Multiple Regression
Part 2

STAT 441/541 Statistical Methods II

Objectives of Part 2

- Assess the goodness of fit of the regression model
- Check for multicollinearity among the independent variables
- Predict new *y* values using the estimated multiple regression model
- Check assumptions for regression analysis
- Check for potential outliers by identifying observations with high leverage or influence

Sections Covered for Part 2

Chapter 12

- 12.6 Forecasting Using Multiple Regression

Chapter 13

- 13.4 Checking Model Assumptions

Model Standard Deviation

- A measure of how well the multiple regression model fits the data
- It is important to estimate the model standard deviation, denoted by σ_ε
- Residuals, e_i, are used to estimate σ_ε
- We can find the estimate of the model standard deviation s_ε in R output. It is labeled "Residual standard error:"

Residual: $e_i = y_i - \hat{y}_i$

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-(k+1)}}$$

Adjusted R-Squared

- The Adjusted R-Squared value is also a measure of how well the multiple regression model fits the data
- It is labeled as “Adjusted R-squared” in R output
- Values of Adjusted R-squared are between 0 and 1
- We can interpret Adjusted R-squared as the proportion of total variation in the response variable that is explained by the model
- Values are often reported as a percentage
- For example, “Adjusted R-squared: 0.7275” indicates that the multiple regression model explains over 70% of the total variation in the response variable

Effect of Multicollinearity

- If the independent variable x_j is highly correlated with one or more other independent variables, than the parameter estimates are inaccurate and have large standard errors
- The variance inflation factor (VIF) measures how much the variance of a coefficient is increased because of multicollinearity
  - If VIF=1, there is no multicollinearity
  - If VIF>10, there may be a serious problem
- A VIF value is calculated for each independent variable

Section 12.6 Forecasting using Multiple Regression

- One of the major uses for multiple regression models is in forecasting a y-value given certain values of the independent x variables
- The best forecast is substituting the specified x-values into the estimated regression model
- The standard error of a forecast depends on the interpretation of the forecast

Two Interpretations for Forecasts

- The forecast of y for given x-values can be interpreted two ways
    1. As the estimate for $E(y)$, the long-run average y-values from averaging many observations of y when the $x's$ have the specified values (confidence interval)
    2. The predicted y value for one individual case having the given x-values (prediction interval)
- We will use software to calculate confidence and prediction intervals

Section 13.4 Checking Model Assumptions

- It is always important to check assumptions for any statistical method
- For multiple regression, we will use graphical and numerical methods. These include:
  - Diagnostic plots
  - Shapiro-Wilk Normality Test
  - Breusch-Pagan test for a common variance

Assumptions for Multiple Regression

- The model has been properly specified
- The variance of the errors is σ_ε^2 for all observations
- The errors are independent
- The errors are normally distributed and there are no outliers
- More formally, in statistical notation, these are:
  - Zero expectation: $E(\varepsilon_i) = 0$ for all observations
  - Constant variance: $V(\varepsilon_i) = \sigma_\varepsilon^2$ for all observations
  - Normality: ε_i is normally distributed
  - Independence: The ε_i are independent

