Comparing Two Population Variances, Experimental Design, and One-Way ANOVA

Chapter Nine: Statistical Inferences Based on Two Samples and Chapter Ten: Experimental Design and Analysis of Variance

9.10: F Tests about a Difference in Population Variances

- Since our hypothesis tests will always be in the form $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a: \sigma_1^2 > \sigma_2^2$, we can rearrange these in the form: $\boldsymbol{H_0}: \boldsymbol{\sigma_1^2 / \sigma_2^2 = 1}$ **versus** $\boldsymbol{H_a}: \boldsymbol{\sigma_1^2 / \sigma_2^2 > 1}$
- Intuitively, we reject the null hypothesis if the sample variances divided by one another are significantly larger than 1; the population of all possible values of s_1^2 / s_2^2 is described by the ***F distribution***, which is skewed to the right, with the exact shape depending on the ***numerator degrees of freedom (df_1)*** and the ***denominator degrees of freedom (df_2)***
  - The numerator degrees of freedom is equal to ***$(n_1 - 1)$*** while the denominator degrees of freedom is equal to ***$(n_2 - 1)$***
- In order to use this distribution, we employ an ***F point, F_α***, such that the area under the curve of the F distribution to the right of the F point is equal to α
- We then follow these steps to test the equality of population variances (one sided):
  - **Define the test statistic F =** $\boldsymbol{s_1^2 / s_2^2}$ **and compute the p value of the area under the curve of the F distribution with df_1 and df_2**
  - **If we are saying that the variance of one population is greater than the variance of the second population, then we can reject H_0 if $F > F_\alpha$ or, equivalently, if p value $< \alpha$**
  - **If we are saying that the variance of one population is less than the variance of the second population, we use the same process but define the test statistic as F =** $\boldsymbol{s_2^2 / s_1^2}$
- We then follow these steps to test the equality of population variances (two-sided)
  - **Define the test statistic F =** $\boldsymbol{\text{larger of } s_1^2 \text{ and } s_2^2 / \text{smaller of } s_2^2 \text{ and } s_1^2}$ **with df_1 = {the size of the sample with the largest variance – 1} and df_2 = {the size of the sample with the smaller variance – 1}, and compute the p value as 2 times the area under the F curve to the right of F**
  - **We can reject H_0 if $F > F_{\alpha/2}$ or, equivalently, if p value $< \alpha$**

10.1: Basic Concepts of Experimental Design

- Researchers are interested in how ***independent variables (IV)*** influence ***dependent variables (DV, response variable)***
- If we cannot control the independent variables/factors being studied, we say that the data obtained is ***observational***
- If we can control the factors being studied, we say that the data are ***experimental***, and the values or levels of the factor (or combination of factors) are called ***treatments***
  - The purpose of most experiments is to compare and estimate the effects of the different treatments (IVs) on the response variable (DV)
  - For example, a company that wishes to see how three gasoline types (A, B, C) affect fuel efficiency would be conducting an experiment with 1 factor (gasoline type), one DV (fuel efficiency), and three treatments (levels of the factor) – A, B and C (this is described as a 1 X 3 experiment because there is one factor with three levels)
- ***Completely Randomized Experiment Design*** have independent random samples of experimental units (objects to which treatments are assigned) are assigned to the treatments

10.2: One-Way Analysis of Variance (ANOVA)

- The goal of one-way ANOVA is to estimate and compare the effects of the different treatments on the response variable
- If we consider a particular treatment $\boldsymbol{i}$, then we can define $\mu_i, \sigma_i, n_i, \overline{x_i}, s_i, and\ x_{ij}$ to be the treatment mean, the treatment standard deviation, the size of the of the sample that has been chosen for treatment i, the point estimate of the mean, the point estimate of the standard deviation, and the j^th^ value of the response variable that is observed using treatment i, respectively
- One-way ANOVA allows us to test for significant differences between treatment means, and depends on:
  - Constant variance (all populations of values for all treatments have equal variances) – this is not very important if all the samples are approximately the same size (the one-way ANOVA will be approximately equal if the **largest sample standard deviation is no more than twice the smallest sample standard deviation)**
  - Normality (all populations associated with all treatments have normal distributions)
  - Independence (all samples associated with the treatments are randomly selected, independent samples

