MSCA 602: Applied Linear Statistical Models

The course deals with systematic treatments of linear statistical models for regression, analysis of variance and experimental design, elements of logistic regression and time series analysis with special focus on the application in business research.

Topics covered

- Linear regression analysis - inference, diagnostics and remedial measures; model building techniques
- Time series regression - Modeling trend and seasonal variations; detection and remedial methods for autocorrelation; forecasting with autocorrelated errors
- Logistic regression: inference and diagnostics
- Analysis of variance models and elements of experimental designs

LINEAR MODELS

- **Content**
  The general heading covers areas such as regression analysis, the analysis of variance, multivariate analysis and time series analysis. In this course we consider only the first two major subsets.

- **Significance**
  Major portion of statistics that is used in applied problems is directly or peripherally related to linear statistical models.

Regression Analysis

Regression analysis – used as a tool for evaluating the relationships of one or more independent variables $X_1, X_2, ..., X_K$ $(K \geq 1)$, to a single continuous dependent variable Y.

Objectives

- Understand the fundamental ideas behind regression analysis including
  a. the importance and the scope of regression modeling of data
  b. identify situations when use of linear regression modeling is appropriate
  c. build correct linear regression models in particular the modeling of business data that are inherently linear in structure

Practical Applications – for which regression analysis would be appropriate

a. Characterize the relationship between the independent variables, $X_1, X_2, ..., X_K$ $(K \geq 1)$, and the dependent variable Y – determine the extent, direction, and strength of association among these variables.
b. Providing a quantitative formula or equation to describe (e.g. predict) the dependent variable Y as a function of the independent variables $X_1, X_2, ..., X_K$.
c. Describe quantitatively or qualitatively the relationships between $X_1, X_2, ..., X_K$ and Y, controlling for effects of other variables $Z_1, Z_2, ..., Z_L$, which are assumed to have important relationship with the dependent variable Y.
d. Provide means for variable selection among the several independent variables $X_1, X_2, ..., X_K$.
e. Provide the best interpretive mathematical model for describing the relationship between a dependent variable and one or more independent variables.
f. Provide means to compare several derived regression relations.
g. Assess interactive effects of $K \geq 2$ variables
h. etc. …

SIMPLE LINEAR REGRESSION

Random sample of n-pairs (independent) of observations on variables X and Y:

$$(X_1, Y_1), (X_2, Y_2), \ \ldots\ , (X_n, Y_n)$$

Objective

To develop a regression model where the values of the independent variable X are used to estimate or predict the values of the dependent variable Y.

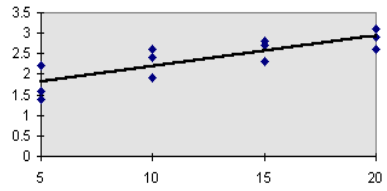
Scatter Diagram

Scatter diagram is a plot of Y versus X, used to see or identify the underlying relationship between X and Y.

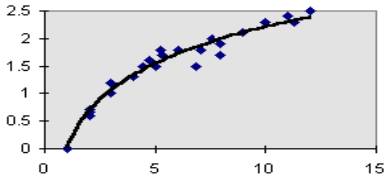
Types of scatter plots

- Positive linear (weak, strong, perfect)
- Negative linear (weak, strong, perfect)
- Nonlinear (quadratic, cubic, curvilinear, etc)
- No relationship

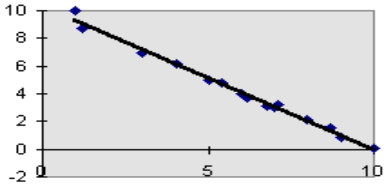
Positive Linear Relationship



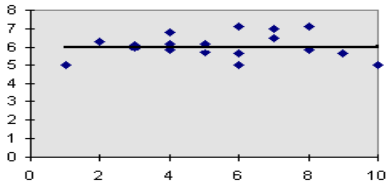
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Types of regression models

- Functionally related models

  $y_i = f(x_i; \varepsilon_i) = g(x_i) + \varepsilon_i$, (assuming *f* is additively separable);

  g(x) is linear function in parameters (not necessarily in *X*);

  ε is a random factor with a zero mean and a finite variance;

  Thus, y may be written as

  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1(1)n,$$

  such that

  Y_i *= i-th observation on observable random variable Y;*

  X_i *= mathematical variable assumed fixed in a given range from sample to sample;*

  ε_ι = unobservable random term reflecting errors and unexplained variations excluded from X.

- Conditional regression models

  The form of the model is the same;

  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1(1)n.$$

  However, here both X and Y are random variables that take their values from observed data. Inferences concerning the distribution of Y are made conditional on X taking values equal to observed value x_i, or other values in the same general range.

  Remarks:

  - Assumptions about ε_i and β_0, β_1 are the same for both types of models.
  - Both models are called error-in-equation models.

- Error-in-variables models

  Here neither Y nor X is directly observable without error:

  Let $V_i = \beta_0 + \beta_1 U_i$, $V_i = Y_i + \eta_i$ and $U_i = X_i + \tau_i$, then

  $$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

  where $\varepsilon_i = \eta_i + \beta_1 \tau_i$.

  This is a difficult model to handle since there are too many unknown parameters to estimate for a given data set.

- **In this course, we deal with the error-in-equation models.**

Simple Linear regression (SLR) Population Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1(1)n,$$

wehere

Y_i *= i-th observation on an observable random variable Y;*

X_i *= mathematical variable assumed fixed in a given range from sample to sample;*

ε_i = unobservable random term reflecting errors and unexplained variations excluded from X.

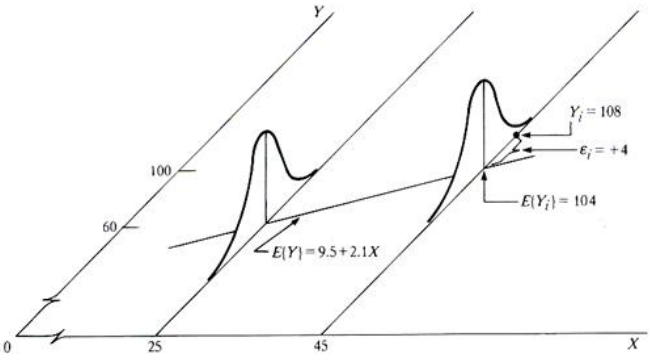
Assumptions of SLR Model

- Existence
  For any fixed values of the independent variable X, the dependent variable Y is a random variable with certain probability distribution. We denote the mean by of Y for a given X by $E(Y \mid X) = \mu_{y|x}$ and its variance by $V(Y \mid X) = \sigma^2_{y|x} = \sigma^2$. It is usual to assume this conditional probability distribution to be Normal. Validity of this assumption is verified as part of routine procedure in regression analysis.

