PSTAT 126 - Regression Analysis – Spring 2017

Lab 5 Handout

Data Exploration and Diagnosing Violations of Model Assumptions

**Note:** These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service.

Goals for this Lab

* Learn how to produce residuals plots and QQ normal plots in R.
* Learn how to use data displays to explore the data and diagnose any violations of the following major assumptions of the normal error linear regression model:
  + Linearity
  + Constant Variance
  + Independence
  + Normal Distribution
* Note: this lab focuses on producing and interpreting the plots. How to evaluate the assumptions will be discussed in more detail in lecture.

Lab Exercise – Producing Residuals Plots and QQ Normal Plots

The **USArrests** dataset of the **datasets** package contains crime statistics and population composition for each of the 50 US states from the year 1973. For each state, the dataset contains the number of arrests (per 100,000 in population size) for **Murder, Assault** and **Rape**, along with the variable **UrbanPop** - the percent of the state’s population that lives in an urban area. We will use this dataset to learn how to produce diagnostic plots.

1. Open and attach the **USArrests** dataset in the **faraway** package. Fit a linear model that predicts **Assault** from **UrbanPop** (name this linear model **fit1**).

> data(USArrests,package="datasets")

> attach(USArrests)

> fit1=lm(Assault~UrbanPop)

* 1. Linearity - Create a scatterplot with a regression line that shows **UrbanPop** on the X axis and **Assault** on the Y axis. Does there appear to be a linear relationship between Urban Population and Assaults?

> plot(UrbanPop,Assault)

> abline(fit1)

* 1. Constant Variance - Now create a residuals plot that has the predicted values of Y on the X axis and the residuals (Y-Y’) on the Y axis, and draw a horizontal line at Y=0, using the following R commands. Does there appear to be constant variance?

> plot(fitted(fit1),residuals(fit1)

> abline(h=0)

Now repeat these steps using **Rape** as the outcome variable. Fit a linear model predicting **rape** from **UrbanPop** (name this model **fit2**).

* 1. Linearity - Create a scatterplot and regression line with **UrbanPop** on the X axis and **Rape** as the Y axis.

Constant Variance - Create a residuals plot as you did in Part 1.b. How is this residuals plot different from Part 2.b? Are there any unusual observations (i.e., outliers)?  
  
> fit2=lm(Rape~UrbanPop)

> plot(UrbanPop,Rapte)

> abline(fit2)

> plot(fitted(fit2),residuals(fit2)

> abline(h=0)

1. Independence of Errors – If errors are independent, then one observation does not influence another observation in the dataset. Dependent (or correlated) errors often occur when observations are taken physically near each other, or close together in time. One method of testing this is to plot the residuals according to the order that the observations appear in the dataset. If data points that are next to each other in the plot have similar errors, then they are not independent (also called correlated errors)
   1. Plot the errors from the **Assault** linear model in the order that the observations appear in the dataset using the following commands. Repeat this for the **Rape** mode. Do any of the observations near to each other appear to be correlated?

> plot(1:50,residuals(fit1))

> plot(1:50,residuals(fit2))

1. Normal Distribution of Errors - We use a QQ Normal plot to evaluate the assumption of a normal distribution of errors.
   1. Create a QQ plot for the **Assault** model, and QQ line using the following R commands. In a normal distribution, the residuals will follow the QQ line. Repeat for the **Rape** model. Do the residuals in this plot deviate from the QQ line?

> qqnorm(residuals(fit1))

> qqline(residuals(fit2))

* 1. If the residuals deviate from the QQ line, we can use a histogram to look directly at the shape of the distribution. Create histograms of the residuals. Are either of the distribution skewed, or heavy in the tails?

> hist(residuals(fit1))

> hist(residuals(fit2))