- To test if there are any significant differences between treatment means, we set up the null hypothesis (where each of the treatment means is set equal to one another)
- The alternative hypothesis in this case is that at least two of the treatment means differ
- To carry out the test, we compare the ***between-treatment variability*** (the variability between means from different treatments) and the ***within-treatment variability*** (the variability of observations within specific treatments)
    - If the between-treatment variability is large compared to the within-treatment variability, we can reject the null hypothesis
- The *treatment sum of squares (SST)* is defined as $\boldsymbol{SST = \sum n_i(\overline{x_i} - \overline{x})^2}$**, where $\overline{x}$ is the over-all mean of all observed values of the response variable, and n_i is the number of observations for that particular treatment**
- The *error sum of squares (SSE)* is defined as $\boldsymbol{SSE = \sum(x_{1j} - \overline{x_1})^2 + \sum(x_{2j} - \overline{x_2})^2 + \cdots + \sum(x_{pj} - \overline{x_p})^2}$ **- that is, we calculate the squared differences between each observed value of the response and its corresponding treatment and by summing these squared differences over all the observations in the experiment**
- The *total sum of squares (SSTO)* is defined as $SSTO = SST + SSE$
- The ***treatment mean square*** *(MST)* is defined as $MST = SST/(p-1)$, where p is the number of treatments
- The ***error mean square*** *(MSE)* is defined as $MSE = SSE/(n-p)$, where p is the number of treatments
- We can then use this information to test for differences between treatment means as follows:
    - **Define the F statistic $F = \frac{MST}{MSE} = \frac{SST/(p-1)}{SSE/(n-p)}$ and compute its p value as the area under the F curve to the right of F with p – 1 (numerator) and n – p (denominator) degrees of freedom**
    - **We can reject the null hypothesis if F > F_α or, equivalently, if p value < α**
- If the ANOVA test shows that at least two treatment means differ, we investigate which ones differ and estimate how large these differences are using ***pairwise comparisons***
- One way to do this is the compute point estimates and confidence intervals for ***pairwise differences*** – the differences between each set of two treatment means
    - We can do so by computing individual confidence intervals, which gives us a ***comparisonwise error rate*** that we determine when setting the intervals
    - However, we are less than the (comparisonwise error rate)% sure that all of the pairwise differences are simultaneously contained in their respective intervals
    - To do this we take intervals such that we are 100(1 – α) certain that all treatment means are in their respective intervals, and we refer to **α as the *experimentwise error rate***
- We can use the ***Tukey Formula*** (which yields the most precise (shortest) simultaneous confidence intervals) to accomplish estimation in one-way ANOVA as follows:
    - **Consider the pairwise difference between two treatment means, which can be interpreted to be the change in the mean value of the response variable associated with changing from using the second treatment to the first treatment, thus giving us a point estimate of the difference $\mu_i - \mu_h$ as $\overline{x_i} - \overline{x_h}$**
    - **An individual 100(1 – α) percent confidence interval for the pairwise difference is** $\left[(\overline{x_i} - \overline{x_h}) \pm t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_h}\right)}\right]$, where t is based on n – p degrees of freedom
    - **A Tukey simultaneous 100(1 – α) percent confidence interval for the pairwise difference is** $\left[(\overline{x_i} - \overline{x_h}) \pm q_\alpha\sqrt{\frac{MSE}{m}}\right]$, where q is based on the table of percentage points of the studentized range, where q is listed corresponding to values of p and n – p, and we assume that the two sample sizes are the same; if they are not, we replace the above formula with $\left[(\overline{x_i} - \overline{x_h}) \pm \left(\frac{q_\alpha}{\sqrt{2}}\right)\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_h}\right)}\right]$
    - **An individual 100(1 – α) percent confidence interval for a given treatment mean is** $\left[\overline{x_1} \pm t_{\alpha/2}\sqrt{\frac{MSE}{n_i}}\right]$

Comparing Two Population Variances, Experimental Design, and One-Way ANOVA

Chapter Ten - Randomized Block Design and Chapter Thirteen - Kruskal-Wallis and Friedman Tests

10.3 – The Randomized Block Design

- We use randomized block design to compare *p* treatments using *b* blocks, using each block once per treatment to measure its effects
- We often use matched sets of experimental units (ie, same set of students, with same GPA, university, and year, doing a series of tests to isolate the effects of the tests, and not any other extraneous variables)
- We use the following symbols: $x_{ij}, \overline{x_{i\circ}}, \overline{x_{\circ j}}, \overline{x}$ to represent the value of the response variable observed when block j uses treatment i, the mean of the b values of the response variable observed when using treatment i, the mean of the p values of the response variable observed when using block j, and the mean of the total of the bp values of the response variable that we have observed, respectively
- The ANOVA procesure when using block design partitions ***total sum of squares (SSTO)*** into ***treatment sum of squares (SST), block sum of squares (SSB), and error sum of squares (SSE)***, such that $SSTO = SST + SSB + SSE$
- We then calculate these numbers using the following equations:
  - $\boldsymbol{SST = b\sum_{i=1}^{p}(\overline{x_{i\circ}} - \overline{x})^2}$
  - $\boldsymbol{SSB = p\sum_{i=1}^{b}\left(\overline{x_{\circ j}} - \overline{x}\right)^2}$
  - $\boldsymbol{SSTO = \sum_{i=1}^{p}\sum_{i=1}^{b}\left(\overline{x_{\circ j}} - \overline{x}\right)^2}$
  - $\boldsymbol{SSE = SSTO - SSB - SST}$
- We then calculate the MST, MSB and MSE, and calculate the F scores for both the treatments and the blocks, using the numerator degrees of freedom corresponding to the numerator of the F-score (MST = df of (p – 1) for F(treatments), MSB = df of (b – 1) for F(blocks) and the denominator degrees of freedom corresponding to the numerator value

TABLE 10.8 ANOVA Table for the Randomized Block Design with p Treatments and b Blocks

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Treatments	$p - 1$	SST	$MST = \frac{SST}{p-1}$	$F(\text{treatments}) = \frac{MST}{MSE}$
Blocks	$b - 1$	SSB	$MSB = \frac{SSB}{b-1}$	$F(\text{blocks}) = \frac{MSB}{MSE}$
Error	$(p-1)(b-1)$	SSE	$MSE = \frac{SSE}{(p-1)(b-1)}$	
Total	$pb - 1$	$SSTO$		

- We can compare these to the required level of significance to see if 1) the F(treatment) value suggests that there is a difference between treatment effects on the dependent variable and 2) the F(blocks) value suggests that there is a difference between the effects of blocks on the dependent variable

Point Estimates and Confidence Intervals in a Randomized Block ANOVA

Consider the **difference between the effects of treatments *i* and *h* on the mean value of the response variable**.

1. A **point estimate** of this difference is $\bar{x}_{i\cdot} - \bar{x}_{h\cdot}$.
2. An **individual 100(1 − α) percent confidence interval** for this difference is

$$\left[(\bar{x}_{i\cdot} - \bar{x}_{h\cdot}) \pm t_{\alpha/2}\, s\sqrt{\frac{2}{b}}\right].$$

Here $t_{\alpha/2}$ is based on $(p-1)(b-1)$ degrees of freedom, and s is the square root of the *MSE* found in the randomized block ANOVA table.

