{0.025} = 2.306$ and $F_{0.05} = 5.32$

Since $2.3062^2 = 5.32 \rightarrow t^2_{0.025} = F_{0.05}$

NOTE: For t, d$f = 8$; for F, $df_{num} = 1$, $df_{den} = 8$

One-Sided Tests

Test using the one-sided alternative:
$H_1: \beta_1 > 0$ for a **direct** relationship between X and Y.

Test using the one-sided alternative:
$H_1: \beta_1 < 0$ for an **inverse** relationship between X and Y.

Confidence Interval Estimate for β_1

$$\hat{\beta}_1 - t_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_1} \leq \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_1}$$

If 0 is contained in the above confidence interval then conclude DNR $H_0: \beta_1 = 0$

Example of CIE for β_1

Construct a 95% confidence interval estimate for β_1.

$$0.715 - 2.306(0.1859) \leq \beta_1 \leq 0.715 + 2.306(0.1859)$$

$$0.2863 \leq \beta_1 \leq 1.1437$$

Pearson Coefficient of Correlation

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{(x-\bar{x})^2(y-\bar{y})^2}}$$

The Pearson coefficient measures the strength of the **linear** relationship between two variables x and y.

Testing the Hypothesis for Pearson Coefficient of Correlation ρ

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$
$$TS: r = 0.806$$

p-value $= 0.005$

Conclusion: Reject H_0

Since the p-value < 0.05 we reject the null hypothesis and conclude that there is a significant linear correlation between profit and advertising.

Properties of the Coefficient of Correlation r

1. The value of r will always be between -1 and $+1$, inclusive.
2. $r = 1$ indicates a perfect linear association between X and Y.
3. $r = 0$ indicates no linear relationship between X and Y.
4. $r > 0$ implies that X and Y move in the same direction (direct relationship).
5. $r < 0$ implies that X and Y move in opposite directions (inverse relationship).
6. The closer r is to 1 the stronger the linear relationship and the closer the points will be to the estimated regression line.
7. In Excel, r is called "multiple r".

Test of hypothesis for population correlation

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

$$TS: t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$CV: \pm t_{\frac{\alpha}{2}}$$

Example

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

$$TS: t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8056\sqrt{8}}{\sqrt{1-0.8056^2}} = 3.846$$

$$CV: \pm t_{\frac{\alpha}{2}} = \pm 2.306$$

DR: Do not reject H_0 if $-2.306 \leq t \leq 2.306$

Conclusion: Reject H_0

Equivalence of tests for β_1 and ρ

It can be shown that:

$$r = \beta_1 \sqrt{\frac{SS_{xx}}{SS_{yy}}} = \hat{\beta}_1 \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$$

So that the test

$H_0: \beta_1 = 0$ equivalent to the test $H_0: \rho = 0$

Note that both tests result in the t-value 3.846.

Determination

The coefficient of determination r^2 measures the percent of variation in the dependent variable that is explained by the independent variable.

Properties of the Coefficient of Determination

1. r^2 will always be between 0 and 1.
2. The closer r^2 is to 1, the better the regression model.
3. $r^2 = \frac{SSR}{SSTO}$

Formulas for Coefficient of Determination r^2

$$r^2 = \frac{SSR}{SSTO}$$

Alternatively, we can write:

$$r^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}}$$

Example

$$r^2 = \frac{SSR}{SSTO} = \frac{10.2245}{15.7560} = 0.6489$$

Interpretation

64.89% of the variation in Y is explained by X.

Using the model for estimation and prediction of the response variable

In the next slides, we will show how to use the regression analysis to construct:

1. Confidence intervals for the mean value of Y, denoted $E(Y)$, when $X = Xp$
2. Prediction intervals for an individual value of Y, denoted Y_{new}, when $X = Xp$

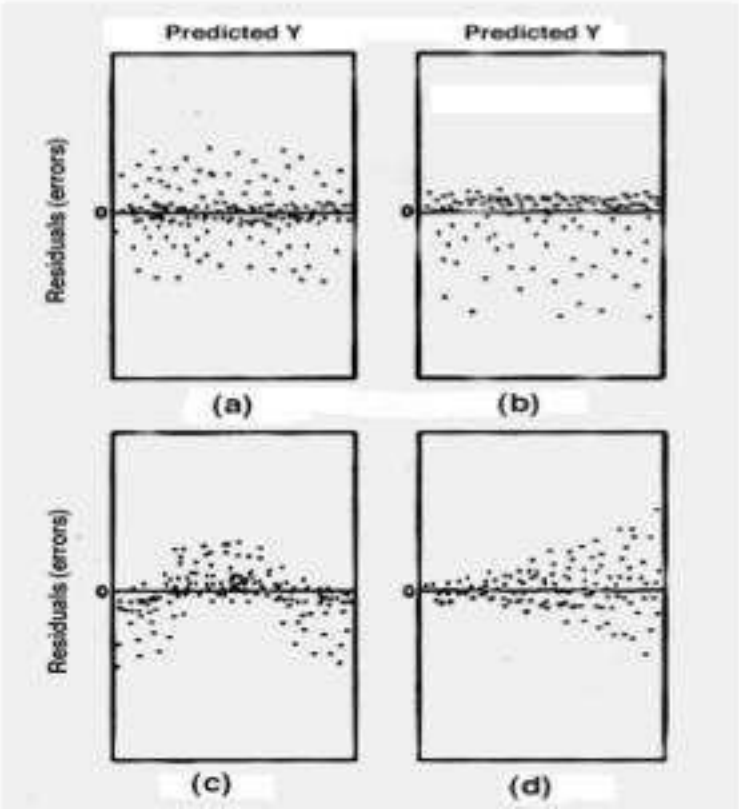
Residual Plots

A simple technique to discover whether or not any of the regression assumptions are violated is to examine a plot of all the residuals against the independent variable X. See the next slide.

If the scatter diagram of the residuals form a **horizontal band** around zero and roughly obeys the empirical rule then there is no evidence that any of the assumptions have been violated. See Panel A on the next slide.

Panel B shows a lack of normality, Panel C shows a non-linear relationship, and Panel D exhibits a violation of the constant variance assumption (homoscedasticity).

The plots demonstrate, (a) assumptions met, (b) failure of normality, (c) nonlinearity, and (d) heteroscedasticity.



Warning

While graphical inspection of residual plots is a useful guideline for detecting potential violations of regression assumptions, it is subjective and requires the statistician to use personal judgment in assessing the seriousness of the violation and whether they are significant enough to invalidate the underlying regression models.

Later in the course we will introduce tests of hypotheses to formally test the assumptions.

Causal Relationships

Although we will interpret our model as a linear **statistical** relationship between X and Y, regression analysis does not prove a **causal** relationship.

Example: Causality

Let $Y =$ an industry's productivity and $X =$ the percentage of the industry's workforce that is unionized.

A statistical relationship with a strong correlation in which increasing X results in an increase in Y might indicate:

1. Unionization causes higher productivity (causation).
2. Industries with high productivity encourage unionization (reverse causation).
3. Unionization causes higher wages that attract more production workers (intervening or lurking variable).
4. Industries with high profits encourage both unionization and productivity (common response).

Note on correlation and causality

Remember that a strong correlation does not necessarily imply causality.

A causes B: Direct causation

B causes A: Reverse causation

A causes X, X causes B: Lurking variable

X causes A and B: Common response

Analysis of relationship between Ice Cream Sales and Drowning

Causation: Ice cream sales cause drowning.

Reverse causation: Drowning causes ice cream sales.

Intervening variable: Ice cream sales causes changes in Variable Z and Variable Z causes changes in Drowning.

Common response: Seasonal variation causes changing patterns of both ice cream sales and drowning.

Multiple Regression

Multiple Regression

Multiple regression is regression analysis with more than one independent variable.

True regression model:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon_i$, where ϵ are independent $N(0, \sigma^2)$

Or $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_K X_K$

Note: The assumptions on ϵ imply that the Y_i are independent N variables with constant variance i.e. Y_i: $N(E(Y), \sigma^2)$.

Estimated regression equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \cdots + \hat{\beta}_K X_K$$

Definition of Terms

$X_i =$ the independent variables, $i = 1,2, \dots, K$.

$\beta_i =$ the true regression parameters, $i = 1,2, \dots, K$.

$\hat{\beta}_i =$ the estimated regression coefficients, $i = 1,2, \dots, K$.

$\epsilon_i =$ the error terms, $i = 1,2, \dots, K$.

$\beta_i =$ the change in Y (the dependent variable) per unit increase in X_i with **all other independent variables held fixed**.

Notation

Our text uses the notation $\hat{\beta}_i$ to represent the estimated regression coefficients.

Many textbooks use b_i instead. I will primarily use the $\hat{\beta}_i$ notation but you should be aware when reading from other sources that: $b_i = \hat{\beta}_i$

F-tests

Is the overall MR Model significant?

Commonly used to test for the significance of all independent variables combined i.e. the overall efficacy of the model. Use the ANOVA table to test.

$$\mathrm{H}_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

H_1: Not all $\beta = 0$ (regression equation is significant)

$$TS: F^* = \frac{MSR}{MSE}$$

$CV: F_\alpha$ with $df_{num} = k$ and $df_{den} = n - (k + 1)$

DR: Do not reject H_0 if $F^* \leq F_\alpha$, reject H_0 if $F^* > F_\alpha$

Remember: $k =$ number of independent variables in the model.

t-tests

Are the individual independent variables significant in the model?

$$H_0: \beta_j = 0$$

$H_1: \beta_j \neq 0$ (**significant** relationship between X_j and Y)

$$TS: t_j^* = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}(\hat{\beta}_j)}$$

$CV: \pm t_{\frac{\alpha}{2}}$ with $df = n - (k+1)$

DR: Do not reject H_0 if $-t_{\frac{\alpha}{2}} \leq t_j^* \leq +t_{\frac{\alpha}{2}}$, reject H_0 otherwise.

t-tests using the b_j notation

$$H_0: \beta_j = 0$$

$H_0: \beta_j \neq 0$ (**significant** relationship between X_j and Y)

$$TS: t_j^* = \frac{b_j - \beta_j}{s(b_j)}$$

$CV: \pm t_{\frac{\alpha}{2}}$ with $df = n - (k+1)$

DR: Do not reject H_0 if $-t_{\frac{\alpha}{2}} \leq t_j^* \leq +t_{\frac{\alpha}{2}}$, reject H_0 otherwise.

Example: t-tests for Model 2

For variable $X_1: t = \frac{0.7150}{0.1047} = 6.83$ (p-value $= 0.000$)

For variable $X_2: t = \frac{0.3261}{0.0763} = 4.27$ (p-value $= 0.004$)

Alternatively, if you are given a fixed α say 0.05, the critical value of t, assuming a two-tail test with $df = 7$:

$$\pm t_{0.025} = \pm 2.365$$

Conclusion: Reject $H_0: \beta_1 = 0$ and reject $H_0: \beta_2 = 0$ at $\alpha = 0.05$

Therefore, variable X_1 and X_2 are both significant in the model.

Confidence Interval Estimates for β_1

$$\hat{\beta}_j - t_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)$$

Note: If 0 is contained in the above confidence interval then there is no evidence to reject $H_0: \beta_j = 0$.

Example: CI for β_1 in Model 2

$$\hat{\beta}_1 = 0.715, \hat{\sigma}(\hat{\beta}_1) = 0.1047, t_{\frac{\alpha}{2},7} = 2.365$$

CI for regression coefficients are of form:

$$\hat{\beta}_j - t_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)$$

$$0.715 - 2.365(0.1047) \leq \beta_1 \leq 0.715 + 2.365(0.1047)$$

$$0.467 \leq \beta_1 \leq 0.963$$

Comments

1. The independent variables can be ranked in order of importance (in terms of explaining variation in Y) by the absolute values of the t-statistics.

2. Test using the one-sided alternative $H_1: \beta_i > 0$ for a **direct** (positive) linear relationship and $H_1: \beta_i < 0$ for a **inverse** (negative) linear relationship.

Independence of x_1 and x_2

In this model it is important to note that the two variables x_1 and x_2 are statistically independent of each other, hence $r_{12} = 0$. In fact, the variables x_1 and x_2 are experimental variables whose values were deliberately assigned to ensure that the variables are not correlated.

Therefore changing the value of x_1 does not cause a change in the value of x_2 and vice versa.

Interpretation of coefficients

A $1000 increase in ADVERTISING results in a $715 increase in profit, independent of the values of specials.

A one unit increase in SPECIALS results in a $326 increase in profit, independent of the value of advertising.