Checking Assumptions Using Residuals

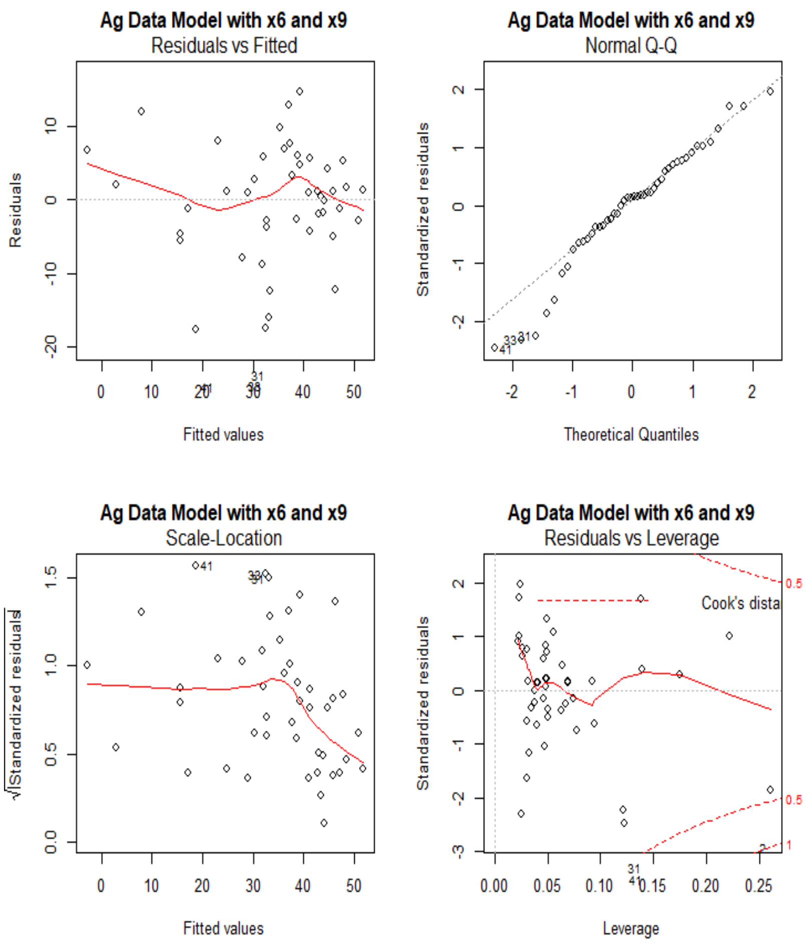
- Residuals are defined as the difference between observed and predicted values of the dependent variable for each observation
- Residuals, e_i, are used to estimate random error, ε_i

$$e_i = y_i - \hat{y}_i$$

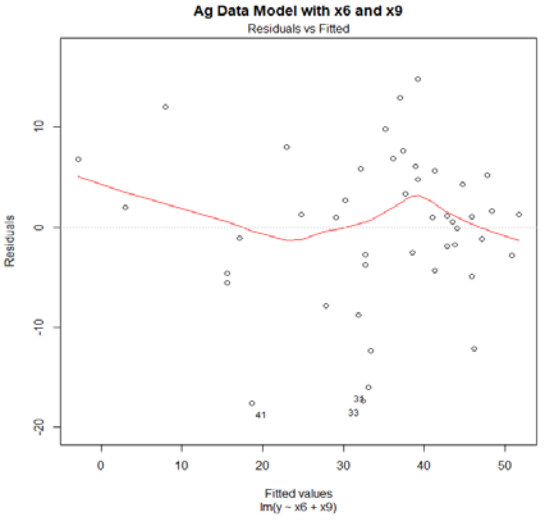
- We will use residuals to check assumptions for multiple regression

Diagnostic Plots

R provides four diagnostic plots based on residuals that help us visually check assumptions