3. A Tukey simultaneous 100(1 − α) percent confidence interval for this difference is

$$\left[(\bar{x}_{i\cdot} - \bar{x}_{h\cdot}) \pm q_{\alpha}\frac{s}{\sqrt{b}}\right].$$

Here the value q_α is obtained from Table A.9, which is a table of percentage points of the studentized range. In this table, q_α is listed corresponding to values of p and $(p-1)(b-1)$.

13.4 – Comparing Several Populations Using the Kruskal-Wallis H Test

- We use this type of test when the normality and/or equal variances assumptions for one-way ANOVA do not hold
- To do so, we arrange all observations from all random samples for each treatment from smallest to largest, and assign each value a rank based on its position in this ordering (1, 2, …)
  - If two observations are equal values, we assign to them an equal value equal to the average of the rank that would have been assigned to them
- We then sum up the total values (T_i) for each individual treatment
- We can reject the null hypothesis in favor of the alternative at the *alpha* level of significance if the ***Kruskal-Wallis H statistic***

$$\boldsymbol{H = \frac{12}{n(n+1)} \sum_{i=1}^{p} \frac{T_i^2}{n_i} - 3(n+1)}$$

is greater than the X_α^2 point based on (p – 1) degrees of freedom, as long as there are five or more observations in each sample and the number of ties is small relative to the total number of observations

11.1 – Correlation Coefficient

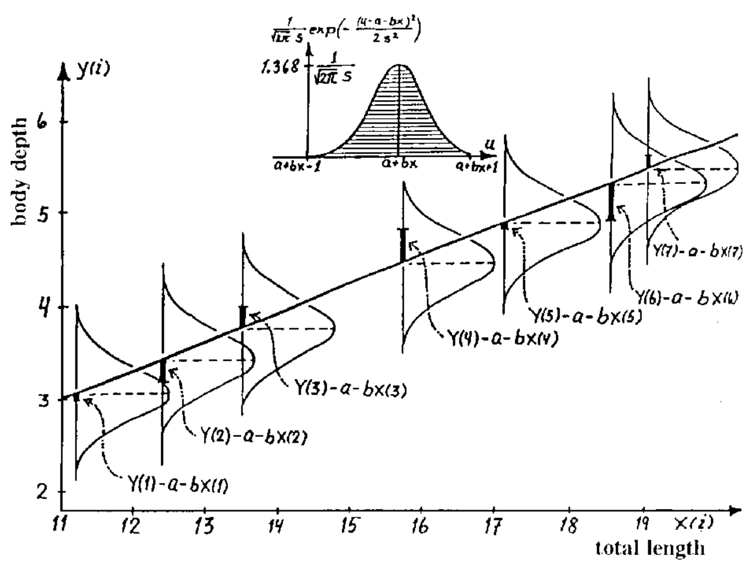
- If we assume a linear relationship between a dependent and independent variable (x and y), we measure the strength of the linear relationship using ***covariance***
    - $\boldsymbol{Sample\ Covarriance} = \boldsymbol{s_{xy}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$
    - If we draw lines on a scatterplot with all the pairs at the mean x and y values, we split the area into four quadrants; positive covariance goes from the bottom-left to top-right quadrant and vice versa
- An alternate measure that is not dependent on units is the ***correlation coefficient***, which is calculated as $\boldsymbol{r} = \frac{s_{xy}}{s_x s_y}$
    - This value ranges from -1 to 1, with those boundaries indicating increasingly strong linear relationships between the variables (0 does not necessarily mean no relationship, but instead means no *linear* relationship

11.2 – The Simple Linear Regression Model

- The simple linear regression model is defined by $\boldsymbol{y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon}$, where $\boldsymbol{\mu_{y|x} = \beta_0 + \beta_1 x}$ is the mean value of the dependent variable given a preset value of the independent variable, and ε is the error term that accounts for natural variance and any variables not defined within the model

11.3 – The Least Squares Estimates, Point Estimation and Prediction

- We can use the simple linear regression model to estimate the value of y based on a given x using estimates for the betas
- The closeness of our estimate is measured by the ***residual*** – it is equal to the actual observed value at that independent variable (y) less the estimate (y hat)
- To find the ***line of best fit/ least squares regression line/ least squares prediction equation***, we must find the equation of the line that minimizes the sum of the squared residuals (***sum of squared residuals/ sum of squared errors/*** SSE)
- The general form of the line of best fit is given by $\hat{y} = b_0 + b_1 x$, where:
    - $\boldsymbol{SSE = \sum(y_i - (b_0 + b_1 x))^2}$
    - $\boldsymbol{b_1 = \frac{SS_{xy}}{SS_{xx}}, SS_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}}$
    - $\boldsymbol{b_0 = \bar{y} - b_1 \bar{x}}$
- Using this equation gives us two values when we sub in a given x-value: 1) the point estimate of the mean values of the dependent variable y when the x value is the given value, and 2) the point prediction of the dependent variable's value in a single observation when the independent variable is the given value
- When using a least squares line, we should not estimate a mean value or predict an individual value when the corresponding value of x is outside of the ***experimental range*** – the range of previously observed values of x



11.4 – Model Assumptions and the Standard Error

- The ***regression assumptions*** are as follows: at any given point of x the population of potential error values has a mean of 0, at any value x, the population of potential error terms has a variance that does not depend on the value of x, and that these terms are normally distributed, and that any one error term is statistically independent of any other error term
- This implies that the population of all possible y values observed at a given x value are normally distributed with a mean of $\beta_0 + \beta_1 x + \varepsilon$ with a variance that is independent of the x value
- Mild departure from these assumptions does not generally lead to incorrect analysis
- The point estimate of the constant variance is the ***mean square error:*** $\boldsymbol{s^2} = \frac{SSE}{n-2}$ and the point estimate of the constant standard deviation is the ***standard error:*** $\boldsymbol{s} = \sqrt{\frac{SSE}{n-2}}$