Additivity of Coefficients of Determination

Recall that:

$$R_{y_1} = 0.8056 \rightarrow R^2_{y_1} = 0.6490 \cong 0.65$$
$$R_{y_2} = 0.5037 \rightarrow R^2_{y_2} = 0.2537 \cong 0.25$$

Also, $r_{12} = 0$, so that variables x_1 and x_2 are independent.

In Model 2, we see that $R^2_{y_{12}} = 0.9026 \cong 0.90$ so that Model 2 explains approximately 90% of variation in Y.

Note that $0.90 = 0.65 + 0.25$.

This additivity of the coefficients of determination is expected when the correlation between variables x_1 and x_2 is zero.

Anomalies in Model 3

$$\hat{Y} = 4.283 + 0.516x_1 + 0.230x_2 + 0.097x_3$$

In Model 1 and 2, the coefficient of x_1 is 0.715.

Question: Why has it changed to 0.516 in Model 3?

In Model 2, the coefficients of x_1 and x_2 are both significant (p-value < 0.05).

Question: Why are the coefficients of all three variables x_1, x_2, and x_3 not significant in 3?

Question: In Model 3, the F-test is highly significant (p-value=0.001) while all three t-tests are not significant. Is this a contradiction?

Multiple Regression

Multicollinearity

All anomalies mentioned on the previous slide are explained by the fact that the variable x_3 is highly correlated with variable x_1 ($r_{13} = 0.7577$) and variable x_2 ($r_{23} = 0.5014$).

Multicollinearity: A condition that occurs when two or more independent variables are highly correlated.

It is impossible to indicate the 84% of variability explained by x_3 without significantly overlapping x_1 and x_2. In other words, x_1 and x_2 already explain most of the 84% explained by x_3 alone.

Should we add x_3 to the model containing x_1 and x_2?

We see that x_3 causes the three anomalies discussed above while only explaining an extra 1.79% of total variability.

Intuitive conclusion: Do not add x_3 to this model.

Statistical conclusion: Arrived at via a t-test.

Note: In the MR context, t-tests are partial tests in that they are testing for the significance of a particular variable **in the presence of other independent variables**. In other words, the t-tests are measuring the **marginal contributions** of the independent variables.

Testing $H_0: \beta_3 = 0, t = 1.055, p = 0.3322$. We conclude that x_3 does not make a significant contribution to the model already containing x_1 and x_2 ($p > 0.05$).

Interpretation of Coefficients in Model 3

Because x_3 is correlated with x_1 and x_2, the interpretation of the coefficients of any one of the three variables is dependent on the values of the other variables in the model.

In Model 3:

$b_1 = 0.516$ implies that a one unit ($1000) increase in x_1 (advertising) will result in an average increase of $516 in profit **in stores that have the same number of specials and the same size**.

$b_2 = 0.230$ implies that a one unit increase in x_2 (specials) will result in an average increase of $230 in profit **in stores that have the same advertising expenditure and the same size**.

$b_3 = 0.097$ implies that a one unit (1000 sq. ft.) increase in x_3 (size) will result in an average increase in profit of $97 **in stores that have the same advertising expenditure and the same number of specials**.

Model 4: Profit on Advert, Special, Size, Place

Model 4 is not a good model.

In Model 4, none of the four variables has a significant t-value. Also the increase in R^2 over Model 2 or Model 3 is minimal.

Although the F-test is significant, this model has little to commend it over Model 2 which is free of multicollinearity and has an R^2 value greater than 0.90.

Parsimony

Using the principle of parsimony (keep it simple), Model 2 is preferable to Model 4.

Model 5: A different approach

Looking back at the simple correlation matrix we see that the independent variable that has the strongest linear relationship with y is variable x_3 (size). Let us examine the simple regression model:

$$y = \beta_0 + \beta_3 x_3 + \epsilon$$

As we see from the next slide, the estimated regression equation is:

$$\hat{y} = -0.1 + 0.3x_3$$

We observe that this model has an R^2 value of 0.8363, so that variable x_3 alone accounts for approximately 84% of variation in y.

Model 6: Adding variable x_4 to the model

$$\hat{y} = 4.231 + 0.201x_3 - 1.033x_4$$

This model has $R^2 = 0.9150$ so that x_3 and x_4 together explain 91.5% of variation in y.

Further, we see (on the next slide) that the coefficients of both x_3 and x_4 are statistically significantly (p-value < 0.05).

We conclude that the variables size x_3 and place x_4 result in a successful model for predicting store profit y.

Interpreting the model

$\hat{\beta}_1 = 0.201$ implies that a one unit (1000 sq. ft.) increase in size will correspond to an average increase of $201 in profit, assuming x_4 is constant. In other words, the increase of $201 is true for stores in Montreal and in Toronto.

$\hat{\beta}_2 = -1.033$ implies that when stores are in Montreal ($x_4 = 1$), the average profit will be $1033 less than for **stores of the same size** located in Toronto.

Multiple Coefficient of Determination R^2

Measures the percentage of variation in the dependent variable that is explained by the set of **all** independent variables in the model.

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Note: $SS_{yy} = SSTO$

Where $SSE = \sum(y_i - \hat{y}_i)^2, SS_{yy} = \sum(y_i - \bar{y})^2$

Example: Model 6

$$SSR = 14.4161$$
$$SSE = 1.3399$$
$$SS_{yy} = 15.756$$

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{1.3399}{15.716} = 0.915$$

Adjusted R^2

$$R_a^2 = 1 - \frac{(n-1)}{[n-(k+1)]} \times \frac{SSE}{SS_{yy}} = 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1 - R^2)$$

Comparison between R^2 and R_a^2

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$

$$R_a^2 = 1 - \frac{(n-1)}{[n-(k+1)]} \times \frac{SSE}{SS_{yy}}$$

Another view of R_a^2

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)} \times \frac{SSE}{SS_{yy}} = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{SS_{yy}}{n-1}} = 1 - \frac{s_e^2}{s_y^2}$$

This important formula shows us that the R_a^2 value measures the reduction in **variance**, not just the reduction in sums of squares.

R^2 overestimates explained variability

For Model 6:

$$R^2 = 0.9150$$
$$R_a^2 = 0.8907$$

We see that R^2 provides an overly optimistic estimate of the amount of variability explained by the model.

Adjusted R^2 - Rationale

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)} \times \frac{SSE}{SS_{yy}}$$

R^2 is adjusted to take into account the sample size and number of independent variables.

Note 1: If the number of independent variables is large, the values of R^2 and R_a^2 may differ substantially. In general, R^2 will overestimate the explained variability.

Note 2: R^2 cannot decrease when a new independent variable is added.

Note 3: R_a^2 can decrease when a new independent variable is added, indicating that there is no value to adding the new variable to the model.

Computation of F in ANOVA

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}}$$

Since $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$ we can easily show that

$$F = \frac{n-(k+1)}{k} \times \frac{R^2}{1-R^2}$$

Multiple Regression

Coefficient of Partial Determination

The coefficient of partial determination is defined as the percentage of variation left unexplained in the reduced model which is explained by adding variables to obtain the complete model.

Alternatively, we say that the coefficient of partial determination indicates the marginal effect of reducing the variability in Y when new variables are added to a model.

For example, the coefficient of partial determination when x_2 is added to the model already containing x_1 is given by:

$$R^2_{Y_{2.1}} = \frac{SSE(x_1) - SSE(x_1, x_2)}{SSE(x_1)}$$

$R^2_{Y_{2.134}}$ is the percentage of variation left unexplained in the model containing x_1, x_3, and x_4 which becomes explained by adding x_2 to the model.

$R^2_{Y_{34.12}}$ is the percentage of variation left unexplained in the model containing x_1 and x_2 when variables x_3 and x_4 are added to the model.

Note: The larger model containing all variables is called the **complete model**, and the smaller model is called the **reduced model**.

Hence, in general, the coefficient of partial determination may be expressed as:

$$\frac{SSE_R - SSE_C}{SSE_R}$$

Example: Partial Determination

$$R^2_{Y_{2.1}} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$

$$= \frac{5.5315 - 1.534}{5.5315} = \frac{3.9975}{5.5315} = 0.7227$$

Thus, introducing X_2 into the model already containing X_1 explains 72.27% of the variation left unexplained by variable X_1.

Example: Partial Determination

$$R^2_{Y_{3.12}} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)}$$

$$= \frac{1.534 - 1.2941}{1.534} = \frac{0.2399}{1.534} = 0.1564$$

Thus, introducing X_3 into the model already containing X_1 and X_2 explains 15.64% of the variation left unexplained by variables X_1 and X_2.

Nested Models

Two models are said to be **nested** if one model contains all the variables of the other model, plus at least one extra variable.

The larger of the two models is called the **complete** or **full** model.

The smaller model is the **reduced** or **restricted** model.

We perform a test of hypothesis to determine if the extra variables should be added to the reduced model to construct the complete model. We add the extra variables only if they make a significant contribution to the mode.

Testing a Hypothesis

For example, assume we have a reduced model:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

And we are considering the complete model:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

To test whether the two variables X_4 and X_5 should be added to the model, we perform the test of hypothesis:

$$H_0: \beta_4 = \beta_5 = 0$$

H_1: At least one of β_4 or β_5 differs from 0

Calculation of the F-statistic

$$F = \frac{\frac{SSE_R - SSE_C}{k - g}}{\frac{SSE_C}{n - (k + 1)}} = \frac{\frac{SSE_R - SSE_C}{\text{Number of β tested}}}{MSE_C}$$

Where SSE_R = sum of squared errors for the reduced model
SSE_C = sum of squared errors for the complete model
MSE_C = mean square error for the complete model

$k =$ number of independent variables in the complete model
$g =$ number of independent variables in the reduced model

Testing the value of adding a new variable

The question we want to address is whether it is worthwhile adding variable x_2 to the model already containing x_1.

To do this we will perform an F-test for comparing nested models.

Example: Comparing Model 1 and Model 2

Calculation of the F-statistic

$$F = \frac{\frac{SSE_R - SSE_C}{k - g}}{\frac{SSE_C}{n - (k + 1)}} = \frac{\frac{5.5315 - 1.5340}{1}}{\frac{1.5340}{7}} = 18.2415$$

Test of Hypothesis for Comparing Nested Models

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$
$$TS: F = 18.2415$$
$$CV: F_{0.05;1,7} = 5.59$$

Conclusion: Reject H_0 i.e. variable x_2 should be included in the model.

Recall, $R^2_{Y_{2.1}} = 0.7227$

Indicator Variables

An indicator (dummy) variable is a variable that can assume either of only two values (0 or 1), where one value indicates the existence of a certain condition and the other value indicates that the condition does not hold.

In general, to represent n possible conditions we must create $n - 1$ indicator variables.

Example: Indicator Variables

$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ (Model 6)

$x_4 = 0$ if the store is located in Toronto
$x_4 = 1$ if the store is located in Montreal

$Y = \beta_0 + \beta_3 X_3 + \beta_4 + \epsilon$ (Montreal)

$Y = \beta_0 + \beta_3 X_3 + \epsilon$ (Toronto)

Test of Hypothesis for Indicator Variables

To test for a difference in profit between stores in Montreal and Toronto perform the follow t-test:

$H_0: \beta_4 = 0$ (no difference)
$H_1: \beta_4 \neq 0$ (difference)

Significance of x_4

Since the p-value for $\hat{\beta}_4$ is 0.038, which is less than 0.05, we conclude that the indicator variable x_4 is significant, so we retain it in the model.

Dummy Variables for Seasonal Differences

In general, to represent n possible conditions we must create $n - 1$ indicator variables. For example to test for seasonality in quarterly data where there are 4 seasons: winter, spring, summer, and fall we introduce 3 indicator variables as shown:

Estimated profit vs. Advertising:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$X_1 =$ advertising expenditures
$X_2 = 1$ (if the season is winter), $= 0$ (if not).
$X_3 = 1$ (if the season is spring), $= 0$ (if not).
$X_4 = 1$ (if the season is summer), $= 0$ (if not).

Note: If $X_1 = X_2 = X_3 = 0$ it must be fall, therefore we do not require a dummy variable to indicate that it is fall.

Quadratic Example

Conclusion

The model including years and gender has an R_a^2 value of 66.5%.

The model including years, years squared, and gender has an R_a^2 value of 78.9%.

Therefore, the quadratic term should be included.

Interaction

Statistical interaction refers to the extent to which the effect of one independent variable on the dependent variable depends on the value of other independent variables in the mode.

Example

A drug X might be desirable for treating a medical condition, but you have to be careful if you are taking drug Y, because if you do take drugs X and Y together there may be a drug "interaction". That is, the effect of drug X may be dependent on the level of drug Y in the blood.

Adding an interaction term to the model

To test for the interaction term between two variables in a model we add a new variable which is the product of the two existing variables.

Model without interaction term:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Model with interaction term:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_x + \beta_3 x_3$$

Where $x_3 = x_1 * x_2$

Example: Interaction

Interpreting the model

We see that this model is highly significant.

