Author: Brenda Gunderson, Ph.D., 2015

Statistics 250
Lab Workbook

Fall 2015

Weekly Labs and Supplements
Used in all lab sections of Stat 250

Dr. Brenda Gunderson
Department of Statistics
University of Michigan

Table of Contents

Note to Students

Welcome to Statistics 250 at the University of Michigan!

This is the first summer term in which R and R Commander will be used as the software package for Stats 250. Some of the reasons why we made this switch are:

- The ability to use R is a valuable skill recognized by employers.
- Other Statistics courses use R and this will make for an easier transition into these next courses.
- R is a free, open source software that can be downloaded onto student machines, so students can have access to it any time on their personal devices and won't have to use Virtual Sites.

This lab workbook is designed for you to use in lab and as extra preparation for exams. In the workbook, you will find the following materials:

Supplemental Material – great summaries for reference throughout the term:

1. R Commands Reference
2. Notation Sheet
3. Name That Scenario
4. Interpretation Examples
5. Summary of T-tests
6. Regression Output in R

Weekly Labs (numbered 1 to 12) – each lab contains the follow parts:

- **Lab Background** – objective and brief overview material, which is good to take a couple minutes to read *before* you come to lab each week.
- **Warm-Up Activity** – quick questions for you to do before the In-Lab Project, usually a quick review of concepts you have seen in lecture.
- **ILP (In-Lab Project)** – one or more activities you will work on in lab, in groups.
- **Cool-Down Activity** –questions for you to do after the ILP for further reflection and application of the concepts covered in the ILP.

The Labs are designed to be interactive and to provide you with a complete example for each concept. Completing the corresponding PreLab assignment (a link to video instructions for PreLabs will be on Canvas and the Stat 250 YouTube channel) and reading the upcoming lab background overview before lab each week is a good way to prepare for the various lab activities.

Good luck in Statistics 250!
-- *The Stat 250 Instructors and GSIs*

Special Thanks to the Statistics Graduate Students
Kit Clement
Sean Pikosz
Daniel Walter

For their substantial contributions to transition and modernize the Lab Materials to the Awesome R computing package

Supplement 1: R Commands Summary

By Lab – For Quick Reference

Lab 1 – Bar Charts, Histograms, Numerical Summaries, Boxplots

Open a data file after loading R Commander: Data > Load data set

To produce a **Histogram:** Graphs > Histogram

To generate **Descriptive Statistics:** Statistics > Summaries > Numerical summaries

To produce a **Bar Chart:** Graphs > Bar Graph

To produce a **Boxplot**: Graphs > Boxplot

Lab 5 – Time Plots, QQ Plots

To produce a **Sequence or Time Plot** for the variable named “VARIABLE” in the data set “DATA” you must type these two lines of code into the **R Script** box:

```
plot(DATA$VARIABLE,  type  =”l”,  main="Normal  QQ  Plot  of
variable by name")
```

Note that you can find the dataset name in blue text at the top. To find variable names, click **View data set** and look at the top row. To create the plot, highlight the above code and click the **Submit** button.

To produce a **QQ Plot:** you can use the built in option under **Graphs > Quantile-comparison plot** Or you can make a QQ plot for the variable “VARIABLE” in the data set “DATA” by typing these two lines of code into the **R Script** box:

```
qqnorm(DATA$VARIABLE, main="Normal QQ Plot of variable
by name")
qqline(DATA$VARIABLE)
```

Then highlight this code and click the **Submit** button.

Lab 6 – One-Sample t Procedures for a Population Mean

To perform a **One-Sample T Test for a population mean and obtain a confidence interval:** Statistics > Means > Single-sample t-test

Lab 7 – Paired t Procedures

To perform a **Paired T Test and obtain a confidence interval:** Statistics > Means > Paired t-test

To **compute Differences:** Data > Manage variables in active data set > Compute new variable.