11.5 – Testing the Significance of the Slope and y Intercept

- Singe regression analysis is not useful if there is not a significant relationship between y and x, we must test the significance of this relationship (the slope) using the null hypothesis: $H_0: \beta_1 = 0$ and the alternative, $H_a: \beta_1 \neq 0$
- The population of all possible b_1 values is normally distributed with a mean of β_1 and a standard deviation $\sigma_{b_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$ which can be estimated as $s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$ (*standard error of the estimate b_1*)
- We can then define the test stat $\boldsymbol{t = \frac{b_1}{s_{b_1}}}$ ***with n – 2 degrees*** of freedom and compare it to the level of significance using the following rejection point rules: $\beta_1 \neq 0, |t| > t_{\alpha/2}; \beta_1 > 0, t > t_\alpha; \beta_1 < 0, t < -t_\alpha$
- If we reject the null hypothesis we can conclude that the slope (the regression analysis) is significant at the level of significance specified
- If the regression assumption holds, a $100(1-\alpha)$ confidence interval for the true slope (beta one) is $[\boldsymbol{b_1 \pm t_{\alpha/2} s_{b_1}}]$ ***with (n – 2) df***
- To test for the significance of the y-intercept (b_0), we use the same rejection point rules and degrees of freedom with the following test stat: $\boldsymbol{t = \frac{b_0}{s_{b_0}}, where\ s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}}$

11.6 – Confidence and Prediction Intervals

- Unless we are very lucky, the value we estimate (y hat) will not get either the true mean y value at that x value, nor the observed value of y at that point
- We therefore need to calculate both a ***confidence interval for the mean value of y*** and a ***prediction interval for an individual value of y***
- We employ the $\boldsymbol{distance\ value} = \boldsymbol{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$
- If the regression assumption holds, a confidence interval for the mean value of y when we set a given x value is $[\boldsymbol{\hat{y} \pm t_{\alpha/2} s \sqrt{Distance\ Value}}]$, $based\ on\ n-2\ degrees\ of\ freedom$
- If the regression assumption holds, a prediction interval for an individual value of y when we set a given x value is $[\boldsymbol{\hat{y} \pm t_{\alpha/2} s \sqrt{1 + Distance\ Value}}]$, $based\ on\ n-2\ degrees\ of\ freedom$

11.7 – Simple Coefficients of Determination and Correlation

- The ***simple coefficient of determination (r^2 or eta^2)*** is a measure of the usefulness of a simple linear regression model
- If we do not employ the predictor variable x, our best estimate of y would be y bar (average of all observed y values); by using the predictor variable x we can more accurately estimate a particular y value, extra accuracy defined as $(y_i - \bar{y}) - (y_i - \hat{y}_i) = (\hat{y}_i - \bar{y})$ - the first section is the difference between a y value and the average, the second is the difference between a y value and the estimated value
- It can be shown that $\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2 = \sum(\hat{y}_i - \bar{y})^2$; the first section is the ***total variation***, the second is the ***unexplained variation/SSE***, and the third is the ***explained variation;*** $Total\ Variation = Explained + Unexplained$
- $eta^2 = r^2 = \frac{Explained\ Variation}{Total\ Variation}$ (Always greater than 0, less than 1 – the closer to 1, the more total variation that is explained by the predictor variable and the regression model)

11.9 – An F Test for the Model

- We can again test the significance of the linear regression model (null hypothesis: $H_0: \beta_1 = 0$ and the alternative, $H_a: \beta_1 \neq 0$) using the F stat: $\boldsymbol{F(model) = \frac{Explained\ Variation}{(Unexplained\ Variation)/(n-2)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{[\sum(y_i - \hat{y}_i)^2]/(n-2)}}$ with 1 numerator and (n – 2) denominator degrees of freedom
- We can reject the null (accept the significance of the linear regression model) if $F(model) > F_\alpha$

Multiple Linear Regression

Chapter Twelve – Multiple Regression and Model Building

12.1 – The Multiple Regression Model

- The only additional step in setting up the model is to assign a beta (1, 2, 3…) to each predictor variable we wish to include in the model
- The interpretation of beta zero remains the same, but the interpretation of the betas which are coefficients changes to become "the increase in the mean value of the response variable to a one-unit increase in the predictor variable, all other variables being held constant"
- The general form of the multiple regression model is $y = \mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

12.2 – The Least Squares Estimates, and Point Estimation and Prediction

- We do not know the true betas, and so write our ***least squares prediction equation*** as the general form of the multiple regression model with y replaced by (y hat) and with the true betas replaced by 'b's' – point estimates of the true betas that minimize SSE
- $\boldsymbol{SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- Using the least squares point estimates of the true betas, we will obtain, when subbing in a set of independent variable values, the point estimate of the dependent variable, which is also the point prediction of an individual value of the dependent variable given those select values of the independent variables

12.3 – Model Assumptions and the Standard Error

- There are four assumptions of the error term ε: that the population of potential error terms has a mean of 0, that any given combination of IV values has a constant variance that is independent of the combination of IV values (denoted σ^2), that any combination of IV values yields a population of potential error terms that is normally distributed, and that any one error term is statistically independent of any other error term
- A point estimate of the standard error is $\boldsymbol{s = \sqrt{\frac{SSE}{n-(k+1)}}}$ (where k is the number of IV we have) and the mean square error is $\boldsymbol{s^2 = MSE = \frac{SSE}{n-(k+1)}}$
    - The number of degrees of freedom associated with the standard error is **n – (k + 1)**