A one-year increase in age results in an average increase in price of $12.74 for auctions with the same number of bidders.

An increase of one bidder results in an average increase in price of $85.95 for grandfather clocks of the same age.

Minitab

Add an interaction variable $\text{Age} * \text{Numbid}$.

We see that there is a significant interaction between Age and NumBid.

Test of Hypothesis for Interaction

From the Minitab output we see that the coefficient of the interactive term $\text{Age} * \text{NumBid}$ is 1.30 with a t-value of 6.11 and an associated p-value of 0.000 (often expressed as p-value < 0.001).

Therefore, we conclude that the interaction term is significant (i.e. reject the hypothesis $H_0: \beta_3 = 0$).

Interaction: Slope of Age depends on the value of NumBid

$$\text{Price} = 320 + 0.80\ \text{Age} - 93.3\ \text{NumBid} + 1.30\ \text{Age} \times \text{NumBid}$$

If NumBid $= 5$, Price $= -146.5 + 7.3$ Age

If NumBid $= 10$, Price $= -613 + 13.8$ Age

If NumBid $= 15$, Price $= -1079.5 + 20.3$ Age

Interpretation of Interaction

For this example, we conclude that there is a significant interaction between Age and NumBid.

This means that the response of Price to Age depends on the number of bidders.

We see that more bidders always results in a higher auction price. **Furthermore, as the number of bidders increases, the rate of increase of Price as a function of Age also increases.**

In practical terms, this implies that the antique dealer should reserve older clocks for larger auctions, as each extra year of age results in a large increase in price when there is a larger number of bidders.

Warning about predictions

Assume a clock aged 150 years and an auction with 5 bidders.

Model ignoring interaction:

$$P = -133.9 + 12.74(150) + 85.99(5) = \$2206.85$$

Model including interaction:

$$P = 320 + 0.88(150) - 93.3(5) + 1.3(150)(5) = \$960.5$$

If we ignore interaction in this model, the predicted price will be seriously overestimated.

Caution on Interaction

If the interaction term $x_3 = x_1x_2$ in the model

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$$

Is significant, you should not conduct t-tests on the coefficients of the first-order terms x_1 and x_2.

These terms must be retained in the model regardless of their associated p-values.

Note: Interaction $\neq$ Multicollinearity

The example of the next slide shows a data set with two variables that are chosen to have no multicollinearity but do exhibit interaction.

Logarithmic Transformations

	X	
Y	X	logX
Y	*linear* $\hat{Y}_i = \alpha + \beta X_i$	*linear-log* $\hat{Y}_i = \alpha + \beta \log X_i$
logY	*log-linear* $\log\hat{Y}_i = \alpha + \beta X_i$	*log-log* $\log\hat{Y}_i = \alpha + \beta \log X_i$

Table 1: Four varieties of logarithmic transformations

Example: Number of cell phones in new rural community

The linear function

$$NumCells = -12471 + 4878t$$

Forecast for 2015 ($t = 10$):

$$NumCells = -12471 + 4878(10) = 36309$$

The log-linear function

$$\ln(NumCells) = 5.524 + 0.5835t$$
$$NumCells = e^{5.524+0.5835t} = 250.64e^{0.5835t}$$

Forecast for 2015 ($t = 10$):

$$NumCells = 250.64e^{0.5835(10)} = 85735$$

Note on model fit

Note that the linear model seriously underestimates the predicted number of cell phones for 2015 **assuming the historical exponential rate of growth continues at the same rate.**

Log-log function

$$\ln Y = \beta_0 + \beta_1 \ln(t)$$

$$\ln Y = \beta_0 + \ln(t)^{\beta_1}$$

$$\ln Y - \ln t^{\beta_1} = \beta_0$$

$$\ln\left(\frac{Y}{t^{\beta_1}}\right) = \beta_0$$

$$\frac{Y}{t^{\beta_1}} = e^{\beta_0}$$

$Y = Ct^{\beta_1}$, where $C = e^{\beta_0}$

Lack of Fit Tests

A regression model exhibits lack-of-fit when it fails to adequately describe the functional relationship between the experimental factors and the response variable. Lack-of-fit can occur if important terms from the model such as interactions or quadratic terms are not included. It can also occur if several, unusually large residuals result from fitting the model.

Lack of Fit Test in Minitab

Minitab displays the lack-of-fit test when your data contain replicates (multiple observations with identical x-values). Replicates represent "pure error" because only random variation can cause differences between the observed response values.

To determine whether the model accurately fits the data, compare the p-value to your significant level. Usually, a significance level (also called alpha or α) of 0.05 works well. An α of 0.05 means that your chance of concluding that the model does not fit the data when it really does it only 5%.

P-value $< \alpha$: The model does not fit the data

If the p-value is less than or equal to α, you conclude that the model does not accurately fit the data. To get a better model, you may need to add terms or transform your data.

P-value $> \alpha$: There is no evidence that the model does not fit the data

If the p-value is larger than α, you cannot conclude that the model does not fit the data well.

Example: Conclusion

From this example we see that the linear model is not a good fit to the data (p-value $= 0.038$), whereas the quadratic model is a good fit (p-value $= 0.558$).

Variable Screening Techniques

Stepwise Regression

Variables are entered into the model in order of significance (based on p-values or F-values) until the new value being added is not significant.

If introducing a new variable causes an existing variable to become non-significant it is dropped from the model.

Data

STORE	Y PROFIT	X_1 ADVERT	X_2 SPECIAL	X_3 SIZE	X_4 PLACE
1	9.4	3	1	30	1
2	10.3	3	5	37	1
3	10.9	4	5	38	1
4	9.9	4	2	35	1
5	12.9	5	6	40	0
6	11.8	5	6	40	0
7	11.5	6	2	39	1
8	13.2	6	5	45	0
9	12.8	7	5	41	0
10	12.1	7	1	41	0

We reproduce the simple correlation matrix here for convenience:

	Profit	Advert	Special	Size	Place
Profit	1				
Advert	.8056	1			
Special	.5037	0	1		
Size	.9145	.7577	.5014	1	
Place	-.8604	-.7071	-.4126	-.7318	1

Regression Output

Select "Stepwise" from the Regression Screen.

Step 1 produces the model:

$$\hat{y} = -0.10 + 0.3x_3$$

With $R_a^2 = 81.58\%$

Step 2 produces the model:

$$\hat{y} = 4.23 + 0.2012x_3 - 1.033x_4$$

With $R_a^2 = 89.07\%$

Minitab's Stepwise Regression Results

Minitab selected Model 6

$$y = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

$$\hat{Y} = 4.231 + 0.201x_3 - 1.033x_4$$

This model explains Profit (Y) in terms of variables Size (x_3) and Place (x_4).

Forcing a variable into the model

Suppose that the manager in charge of advertising budgets insists she needs a model that contains the variable Advert (x_1).

In the next slide we show how to force the variable Advert into the model.

Advert will now be forced into the model – note the letter E to the left of ADVERT.

Stepwise Regression with Advert forced into model

Regression Analysis: PROFIT versus ADVERT, SPECIAL, SIZE, PLACE

```
Stepwise Selection of Terms

Candidate terms: ADVERT, SPECIAL, SIZE, PLACE

             -----Step 1----     -----Step 2-----
              Coef          P      Coef          P
Constant     7.905                6.666
ADVERT       0.715      0.005     0.715      0.000
SPECIAL                           0.3261     0.004

S                     0.831527             0.468120
R-sq                    64.89%               90.26%
R-sq(adj)               60.50%               87.48%
R-sq(pred)              48.61%               81.79%
Mallows' Cp              23.47                 4.17

α to enter = 0.15, α to remove = 0.15
At your request, the stepwise procedure included these terms in every model:
     ADVERT
```

Stepwise Regression with x_1 = Advert forced into the model

Stepwise regression has selected Model 2.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\hat{Y} = 6.666 + 0.715 x_1 + 0.326 x_2$$

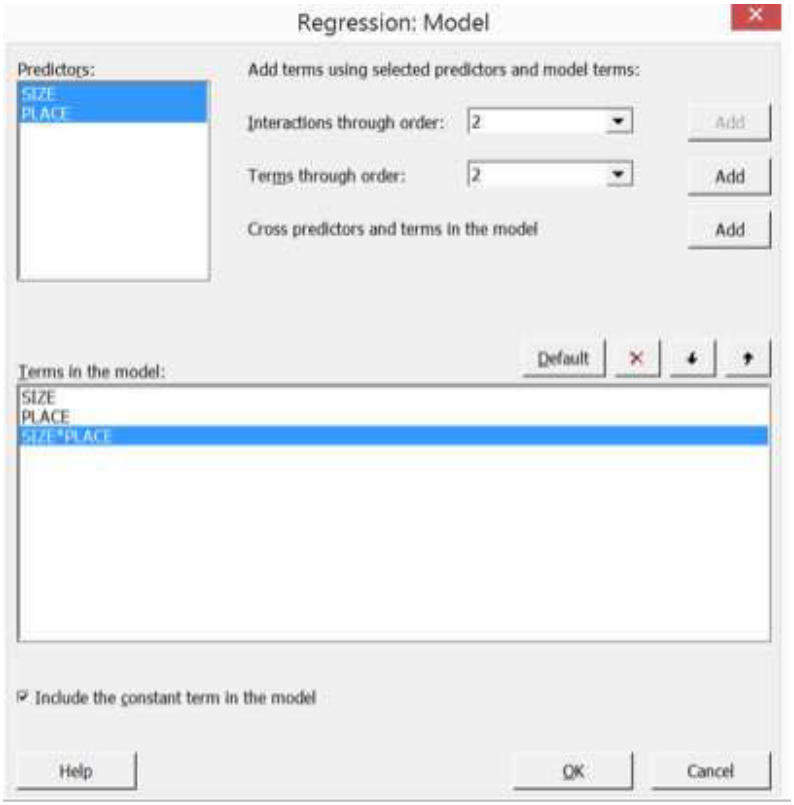
Advert (x_1) and Special (x_2) are the selected variables.

Summary

We see that Stepwise Regression can help the analyst find a suitable model with important variables included as required for a specific application.

Once we have decided upon the independent variables to include in a model it is important to check for possible interaction among the variables.

Specifying a model with interaction



No interaction between Size and Place

```
Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.469389  91.61%     87.41%       66.41%


Coefficients

Term          Coef  SE Coef  T-Value  P-Value     VIF
Constant      5.39     4.69     1.15    0.294
SIZE         0.173    0.113     1.53    0.177    8.51
PLACE        -2.53     5.25    -0.48    0.647  313.19
SIZE*PLACE   0.037    0.131     0.29    0.785  253.31


Regression Equation

PROFIT = 5.39 + 0.173 SIZE - 2.53 PLACE + 0.037 SIZE*PLACE
```

Selecting a stepwise regression method

Method: Specify the method that Minitab uses to fit the model.

None: Choose to fit the model with all of the terms that you specify in the Model dialog.

Stepwise: By default, this procedure starts with an empty model and then adds or removes a term for each step. You can specify terms to include in the initial model or to force into every model.

Forward selection: By default, the procedure starts with an empty model and adds the most significant term for each step. You can specify terms to force into every model.

Backward selection: This procedure starts off with all potential terms in the model and removes the least significant term for each step. You can specify terms to force into every model.

Forward selection

Starts with no variables and introduces variables into model in order of decreasing significance. Unlike stepwise, it never drops variables already in model.

Forward selection output

Regression Analysis: PROFIT versus ADVERT, SPECIAL, SIZE, PLACE

```
Forward Selection of Terms

Candidate terms: ADVERT, SPECIAL, SIZE, PLACE

             -----Step 1-----     -----Step 2-----
               Coef          P      Coef          P
Constant      -0.10                 4.23
SIZE         0.3000      0.000    0.2012      0.007
PLACE                             -1.033      0.038

S                        0.567891             0.437510
R-sq                       83.63%               91.50%
R-sq(adj)                  81.58%               89.07%
R-sq(pred)                 70.09%               83.40%
Mallows' Cp                  7.74                 3.14

α to enter = 0.25
```

The selected model is Profit $= 4.23 + 0.2012$ Size -1.033 Place

Backward elimination

Starts with all k variables in model and first eliminates the least significant variable. It reruns model on $k - 1$ variable and again eliminates the least significant variable. Repeats process until all remaining variables are significant.