Lab 8 – Independent Samples t Procedures

To perform **Levene's Test**: Statistics > Variances > Levene's Test

To perform a **Two-Samples T Test and obtain a confidence interval**: Statistics > Means > Independent samples t-test

Lab 9 – One-way Analysis of Variance (ANOVA)

To perform an **ANOVA:** Statistics > Means > One-Way ANOVA

Lab 10 and 11 – Linear Regression

To produce the **correlation (R)** for all pairs of variables: Statistics > Summaries > Correlation matrix

To produce a **Scatterplot:** Graphs > Scatterplot

To perform a **Linear Regression:** Statistics > Fit models > Linear regression

To produce a **Residual plot and QQ Plot of residuals**, first make sure you have the correct model selected, then follow**:** Models > Graphs > Basic diagnostic plots

Lab 12 – Chi-Square Tests

To perform a **Goodness of Fit Test:** Statistics > Summaries > Frequency distributions. Make sure to check the box to run a goodness of fit test, and then you can specify the null probabilities.

To perform a **Test of Independence:** Statistics > Contingency tables > Two-way table

To perform a **Test of Homogeneity**: Statistics > Contingency tables > Two-way table

Supplement 2: Notation Sheet

The table below defines important notations, including that used by R, which you will come across in the course. This is not an exhaustive list, but it is a fairly comprehensive overview of the “strange letters” used in the course.

Note: Blank cells mean there is no corresponding notation.

Name	Population Notation	Sample Notation	Notation used in R Commander
Summary Measures			
Mean	μ (read as “mu”)	$\bar{x}$ (*x*-bar)	Mean
Proportion	*p*	$\hat{p}$ (*p*-hat)	
Standard deviation	σ (sigma)	*s*	Varies, often “sd”
Variance	σ^2	s^2	Variance
Sample size		*n*	*n (sometimes N)*
Confidence Intervals			
Multipliers		*z** (*z*-star)	
		*t** (*t*-star)	
Margin of error		m, m.e.	
Hypothesis Testing			
Test statistics Note: *t*, *F*, and χ^2 statistics have degrees of freedom (abbreviated df) associated with them. Look for these on your Formula Card.		*z*	
		t	*t*
		F	*F*
		χ^2 (chi-square)	Chi-square
Significance level		α (alpha)	
***p*-value**		*p*-value	Pr(*) (the star will depend on what test is being used)

Name	Population Notation	Sample Notation	Notation used in R
Analysis of Variance (abbreviated ANOVA)			
Sum of squares for groups		SSG	Row labeled with the grouping variable, column labeled Sum Sq
Sum of squares for error		SSE	Row labeled Residuals, column labeled Sum Sq
Mean square for groups		MSG	Row labeled with the grouping variable, column labeled Mean Sq
Mean square error		MSE	Row labeled Residuals, column labeled Mean Sq
Regression			
Response (dependent) variable	y	y	(given by name of *y*-variable)
Predicted (estimated) response	E*(y)* (expected value of *y*)	$\hat{y}$ (*y*-hat)	
Explanatory (independent) variable	x	x	(given by name of *x*-variable)
y-intercept	β_o (beta-not)	b_o	B (look in the row labeled (Intercept))
Slope	β_1 (beta-one)	b_1	B (look in the row labeled with the name of the x-variable)
Coefficient of correlation		r	Values in Correlation Matrix
Coefficient of determination		r^2	Multiple-R Squared
Error terms vs Residuals	ε (error terms)	*e* (residuals)	Unstandardized residuals

Supplement 3: Name That Scenario

The first thing to do in any research inference problem is determine what type of inference problem it is. This will help in deciding what procedure/formulas are appropriate to use. The following questions can help you determine the data scenario you are working with.

Please note, when answering, "How many variables are there?" do not count the variable which defines the populations (if there is more than one population).

□ How many populations are there?

One **Two** **More than two**

□ How many variables are there?

One **Two**

□ What type of variable(s)?

Categorical **Quantitative**

□ Then use the following table to determine which type of inference would be appropriate for this scenario.

Note the corresponding parameter is in parentheses, where appropriate.

		Number of Populations		
Number of Variables and Type		**One**	**Two**	**More Than Two**
One	**Categorical**	□ 1-sample inference for population proportion (p) (Labs 3-4) □ Chi-square: Goodness of Fit (Lab 12)	□ 2 indep. samples inference for the difference between 2 population proportions ($p_1 – p_2$) □ Chi-square: Homogeneity (Lab 12)	□ Chi-square: Homogeneity (Lab 12)
	Quantitative	□ 1-sample inference for population mean (μ) (Lab 6) □ Paired samples inference for a population mean difference (μ_D) (Lab 7)	□ 2 indep. samples inference for the difference between 2 population means (μ_1 - μ_2) (Lab 8)	□ ANOVA (μ_i – where there is one μ_i for each population) (Lab 9)
Two	**Categorical** (relationship)	□ Chi-square: Independence (Lab 12)		
	Quantitative (relationship)	□ Regression (β_1) (Labs 10-11)		