12.4 – R^2 and Adjusted R^2

- $\boldsymbol{Total\ Variation = SSTO = \sum(y_i - \bar{y})^2}$
- $\boldsymbol{Explained\ Variation = SSR = \sum(\hat{y}_i - \bar{y})^2}$
- $\boldsymbol{Unexplained\ Variation = SSE = \sum(y_i - \hat{y}_i)^2}$
- $\boldsymbol{SSTO = SSE + SSR}$
- $\boldsymbol{Multiple\ Coefficient\ of\ Determination = R^2 = \frac{SSR}{SSTO}}$ (Proportion of the total variation in the n observed values of the dependent variable that is explained by the overall regression model
- $Multiple\ Correlation\ Coefficient = R$
- To avoid overestimating the importance of IV (since adding an IV will always increase SSR, regardless of its relationship with the dependent variable), we computer the ***adjusted*** R^2: $\boldsymbol{\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right)}$
    - This number is useful primarily in comparing the usability of different models with different numbers of IV

12.5 – The Overall F Test

- If we wish to test whether *any* of the IV are significantly related to y (the regression relationship is significant), we test $H_0: \beta_1 + \cdots + \beta_k = 0$ versus $H_a: at\ least\ one\ of\ \beta_1, \ldots, \beta_k\ does\ not\ equal\ 0$
- The $Overall\ F\ Statistic$ is $\boldsymbol{F(model) = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} = \frac{MSR}{MSE}}$; we can reject the null in favor of the alternative (accept that the regression relationship is significant) if F(model) > F_α or, equivalently, if p values < α

12.6 – Testing the Significance of an Independent Variable

- We will test whether a particular IV is significant *in the model under consideration* (its significance is dependent on the other variables within the model) by testing $H_o: \beta_j = 0\ versus\ \beta_j \neq 0$
- We first need the ***standard error of the estimate*** b_j, which is shown on Minitab output under the Standard Error column, and we then compute the t-statistic $\boldsymbol{t = b_j / s_{b_j}}$

- We then test the significance of the IV in question using the following rules:

Alternative Hypothesis	Rejection Point Rule: Reject H_0 If:	p Value (Reject H_0 if p Value < α)
$H_a: \beta_j > 0$	$t > t_\alpha$	The area under the curve to the right of t
$H_a: \beta_j < 0$	$t < -t_\alpha$	The area under the curve to the left of t
$H_a: \beta_j \neq 0$	$t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$	Twice the area under the curve to the right of $\lvert t \rvert$

- Generally speaking, if we reject the null hypothesis, we have strong evidence that the IV is significantly related to the dependent variable in the current regression model

12.7 – Confidence and Prediction Intervals

- $s_{\hat{y}} = s\sqrt{Distance\ Value}$ – This is the ***standard error of the estimate (y hat)*** and is given on the Minitab output as "SE Fit" and, since we are also given s, we can find the distance value as $\left(\frac{s_{\hat{y}}}{s}\right)^2$
- $s_{y-\hat{y}} = s\sqrt{1 + Distance\ Value}$ – This is the ***standard error of the prediction error***
- $\boldsymbol{Confidence\ Interval\ for\ the\ mean\ value\ of\ Y = \left[\hat{y} \pm t_{\alpha/2} s\sqrt{Distance\ Value}\right]}$
- $\boldsymbol{Prediction\ Interval\ for\ an\ individual\ value\ of\ y = \left[\hat{y} \pm t_{\alpha/2} s\sqrt{1 + Distance\ Value}\right]}$

Multiple Linear Regression
Chapter Twelve – Multiple Regression and Model Building

12.8 – The Quadratic Regression Model

- This model takes the form $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, which is the equation of a parabola (can shown on a graph as any section of the parabola, depending on the exact equation
- Beta 0 is the intercept, Beta 1 is the ***shift parameter*** (moves it right or left) and Beta 2 is the ***rate of curvature*** (can be positive or negative)
- This can be used in combination with a linear model if there are multiple variables and one (or more) appear to be parabolic in nature

12.9 - Interaction

- ***Interaction Variables*** are variables calculated by multiplying one variable by another
- It is appropriate to use this if the mean value of y and one of the independent variables is dependent on the value of another independent variable (For example, the effects of tenure on salary may be well served by employing an interaction variable of tenure and gender
- Interaction leads to different slopes when and/or intercepts for the linear regression relating one independent variable to y, depending on the specific value of the other independent variable
- This model takes the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$, which can be rearranged as $y = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1)x_2$; this implies that the slope of the line relating y to x_2, $(\beta_2 + \beta_3 x_1)x_2$, will be different for different values of x_1 (a similar rearrangement can be done for x_1)
- If the interaction term is found to be significant in the model, it is practice to retain their corresponding linear terms, regardless of their p-values

12.10 – Using Dummy Variables to Model Qualitative Independent Variables

- ***Dummy Variables*** help us model qualitative values, and usually either take the value of 0 or 1
- If we wish to model the effect of a qualitative IV with *a* levels, we need *a – 1* dummy variables

12.11 – The Partial F Test: Testing the Significance of a Portion of a Regression Model

- We can test a set of IV for their significance by testing the null hypothesis that their respective Betas equal 0 versus the alternative, that at least one of those Betas does not equal 0
- We first define SSE_C and SSE_R, the unexplained variation for the complete model and the unexplained variation for the reduced (variables under consideration removed) model, respectively
    - The difference between these two is called the ***drop in the unexplained variation attributable to the independent variables (variables under consideration)***
- We define the test statistic $F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]}$, based on k – g numerator degrees of freedom (k is the number of IV in the complete model while g is the number in the reduced model), and n – (k + 1) denominator degrees of freedom, and compare it in the typical way to the desired level of significance