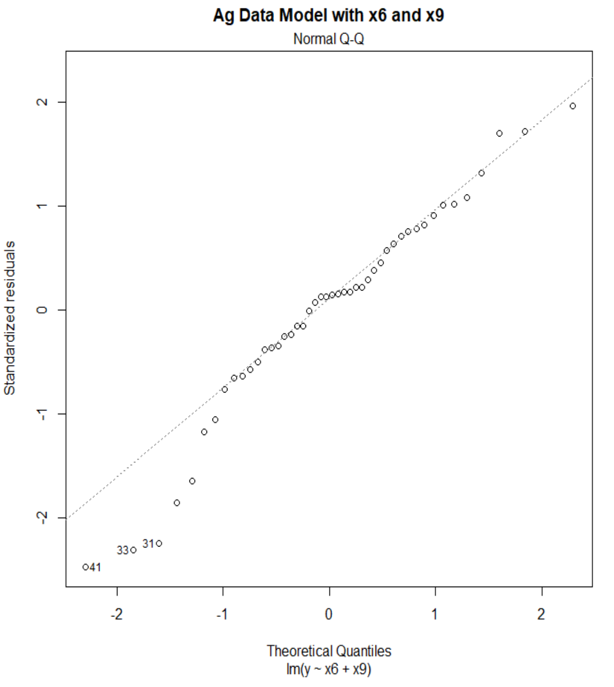
Residuals vs Fitted



- Fitted, or predicted, values are on the horizontal axis and residuals are on the vertical axis
- If the solid red line is horizontal or almost horizontal, then the model has been properly specified
- Deviations from a horizontal line or nonrandom patterns of points may indicate the model needs to modified or the data transformed
  - Scan the plot from left to right to see if the residuals remain close to zero
  - Scan the plot from left to right to see if the spread of the residuals remain approximately constant
  - Check for points that are labeled and identify these as potential outliers

Normal Q-Q

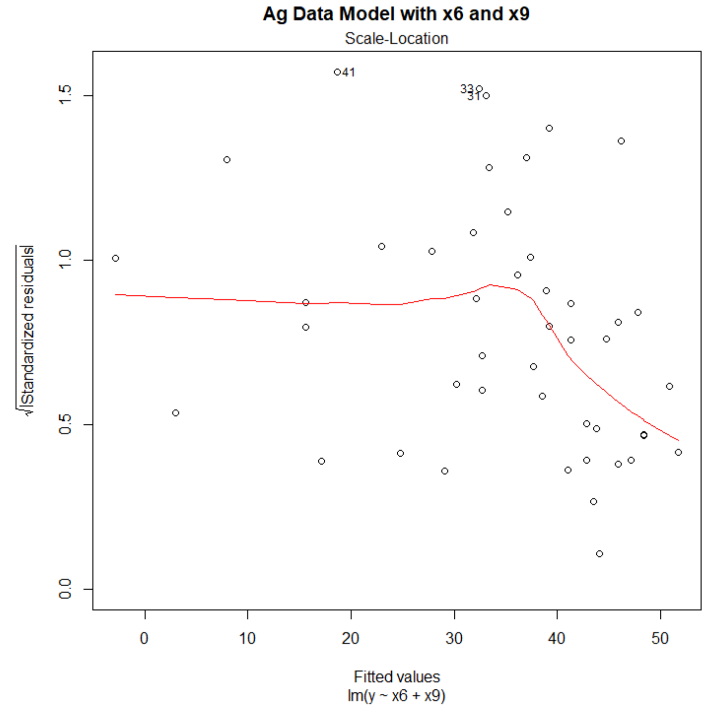
- Theoretical normal quantiles are on the horizontal axis and studentized residuals are on the vertical axis
- If the points tend to follow the straight dashed line then the errors are normally distributed



- Deviations from the straight dashed line or nonrandom patterns may indicate the model needs to modified or the data transformed
    - Check for points that are labeled and identify these as potential outliers

Scale-Location

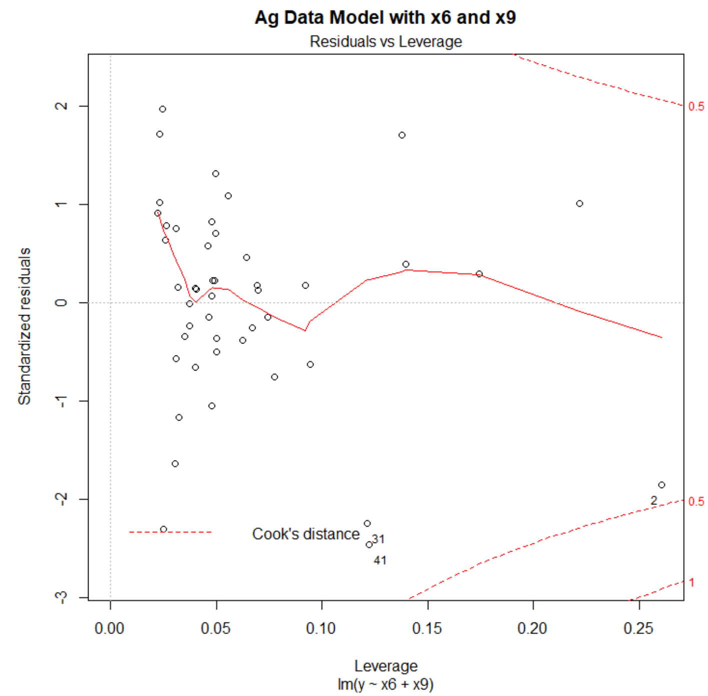
- Fitted, or predicted, values are on the horizontal axis and the square root of the absolute values of studentized residuals are on the vertical axis
- If the solid red line is horizontal or almost horizontal, then there is a constant variance



