BSTA 378 Notes

Chapter 3: Simple Linear Regression

Regression Model: $Y = B0 + B1X + e$
Regression Equation: $y = b0 + b1X$

The regression equation is used to estimate the regression model. The "e" represents the random error term that always occurs when building a model.

Hypothesis Testing

Steps for conducting a Hypothesis Test:

1) **State Hypothesis**
   Ho: This is what we assume to be true. This statement contains the = sign.
   Ha: This is what we are trying to prove. This statement never contains the = sign.
2) **State Critical Value/Decision Rule**
   Use a) *the alpha level* and b) *the degrees of freedom* to look up the value used as a comparison point in the appropriate table. This value is the Critical Value.
3) **Calculate Test Statistic**
   Use the information in the problem to calculate the Test Statistic value for the test by using the appropriate formula.
4) **Calculate the p-value**
   Find the probability region beyond the calculated test value.
   OR Find the Test Statistic in the table while looking only at the corresponding df row. Identify the two values within which the Test Statistic falls in between and look up at the top of the table to identify the the two corresponding alpha level. Express the p-value as a range.
5) **Conclude**
   Show how the Test Statistic compares to the Critical Value and draw conclusions.

There are three possible hypothesis testing questions for simple linear regression:

1) **Is the model significant/valid?**
   Use the T-test to test the slope of the regression equation.

   Ho: $B1 = 0$
   Ha: $B1 \neq 0$
   CV = T* (alpha/2, df) = T* (alpha/2, n – 2) where "n" is the number of observations
   TS = T = (b1 – 0) / sb1
   Test stat: $T_{stat} = \frac{b1 - 0}{Sb_1}$

   $T_{critical} = T^*\left(\frac{\alpha}{2}, n-2\right)$

2) **Is the relationship positive/negative?**
Use the T-test to test if the slope of the regression equation is bigger/smaller than zero.

Ho: $B1 \leq 0$ OR $B1 \geq 0$
Ha: $B1 > 0$ OR $B1 < 0$
CV = T* (alpha, df) = T* (alpha, n – 2) where "n" is the number of observations
TS = T = (b1 – 0) / sb1
Test stat: $\frac{b_1 - 0}{s_{b_1}}$ $T^*(\alpha, n-2)$

3) **Can we claim that each X is more/less than A?**
Use the T-test to test if if the slope of the regression equation is bigger/smaller than A.

Ho: $B1 \leq A$ OR $B1 \geq A$
Ha: $B1 > A$ OR $B1 < A$
CV = T* (alpha, df) = T* (alpha, n – 2) where "n" is the number of observations
TS = T = (b1 – A) / sb1
Test stat: $\frac{b_1 - A}{s_{b_1}}$ $T^*(\alpha, n-2)$

Chapter 4: Multiple Linear Regression

Regression Model: $Y = B0 + B1X1 + B2X2 + (...) + e$
Regression Equation: $y = b0 + b1X1 + b2X2$

Note: The regression equation is also called the least squares prediction equation

Hypothesis Testing

1) **Is the the model significant?** (Test the overall significance/validity of the model)
Use the F-test and build the ANOVA table to calculate the F Statistic.

Ho: $B1 = B2 = B3 = Bi = 0$
Ha: At least one $Bi \neq 0$
CV = F* (alpha, d1, d2) = F* (alpha, k, n – k – 1) where k is the number of X variables
TS = F = (SSR / k) / (SSE / n – k – 1) Test stat: $F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$ $F^*(\alpha, k, n-k-1)$

2) **Is Xi a significant variable in the model?** (Test the partial slope of Xi)
Use the T-test for the slope of the individual variable and use the associated p-value to determine its significance based on the alpha level.

Ho: $Bi = 0$
Ha: $Bi \neq 0$
CV = T* (alpha/2, df) = T* (alpha/2, n – 2) where "n" is the number of observations
TS = T = (bi – 0) / sbi
If p-value < alpha, the variable is significant and we reject Ho.

$T_{stat} = \frac{b_i - 0}{s_{b_i}}$ $T^*: (\frac{\alpha}{2}, n-2)$

3) **Should we keep Xi and Xj in the model?**

Use the Partial F-Test to determine if we should keep the full model or accept the reduced model.