- Independence
  Values of the dependent variable Y for given values of X are statistically independent of one another.

- Linearity
  The model is linear in the parameters, not necessarily in the variables.

- Homoscedasticity (constant variance)
  The variation about the regression line, $\sigma^2_{y|x} = \sigma^2$, is assumed constant under changes in the values of independent variable X.



Source: KNNL

Inference on a Regression Function

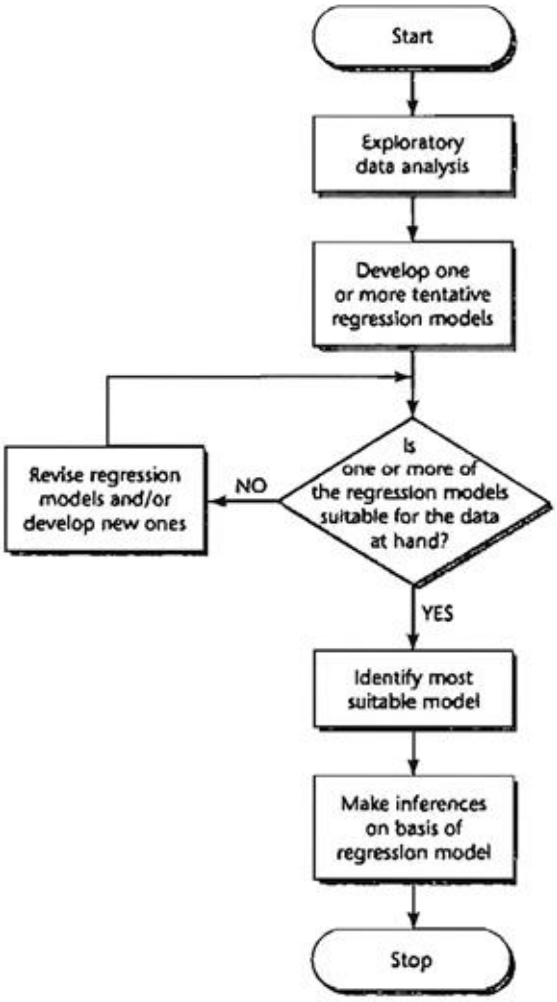
Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Mean of Y for given X: $E(Y \mid X) = \mu_{y|x} = y_i = \beta_0 + \beta_1 x_i$

Variance of Y given X: $V(Y \mid X) = \sigma^2_{y|x} = \sigma^2$

Estimation of the unknown model involves estimating the parameters of the model: β_0, β_1, and $\sigma^2_{y|x} = \sigma^2$.

- **Strategy for Regression Analysis**



Source: KNNL

- **Method of Estimation:** Least Squares Principle

When the distribution of the random error term is not implied, we need to use the method of least squares to get the "best" fitting line to the data.

Let the fitted values of the model be represented by

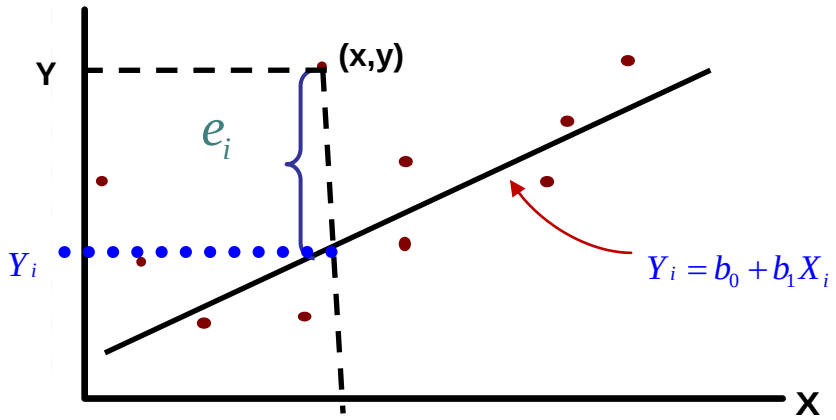
$$\hat{y}_i = b_0 + b_1 x_i .$$

The difference between the actual Y_i and the fit $\hat{Y}_i$ is referred to as the residual (e_i):

$$e_i = Y_i - \hat{Y}_i$$

The Least Squares method states that the "best" fit for the data is obtained by minimizing the sum of squared residuals about the fitted line with respect to b_0 and b_1:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 .$$



The solutions for b_0 and b_1 are obtained by solving the normal equations:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Example: The Market model

The market model assumes that the rate of return on stock is linearly related to the rate of return on the overall market. The model is given as

$$R_i = \beta_0 + \beta_1 R_{mi} + \varepsilon_i,$$

where R is the rate of return on a particular stock (RESPONSE), and R_m is the rate of return on some major stock index (PREDICTOR), such as the TSE or NYSE composite index.

- The coefficient β_1 is the stock's beta coefficient, measuring the sensitivity of the stock's rate of return to changes in the level of the overall market.
- If $\beta_1 > 1,$ the stock's rate of return is more sensitive to changes in the level of the overall market than is the average stock.
- If $b_1 = 2,$ then a 1% increase in the index results in an average increase of 2% in the stock's return. A 1% decrease in the index produces an average 2% decrease in the stock's return. A stock with $\beta_1 > 1$ will tend to be more volatile than the market.
- The coefficient of determination is also useful in measuring the proportion of the total risk that is market related.

• Properties of the sample regression line

1. $\sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} e_i^2$ is at a minimum
3. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
4. $\sum_{i=1}^{n} x_i e_i = 0$
5. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$
6. The point $(\bar{X}, \bar{Y})$ always lies on the regression line

- **The Standard Error of Estimate (Point estimator of $\sigma_{y|x} = \sigma$)**

  - Does the fitted line help in predicting Y? If yes, to what extent?
  - A measure that helps to answer these questions is provided by Residual Sum of Squares or Error Sum of Squares (SSE) given as:

    $$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \quad \text{where } \hat{Y}_i = b_0 + b_1 X_i .$$

    Clearly, if the SSE = 0, the line fits perfectly: $Y_i = \hat{Y}_i$, for each i, and every observed point lies on the line. As fit gets worse, SSE gets large.