Results of backward elimination

Regression Analysis: PROFIT versus ADVERT, SPECIAL, SIZE, PLACE

Backward Elimination of Terms

Candidate terms: ADVERT, SPECIAL, SIZE, PLACE

	-----Step 1-----		-----Step 2-----		-----Step 3-----	
	Coef	P	Coef	P	Coef	P
Constant	5.57		8.03		6.666	
ADVERT	0.341	0.211	0.557	0.013	0.715	0.000
SPECIAL	0.155	0.264	0.2589	0.028	0.3261	0.004
SIZE	0.1022	0.287				
PLACE	-0.657	0.227	-0.632	0.248		
S		0.433259		0.448153		0.468120
R-sq		94.04%		92.35%		90.26%
R-sq(adj)		89.28%		88.53%		87.48%
R-sq(pred)		77.08%		82.38%		81.79%
Mallows' Cp		5.00		4.42		4.17

α to remove = 0.1

The selected model is Profit $= 6.666 + 0.71$ Advert $+0.326$ Special

Variable Screening Technique

Choice of Method

The ability to choose a method gives a statistician flexibility in selecting a regression model.

The Stepwise (forward and backward) method is generally a reliable approach combining the concepts of forward selection and backward elimination.

Best Subsets - All Possible Regressions

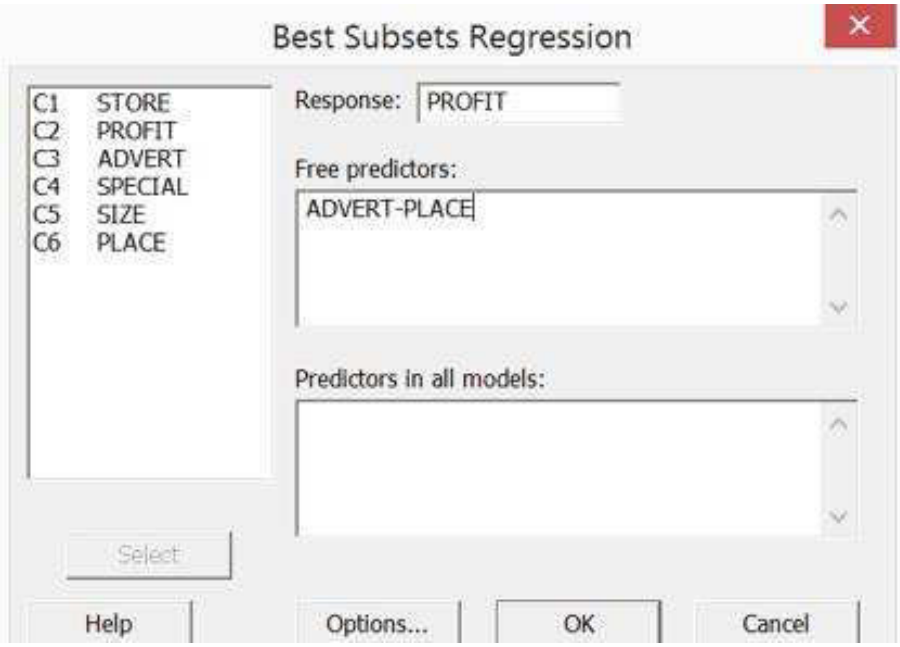
An alternative to Stepwise Regression is the Best Subsets approach, also known as the All Possible Regressions method.

This method produces the output of many alternative regression models and allows the statistician to select a model based on different criteria.

The next screen shows how to run a best subsets regression in Minitab.

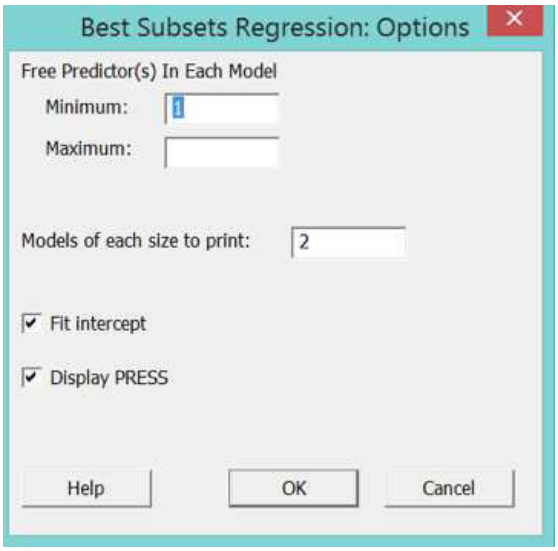
Minitab's Best Subsets option on the Regression Menu

Stat → Regression → Regression → Best subsets



Click Options…

Select box Display PRESS



Best Subsets Regression Output

Best Subsets Regression: PROFIT versus ADVERT, SPECIAL, SIZE, PLACE

Response is PROFIT

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	ADVERT	SPECIAL	SIZE	PLACE
1	83.6	81.6	4.7	70.1	7.7	0.56789			X	
1	74.0	70.8	6.4	59.4	15.8	0.71519				X
2	91.5	89.1	2.6	83.4	3.1	0.43751			X	X
2	90.3	87.5	2.9	81.8	4.2	0.46812	X	X		
3	92.4	88.5	2.8	82.4	4.4	0.44815	X	X		X
3	92.2	88.2	3.2	79.7	4.6	0.45383	X		X	X
4	94.0	89.3	3.6	77.1	5.0	0.43326	X	X	X	X

Summary of Results

The Minitab output shows six measures:

1. R^2
2. R_a^2
3. PRESS
4. R_{pred}^2
5. Mallows C_p
6. S

Any of these measures can be used as a criterion for selecting a "best" model.

We will discuss the advantages and disadvantages of each criterion.

R^2 and s are not good criteria

We have already encountered $R^2 = \frac{SSR}{SSTO}$ and $s = \sqrt{MSE}$

Although these are potential model selection criteria, they are generally not considered to be good criteria because they ignore the relationship between the sample size and the number of variables in a model i.e. they ignore degrees of freedom.

Consequently we will consider the advantages of the other criteria listed in the Minitab output.

Criteria for model selection

Adjusted R-squared and the Mallows C_p statistic are popular model selection criteria.

Mallows C_p is related to adjusted R-squared, but imposes a penalty for increasing the number of independent variables. Consequently C_p is called a **parsimonious** decision criterion.

C_p values are usually positive and **lower values are better**. Occasionally C_p maybe negative and this usually implies a good model.

Although models that result in lower values of C_p tend to be similar to those that give higher values of adjusted R-squared, the exact ranking will often be slightly different.

Sometimes there will be several models with the same or very slightly smaller values of R_a^2 . In such cases Mallow's C_p is often useful in helping to choose among the competing models.

Mallows C_p

Mallows C_p is defined by:

$$C_p = \frac{SSE_p}{MSE_k} + 2(p+1) - n$$

It can be shown that $E(C) \sim p+1$

Where n is the sample size, p is the number of independent variables in the **subset** model, k is the total number of potential independent variables, SSE_p is the SSE for the subset model, and MSE_k is the MSE for the model containing all k independent variables.

The C_p criterion selects a model with (1) small value of C_p and (2) C_p near to $p+1$.

Note that adding the term $2(p+1)$ penalizes (increases the value of) C_p for every additional variable in the subset model.

Note: Because n is subtracted it is possible for C_p to be negative.

Variable Screening Techniques

The PRESS Statistic

PRESS (Predicted Residual Sum of Squares)

PRESS is based on the "leave one out" or "Jackknife" technique in which one fits the model without the i^{th} observation x_i and uses this fitted model to predict the response $\hat{y}_i$ when $x = x_i$. The PRESS residuals are defined as $e_i = \text{y}_i - \hat{y}_i$. The process is repeated for all n observations and the PRESS statistics is computed as:

$$PRESS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

The lower the value of PRESS the better the predictive model.

Rationale for PRESS

The candidate model is fit to the sample data n times, each time leaving out one data point and substituting in its place the value predicted by the remaining $n-1$ data points.

Since small differences $e_i = y_i - \hat{y}_i$ indicate that the model is predicting well, we choose a model with a small value of PRESS.

PRESS is a computationally intensive procedure, but advanced computer packages have options for computing PRESS.

Predicted R-squared R^2_{pred}

The predicted R-squared indicates how well a regression model predicts responses for new observations. This statistic helps you determine when the model fits the original data but is less capable of proving valid predictions for new observations. As in the case of the PRESS statistic, Minitab calculates predicted R-squared by systematically removing each observation from the data set, estimating the regression equation, and determining how well the model predicts the removed observation. Predicted R-squared can be negative and it is always lower than R-squared.

Even if you don't plan to use the model for predictions, the predicted R-squared still provides crucial information.

A key benefit of predicted R-squared is that it can prevent you from overfitting a model. An overfit model contains too many predictors and it starts to model the random noise.

Because it is impossible to predict random noise, the predicted R-squared must drop for an overfit model. If you see a predicted R-squared that is much lower than the regular R-squared, you almost certainly have too many terms in the model.

Relationship between R^2_{pred} and PRESS

$$R^2_{pred} = \left[1 - \frac{PRESS}{SSTO}\right](100)$$

Example

Profit $= 4.23 + 0.2012$ Size $- 1.033$ Place

Regression Analysis: PROFIT versus SIZE, PLACE

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	14.4161	91.50%	14.4161	7.20805	37.66	0.000
SIZE	1	13.1760	83.63%	2.7521	2.75209	14.38	0.007
PLACE	1	1.2401	7.87%	1.2401	1.24009	6.48	0.038
Error	7	1.3399	8.50%	1.3399	0.19142		
Lack-of-Fit	5	0.4899	3.11%	0.4899	0.09798	0.23	0.919
Pure Error	2	0.8500	5.39%	0.8500	0.42500		
Total	9	15.7560	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.437510	91.50%	89.07%	2.61593	83.40%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	4.23	2.21	(-0.98, 9.45)	1.92	0.097	
SIZE	0.2012	0.0531	(0.0757, 0.3266)	3.79	0.007	2.15
PLACE	-1.033	0.406	(-1.993, -0.073)	-2.55	0.038	2.15

Regression Equation

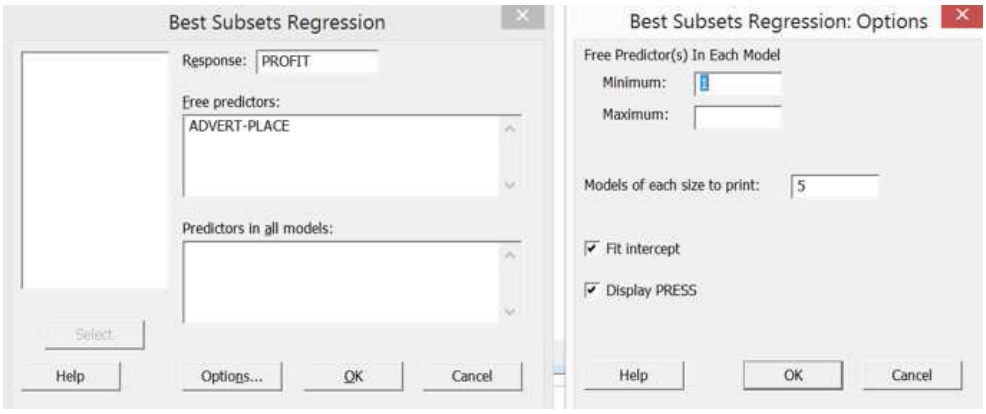
PROFIT = 4.23 + 0.2012 SIZE - 1.033 PLACE

$$R^2_{pred} = \left[1 - \frac{PRESS}{SSTO}\right](100)$$

$$PRESS = 2.61593; SSTO = 15.756$$

$$R^2_{pred} = \left[1 - \frac{2.61593}{15.756}\right](100) = [1 - 0.166](100) = 83.4\%$$

Best subsets models including PRESS



Full set of models

Best Subsets Regression: PROFIT versus ADVERT, SPECIAL, SIZE, PLACE

Response is PROFIT

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	ADVERT	SPECIAL	SIZE	PLACE
1	83.6	81.6	4.7	70.1	7.7	0.56789		X		
1	74.0	70.8	6.4	59.4	15.8	0.71519				X
1	64.9	60.5	8.1	48.6	23.5	0.83153	X			
1	25.4	16.0	19.0	0.0	56.6	1.2124		X		
2	91.5	89.1	2.6	83.4	3.1	0.43751			X	X
2	90.3	87.5	2.9	81.8	4.2	0.46812	X	X		
2	86.6	82.8	4.2	73.1	7.2	0.54908	X		X	
2	83.9	79.3	5.5	64.8	9.5	0.60202		X	X	
2	81.8	76.6	6.1	61.4	11.3	0.63998	X			X
3	92.4	88.5	2.8	82.4	4.4	0.44815	X	X		X
3	92.2	88.2	3.2	79.7	4.6	0.45383	X		X	X
3	91.8	87.7	4.1	74.2	4.9	0.46441	X	X	X	
3	91.6	87.4	4.0	74.8	5.1	0.46996		X	X	X
4	94.0	89.3	3.6	77.1	5.0	0.43326	X	X	X	X

Summary of results

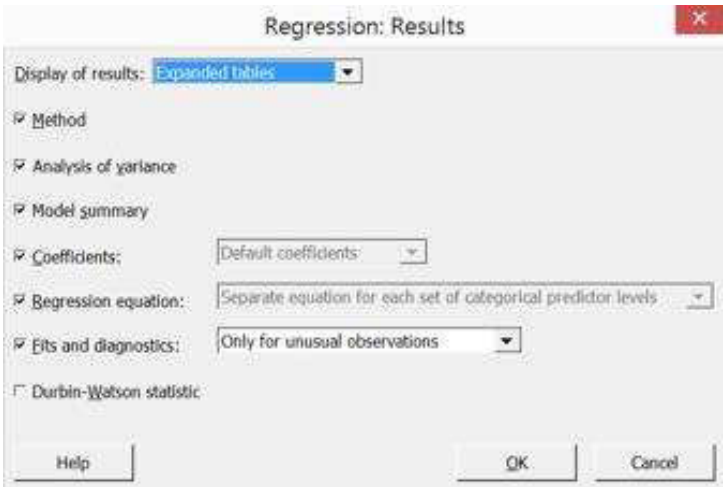
R_a^2 and C_p are generally accepted as excellent measures to use when fitting a model. In this example, it can be seen that C_p attains its minimum value of 3.1 in the fifth model listed, namely the model with independently variables x_3 and x_4.