12.12 – Model Building, and the Effects of Multicollinearity

- ***Multicollinearity*** exists when some of the IV in a regression analysis are related to or dependent on one another (they give duplicate information)
- This is severe if at least one simple correlation between the IVs is at least 0.9 (however even moderate multicollinearity is an issue)
- We can also measure this using VIF (***variance inflation factors***); $VIF = \frac{1}{1-R_j^2}$, Where R_j^2 is the multiple coefficient of determination for a hypothetic regression model relating x_j to all other independent variables (VIF = 1 if the IV in question is not at all related to the other IVs)
    - Generally, multicollinearity is considered severe if the larges VIF is larger than 10 (largest coefficient of determination is greater than 0.9) and/or if the mean VIF is substantially greater than 1
- An independent variable should not be included in a final regression model unless it reduces *s* enough to reduce the length of the prediction intervals for y
- We can use the adjusted R^2, s, and the C_p statistic for comparison of different models, but in the case where multiple models that have very similar results for these measures are found, we should use the model with the fewest variables
- The ***C Statistic*** is found as $C = \frac{SSE}{s_p^2} - [n - 2(k+1)]$, where the denominator is the mean square error for the model using all *p* potential IVs, while SSE is the unexplained variation using only *k* of those variables; the smaller this number, the better (however, it is desirable to have a model for which C = k + 1; anything substantially greater than k + 1 indicates substantial bias in the model)

- ***Stepwise Regression*** involves successively adding IVs to a model and judging whether or not the new variable should stay base don whether the t-statistic p-value is less than a defined α_{entry}, and whether the other variables should stay in the new model based on a defined α_{stay} (lowest p-value stays at each step)
- ***Backwards Elimination*** is a similar method, but in reverse; we start with all IVs in the model, then compare the smallest t-stat/ largest p-value variable to a desired α_{stay}, eliminate if necessary, and repeat

12.13 – Residual Analysis in Multiple Regression

- We can plot residuals against the values of each IV, the predicted value of the dependent variable, and/or the time order in which data have been observed

12.14 – Diagnostics for Detecting Outlying and Influential Observations

- An ***Outlier*** is a value substantially separated from the rest of the data, while an ***Influential Point*** is a an observation that would cause a significant change in the regression model were it to be removed (generally, must be an outlier in both its x- and y-values)
- ***Leverage Values*** are used to identify x-outliers; it is indicated in Minitab as HI1; any leverage value greater than twice the average of all leverage values, which can be found to equal $\frac{2(k+1)}{n}$, is said to be an outlier
- ***Studentized Residuals*** are used to find y-outliers; they are the number of standard deviations from the mean (here, calculated as the residual over its standard error), and if they have an absolute value greater than 2 are said to be outliers
    - ***Studentized Deleted Residuals***: to find the deleted residual, we take the difference between the actual measure y_i and the point estimate $\hat{y}_i$ computed using regression analysis based on all measurements *except* the measurement under consideration, which is then divided by its standard error to obtain the studentized deleted residual
    - This is shown as TRES1 on Minitab, and we compare it to $t_{0.025}$ and $t_{0.005}$ based on n – k – 2 degrees of freedom; if it is larger than the 0.025, we have some evidence of a y-outlier, and if it is larger than the 0.005, we have strong evidence
- ***Cook's D (Distance Measure)*** is a measure of whether an observation is influential, we compare its D value to two F distribution points: $F_{0.80}$ and $F_{0.50}$, based on (k + 1) numerator degrees of freedom and [n – (k + 1)] denominator degrees of freedom; if D is greater than the 0.5, we consider the measurement influential, and if it is between the two numbers we consider it moderately influential (the close to the 0.5 number, the more influential)

Analyzing Categorical Data
Chapter Fourteen – Chi-Square Tests

14.1 – Chi-Square Goodness of Fit Tests

- We use this when we collect count data in order to study how the counts are distributed are distributed among several ***categories or cells***
- To perform statistical inference on these counts, we perform an experiment in which we carry out *n* identical trials and in which there are *k* possible outcomes on each trial, which generates k different probabilities which sum up to 1 (trials are independent) and k different frequencies
    - The null hypothesis will generally compare the probabilities/proportions to some predefined set, while the alternative is that the null is not true
    - We then define the ***chi-square statistic*** by comparing the actual frequency of an observation to its expected frequency based on the null hypothesis; that is, $\boldsymbol{\chi_o^2 = \sum_{i=1}^{k} \frac{(f_i - E_i)^2}{E_i}}$ (where $E_i = np_i$), which we then compare to the chi-squared value at the desired level of significance (with k – 1 degrees of freedom)
- This test is only applicable if all expected cell frequencies are at least 5, or, if this is not true, if the number of groups exceeds 4, the average expected frequency is at least five, *and* the smallest expected value is at least 1
- We can use this procedure for testing that the population has a normal distribution, in which we find the sample mean and sample standard deviation, define k intervals for the test, calculate the expected frequency based on the normality assumption, record the actual frequency, and base the subsequent test on k – 1 – m degrees of freedom (where m is the number of estimates used, here m = 2)

14.2 – A Chi-Square Test for Independence

- We can do a test for independence in a ***two-way cross-classification table*** (the table shows some count for a unit of the population based on them having some other characteristic as well using a chi-square test
- We define f_{ij} as the cell frequency corresponding to row i and column j, r_i as the row total for row i in the contingency table and c_j for the column total, and we will be testing the null hypothesis that states that two classifications are statistically independent versus the alternative that the two classifications are statistically dependent
    - We define the test stat $\boldsymbol{\chi^2 = \sum_{all\ cells} \frac{(f_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}}$, where $\boldsymbol{\hat{E}_{ij} = \frac{r_i c_i}{n}}$ and test against the desired level of significance based on (r – 1)(c – 1) degrees of freedom
- This test is only applicable if all expected cell frequencies are at least 5, or, if this is not true, if the number of groups exceeds 4, the average expected frequency is at least five, *and* the smallest expected value is at least 1