  - Two possible factors contribute to an increase in SSE:
    - Presence of large variation in data: $\sigma_{y|x} = \sigma$ may be large;
    - Wrong model: the assumption of straight line model may be inappropriate.

  - Estimate of $\sigma^2_{y|x} = \sigma^2$ **:**

    - SSE is based on the n residuals. However, it has only (n-2) degrees of freedom associated with it, due to the loss of the 2 degrees of freedom in estimating β_0 and β_1 in order to get estimated means $\hat{Y}_i$.
    - The Unbiased Point Estimator of $\sigma^2_{y|x} = \sigma^2$ is Residual Mean Square or Mean Error Square, denoted by MSE or S^2:

      $$S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

  - The square root of MSE is called the standard deviation of the estimated regression equation or the standard error of estimate: $S = \sqrt{MSE}$.

  - A convenient computational formula for S:

    $$S = \sqrt{\frac{1}{n-2}\left(\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i\right)} = \sqrt{\frac{n-1}{n-2}(s_y^2 - b_1 s_x^2)} ,$$

    where s_y^2 and s_x^2 are sample variances of Y and X, respectively.
    Note that S is measured in units of the dependent variable Y.

Normal Error SLR Model

- The normal error regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1(1)n,$$

where in this case, the unobservable random term ε_i reflecting errors and unexplained variations excluded from X, is assumed to be independently and identically distributed as $N(0,\sigma^2)$.

- The additional assumptions in this model allow us to draw inferences from our data. This means that we can test our parameter estimates and build confidence intervals.

- Method of Estimation**:** Maximum Likelihood Estimation (MLE)
  MLE is used to obtain the estimators of the parameters β_0, β_1, and σ^2 when the functional form of the probability distribution of the error terms is specified. This method chooses as estimates those values of the parameters that are most consistent with the sample data. For more, see text.
  - Estimation of model parameters
  - Properties of MLE

Inference

- Does X help predict Y? If yes, construct confidence/prediction limits and/or conduct hypotheses tests about the regression parameters

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1(1)n.$

Fit: $\hat{y}_i = b_0 + b_1 x_i$.

- The sampling distribution of the estimated regression coefficients

The distributions of b_0 and b_1 are simply linear combinations of the Y_i. Since a linear combination of a normally distributed random variables is also normally distributed, the distributions of the estimated regression coefficients are:

$$b_0 \sim \mathrm{N}(\beta_0, \sigma^2_{b_0}) \text{ and } b_1 \sim \mathrm{N}(\beta_1, \sigma^2_{b_1}),$$

where the estimated variances of b_0 and b_1 easily obtained as

$$s^2_{b_0} = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s^2_x}\right) \text{ and } s^2_{b1} = \frac{s^2}{(n-1)s^2_x}$$

- Why conduct hypotheses tests for β_1?
  - Check the existence of linear relationship between Y and X
  - If it exists, study further by testing its magnitude and direction

- **Testing for the significance of β_1 (one-sided or two-sided tests)**

  1. $\begin{cases} H_0 : \beta_1 = \beta^* \\ H_a : \beta_1 \neq \beta^* \end{cases}$
  2. Test statistic: $t^* = \frac{b_1 - \beta_1^*}{S_{b_1}} \sim t_{n-2}$
  3. Test is found significant if $|t^*| > t_{1-\alpha/2;n-2}$
  4. p-value of the test = $2P(t_{n-2} > t^*)$

Example: Testing for volatility of Nortel's stocks than the TSE, the null and alternative hypotheses

1. $\begin{cases} H_0 : \beta_1 \leq 1 \\ H_a : \beta_1 > 1 \end{cases}$
2. Test statistic: $t^* = \frac{b_1 - 1}{S_{b_1}} \sim t_{n-2}$
3. Test is found significant if $t^* > t_{1-\alpha;n-2}$
4. p-value of the test = $P(t_{n-2} > t^*)$

- **Confidence interval estimator of β_1:**

  The 100(1-α)% confidence interval estimator of β_1:

$$b_1 \pm t_{1-\alpha/2;n-2} S_{b_1}$$

- **Inference for β_0:** Occasionally you may be interested in inference concerning β_0. For example, in a regression through the origin, you do not estimate β_0, but you assume it is theoretically equal to zero.

- **Inference on Mean of Y for given X, E(Y_h)**

The 100(1-α)% confidence interval estimator of the mean value of Y for given X= $X = X_h$:

$$\hat{Y}_h \pm t_{1-\alpha/2;n-2} S_{\hat{Y}_h}, \qquad \text{where } S^2_{\hat{Y}_h} = S^2\left(\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right)$$

- **Prediction of Y for given X**

The 100(1-α)% Prediction interval estimator of Y for given $X = X_h$:

$$\hat{Y}_{h(new)} \pm t_{1-\alpha/2;n-2} S_{Pred}, \quad \text{where } S^2_{Pred} = S^2\left(1 + \frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right)$$

Decomposition of Total Variation in Y

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

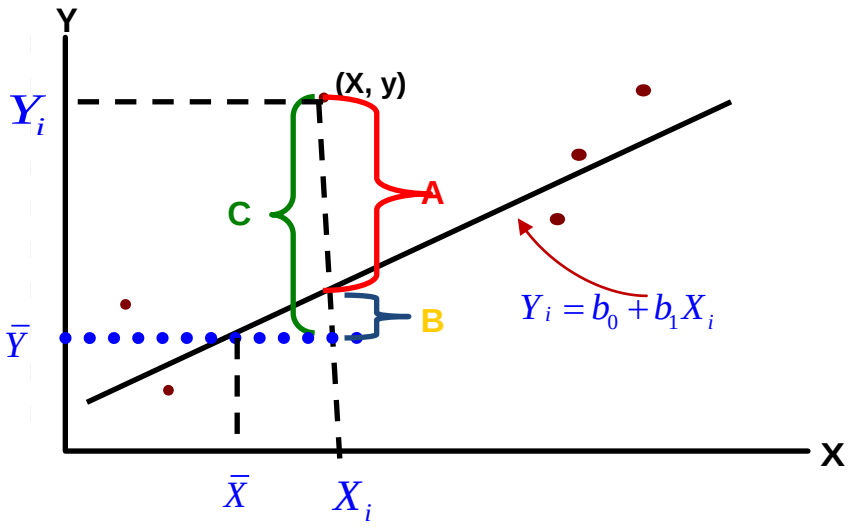
$$\begin{pmatrix} Total \\ Variation \end{pmatrix} = \begin{pmatrix} Explained \\ Variation \end{pmatrix} + \begin{pmatrix} Unexplained \\ Variation \end{pmatrix}$$