Supplement 4: Interpretation Examples

In 1980, Bausch and Lomb Corporation developed a new type of extended-life contact lens made of silicone, which it claimed had a useful life of more than 4 years. During the research and development period, a random sample of 6 contact wearers was asked to wear the new contact lenses and record how long they lasted. The average useful life of the six pairs of lenses was 4.6 years, with a standard deviation of 0.49 years.

a. **Interpretation of the Standard Deviation *s*:**
An **estimate** of the **average distance** of the observed useful lives of these lenses **from their mean** useful life of 4.6 years is ***about*** 0.49 years.

Note: if given the true population standard deviation (σ) this becomes:

The **average distance** of the observed useful lives of these lenses **from their mean** useful life of 4.6 years is ***about*** 0.49 years.

b. Calculate the value of the **standard error of the mean**. $SE(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{0.49}{\sqrt{6}} = 0.200$

Interpretation:
The standard error is an **estimate** of the **average distance** of all the possible sample means from the true population mean (roughly). In context: An **estimate** for the **average distance** of xbar (sample averages of contact life from samples of size 6) from the **population mean useful life**, μ, *is **roughly*** 0.20 years.

c. Construct a **90% confidence interval** for the population mean life of all such silicone-based lenses:

$$4.6 \pm (2.015)(0.200) \Rightarrow (4.197, 5.003)$$

Interpretation of the <u>Interval</u>:
This interval provides a range of reasonable values for the population mean useful life, μ. We would estimate the **population mean** useful life, μ, to be between 4.197 years and 5.003 years, with 90% confidence.

Interpretation of the 90% Confidence <u>Level</u>:
If we repeatedly took new samples of the same size (computing new 90% confidence intervals each time), we would expect 90% of these resulting intervals to contain the population mean life, μ.

d. State the hypotheses to test the claim made by Bausch and Lomb about their new contact lens; that is, test if the population mean useful life is more than 4 years.
$H_0 : \mu = 4$, $H_a : \mu > 4$, with an observed t-test statistic of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.6 - 4}{0.200} \cong 3.00 .$$

The ***p*-value** for this test is the probability of getting a *t*-test statistic at least as extreme as the observed test statistic, **assuming the null hypothesis is true**. So we have the *p*-value = Prob(T >= 3 | H_0=True) found under the *t*(5) distribution. This *p*-value turns out to be equal to 0.015.

Interpretation of the value of the test statistic *t = 3.00* in terms of a distance: The observed sample mean was 3 average distances (i.e. 3 standard errors) above the hypothesized mean of 4. In other words, since the standard error for xbar was .2 it took 3 of them to get from 4 (value under null) to 4.6 (test statistic value)

Interpretation of the resulting *p-value of 0.015*: If the null hypothesis was true (the population mean useful life is just 4 years) and this procedure (study) was repeated many times, we would expect to see a *t*-test statistic value of 3.00 or larger in only 1.5% of the repetitions. Thus are data are somewhat unusual under the null hypothesis theory, providing evidence for the alternative theory that the population mean useful life is greater than 4 years.

e. At a 10% significance level, what is the **decision**?
Reject *H_0* since the *p*-value is less than 0.10.

f. What is the **conclusion**? There is sufficient evidence to conclude that the **population mean** useful life of the new lenses is greater than 4 years.

NOTE: These interpretations can be extended to the any test and confidence interval, adjusting for the different parameters, different directions of extreme, different test statistics, etc.

Supplement 5: Summary of the Main *t*-Tests

The three inference scenarios presented in Labs 6, 7, 8 are: **one-sample *t* procedures, paired *t* procedures, and two independent samples *t* procedures**. Data exploration is always essential to determining whether the model you want to use is appropriate. That is, we need to check the assumptions. (Recall that checking assumptions is the second step in performing a hypothesis test.)

The *t* procedures have the following general assumptions:

1. Each sample is a **random sample** – (the observations can be viewed as realizations of independent and identically distributed random variables). In the paired *t* procedures, the **differences** are assumed a random sample.