Ho: Bi = Bj = 0

Ha: At least one B ≠ 0

CV = F* (alpha, d1, d2) = F* (alpha, f – r, n – k – 1) where f is the number of X variables in the full model and r is the number of variables in the reduced model.

TS = F = (SSEr – SSEf / f – r) / (SSEf / n – k – 1)

$$T_{stat} = F = \frac{\frac{SSE_{Red.} - SSE_{Full}}{f-r}}{\frac{SSE_{Full}}{n-k-1}} \qquad F^*(\alpha, f-r, n-k-1)$$

Multicollinearity

This phenomenon occurs when the X variables are correlated with each other. The prediction ability of the model becomes redundant.

Consequences:

- The standard errors of the slopes (sbi) are larger than they should be
  - ✓ This is apparent when all the individual T-values are near 0 and F is quite large.

- It is hard to get a good estimate for the model's slope coefficients (Bi)
  - ✓ This is apparent when the slope (bi) is the wrong sign or when one of the slopes (bi) changes significantly when another variable is dropped.

How to measure:

VIF = 1 / (1 – R2j)

- If individual VIF > 10, this variable is a problem
- If average VIF > 1, SSE is inflated
- If the VIF for the entire model is larger than an individual VIF, multicollinearity is not a problem

Table of Correlations

Identify the variable pairs with a correlation larger than 0.50, these variables are related in some way and should be investigate.

Chapter 6: Assessing Assumptions

Note: Error = Residual = Disturbance

Assumptions:

1) **The errors are independent**
2) **The errors have constant variance**
3) **The errors are normally distributed**
4) **The average of the errors "e" = 0 and the regression line passes through y-bar**

✳ *If one assumption is not respected, we cannot proceed with the regression model as it is will be biased. The error terms must respect all these conditions.*

How to perform a Residual Analysis: ≠ ANOVA Assumptions

1) **Check for Independence**

   Ho: Residuals are independent
   Ha: Residuals are not independent

   Comment on the distribution of the error terms by looking at the errors plot. If there is no distinct pattern in the plot, we do not reject Ho and assume the errors to be independent.

2) **Check for Constant Variance**

   Ho: Variance is constant
   Ha: Variance is not constant

   Run a Heteroschedasticity test (First & Second Moment Specification).
   If the p-value is smaller than 0.05, do not reject Ho and assume constant variance.

3) **Check for Normality**

   Ho: Studentized Residual is normal
   Ha: Studentized Residual is not normal

   If the majority of the p-values are smaller than 0.05, do not reject Ho and assumer normality of residuals.

4) **Check for Linearity**

   Use the *Pure-Error Lack-of-Fit Test* OR the *Data Subsetting Test*. Unfortunately, only SAS coding yields the first test and only Minitab can provide the p-value table for the second test. Do not test this assumption unless the appropriate table is provided to you.

How to identify outliers in Residual Analysis:

1) **Leverage**
   Flag value if it is larger than $2*(k + 1) / n$

2) **DFIT**
   Flag value if it is larger than $2*SQRT((k + 1)/ n)$

3) **Cook's Distance**
   Flag value if it is larger than $F (k + 1, n - k - 1, alpha)$

4) **Studentized Residual**
   Flag value if it is larger than 2 (absolute)

   *An outlier fails at least 2 tests and must be investigated further.

Chapter 7: Indicators and Interactions

Indicators variables
These are used to introduce qualitative data into the model.

Interaction variables
These are used to try to build a better model by multiplying two already existing variables together into one variable.

$Y = B0 + B1X1 + B2X2 + B3X1X2$
$Y = B0 + (B1 + B3X2) X1 + B2X2$

Find the BEST model
The BEST model is found by eliminating the least significant variables and running the regression again. We can conclude whether the new model is better than the old model by conducting a partial F-test (Chapter 4).

How to identify the least significant variables:
- Variables that have p-values that are greater than alpha
- Variables that have a p-value relatively far from zero

Sum of Squares Type 1
By adding variable X to the model, SSR will increase by the value of SS1 associated with variable X.
- This is a weak assessment because the sequence is set by the order in which the variables were entered into the model.

Sum of Squares Type 2
By adding variable X to the model, in addition to all other variables, SSR will increase by the value of SS2 associated with variable X.
- This is a better assessment because it is what you have gained from adding a variable to the model.