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

SSTO = SSR + SSE

SSR = Variation attributable to relationship between Y and X

SSE = Unexplained variation not accounted by the model



ANALYSIS of VARIANCE (ANOVA) APPROACH to SLR

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- Sum of Squares (SS)

SSTO = $\sum (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum Y_i)^2}{n}$

SSR = $\sum (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = b_1^2 \left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum X_i)^2}{n} \right)$

SSE = SSTO – SSR

- Mean Squares ($MS = \frac{SS}{df}$):

$MSR = \frac{SSR}{1}$, $MSE = \frac{SSE}{n-2}$

E(MSE) = σ^2, *E(MSR)* = $\sigma^2 + b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

- MSE is unbiased estimator of σ^2 whether or not $\beta_1 = 0$

- MSR is biased for σ^2 unless $\beta_1 = 0$:

- Overall test for a Regression (F –Test for testing the existence of SLR):

1. $\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$

2. $F^* = \frac{MSR}{MSE} \sim F_{1-\alpha;1,n-2}$

3. p-value of test = P($F_{1,n-2} > F^*$)

- **ANOVA Table**

Source	Sum of squares	Degrees of freedom	Mean square	F	p-value =P(F ≥ F*)
Regression	SSR	1	MSR	$F = \frac{MSR}{MSE}$	P($F_{1,n-2}$ > F*)
Error	SSE	n-2	MSE		
Total	SSTO	n-1	-	-	-

General Linear Model Approach

Full Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

SSE(F) = $\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_i)^2 = SSE$

Reduced Model under the null hypothesis $H_0 : \beta_1 = 0$

SSE(R) = $\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0)^2 = \sum (Y_i - \bar{Y})^2 = SSTO$

SSE(F) with df_F = (n-2) ≤ SSE(R) with df_R = (n-1)

Test statistic:

$$F^* = \frac{(SSE(R) - SSE(F)]/(df_R - df_F)}{SSE(F)/df_F} \sim F_{1-\alpha; df_R - df_F, df_F}$$

Coefficient of Determination

Measures the proportion of variation in Y that is explained by the independent variable X in the regression model

$$r^2 = \frac{SSR}{SST} = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

$$0 \le r^2 \le 1$$

- The closer r^2 is to 1, the stronger the linear association is between X and Y.

Correlation Coefficient

Measures the strength of an association (linear relationship) between two numerical variables.

$$r = \sqrt{r^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- Only concerned with strength of the relationship
- No causal effect is implied
- Inference on population correlation coefficient ***(ρ):***

i. Hypotheses
H_0: $\rho = 0$ (no correlation) versus H_1: $\rho \neq 0$ (correlation)

ii. Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}, \quad \text{or} \quad F = \frac{r^2}{\frac{1-r^2}{n-2}} \sim F_{1,\, n-2}$$

Fitting a Linear Trend Model to Time-Series Data

- Data gathered on different units at the same point in time are called cross sectional data.

- Data gathered on a single unit (person, firm, etc.) over a sequence of time periods are called time-series data. With this type of data, the primary goal is often building a model that can forecast the future. There are many types of models that attempt to identify patterns of behavior in a time series in order to extrapolate it into the future. Some of these will be examined in Chapter 11, but here we will just employ a simple linear trend model.

  - **The Linear Trend Model**

    - We assume the series displays a steady upward or downward behavior over time that can be described by:

      $$y_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

      where t is the time index ($t = 1$ for the first observation, $t = 2$ for the second, and so forth).

    - The forecast for this model is quite simple: $\hat{y}_T = b_0 + b_1 T.$

Assessing Quality of Prediction

- We use the model's R^2 as a measure of fit ability, but this may overestimate the model's ability to predict.
- The reason for that is that R^2 is optimized by the least squares procedure, for the data in our sample. It is not necessarily optimal for data outside our sample, which is what we are predicting.

Data Splitting

- We can split the data into two pieces. Use the first part to obtain the equation and use it to predict the data in the second part.

- By comparing the actual *y* values in the second part to their corresponding predicted values, you get an idea of how well you predict data that is not in the "fit" sample.
- The biggest drawback to this is that it won't work too well unless we have a lot of data. To be really reliable we should have at least 25 to 30 observations in both samples.

The PRESS Statistic

- Suppose you temporarily deleted observation *i* from the data set, fit a new equation, then used it to predict the y_i value.
- Because the new equation did not use any information from this data point, we get a clearer picture of the model's ability to predict it. The sum of these squared prediction errors is the PRESS statistic (PRESS residuals). PRESS stands for prediction sum of squares defined as

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-1})^2,$$

where $\hat{y}_{i,-1}$ is the prediction from a model estimated with the i-th sample observation deleted.

- You can then compute an R^2- like measure called the prediction R^2:

$$R^2_{PRED} = 1 - \frac{PRESS}{SST}$$

Some Cautions in Interpreting Regression Results

Two common mistakes that are made when using regression analysis are:

1. Association versus Causality

- If you have a model with a high R^2, it does not automatically mean that a change in *x* causes *y* to change in a very predictable way.
- It could be just the opposite, that *y* causes *x* to change. A high correlation goes both ways.
- It could also be that both *y* and *x* are changing in response to a third variable that we don't know about.
- The Third Factor:
  - One example of this third factor is the price and gasoline mileage of automobiles. As price increases, there is a sharp drop in mpg. This is caused by size. Larger cars cost more and get less mileage.
  - Another is mortality rate in a country versus percentage of homes with television. As TV ownership increases, mortality rate drops. This is probably due to better economic conditions improving quality of life and simultaneously allowing for greater ownership.