PRESS also attains its minimum value of 2.6 for this model.

Although R_a^2 also has a favorable value 89.1 for that model, its value in the complete model with all four variables is a little larger 89.3. We recall, however, that the complete model has a serious problem with multicollinearity and adds little to the R^2 or R_a^2 values.

Calculating PRESS

We can also obtain the PRESS values for each model separately by modifying the output to display "Expanded Tables".



The next slide shows how to calculate PRESS for the model with variables x_3 and x_4.

Expanded Tables includes PRESS

Regression Analysis: PROFIT versus SIZE, PLACE

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	14.4161	91.50%	14.4161	7.20805	37.66	0.000
SIZE	1	13.1760	83.63%	2.7521	2.75209	14.38	0.007
PLACE	1	1.2401	7.87%	1.2401	1.24009	6.48	0.038
Error	7	1.3399	8.50%	1.3399	0.19142		
Lack-of-Fit	5	0.4899	3.11%	0.4899	0.09798	0.23	0.919
Pure Error	2	0.8500	5.39%	0.8500	0.42500		
Total	9	15.7560	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.437510	91.50%	89.07%	2.61593	83.40%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	4.23	2.21	(-0.98, 9.45)	1.92	0.097	
SIZE	0.2012	0.0531	(0.0757, 0.3266)	3.79	0.007	2.15
PLACE	-1.033	0.406	(-1.993, -0.073)	-2.55	0.038	2.15

Regression Equation

PROFIT = 4.23 + 0.2012 SIZE - 1.033 PLACE

Regression Pitfalls

Observational vs. Experimental Data

Observation data: Values of independent variable are uncontrolled.

Experimental data: Values of independent variable are controlled via a designed experiment.

Advertising Example

The variables ADVERT, SPECIAL, and PLACE are experimental variables. The general manager decides how much advertising budget to allocate to each store, and how many in-house specials to assign. He also selects stores in Montreal and Toronto.

The variable SIZE is observational – we assume that the stores sampled are already in business and the manager has no control over the size of the random sample of stores selected in the two locations.

Advantages of Using Experimental Variables

The user controls the experiment.

Variable values can be assigned in such a way that independent variables are not correlated so that multicollinearity can be eliminated. For example, we saw that the variables ADVERT and SPECIAL were designed to have zero correlation.

Cause and effect relationships can be inferred.

Randomization can be controlled by assigning a desired range of values to the independent variables. For example, we would not get a useful model if all stores had a advertising budget of between $4500 and $5500. Because of extrapolation, we would not be able to make inferences outside that narrow range.

Deviating from the Assumptions

Recall that the multiple regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

The model assumptions:

1. $E(\epsilon) = 0$
2. ϵ is normally distributed
3. $\sigma_\epsilon^2 = \sigma^2$ (a constant)
4. ϵ_i are statistically independent

In summary, ϵ_i are independent $N(0, \sigma^2)$ variables.

Multicollinearity

Multicollinearity occurs when two independent variables are moderately or highly correlated.

This does not mean that on its own each independent variable does not explain a significant amount of variation in the dependent variable. It simply means that there are variables in the mode which are performing the same job in explaining the variation in the dependent variable and therefore they are all not needed in the regression model together.

The existence of multicollinearity in a multiple regression model affects the reliability of the least squares estimates of the regression coefficients.

How is multicollinearity detected?

1. There exists high correlation between pairs of independent variables in the model. Refer to the correlation matrix.

2. Estimated regression coefficients of variables change when a variable is added or deleted. In extreme cases the coefficient may change sign.

3. There are conflicting results between the F-test and the t-tests.

4. A variance inflation factor (VIF) exceeds 5 (this is somewhat arbitrary but experience supports the idea that if a VIF > 5 the model probably has a problem with multicollinearity; some authors suggest a cut-off value of 10).

Variance Inflation Factor

The variance inflation factor for variable x_i is given by:

$$VIF_i = \frac{1}{1 - R_i^2}, i = 1,2, \dots, k$$

Where R_i^2 is the multiple coefficient of determination in the multiple regression model that express x_i as a function of all variables in the model except x_i, i.e. in the model:

$$E(x_i) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_k x_k$$

Four methods for detection of multicollinearity

1. Using Correlation Matrix

2. Examining changes in regression coefficients as new variables are included in the model

3. Overall F is significant while individual t values are not significant

4. Using VIFs

Correlation Matrix: Method 1

	Profit	Advert	Special	Size	Place
Profit	1				
Advert	.8056	1			
Special	.5037	0	1		
Size	.9145	.7577	.5014	1	
Place	-.8604	-.7071	-.4126	-.7318	1

Coefficients Change: Method 2

Model 1: $\hat{Y} = 7.905 + 0.715X_1$

Model 2: $\hat{Y} = 6.666 + 0.715X_1 + 0.326X_2$

Model 3: $\hat{Y} = 4.283 + 0.516X_1 + 0.230X_2 + 0.097X_3$

In Model 2 we see that the coefficient of X_1 is 0.715 indicating that there is no multicollinearity between variables X_1 and X_2.

However, in Model 3 we note that the coefficient of X_1 has changed to 0.516. This indicates that this model exhibits multicollinearity. We had already seen in the correlation matrix that X_3 is correlated with X_1 and X_2.

Model 3: Profit on ADVERT, SPECIAL, and SIZE: Method 3

Regression Analysis: PROFIT versus ADVERT, SPECIAL, SIZE

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	14.4619	4.8206	22.35	0.001
Total	9	15.7560			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.464414	91.79%	87.68%	74.18%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.28	2.34	1.83	0.117	
ADVERT	0.516	0.215	2.40	0.053	4.29
SPECIAL	0.230	0.118	1.95	0.100	2.44
SIZE	0.0969	0.0919	1.05	0.332	5.73

Regression Equation

PROFIT = 4.28 + 0.516 ADVERT + 0.230 SPECIAL + 0.0969 SIZE

Variance Inflation Factors: Method 4

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.433259	94.04%	89.28%	77.08%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.57	2.37	2.34	0.066	
ADVERT	0.341	0.238	1.44	0.211	6.02
SPECIAL	0.155	0.123	1.26	0.264	3.04
SIZE	0.1022	0.0858	1.19	0.287	5.74
PLACE	-0.657	0.478	-1.38	0.227	3.04

Regression Equation

PROFIT = 5.57 + 0.341 ADVERT + 0.155 SPECIAL + 0.1022 SIZE - 0.657 PLACE

We see that two of the VIF values exceed 5.

Why is multicollinearity a problem?

1. The standard errors of the regression coefficient are inflated, i.e. $\hat{\sigma}_{\hat{\beta}_j}$ are large.

2. Estimated regression coefficients must be interpreted as the average change in the dependent variable per unit change in an independent variable, **when all other variables are held constant**.

3. Inferential statistics on the regression parameters (i.e. t-tests, confidence intervals for β_i) are not reliable.
   - Large standard errors mean large (imprecise) confidence intervals $\hat{\beta}_i \pm t_{\frac{\alpha}{2}}\sigma_{\hat{\beta}_j}$
   - Large standard errors mean small observed test statistics $t = \hat{\beta}_i / \hat{\sigma}_{\hat{\beta}_j}$

The researcher will "accept" too many null hypotheses; high probability of Type II error.

Important Note: Predictions for Y when multicollinearity is present

In the case of multicollinearity, R^2 and the predictive power of the regression line remain unaffected. In other words, a multiple regression model with a relatively high R^2 and a significant F-test can be used for prediction purposes (i.e. confidence intervals for $E(Y)$ and prediction intervals for an individual Y) **even if multicollinearity exists**.

Guidelines for interpreting VIFs

Any individual VIF > 10 suggests severe multicollinearity; thus multicollinearity may be influencing the least-squares estimates of the regression coefficients.

The rule VIF > 10 is very conservative. Start to suspect multicollinearity when VIF > 5.

If all VIFs are less than $1/(1 - R^2)$, where R^2 is the coefficient of determination for the model with all independent variables included, then multicollinearity is not strong enough to affect the coefficients estimates. In this case the independent variables are more strongly related to the Y variable than they are to each other.

Cigarette Data Example

See sample data on next screen.

We run a multiple regression of the variable Carbon Monoxide (CM) on the three independent variables TAR, NICOTINE, and WEIGHT.

We want to investigate multicollinearity to recommend a suitable predictive model.

We start by examining the simple correlation matrix.

	Brand	TAR	NICOTINE	WEIGHT	CM
1	Alpine	14.1	0.86	0.9853	13.6
2	Benson & Hedges	16.0	1.06	1.0938	16.6
3	Bull Durham	29.8	2.03	1.1650	23.5
4	Camel Lights	8.0	0.67	0.9280	10.2
5	Carlton	4.1	0.40	0.9462	5.4
6	Chesterfield	15.0	1.04	0.8885	15.0
7	Golden Lights	8.8	0.76	1.0267	9.0
8	Kent	12.4	0.95	0.9225	12.3
9	Kool	16.6	1.12	0.9372	16.3
10	L&M	14.9	1.02	0.8858	15.4
11	Lark Lights	13.7	1.01	0.9643	13.0
12	Marlboro	15.1	0.90	0.9316	14.4

Minitab: The Correlation Matrix

In Minitab choose the options:
Stat → Basic Statistics → Correlation

The results:

Correlations: TAR, NICOTINE, WEIGHT

	TAR	NICOTINE
NICOTINE	0.977	
	0.000	
WEIGHT	0.491	0.500
	0.013	0.011

Interpreting the correlation matrix

TAR content (x_1) and NICOTINE content (x_2) are highly correlated $(r = 0.977)$.

Weight (x_3) is moderately correlated with TAR content $(r = 0.491)$ and nicotine $(r = 0.500)$.

The small p-values shown under the correlations confirm that the correlations are all significantly different from 0.

This is strong evidence to suggest that a model containing all three variables will have a serious multicollinearity problem.

Minitab: The VIF Values

7/19/2007 10:29:34 PM

Regression Analysis: CM versus TAR, NICOTINE, WEIGHT

The regression equation is
CM = 3.20 + 0.963 TAR - 2.63 NICOTINE - 0.13 WEIGHT

Predictor	Coef	SE Coef	T	P	VIF
Constant	3.202	3.462	0.93	0.365	
TAR	0.9626	0.2422	3.97	0.001	21.6
NICOTINE	-2.632	3.901	-0.67	0.507	21.9
WEIGHT	-0.130	3.885	-0.03	0.974	1.3

S = 1.44573 R-Sq = 91.9% R-Sq(adj) = 90.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	495.26	165.09	78.98	0.000
Residual Error	21	43.89	2.09		
Total	24	539.15			

Multicollinearity in the model

The VIF for TAR is 21.6 and for NICOTINE is 21.9. Since both of these values exceed the threshold value of 10, we conclude that the model has a serious multicollinearity problem.

Note also that $\frac{1}{1-R^2} = \frac{1}{1-0.919} = 12.35$. Both VIFs exceed this number, implying the presence of multicollinearity.

It appears that we should not have a model that contains all three independent variables TAR, NICOTINE, and WEIGHT.