- If the solid red line is not horizontal or nonrandom patterns of points may indicate the model needs to modified or the data transformed
    - Check for points that are labeled and identify these as potential outliers

Residuals vs Leverage

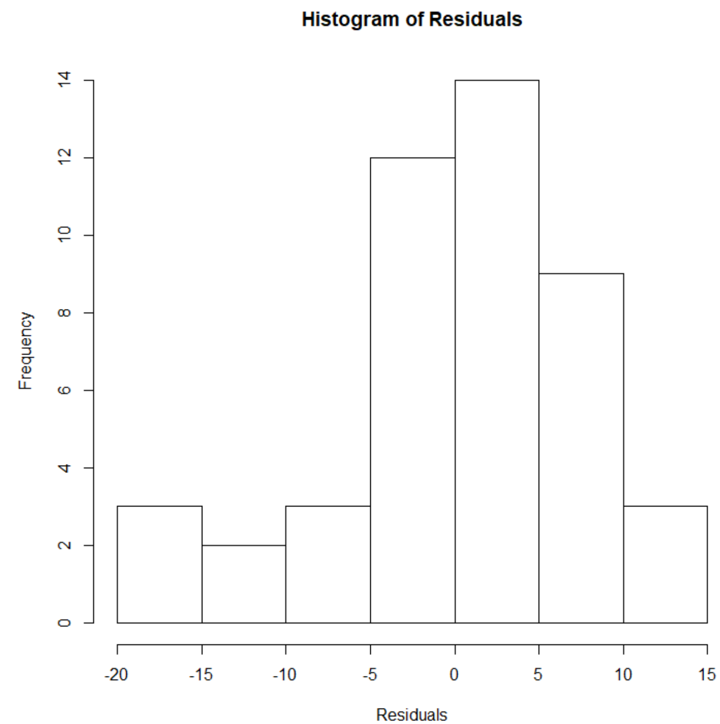
- Leverage values are on the horizontal axis and studentized residuals are on the vertical axis
- Look for points that have Cook's distance greater than one and are plotted beyond the red dashed line labeled "1"



- Check for points that have Cook's distance greater then one and labeled with observation number. These will be in the upper right corner and/or lower right corner of the plot. Identify these as potential outliers since they exhibit undue influence or leverage for estimating model parameters