2. Forecasting Outside the Range of the Explanatory Variable

- When we have a model with a high R^2, it means we know a good deal about the relationship of *y* and *x* for the range of *x* values in our study.
- Think of the Nortel RoR example where the TSE RoR ranged from -0.30 to 0.15. Does our model even hold if we wanted to forecast NTRoR for significantly higher TSERoR of 0.40?
- An Extrapolation Penalty:
  Recall that our prediction intervals were always narrowest when we predicted right in the middle of our data set. As we go farther and farther outside the range of our data, the interval gets wider and wider, implying we know less and less about what is going on.

Assessing the Assumptions: Diagnostics and Remedial Measures

- Model Assumptions

For the SLR model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i,$$

certain assumptions were made about how the errors ε_i behaved. We will now check to see if those assumptions appear reasonable.

(a) We expect the average disturbance ε_i to be zero so the regression line passes through the average value of y.

(b) The disturbances have constant variance σ^2.

(c) The disturbances are normally distributed.

(d) The disturbances are independent.

- The Regression Residuals

  - We cannot check to see if the disturbances ε_i behave correctly because they are unknown. Instead, we work with their sample counterpart, the residuals $e_i = (y_i - \hat{y}_i)$, representing the unexplained variation in the y values.

  - Properties

a. Property 1: The residuals will always average 0 because the least squares estimation procedure makes that happen.

b. Property 2: If assumptions (a), (b) and (d) are true, then the residuals should be randomly distributed around their mean of 0. There should be no systematic pattern in a residual plot.

c. Property 3: If assumptions (a) through (d) hold, the residuals should look like a random sample from a normal distribution, with

$$mean\ \ of\ \ residuals:\ \bar{e}\ =\ 0,$$

$$Stdev\ of\ \ residuals:\ \ S = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE} = Root\ MSE$$

$$Note:\ \ e_i \sim N(0,\ (n-2)\sigma^2)\ and\ e_i\ are\ not\ independent$$

- Standardized residuals or Semi-studentized Residuals

  - The residuals can be standardized by dividing by their standard error:

$$\text{Standardized } residuals\ =\ e_{si}\ =\ \frac{e_i}{\sqrt{MSE}}.$$

$$Mean = \bar{e}_s = 0, \qquad \text{Stdev} = S_{e_s} = 1.$$

    This will not change the pattern in a plot but will affect the vertical scale.

  - Standardized residuals also called 'semistudentized residuals', are scaled so that most are between -2 and +2 as in a standard normal distribution. Consequently, e_{si} are often examined rather than e_i. The only difference is in scale. Later we will look at refinements such as studentized, deleted, and studentized deleted residuals.

  - The study of the patterns of residuals can give us important information about the validity of the regression assumptions as well as model adequacy. Use residuals for examining the departures from SLR model with normal errors which may suggest:

    1. The regression model is not linear
    2. The error terms do not have constant variance
    3. The error terms are not independent
    4. There are outliers in the data affecting the fit of the model
    5. The error terms are not normally distributed
    6. One or more important predictor variables have been omitted from the model

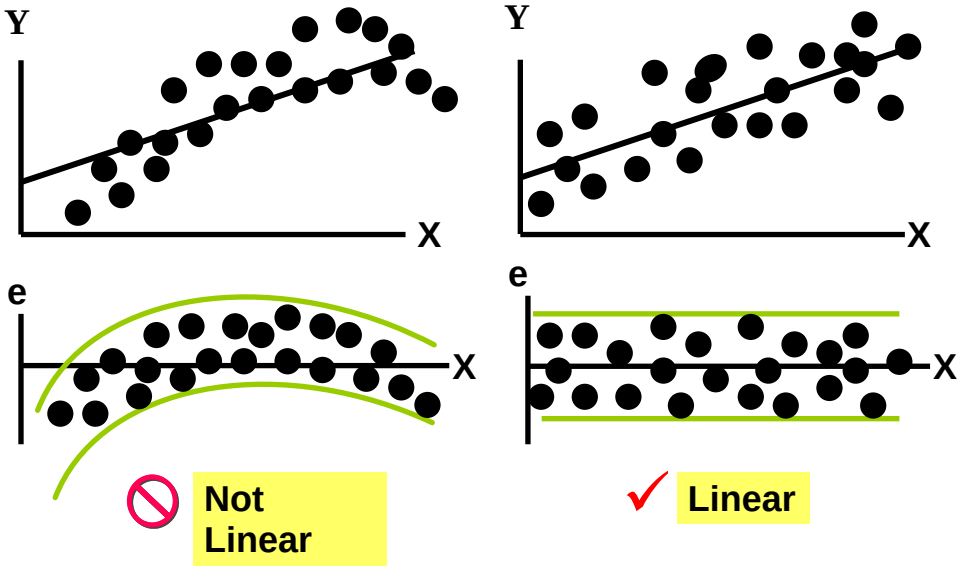
- Residual Diagnostics: Suggested Residual Plots

1. Plot the residuals versus the predictor variable
2. Plot the residuals versus the predicted or fitted values
3. Plot the absolute or squared residuals versus the predictor variable
4. For data collected over time or in any other sequence, plot the residuals versus that sequence.
5. Plot the residuals versus omitted predictor variables
6. Histogram, stem plots and box plots of residuals
7. Normal or half-normal probability plots of residuals

- Checking Linearity

- Although sometimes one can see evidence of nonlinearity in an X-Y scatterplot, in other cases we can only see it in a plot of the residuals versus X.
- If the plot of the residuals versus an X shows any kind of pattern, it both shows a violation and a way to improve the model.