Choosing a model

Correlations: CM, TAR, NICOTINE, WEIGHT

	CM	TAR	NICOTINE
TAR	.957		
NICOTINE	.926	.977	
WEIGHT	.464	.491	.500

Best Subsets Regression: CM versus TAR, NICOTINE, WEIGHT

7/19/2007 10:29:34 PM

Regression Analysis: CM versus TAR, NICOTINE, WEIGHT

The regression equation is
CM = 3.20 + 0.963 TAR - 2.63 NICOTINE - 0.13 WEIGHT

Predictor	Coef	SE Coef	T	P	VIF
Constant	3.202	3.462	0.93	0.365	
TAR	0.9626	0.2422	3.97	0.001	21.6
NICOTINE	-2.632	3.901	-0.67	0.507	21.9
WEIGHT	-0.130	3.885	-0.03	0.974	1.3

S = 1.44573 R-Sq = 91.9% R-Sq(adj) = 90.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	495.26	165.09	78.98	0.000
Residual Error	21	43.89	2.09		
Total	24	539.15			

Best Model

The minimum values of Mallows C_p is 0.5 corresponding to the simple linear regression model with only one independent variable TAR.

From the best subsets regression output we see that

$R^2 = 91.7\%$ and $R_a^2 = 91.3\%$

So that TAR explains over 91% of variability in CM values.

Note that the model with all three independent variables has a smaller R_a^2 value of 90.7%.

The Model

7/19/2007 10:29:34 PM

Regression Analysis: CM versus TAR, NICOTINE, WEIGHT

The regression equation is
CM = 3.20 + 0.963 TAR - 2.63 NICOTINE - 0.13 WEIGHT

Predictor	Coef	SE Coef	T	P	VIF
Constant	3.202	3.462	0.93	0.365	
TAR	0.9626	0.2422	3.97	0.001	21.6
NICOTINE	-2.632	3.901	-0.67	0.507	21.9
WEIGHT	-0.130	3.885	-0.03	0.974	1.3

S = 1.44573 R-Sq = 91.9% R-Sq(adj) = 90.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	495.26	165.09	78.98	0.000
Residual Error	21	43.89	2.09		
Total	24	539.15			

Data Transformations

Both the dependent variable y and independent variables $x_1, x_2, x_3 \ldots, x_k$ can be subject to data transformations.

Transformations of y may be performed:

1. To make the y values satisfy model assumptions.
2. To make the deterministic portion of the model a better approximation to the mean value of the transformed variable.

Transformations of the independent variables are performed only for the second reason above.

Summary of Regression Models

Model	R^2
Linear	0.9682
Quadratic	0.9935
Logarithmic	0.9768
Power	0.9898
Exponential	0.9852

Choosing the "best" model

Based on the table, can we conclude that the quadratic model is the "best" model?

Is there evidence to show that the demand-price relationship is quadratic?

Reciprocal transformation of independent variable

Suppose that consultation with an expert in the area of demand-price relationships informs you that experience has shown that demand depends on the reciprocal of the price. In other words, she suggests the model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

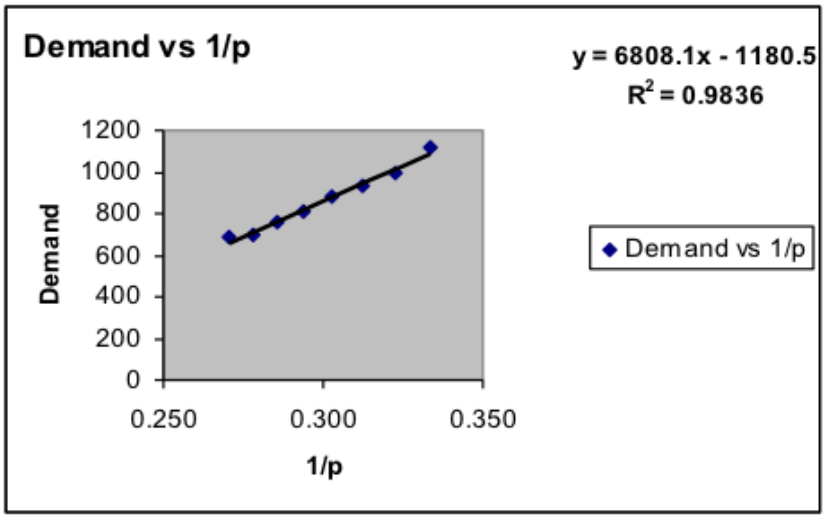
Where $x = \frac{1}{p}$

Or equivalently, $y = \beta_0 + \beta_1 \frac{1}{p} + \epsilon$

Transformed Data $(\boldsymbol{x = \frac{1}{p}})$

Demand (y)	Price (p)	X = 1/p
1120	3.00	0.333
999	3.10	0.323
932	3.20	0.313
884	3.30	0.303
807	3.40	0.294
760	3.50	0.286
701	3.60	0.278
688	3.70	0.270

Demand vs. $\frac{1}{p}$



This model has a high value of R^2 and satisfies the empirical observation that demand is related to the reciprocal of price.

Discussion of efficacy of R^2

Even though the R^2 is a little lower than in the quadratic model, it is important to choose a model that is consistent with the scientific fact associated with the phenomenon under consideration.

Conclusion

Maximizing R^2 is not a sufficient criterion for deciding on a predictive model.

Ultimately, various factors, including the logical association between the variables, must be taken into account in selecting a model. R^2 is just one of those factors.

Residual Analysis I

Residual analysis: Using residuals to detect departure from assumptions.

Residual Plots

Plots of the residuals in a regression analysis can be used to detect departures from the assumption of:

1. Normality of the error terms
2. Constant variance
3. Independence of the error terms

Definition of Residual

Error term in a true multiple regression model:

$$\epsilon_i = y_i - E(y)$$
$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

Residual term in a estimated multiple regression model:

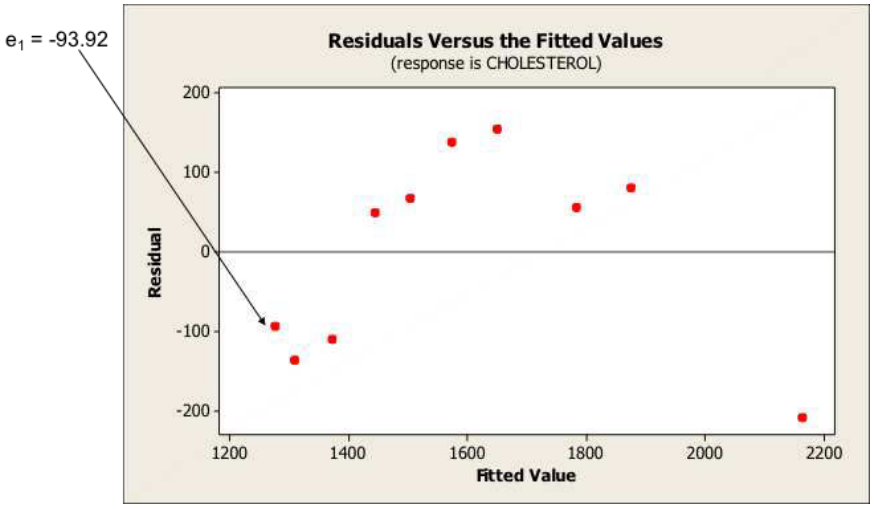
$$e_i = y_i - \hat{y}_i$$
$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$$

Calculating Residuals

ATHLETE	FAT INTAKE	CHOLESTEROL
1	1290	1182
2	1350	1172
3	1470	1264
4	1600	1493
5	1710	1571
6	1840	1711
7	1980	1804
8	2230	1840
9	2400	1956
10	2930	1954

The regression equation is: CHOLESTEROL = 578.9 + 0.5403 FAT INTAKE

$$x_1 = 1290$$
$$\hat{y}_1 = 568.9 + 0.5403(1290) = 1275.92$$
$$e_1 = y_1 - \hat{y}_1 = 1182 - 1275.92 = \mathbf{-93.92}$$



Curvature

See the pattern in the residual plot:

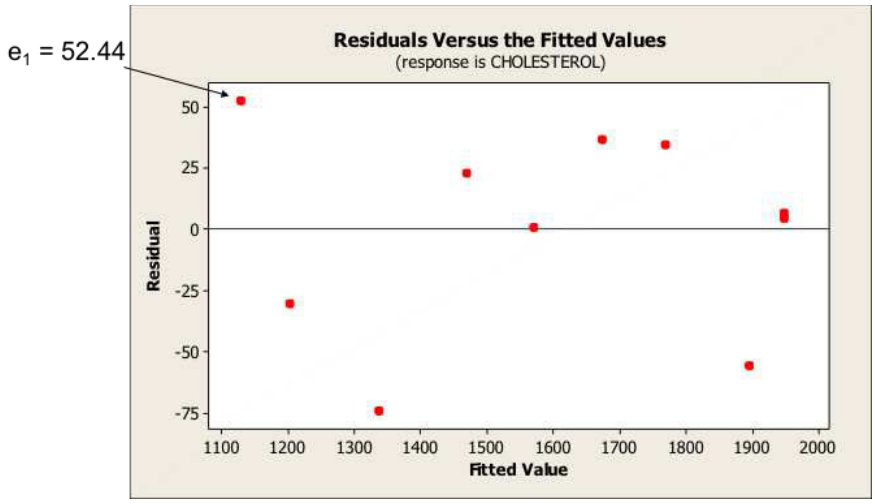
- Residuals are positive for athletes with intermediate levels of fat intake.
- Residuals are negative for athletes with low or high levels of fat intake.

This suggests the use of a second order (quadratic) model.

Quadratic Model

The regression equation is: CHOLESTEROL = -1216.14 + 2.3989 FAT INTAKE - 0.000450 FAT^2

$$x_1 = 1290$$
$$\hat{y}_1 = -1216.14 + 2.3989(1290) - 0.000450(1290)^2 = 1129.56$$
$$e_1 = y_1 - \hat{y}_1 = 1182 - 1129.56 = \mathbf{52.44}$$



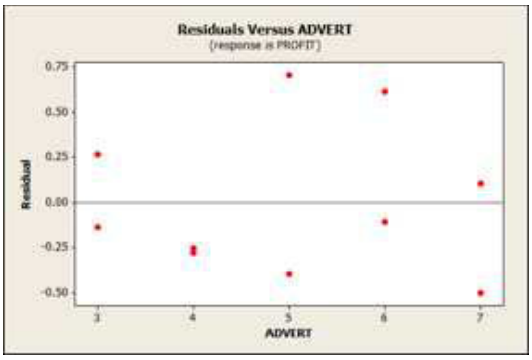
Detecting Lack of Fit

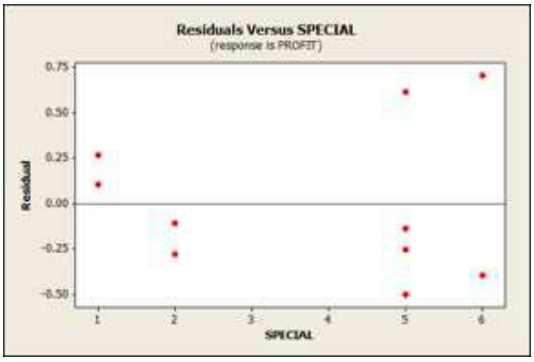
1. Plot residuals e_i on the vertical axis, against each of the independent variables.
2. Plot residuals on the vertical axis against the predicted value $\hat{y}$ on the horizontal axis.
3. In each plot look for trends, dramatic changes in variability, and/or more than 5% of residuals that lie outside 2 std. dev. of 0.

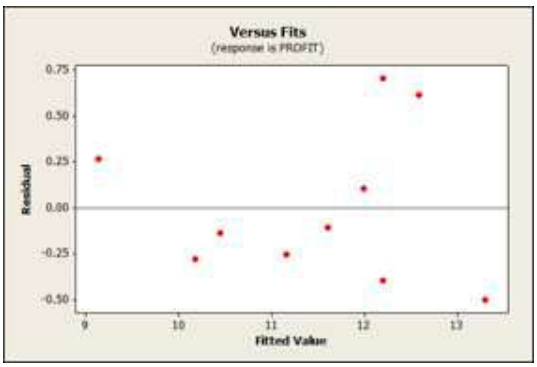
Example of Residual Plots

PROFIT	ADVERT	SPECIAL	SIZE	PLACE
9.4	3	1	30	1
10.3	3	5	37	1
10.9	4	5	38	1
9.9	4	2	35	1
12.9	5	6	40	0
11.8	5	6	40	0
11.5	6	2	39	1
13.2	6	5	45	0
12.8	7	5	41	0
12.1	7	1	41	0

Residual Plots





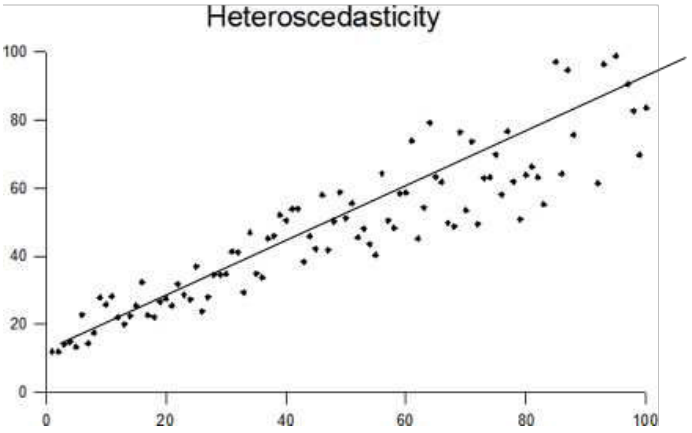


Detecting Unequal Variances

Heteroscedasticity occurs when regression results produce error terms that are of significantly varying degrees across different observations.