2. Each sample is drawn from a **normal population**, that is, the response variable has a normal distribution for each population. In the paired *t* procedures, the population of **differences** is assumed to have a normal distribution. In the two-sample case, both populations of responses are assumed to have normal distributions.

   You need normality of the underlying population for the response in order to have normality for the sample mean. In the case where you do not have a normal population, you can still have normality of the sample mean if you have a large enough sample size (most texts state that a sample size of *at least 25-30* is required). Thus we will accept *at least 25* as large enough to assume CLT holds for non-normal populations.

3. For the two independent samples *t* procedures, we also assume that the **two samples are independent**. We also need to assess whether the **two population variances** can be assumed **equal** in order to decide between the pooled and the unpooled *t* tests.

Graphical tools can be used to check these assumptions (see Labs 1 and 5 for more details about these various graphs).

Time Plots (or Sequence Plots): If your quantitative data have been gathered over time, then a time plot can be used to determine if the underlying process that generated that time dependent data ***appears*** to be stable. For example, in paired design problems we assume our set of differences calculated from the paired observations (d_1, d_2, ..., d_n) are a random sample. To check this, the values should be plotted by time to see if it is plausible that all values randomly came from one parent population. If that was the case the graph would be stable, with no patterns and constant mean/variance.

Remember:

#1 Time or Sequence plots are useful for checking stability only when the data are ordered in some sense. If there is no inherent order to the data, a sequence plot should not be made.

#2 If a Time plot makes sense to be examined and does show evidence of instability, it would not make sense to treat those observations as being a random sample; thus it would not be appropriate to make a histogram, QQ plot, or boxplot of the observations. No statistical procedure taught in this course is appropriate for non-stable data.

Histograms: Histograms are especially useful for displaying the distribution of a quantitative response variable. You could make a histogram of the observations in a one-sample problem, of the differences in a matched pairs design, and of each of the two samples separately in the independent samples design. Examine the histogram for evidence of strong departures from normality, such as bimodality or extreme outliers. Since you are just plotting data (just a sample and not the entire population of responses), your histogram may not look *perfectly* bell-shaped or normal.

QQ plots: QQ plots (or quantile plots or normal probability plots) are generally better than histograms for assessing if a normal model is appropriate. If the points in a QQ plot fall *approximately* in a straight line (with a positive slope) then the normal model assumption is reasonable.

Boxplots: Boxplots are most useful for assessing the validity of the assumption of **equality of population variances** in **the two independent samples design**. We would see if the IQRs (shown graphically by the length of the boxes) are comparable, and also compare the overall ranges. If they do have comparable lengths or sizes (they do not need to be lined up), then we have support that the equality of population variances assumption is reasonable. We would also want to compare the two sample standard deviations themselves, and Levene's test of equality of the two population variances may also be available.

Name that Scenario Practice for the Three T Tests:

Having just reviewed the three main *t*-test inference scenarios, you should understand the testing procedures and be able to interpret the results of a test. However, it is important to know when each scenario applies. Read each of the following inference scenarios and determine which of the three *t*-test procedures would be most appropriate: the one-sample *t*-test, the paired *t*-test, or the two-independent samples *t*-test.

1. A researcher is studying the effect of a new teaching technique for middle school students. One class of 30 students is taught using the new technique and their mean score on a standardized test is compared to the mean score of another class of 27 students who were taught using the old technique.

2. A company claims that the economy size version of their product contains 32 ounces. A consumer group decides to test the claim by examining a random sample of 100 economy size boxes of the product, since they have received reports that the boxes contain less than the 32 ounces claimed.

3. At some universities, athletic departments have come under fire for low academic achievement among their athletes. An athletic director decides to test whether or not athletes do in fact have lower GPAs. A random sample of 200 student athletes and a random sample of 500 non-athlete students are taken and their GPAs are recorded.

4. As part of a biology project, some high school students compare heart rates of 40 of their classmates before and after running a mile. They want to see if the heart rate of students their age is faster after running a mile than before, on average.

5. A hospital is studying patient costs; they decide to follow 500 surgery patients' hospital and medical bills for a year after surgery, and compare them to the estimated costs provided to the patients before surgery. They want to see if the estimated and actual costs are comparable on average.