Prototype Residual Plots



Corrections for nonlinearity

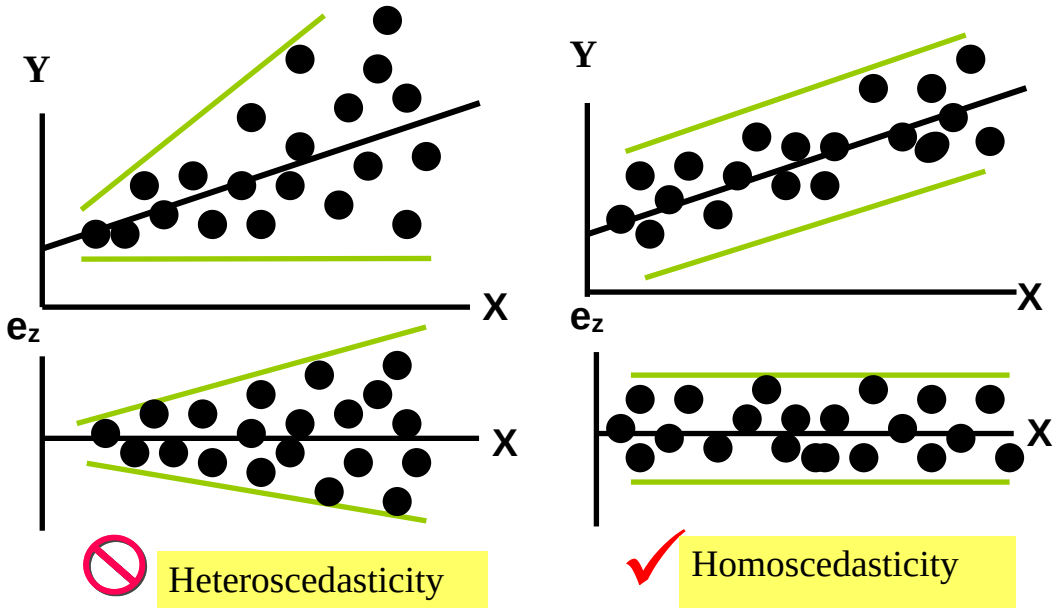
- If the linearity assumption is violated, the appropriate correction is not always obvious. Alternative models must be explored.

Nonconstancy of Error Variance or Heteroscedasticity

Checking for Constancy of Error Variance

- The assumption $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2$, *for all* $i = 1(1)n$ states that the errors ε_i should have the same variance everywhere. This implies that if residuals are plotted against an explanatory variable, the scatter should be the same at each value of the X variable.
- In economic data, however, it is fairly common to see that a variable that increases in value often will also increase in scatter.

Residual Analysis for Homoscedasticity



Remarks

- When the errors ε_i do not have a constant variance, the usual statistical properties of the least squares estimates may not hold.
- In particular, the hypothesis tests on the model may provide misleading results.

Tests for Normality

There are many tools available to check the assumption that the disturbances are normally distributed.

- If the assumption holds, for reasonably large number of cases, the standardized residuals should behave like they came from a standard normal distribution:
  - about 68% of the residuals e_i fall between $\pm\sqrt{MSE}$
  - about 95% of the residuals e_i fall between $\pm 2\sqrt{MSE}$
  - about 99% of the residuals e_i fall between $\pm 3\sqrt{MSE}$

- When sample size is moderately large, using the t distribution, it is expected under normality:
  - about 68% of the residuals e_i fall between $\pm t_{0.16,n-k-1}\sqrt{MSE}$
  - about 95% of the residuals e_i fall between $\pm t_{0.025,n-k-1}\sqrt{MSE}$
  - about 99% of the residuals e_i fall between $\pm t_{0.0.005,n-k-1}\sqrt{MSE}$

- Using Plots to Assess Normality:
  - You can plot the standardized residuals versus fitted values and count how many are beyond -2 and +2; about 1 in 20 would be the unusual case.

  - Use a Normal Probability plot to check for normality. A normal probability plot is a plot of ordered residuals against their expected values under normality. A plot that is approximately linear suggests agreement with normality. Departure from linearity is an indication of a nonnormal error distribution.

  - Use a histogram (perhaps with a superimposed normal curve) to look at shape. Dot plots or stem plots of the residuals based on large data set can be helpful for detecting gross departure from normality.

  - Use a Boxplot of the residuals for getting summary information about the symmetry of the residuals and about possible outliers.

Normal Probability plots

(probability-probability (P-P) and quantile-quantile (Q-Q) plot)

- Both probability-probability (P-P) plot and quantile-quantile (Q-Q) plot are used to see if a given set of data follows some specified distribution (in our case normal distribution). Both should be **approximately linear** if the specified distributions are the correct model.

- The probability-probability (P-P) plot is constructed using the theoretical cumulative distribution function, F(x), of the specified model. The values in the sample of data, in order from smallest to largest, are denoted $X_{(1)}$, $X_{(2)}$, ...,$X_{(n)}$. For i = 1, 2,…, n, $F(x_{(i)})$ is plotted against (i-0.5)/n.

- The quantile-quantile (Q-Q) plot is constructed using the theoretical cumulative distribution function, F(x), of the specified model. The values in the sample of data, in order from smallest to largest, are denoted $X_{(1)}$, $X_{(2)}$, ...,$X_{(n)}$. For i = 1, 2,…,n,. For i = 1, 2, ..., n, $X_{(i)}$ is plotted against $F_{-1}((i-0.5)/n)$.

Tests for Constancy of Error Variance

Breusch-Pagan Test

- This test is suitable for large sample
- Assumes the error terms to be independent and normally distributed.
- The variance of the error term ε_i, denoted by σ_i^2, is related to level of X as follows:

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 x_i,$$

suggesting that σ_i^2 increases or decreases with the level of X, depending on the sign of γ_1.

- Constancy of variance corresponds to γ_1=0.

- Hypothesis Test: $H_o : \gamma_1 = 0$ versus $H_o : \gamma_1 \neq 0,$ is carried out as follows:

1. Regress Y on X and save the residuals (e), and the error sum of squares (SSE).
2. Square the residuals (e^2).
3. Regress the squared residuals (e^2) on X, and obtain the regression sum of squares SSR^* from this regression.
4. The Breusch-Pagan (X^2_{BP}) test statistic is obtained as

$$X^2_{BP} = \frac{SSR^*}{2} \div \left(\frac{SSE}{n} \right)^2 \sim \chi^2_1$$

5. The BP test can be modified to allow for different relationships between the error variance and the level of X than the one given above.
6. The test is carried out exactly the same for multiple regression when the error variance increases or decreases with one or more predictor variables. The steps are the same except now, when q predictors are involved, the degrees of freedom for the Chi-square test becomes q=the number of predictors.
7. SAS also computes the BP statistic with p-values as an option under proc **model**.

Brown-Forsythe Test

- For the SLR model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i,$$

it was assumed that the disturbances have constant variance σ^2, *i.e.*, $\varepsilon_i \sim iid\ N(0, \sigma^2)$ for all $i=1(1)n$.