		HIGH	MEDIUM	LOW	Total
BOND	**Observed**	15	12	3	30
	% of row	50.0%	40.0%	10.0%	100.0%
	% of column	37.5%	30.0%	15.0%	30.0%
	% of total	15.0%	12.0%	3.0%	30.0%
STOCK	**Observed**	24	4	2	…
	% of row	80.0%	13.3%	6.7%	
	% of column	60.0%	10.0%	10.0%	
	% of total	24.0%	4.0%	2.0%	
TAXDEF	**Observed**	1	…	…	…
	% of row	2.5%			
	% of column	2.5%			
	% of total	1.0%			
Total	**Observed**	…	…	…	…
	% of row				
	% of column				
	% of total				

16.1 – Times Series Components and Models

- $Trends$ are the upwards or downwards movement that characterize a time series over time (long-run)
- $Cycles$ are recurring up-and-down movements around trend levels that take longer than one year
- $Season\ Variations$ are recurring movements the complete themselves within a calendar year
- $Irregular\ Fluctuations$ are erratic time series movements the follow no recognizable pattern

16.2 – Times Series Regression: Basic Models

- We can model time series as regression with any level of complexity (linear, quadratic, etc)
- Seasonal trends are accounted for by using dummy variable (one for each season) that take on a value of 1 if we are in that season and 0 if not (we need n – 1 dummy variables, and the coefficients for these variables indicate the amount greater/lower that season is with regards to the baseline season)
- This only works for constant seasonal variation – increasing/decreasing seasonal variation must either be turned into constant seasonal variation using fractional power transformation or using the multiplicative model
- Cannot be used to extrapolate or to estimate prediction intervals because it is based on a model that does not hold true in times series (the error terms are likely dependent), but computing a point estimate from this is fine

16.4 – Multiplicative Decomposition

- The model we use is $y_t = TR_t \times SN_t \times CL_t \times IR_t$, where TR is the trend component, SN the seasonal variation, CL the cycle, and IR the irregular variation

- Compute n-period moving averages, where n is the level of the data collected (12 for monthly, 4 for quarterly; if the data has an uneven number of period we can skip the next step)
- If the data has an even number of observations, compute the estimate $tr_t \times cl_t$ as the centered moving average of the surrounding n-period moving averages
- Since $sn_t \times ir_t = \frac{y_t}{tr_t \times cl_t}$, we can calculate $sn_t \times ir_t$ as y divided by the number found in whichever of the above two steps applied
- We take all $sn_t \times ir_t$ that we have and group by the period we are looking at (month, quarter), calculate an average for that period ($\overline{sn_t}$), then divide n by the sum of all averages ($\frac{n}{\sum sn_t}$) to normalize them, then multiply this normalization by each period's average to get the sn_t term to use for each occurrence of that month
- Divide the y observation by its seasonal term to get the deseasonalized observation ($d_t = \frac{y_t}{sn_t}$)
- Calculate the trend term as the regression model for the deseasonalized observations ($tr_t = d_t = \beta_0 + \beta_1 t$, if linear)
- We can the calculate $cl_t \times ir_t = \frac{y_t}{tr_t \times sn_t}$ since we have $tr_t \times sn_t$
- We then calculate the 3 period moving average of $cl_t \times ir_t$ to find cl_t, and thus find $ir_t = \frac{y_t}{tr_t \times sn_t \times cl_t}$
- If there is insufficient periods to find a cycle/ most values are close to 1, we can exclude the cycle term from our model
- If most values of ir are close to 1, we can exclude it from the model as well
- To compute a prediction interval for a given period, we use the point estimate generated by our model and use half the length of the equivalent prediction interval for the deseasonalized observation at that same period on both sides of the point estimate

		$sn_t \times ir_t = y_t / (tr_t \times cl_t)$ Year 1	Year 2	$\overline{sn}_t$	$sn_t = 1.0008758(\overline{sn}_t)$
1	Jan.	0.495	0.490	0.4925	0.493
2	Feb.	0.583	0.607	0.595	0.596
3	Mar.	0.608	0.582	0.595	0.595
4	Apr.	0.684	0.675	0.6795	0.680
5	May	0.567	0.561	0.564	0.564
6	Jun.	0.991	0.978	0.9845	0.986
7	Jul.	1.466	1.465	1.4655	1.467
8	Aug.	1.707	1.676	1.6915	1.693
9	Sep.	1.985	1.992	1.9885	1.990
10	Oct.	1.312	1.300	1.306	1.307
11	Nov.	1.026	1.030	1.028	1.029
12	Dec.	0.577	0.622	0.5995	0.600

