Checking Assumptions (Ch 8)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assumption | Graph/Statistics | Violate When | How to correct | Ch Ref |
| “Mean zero” = lack of fit | Residual vs x [Residual Plot in SAS] | Clear trend in plot | Fit a different model (transform independent variables) | Ch 8.3 |
| Equal Variance = homoscedastic  True variation = sigma ^2 | Residual vs X  Or  Residual vs predicted Y | Clear pattern “fanning in /out” | Transform y variable | 8.4 |
| Normality | QQplot or Histogram of residual | -off the line  -Skewed histogram | Transform y variable | 8.5 |
| Non-correlated  (one error isn’t dependent on the previous one)  Time series goes against this | Residual vs X and Durbin-Watson Test | -see positive or negative residual pattern  -P-value is less than 0.05 | Add auto-correlation term to model (Ch10) | 8.7 |

Outliers and Influential points Ch 8.6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistics | What does it tell me? | Name in SAS | Cut off Value | Page# |
| Standardized Residua;  -What points might be outliers? | General Outliers | Not Listed | +- 3  (plus or minus 3 of the mean) | 412-413 |
| Studentized Residual | Outliers in y or general outliers | Student Residual | +- 2 | 413 |
| Leverage | Outliers in x | Hat Diag H | h\_i > 2(k+1)/n | 418 |
| Cooks Distance | Overall Influence | Cooks D | F[a = 0.05, v\_1 = k+1, v\_2 = n-(k+1)] OR SAS uses 4/n | 419-420 |
| Studentized Deleted Residual | Overall Influence | RStudent | T(a = 0.05, df = n-k+2) OR SAS uses +-2 | 422 |

Observation – Predicted = Residual

DF = model = 2 which is K (number of predictors)

Q 26)

Mean square of the regression (Mean Square) = Sum squares / degree of freedom

* \*\* Everything in row stays together

Error Degree of freedom

* Model Degree of Freedom + Error Degree of Freedom = Total Degree of freedom
* N-k-1

Mean Square error is an estimate of variance (sigma ^2)

Calculate S (estimate of Sigma)

* Square root?

How many observations were used for prediction? 15

Total Sum of squares = add them together

Calculated R-Square

* Proportion of variation in y that is explained by my model

\*\*Anytime about significance we need to do a hypothesis test that has a P-value \*\*

F-value

Mean square regression / mean square of error

If there is a variable x^2 (use adjusted r squared)

If there is a child then you need all the children

* If u have a s cubed term
  + You need all x^2 and x variables

Prediction vs confidence

* Prediction interval is wider
* Center is the same