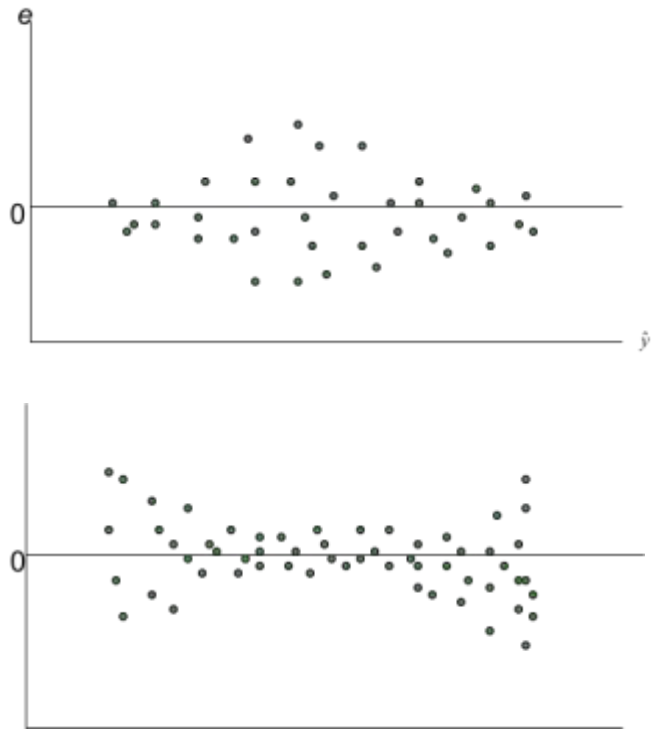
One example is when your independent variable gets larger, so does the distribution of error terms around it.

Scatter Diagram of Heteroscedasticity



$\hat{y} = b_0 + b_1 x$

The terms are more spread out on the right side of the chart when compared to the left hand side of the chart: heteroscedasticity.



Stabilizing Transformations

Possible transformation to help eliminate or reduce heteroscedasticity are:

$\sqrt{y}$ or $\ln(y)$

Comment on Transformation of Data

There are many possible transformation of the form Y^k where k can be any real exponent.

Examples

$$y^{\frac{1}{2}} = \sqrt{y}; y^2; y^{-2} = \frac{1}{y^2}; y^{-1} = \frac{1}{y}; y^{\frac{1}{3}}; y^{-\frac{2}{3}}; etc.$$

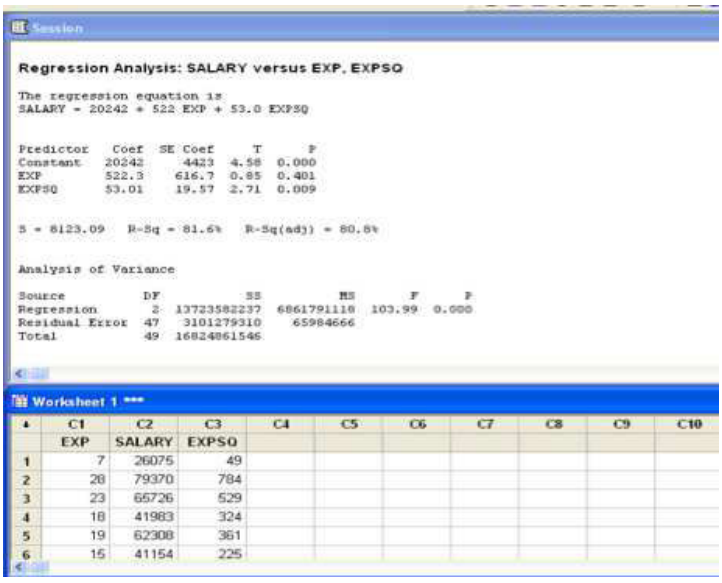
Example

Salary vs. Years of Experience for a random sample of 50 social workers.

Exp (X)	Salary (Y)
7	26075
28	79370
23	65726
18	41983
19	62308
15	41154
.	.

Second Order Model

Minitab Output: Second Order Model



Session

Regression Analysis: SALARY versus EXP, EXPSQ

The regression equation is
SALARY = 20242 + 522 EXP + 53.0 EXPSQ

Predictor	Coef	SE Coef	T	P
Constant	20242	4423	4.58	0.000
EXP	522.3	616.7	0.85	0.401
EXPSQ	53.01	19.57	2.71	0.009

S = 8123.09 R-Sq = 81.6% R-Sq(adj) = 80.8%

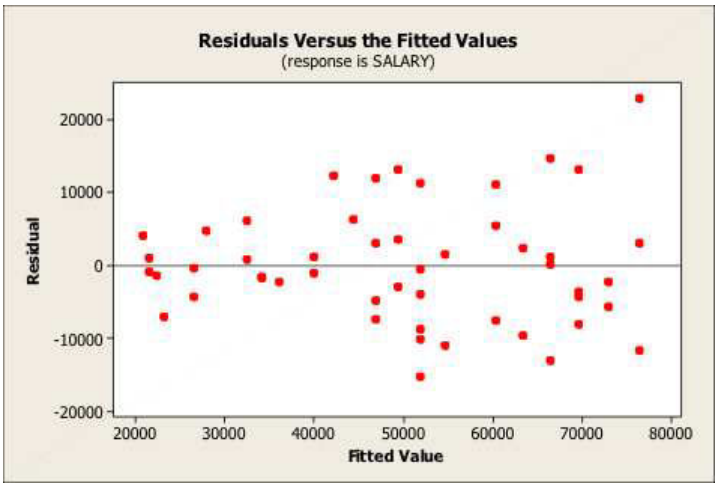
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	13723582237	6861791118	103.99	0.000
Residual Error	47	3101279310	65984666		
Total	49	16824861546			

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	EXP	SALARY	EXPSQ							
1	7	26075	49							
2	28	79370	784							
3	23	65726	529							
4	18	41983	324							
5	19	62308	361							
6	15	41154	225							

Residual Plot: Evidence of Heteroscedasticity



Summary of Second Order Model

Regression equation: SALARY = 20242 + 522EXP + 53.0EXPSQ

$$R^2 = 81.6\%$$

But the residual plot shows signs of heteroscedasticity.

Second Order Model with Logarithmic Transformation

Minitab Output: Model Including Logarithmic Transformation and Quadratic Term

Regression Analysis: Salary versus EXP

Method

Box-Cox transformation λ = 0

Analysis of Variance for Transformed Response

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	7.21216	3.60608	148.67	0.000
EXP	1	0.42844	0.42844	17.66	0.000
EXP*EXP	1	0.00002	0.00002	0.00	0.980
Error	47	1.14004	0.02426		
Lack-of-Fit	19	0.46015	0.02422	1.00	0.492
Pure Error	28	0.67989	0.02428		
Total	49	8.35220			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.155744	86.35%	85.77%	84.31%

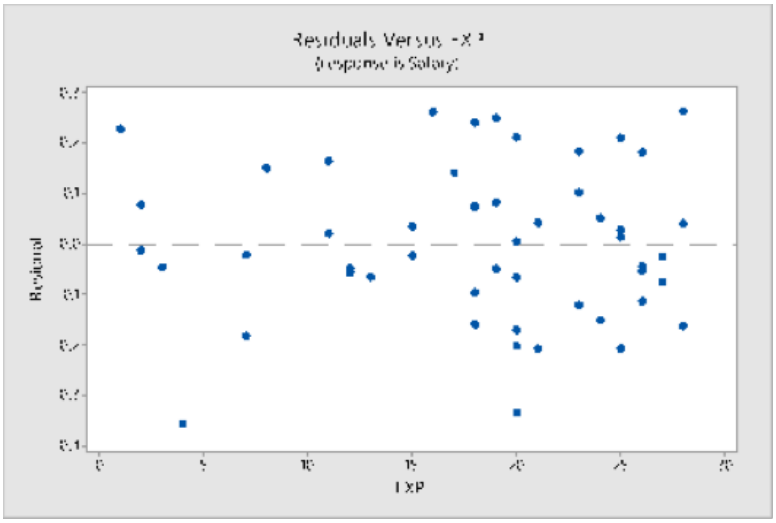
Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9.8429	0.0848	116.08	0.000	
EXP	0.0497	0.0118	4.20	0.000	16.64
EXP*EXP	0.000009	0.000375	0.03	0.980	16.64

Regression Equation

ln(Salary) = 9.8429 + 0.0497 EXP + 0.000009 EXP*EXP

Residual Plot: No Evidence of Heteroscedasticity



Summary of Second Order Model ln(Salary) vs. EXP and EXPSQ

Regression equation: ln(SALARY) = 9.84 + 0.0497EXP + 0.000009EXPSQ

$$R^2 = 86.4\%$$
$$R_a^2 = 85.8\%$$

The residual plot shows that the log transformation has significantly reduce heteroscedasticity.

But the coefficient of EXPSQ is not significant (p-value = 0.98).

First Order Model with Logarithmic Transformation

Minitab Output

Regression Analysis: Salary versus EXP

Method

Box-Cox transformation λ = 0

Analysis of Variance for Transformed Response

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7.2121	7.21215	303.65	0.000
EXP	1	7.2121	7.21215	303.65	0.000
Error	48	1.1401	0.02375		
Lack-of-Fit	20	0.4602	0.02301	0.95	0.542
Pure Error	28	0.6799	0.02428		
Total	49	8.3522			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.154114	86.35%	86.07%	85.09%

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9.8413	0.0564	174.63	0.000	
EXP	0.04998	0.00287	17.43	0.000	1.00

Regression Equation

ln(Salary) = 9.8413 + 0.04998 EXP

Note: Same R^2 value and higher R_a^2 value.

Interpreting the Model

$$\ln(\hat{y}) = 9.84 + 0.05x$$

$$\hat{y} = e^{9.84+0.05x} = e^{9.84}e^{0.05x} = 18769.72e^{0.05x}$$

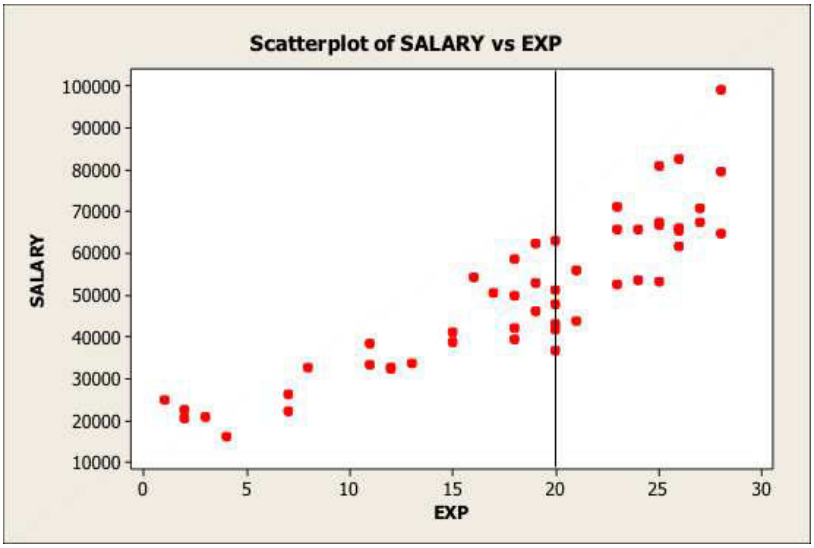
Experience	Predicted Salary
0	$18769.72
5	$24100.80
10	$30946.04

15	$39735.50
20	$51021.39

A Test for Heteroscedasticity

Divide the sample observations based on the values of $\hat{y}$ or equivalently, in this example, the value of x (since for the fitted model $\hat{y}$ increases as x increases).

Examination of the data shows that approximately one-half of the 50 observations fall below $x = 20$.



Testing for Equal Variances

We next calculate the variances of the observations in subgroups 1 and 2 and perform a test of hypothesis for the ratio of the variances.