- The test is applicable if $\sigma^2 = Var(\varepsilon_i)$ changes with the values of the predictor variable. We will now test to see if this assumption is reasonable for a given data set based on variability of the residuals. Larger σ^2 imply larger variability in residual values.
- The Brown-Forsythe (BF) test does not depend on the errors being independent. However, when testing based on residuals, the sample size is required to be sufficiently large to discount for independence.
- The Brown-Forsythe (BF) test does not depend on the errors being Normal. However, the test is efficient if errors are Normally distributed or approximately so.

- Homogeneity of Variances Test (HOVTest) using the BF Method
  1. Divide the data into two groups according to the levels of X: Group1 and Group 2 will include, n_1 comparatively low values and n_2 comparatively high values, respectively, of the predictor variable. If $\sigma^2 = Var(\varepsilon_i)$ changes with values of the predictor, residuals in one group will tend to be more variable than those in the other group.
  2. For each group find the absolute deviations d_{ig} of the residuals around their group median:

$$d_{ig} = | e_{ig} - \tilde{e}_g |, \quad g = 1, 2.$$

Then the mean absolute deviations will be $\bar{d}_1 = \frac{\sum d_{i1}}{n_1}, \quad \bar{d}_2 = \frac{\sum d_{i2}}{n_1}.$

If $\sigma^2 = Var(\varepsilon_i)$ changes with values of the predictor, then $\bar{d}_1 \neq \bar{d}_2$.

  3. The BF test is a two-sample t test based on test statistic $\bar{d}_1$ *and* $\bar{d}_2$:

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2},$$

with the pooled variance $s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + (d_{i2} - \bar{d}_2)^2}{n_1 + n_2 - 2}.$

  4. Note that the Brown Forsythe test for just two groups using a t-test is equivalent to one using the F-test. It may be done in terms of the F-test as this automatically handles more than two groups.

5. The BF (or MODIFIED LEVENE TEST) IN SAS

   SAS has a Brown-Forsythe option in **PROC ANOVA** and also in **PROC GLM**.
   If you run an ANOVA of the residuals as the response and with **hovtest=bf** you get the same test that was developed by finding the deviations of residuals from the median and running a two-sample t test of the difference in mean absolute deviations.

6. To illustrate consider the Toluca Company data from the course text (p.19). The BF test using two-sample t is done pages116-118. We use here the PROC GLM and PROC ANOVA to do the test.

7. Both tables are testing the equality of variances of the residuals with Brown-Forsythe option with results similar to the two-sample t test given in the book.

Tests for lack of fit

- The residuals contain the variation in the sample of Y values that is not explained by the $\hat{y}$ equation.
- This variation can be attributed to many things, including:
  - natural variation (random error)
  - omitted explanatory variables
  - incorrect form of model

Lack of fit test: If nonlinearity is suspected, there are tests available for *lack of fit*.

- There are various versions of this test, one requiring there to be repeated observations at the same X values. There are options under proc reg for such tests.

The pure error lack of fit test

- In the 20 observations for the telemarketing data, there are two at 10, 20 and 22 months, and four at 25 months.
- These replicates allow the SSE to be decomposed into two portions, "pure error" and "lack of fit".

Full model: $y_{ij} = \mu_j + \varepsilon_{ij}$, with parameters μ_j (j=1(1)c), ε_{ij} ~N(o, σ^2).

The error sum of squares for the full model is

$SSE(F) = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2 = SSPE,$ with degrees of freedom of (n-c).

The pure error sum of squares (SSPE) is the full model SSE in the context of test for lack of fit.

$$\begin{cases} H_0: \textit{The relationship is linear} \\ H_a: \textit{The relationship is not linear} \end{cases} \Leftrightarrow \begin{cases} H_0: E(y) = \beta_0 + \beta_1 X \\ H_a: E(y) \neq \beta_0 + \beta_1 X \end{cases}$$

- The reduced model under the null gives the error sum of squares

  $SSE(R) = \sum \sum (y_{ij} - \hat{y}_{ij})^2 = SSE,$ with degrees of freedom of (n-2).

  The lack of fit sum of squares: SSLF= SSE – SSPE,
  with degrees of freedom of (n-2)-(n-c)=c-2

The test statistic: $F = \dfrac{SSLF/(c-2)}{SSPE/(n-c)} = \dfrac{MSLF}{MSPE} \sim F_{c-2,n-c}$,

c = number of distinct levels of X, n = 20 and there were 6 replicates so c = 14.

Remedial Measures

Transformations

- Normality is not necessary for making inference with large samples. It is required for inference with small samples.
- The remedies are similar to those used to correct for nonconstant variance.
- Commonly used forms as remedial measures to correct nonlinearity of regression functions, unequal error variance and skewness of the distribution of error are shown below.
- A transformation is appropriate provided the re-estimated model satisfies the assumptions of the model.

Transformation on response ***y***	Goal of transformation
$y^* = \log_e y, \quad y > 0,$	Stabilize variance (when $\sigma^2 = \sigma^2_{y\|x}$ increases with ***y*** for given ***x***); Normalize ***y*** given ***x*** (when the residuals are positively skewed)
$y^* = \sqrt{y}, \quad y > 0,$	Stabilize variance (when $\sigma^2 = \sigma^2_{y\|x} \approx E(y \mid x) = \mu_{y\|x}$)
$y^* = \frac{1}{y}, \quad y \neq 0,$	Stabilize variance (when $\sigma^2 = \sigma^2_{y\|x} \approx [E(y \mid x)]^4 = [\mu_{y\|x}]^4$)
$y^* = y^2$	Stabilize variance (when $\sigma^2 = \sigma^2_{y\|x}$ decresess with decreasing $\mu_{y\|x}$) Normalize ***y*** given ***x*** (when the residuals are negatively skewed)

Box-Cox: Generalization of the above transformations

$$y^*(\lambda) = \begin{cases} G^{1-\lambda}\left(\frac{y_i^\lambda - 1}{\lambda}\right), & \lambda \neq 0 \\ G(\log_e y_i), & \lambda = 0 \end{cases}$$

where $G = \left(\prod_{i=1}^{n} y_i\right)$, *is the geometric mean of the* y_i *observations.*

The Box-Cox transformation

- The Box-Cox family of power transformations has the form

$$y' = y^{\lambda},$$

where λ is a parameter to be determined from the data.