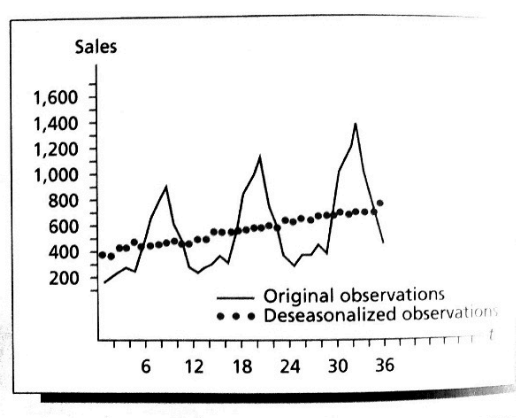


TABLE 16.11 Tasty Cola Sales and the Multiplicative Decomposition Method TastyCola

t, Time Period	y_t, Tasty Cola Sales	First Step, 12-Period Moving Average	$tr_t \times cl_t$, Centred Moving Average	$sn_t \times ir_t$, $\frac{y_t}{tr_t \times cl_t}$	sn_t, Table 16.12	d_t, $\frac{y_t}{sn_t}$	tr_t, $380.163 + 9.489t$	$tr_t \times sn_t$, Multiply tr_t by sn_t	$cl_t \times ir_t$, $\frac{y_t}{tr_t \times sn_t}$	cl_t, 3-Period Moving Average	ir_t, $\frac{y_t}{tr_t \times sn_t \times cl_t}$
1 (Jan.)	189				0.493	383.37	389.652	192.10	0.9839		
2	229				0.596	384.23	399.141	237.89	0.9626	0.9902	0.9721
3	249				0.595	418.49	408.630	243.13	1.0241	1.0010	1.0231
4	289				0.680	425	418.119	284.32	1.0165	1.0396	0.9778
5	260				0.564	460.99	427.608	241.17	1.0781	1.0315	1.0452
6	431	447.833			0.986	437.12	437.097	430.98	1.0000	1.0285	0.9723
7	660	452.417	450.125	1.466	1.467	449.9	446.586	655.14	1.0074	1.0046	1.0028
8	777	458	455.2085	1.707	1.693	458.95	456.075	772.13	1.0063	1.0004	1.0059
9	915	563.833	460.9165	1.985	1.990	459.79	465.564	926.47	0.9876	0.9937	0.9939
10	613	470.583	467.208	1.312	1.307	469.01	475.053	620.89	0.9873	0.9825	1.0049
11	485	475	472.7915	1.026	1.029	471.33	489.542	498.59	0.9727	0.9648	1.0082
12	277	485.417	480.2085	0.577	0.600	461.67	494.031	296.42	0.9345	0.9634	0.9700
13 (Jan.)	244	499.667	492.542	0.495	0.493	494.97	503.520	248.24	0.9829	0.9618	1.0219
14	296	514.917	507.292	0.583	0.596	496.64	513.009	305.75	0.9681	0.9924	0.9755
15	319	534.667	524.792	0.608	0.595	536.13	522.498	310.89	1.0261	1.0057	1.0203
16	370	546.833	540.75	0.684	0.680	544.12	531.987	361.75	1.0228	1.0246	0.9982
17	313	557	551.9165	0.567	0.564	554.97	541.476	305.39	1.0249	1.0237	1.0012
18	556	564.833	560.9165	0.991	0.986	563.89	550.965	543.25	1.0235	1.0197	1.0037
19	831	569.333	567.083	1.465	1.467	566.46	560.454	822.19	1.0107	1.0097	1.0010
20	960	576.167	572.75	1.676	1.693	567.04	569.943	964.91	0.9949	1.0016	0.9933
21	1,152	580.667	578.417	1.992	1.990	578.89	579.432	1,153.07	0.9991	0.9934	1.0057
22	759	586.75	583.7085	1.300	1.307	580.72	588.921	769.72	0.9861	0.9903	0.9958
23	607	591.833	589.2915	1.030	1.029	589.89	598.410	615.76	0.9858	0.9964	0.9894
24	371	600.5	596.1665	0.622	0.600	618.33	607.899	364.74	1.0172	0.9940	1.0233
25 (Jan.)	298	614.917	607.7085	0.490	0.493	604.46	617.388	304.37	0.9791	1.0027	0.9765
26	378	631	622.9585	0.607	0.596	634.23	626.877	373.62	1.0117	0.9920	1.0199
27	373	650.667	640.8335	0.582	0.595	626.89	636.366	378.64	0.9851	1.0018	0.9833
28	443	662.75	656.7085	0.675	0.680	651.47	645.855	439.18	1.0087	1.0030	1.0057
29	374	671.75	667.25	0.561	0.564	663.12	655.344	369.61	1.0119	1.0091	1.0028
30	660	677.583	674.6665	0.978	0.986	669.37	664.833	655.53	1.0068	1.0112	0.9956
31	1,004				1.467	684.39	674.322	989.23	1.0149	1.0059	1.0089
32	1,153				1.693	681.04	683.811	1,157.69	0.9959	1.0053	0.9906
33	1,388				1.990	697.49	693.300	1,379.67	1.0060	0.9954	1.0106
34	904				1.307	691.66	702.789	918.55	0.9842	0.9886	0.9955
35	715				1.029	694.85	712.278	732.93	0.9755	0.9927	0.9827
36	441				0.600	735	721.707	433.06	1.0183		

16.5 – Exponential Smoothing

- We must update our equation over time to make reasonable estimates, and we do this using exponential smoothing which applies a smoothing constant to assign more weight to more recent observations
- Our initial model should be based on at least six observations (if simple exponential smoothing with no trend), or at least half of historical data
- For a simple, no-trend data set, we can update our previous equation ($B_{T-1} = S_{T-1}$) to the new equation after a new observation by using this equation: $S_T = \alpha y_t + (1-\alpha)S_{T-1}, \alpha = smoothing\ constant\ between\ 0\ and\ 1$
- We can use the updated model to forecast one period in advance and, when we actually observe that period's value, we can compute a forecast error (difference between actual and forecast) and a squared forecast error
- The estimate is heavily influenced by recent observations and the weight of observations decreases exponentially; $S_T = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots + \alpha(1-\alpha)^{T-1}y_1 + (1-\alpha)^T S_0$ (remote observations are ***dampened out***)
- The smaller alpha is, the more slowly observations are dampened out