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Subgroups 1 and 2

EXP1	SAL1	EXP2	SAL2
1	24833	20	43076
2	22444	20	63022
2	20521	20	47780
3	20844	20	51301
4	16105	20	41721
7	26075	20	36530
7	22210	21	56000
8	32562	21	43628
11	33233	23	65726
11	38371	23	52624
12	32586	23	71235
12	32303	24	53610
13	33697	24	65644
15	41154	25	66537
15	38853	25	67477
16	54288	25	53272
17	50594	25	80931
18	41983	26	61581
18	58667	26	65929
18	49727	26	82641
18	39346	26	65343
19	62308	27	70678
19	46216	27	67282
19	52745	28	79370
		28	64785
		28	99139

Calculating s_1^2 and s_2^2 for SAL vs. EXP, EXPSQ

$$MSE_1 = s_1^2$$

$$MSE_2 = s_2^2$$

Testing the Hypothesis of Equal Variances

$$H_0: \frac{\sigma_{larger}^2}{\sigma_{smaller}^2} = 1$$

$$H_1: \frac{\sigma_{larger}^2}{\sigma_{smaller}^2} \neq 1$$

$$TS: F = \frac{s_{larger}^2}{s_{smaller}^2} = \frac{MSE_{larger}}{MSE_{smaller}} = \frac{94711023}{31576998} = 2.99937$$

$$CV: F_{0.025,23,21} \cong 2.37$$

Accept H_0 if $F \leq 2.37$

Reject H_0 if $F < 2.37$

Since $F = 2.99937$ is greater than 2.37 we reject the hypothesis of equal variances. Therefore, the quadratic model has residuals that exhibit heteroscedasticity.

Checking the Normality Assumption

Important note

Moderate departures from the normality assumption will generally not invalidate the results of a regression analysis. We can say that regression analysis is robust with regard to the normality assumption.

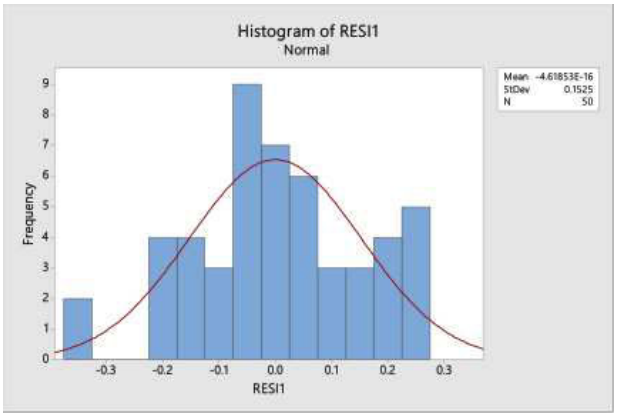
If a graphical display of the data (stem-and-leaf plot, histogram, etc.) is not badly skewed, and has one major central peak, we can be confident in using the model.

Checking for Normality

We will use the model:

$$\ln(Salary) = 9.84 + 0.05Exp$$

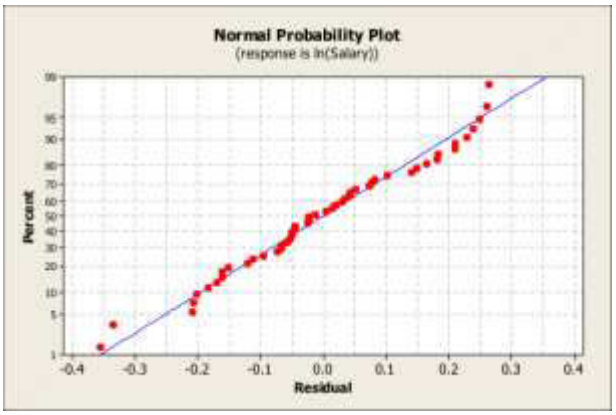
The histogram of the residuals of the residuals shows that the distribution is mound-shaped and reasonable symmetric. Therefore, we suspect that the normality assumption is satisfied. However, this claim is subjective so we need a more formal approach.

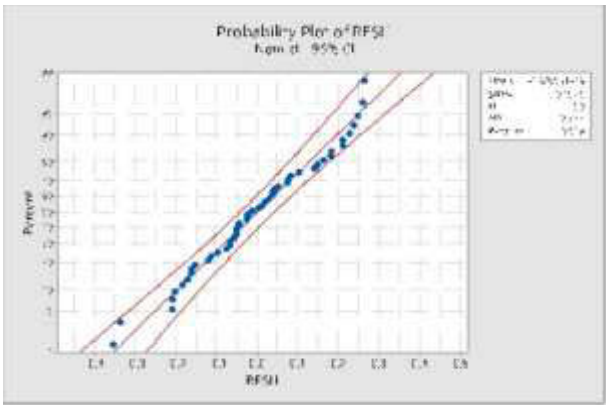


Normal Probability Plot

The normal probability plot graphs the residuals against the expected values of the residuals under the assumption of normality.

If the assumption of normality is true then a residual value should approximately equal its expected value, resulting in a straight line graph.





The Anderson-Darling Statistic

The AD statistic is used to test the hypothesis.

H_0: Distribution is normal
H_1: Distribution is not normal

If the p-value for the AD is ≥ 0.05, there is no reason to conclude that the distribution is not approximately normal.

Conclusion

This graph shows that the points lie within 95% confidence limits of the straight line that represents a normal distribution.

In addition, the AD statistic has a p-value of 0.519.

Therefore, we are confident that the distribution is reasonably well approximated by a normal distribution.

Standardized Residuals

The standardized residual, denoted z_i, for the ith observation is the residual for the observation e_i divided by the standard error of the estimate s.

$$z_i = \frac{e_i}{s} = \frac{y_i - \hat{y}_i}{\sqrt{MSE}}$$

Example of Outliers

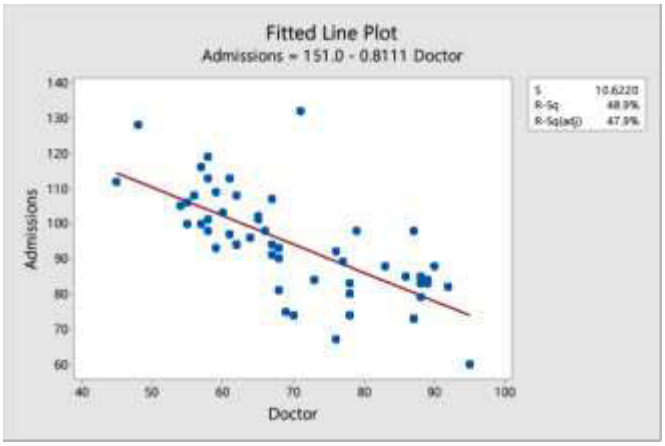
Hospital Admissions vs. Number of Doctors

↓	C1	C2
	Admissions	Doctor
1	80.000	78
2	93.000	68
3	75.000	69
4	89.000	77
5	102.000	65
6	83.000	89
7	88.000	83
8	88.000	90
9	107.000	67
10	132.000	71
11	85.000	86
12	83.000	88

These data show the number of weekly admissions to city hospitals vs. the number of doctors per thousand of population in the city.

Sample size is 53.

Scatter diagram of Admissions vs. Doctors



Interpreting the regression

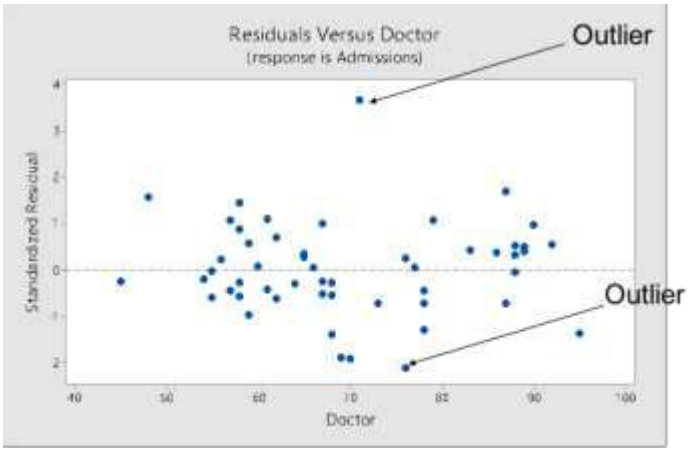
The regression equation is:

$$Admissions = 151.0 - 0.8111 Doctors$$

The negative slope implies that as the number of doctors increases, the number of admission decreases. For each extra doctor, the number of admissions decreases by 0.8111 so that , for example, 10 extra doctors would result in approximately 8 few admissions.

This may be explained by the fact that if patients have a family doctor they may be diagnosed and cured without requiring a hospital admission.

Plot of standardized residuals vs. Doctors



Residuals and standardized residuals are in C3 and C4

	C1	C2	C3	C4
	Admissions	Doctor	RESI1	SRES1
1	80.000	78	-7.7647	-0.74107
2	93.000	68	-2.8762	-0.27342
3	75.000	69	-20.0650	-1.90715
4	89.000	77	0.4241	0.04044
5	102.000	65	3.6904	0.35124
6	83.000	89	4.1578	0.40441
7	88.000	83	4.2910	0.41226
8	88.000	90	9.9690	0.97207
9	107.000	67	10.3127	0.98062
10	132.000	71	38.5573	3.66502
11	85.000	86	3.7244	0.35982
12	83.000	88	3.3467	0.32474

Minitab identifies outliers

Fits and Diagnostics for Unusual Observations

Obs	Admissions	Fit	Resid	Std Resid	
10	132.00	93.44	38.56	3.67	R
53	67.00	89.39	-22.39	-2.13	R

R Large residual

The outliers

Observation 10 has a residual of 38.56 and a standardized residual of 3.665.

This means that value of the observation (132) is 38.56 units above the value predicted by the fitted regression line. It is hard to interpret the value 38.56 because it depends on the unit of measurement. However, the standardized residual is 3.665 indicating that is 3.665 standard deviations above the regression line.

Minitab tags it as an unusual observation because it is more than 2 standard deviations away from the regression line.

Another outlier

Observation 53 has a negative residual of -22.39. Again, it is hard to interpret this value, but we see that the standardized residual is -2.13.

Since it is 2.13 standard deviations below the regression line, it is tagged as an outlier.

Outliers

An **outlier** is often defined to be an observation with a residual greater than 3s, or equivalently, with a standardized residual that is greater than 3.

Minitab takes a conservative approach and defines an outlier to be an observation with a residual that is greater than 2s or a standardized residual greater than 2.

Outliers are often caused by an invalid measurement.

Residual Analysis II

More on outliers

As we saw, an outlier may be due to an error in measurement of data entry.

When an outlier is **not** due to an error but represents an accurate value, we have to investigate the following possible causes in the model itself such as:

- Omission of important variables in a model

- Omission of higher-order terms

If the outlier represents a real and plausible value in your data set (e.g. the annual income of a multimillionaire in a sample of people who work in the computer industry) it is necessary to decide how to handle the data. One possibility is to exclude the data from the statistical analysis and write an exception report to explain the absence of the extreme value from the data set.

Influential Observations and Leverage

Influential observation: One whose removal would substantially affect the regression equation.

Leverage: A measure of how influential an observation is: the larger the leverage value the more influence the observed y value has on its predicted value.

Computing leverage

In regression analysis it is known that the predicted value for the ith observation, $\hat{y}_i$, can be written as a linear combination of the n observed values $y_1, y_2, \dots, y_n$.

Thus, for each value y_i, $i = 1$ to n, we have the equation:

$$\hat{y_i} = \mathrm{h}_1 y_1 + \mathrm{h}_2 y_2 + \mathrm{h}_3 y_3 + \cdots + \mathrm{h}_i y_i + \cdots + \mathrm{h}_n y_n$$

For $i = 1,2, \dots, n$

Note: $0 < \mathrm{h}_i$ and $\sum \mathrm{h}_i = k + 1$

The coefficient h_i measures influence of the observed y_i value on its own predicted value $\hat{y}_i$.

This value h_i is called the leverage of the ith observation.

Minitab says…

Leverage values provide information about whether an observation has unusual predictor values compared to the rest of the data.

Leverages are a measure of the distance between the x-values for an observation and the mean of x-values for all observations. A large leverage value indicates that the x-value of an observation are far from the center of x-values for all observations.

Observations with large leverage may exert considerable influence on the fitted value, and thus the regression model.

Detecting influence with leverage

Leverage values fall between 0 and 1.

According to Minitab a leverage value greater than $\frac{3(k+1)}{n}$, where k is the number of predictors (independent variables) and n is the number of observations, is considered large and should be examined. Most other sources compare to $\frac{2(k+1)}{n}$.

Minitab identifies observations with leverage over $\frac{3(k+1)}{n}$ or 0.99, whichever is smaller, with an X in the table of unusual observations.