- The following are simple transformations from the Box-Cox family of power transformations:

$\lambda = 2,$ $\quad y' = y^2$

$\lambda = .5,$ $\quad y' = y^{.5} = \sqrt{y}$

$\lambda = 0,$ $\quad y' = \log_e y$ (*by definition*)

$\lambda = -.5,$ $\quad y' = y^{-.5} = \dfrac{1}{\sqrt{y}}$

$\lambda = -1.0,$ $\quad y' = y^{-1} = \dfrac{1}{y}$

- The regression model with the response variable a member of the family of power transformations is given as:

$$y_i^{\lambda} = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

The model includes the additional parameter λ which will be estimated using the method of maximum likelihood (ML), along with the regression coefficients $\beta_0, \beta_1,$ and σ^2. The estimated value of λ is used in the power transformation.

- Some statistical software do not automatically give the ML estimator of λ for the power transformation. Simple iterative schemes involving numerical search and using the standard regression software may be used instead to estimate λ.
- In order to make SST, SSR and SSE independent of λ, the observations are standardized first:

$$y^*(\lambda) = \begin{cases} G^{1-\lambda}\left(\dfrac{y_i^{\lambda}-1}{\lambda}\right), & \lambda \neq 0 \\ G(\log_e y_i), & \lambda = 0 \end{cases},$$

where $G = \left(\prod_{i=1}^{n} y_i\right)$, *is the geometric mean of the* y_i *observations.*

- For a given λ value, create the $y^*(\lambda)$.
- Regress $y^*(\lambda)$ on the predictor variable(s). Obtain the error sum of squares SSE.
- The ML estimate $\hat{\lambda}$ is the value of λ for which SSE is minimum.
- A finer search may be done in the region of λ value for which SSE is minimum.

Simultaneous Confidence Intervals in Regression

- **Bonferroni Confidence Intervals for β_0 and β_1:**

The (1-α)100% family CI for β_0 and β_1:

$$b_0 \pm (t_{1-\alpha/4,n-2})S_{b_0}, \qquad b_1 \pm (t_{1-\alpha/4,n-2})S_{b_1},$$

- **Bonferroni Confidence Intervals for Mean Response for g levels of X_h: $E(Y_h)$**

The (1-α)100% family CI for $E(Y_h)$:

$$\hat{Y}_h \pm (t_{1-\alpha/2g,n-2})S_{\hat{Y}_h},$$

- **Bonferroni Confidence Intervals for new observation for g levels of X_h: Y_i**

The (1-α)100% family CI for Y_i:

$$\hat{Y}_h \pm (t_{1-\alpha/2g,n-2})S_{Y_i},$$

Regression through Origin

- The regression function is assumed to be linear and go through the origin at (0, 0).
- Examples: X= unit of output, Y=variable cost, so that Y=0, by definition when X=0.

  X=number of items of product in stock, Y= volume of sales of this product, so that Y=0, by definition when nothing of this product is in stock.

- The normal error model is

  $y_i = \beta_1 x_i + \varepsilon_i,$

  where β_1 is a parameter, X_i are known constants,

  and ε_i are independent N(0, σ^2).

- The regression function for the model: $E(y_i) = \beta_1 x_i,$ is a straight line through the origin with a slope β_1.

- The least squares estimates:

$$\text{Fit:} \qquad \hat{y}_i = b_1 x_i,$$

$$\text{Re}\,sidual: \quad e_i = (y_i - \hat{y}_i),$$

$$SSE = \sum_{i=1}^{n} e_i^2, \quad \text{MSE} = \frac{\text{SSE}}{\text{n-1}}$$

Confidence limits for SLR through origin

Estimate of	Estimated variance	Confidence limits
β_1	$s_{b_1} = \frac{MSE}{\sum_{i=1}^{n} X_i^2}$	$b_1 \pm ts_{b_1}$
$E(y_h)$	$s^2_{\hat{y}_h} = \frac{X_h^2 MSE}{\sum_{i=1}^{n} X_i^2}$	$\hat{y}_h \pm \text{ts}_{\hat{y}_h}$
$y_{h_{new}}$	$s^2_{pred} = MSE\left(1 + \frac{X_h^2}{\sum_{i=1}^{n} X_i^2}\right)$	$\hat{y}_h \pm \text{ts}_{pred}$

$$where \quad t = t_{\alpha/2;n-1}$$

Remarks:

- The residuals e_i may not sum to zero. That is e_i may not balance around e=0 line when plotted.
- When a nonlinear pattern or linear with β_0 far from the origin prevails, SSE may exceed SST (the total sum of squares of Y), resulting in a negative r^2.
- It is safe practice not to use such models.

Inverse prediction

Inverse prediction is a process where a regression model of Y on X is used to make a prediction of the value of X which gave rise to a new observation Y.

Example: consider two competing companies -

A: Y=selling price of a product regressed on its cost X is available.

B: selling price $Y_{h(new)}$ is known, you wish to estimate the cost $X_{h(new)}$ for this company on basis of results for Company A.

- For inverse prediction the regression model is assumed as before: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$
- The estimated regression function is found as usual: $\hat{y}_i = b_0 + b_1 x_i.$
- For a new observation $Y_{h(new)}$, we want to estimate the level of $X_{h(new)}$ that gave rise to this new observation. The point estimator is found by solving for X given $Y_{h(new)}$:

$$\hat{X}_{h(new)} = \left(Y_{h(new)} - b_0\right)/b_1,\ b_1 \neq 0,$$

where $\hat{X}_{h(new)}$ is a point estimator of the new level $X_{h(new)}$.

- Approximate (1-α)100% confidence limits for $X_{h(new)}$:

$$\hat{X}_{h(new)} \pm (t_{\alpha/2;n-2})(s_{pred\ X}), \quad where, \quad s^2_{pred\ X} = \frac{MSE}{b_1^2}\left(1 + \frac{1}{n} + \frac{(\hat{X}_{h(new)} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).$$

Remarks:

- The inverse prediction problem is also known as a '*calibration*' problem. It is used when less costly, quick and approximate values of Y are related to precise and often costly and time-consuming values of X based on n observations.
- The approximate confidence interval is appropriate if the quantity

$$\left(t_{\alpha/2;n-2} MSE\right)/b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ is small, say less than 0.1}$$

- Procedure is controversial among statisticians, who suggest that inverse regression should be made in direct fashion by regressing X on Y.