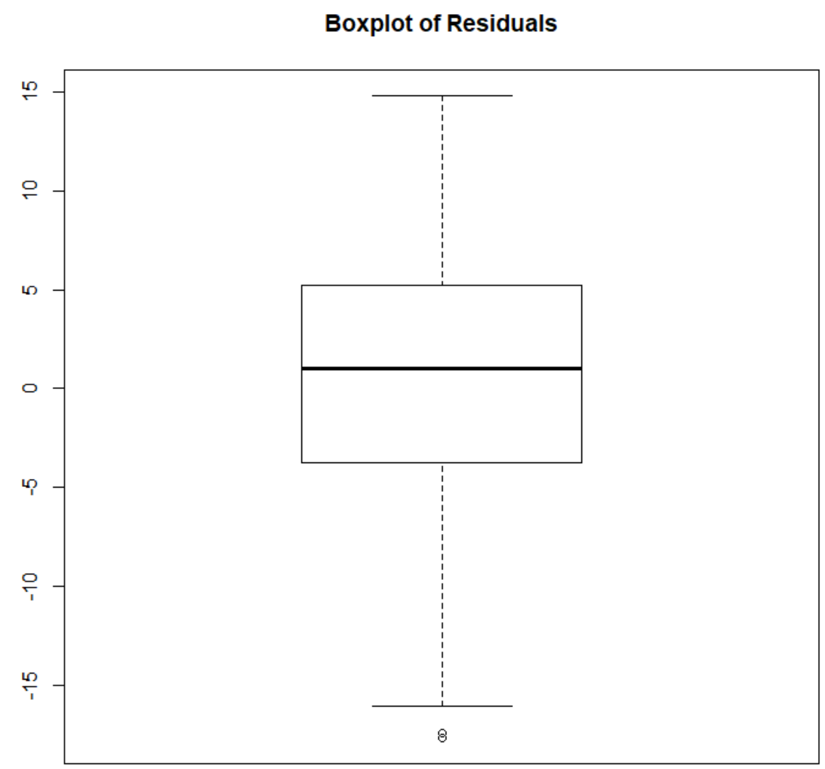
Histogram of Residuals

- Look for a shape that is bell-shaped, symmetrical, no gaps, and no outliers
- Should keep in mind that a histogram's appearance is influenced by a small number of observations



Boxplot of Residuals

- Look for whiskers that are similar in length
- Look for the solid line in the box (the median) to be near the center of the box
- Look for points that are plotted with open circles, these are potential outliers



Shapiro-Wilk Test for Normality

Used to test the assumption that the errors are normally distributed based on following hypotheses:

$$H_0: The\ errors\ are\ from\ a\ normal\ distribution$$

$$H_a: The\ errors\ are\ not\ from\ a\ normal\ distribution$$

```
> # Shapiro-Wilk test for normality of errors and use alpha=0.01
> shapiro.test(resid(model))

        Shapiro-Wilk normality test

data:  resid(model)
W = 0.96418, p-value = 0.1665
```

Breusch-Pagan Test for Constant Variance

Used to test the assumption that the variance is the same for all observations based on the following hypotheses:

H_0: *The variance is constant*

H_a: *The variance is not constant*

```
> # Breusch-Pagan Test for a common error variance and use alpha=0.01
> bptest(model)

	studentized Breusch-Pagan test

data:  model
BP = 4.6148, df = 2, p-value = 0.09952
```

R Function influence.measures

This function uses several methods to identify observations that have high leverage or influence. An “*” marks those observations that have an impact on the multiple regression model

- dfbetas
- dffit
- cov.r
- cook.d
- hat

```
> influence.measures(model)
Influence measures of
         lm(formula = y ~ x6 + x9, data = dataobj) :

      dfb.1_   dfb.x6    dfb.x9    dffit cov.r   cook.d    hat inf
1   0.027301 -0.02895 -0.018427  0.03491 1.152 4.16e-04 0.0697
2  -1.073012  0.91723  1.042055 -1.13557 1.130 4.05e-01 0.2608   *
3   0.139330 -0.14036 -0.106024  0.15406 1.235 8.07e-03 0.1398   *
4   0.041139 -0.04640 -0.024258  0.05387 1.180 9.90e-04 0.0923
5   0.097695 -0.08655 -0.088674  0.11841 1.130 4.76e-03 0.0643
6   0.029586 -0.20565  0.220084  0.54005 1.284 9.72e-02 0.2221   *
7   0.026174 -0.06583  0.034510  0.13050 1.292 5.80e-03 0.1745   *
8  -0.274608  0.05972  0.519511  0.69611 1.010 1.54e-01 0.1378
9  -0.143161  0.08664  0.198147  0.26409 1.046 2.32e-02 0.0557
10 -0.046490  0.03927  0.052977  0.12740 1.056 5.46e-03 0.0264
```