6. A chemical process requires that no more than 23 grams of an ingredient be added to a batch before the first hour of the process is complete. An analyst feels that due to current settings more than 23 grams may actually be added. If the analyst is correct, the settings need to be altered and recent batches recalled. A random sample of 25 batches is obtained from the machine that is supposed to add the ingredient. The measurements are used to test the analyst's claim.

Supplement 6: Regression Output in R

There are several different pieces of output for regression. In this example, we will be using the **dentistry.Rdata** data set. In these models, the explanatory x variable is **DNA,** and the response y variable is **PLAQUE**.

In some situations, we may have many potential predictors of our response variable – here there is just one potential explanatory variable, DNA. To analyze the correlation potential predictors to our response variable, we can create a **Scatterplot matrix**. We see the following matrix for our variables here:

```
             DNA
PLAQUE
DNA          1.0000000
0.8557985
PLAQUE       0.8557985
1.0000000
```

This matrix shows us the **correlation coefficient**, r, for all pairs of variables. The correlation coefficient measures the strength of the linear association between the two variables. The closer it is to +1 or -1, the stronger the linear association.

We choose a pair by picking a column and row for each variable, and checking the value for that column and row pair. We can see that each pair is listed twice in the matrix (DNA-PLAQUE and PLAQUE-DNA), and that each variable is perfectly correlated with itself (r = 1). The main information here that we gather here is that the correlation of our model for predicting plaque using DNA is 0.856.

Next, we can generate our model, and R will give us a summary of the model, which looks like this:

```
Call:
lm(formula = PLAQUE ~ DNA, data = dentistry)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7639 -3.5107 -0.9454  4.0531  6.2532
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.54830    8.19299  -0.067  0.94829
DNA          0.16685    0.03566   4.679  0.00158 **
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual  standard  error:  4.851  on  8  degrees  of
freedom
Multiple R-squared:  0.7324,   Adjusted   R-squared:
0.6989
F-statistic: 21.89 on 1 and 8 DF,  p-value: 0.001584
```

The summary starts with a **Call** line – this just tells you what model you are looking at. Here, **PLAQUE ~ DNA** is saying that we are predicting PLAQUE using the explanatory variable DNA. Next, we see that R gives us some quartiles for our residuals. We might find the median residual especially useful, as having a median of -0.9454 here tells us that the majority of residuals are negative.

Next, we see the **Coefficients** table, which gives us a wealth of information. In this section, the least square estimates for the regression line are given. These estimated regression coefficients are found under the column labeled **Estimate**. The estimated slope is next to the independent variable name (in this example it is DNA), and the estimated intercept is next to **(Intercept)**. So, b_0 is the coefficient for the variable (Constant), and b_1 is the coefficient for the independent variable x in the model. The next column heading is **Std. Error**, which provides the corresponding **standard error** of each of the least squares estimates. Also produced in this table, are the ***t*-test statistics** in the column labeled **t value** and **Pr(>|t|)**, which reports the two-sided ***p*-values** for these *t*-test statistics.

In the last few lines of output, we get our standard deviation, R-squared, and F-statistic. The **Residual standard error** gives the value of *s*, the estimate of the population standard deviation σ. The next line gives two values, Multiple R-squared and Adjusted R-squared. We ignore the Adjusted and just look at the Multiple R-squared. This value, which is the square of the correlation has a useful interpretation in regression. It is often called the **coefficient of determination**, or r^2, and measures the proportion of the variation in the response that can be explained by the linear regression of y on x. Thus, it is a measure of how well the linear regression model fits the data. The final line is an F-statistic, which also gives us a way to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, but it only allows for a two-sided test. This F-statistic comes from an ANOVA table, which we can generate separately using **Model > Hypothesis tests > ANOVA table**. Make sure to select the **Type I** option to get the ANOVA table into a familiar format.

```
Analysis of Variance Table

Response: PLAQUE
          Df Sum Sq Mean Sq F value   Pr(>F)
DNA        1 515.14  515.14  21.894 0.001584 **
Residuals  8 188.23   23.53
```

We see that this ANOVA gives us the same F-statistic as before (F = 21.89). It also gives us some measures of variance within our model, the **Regression Sum of Squares** (SSModel = 515.14), and leftover residual variance, or the **Residual Sum of Squares** (SSRes = 188.23). We can use this to calculate r^2, or the proportion of variability in plaque that can be explained by its linear relationship with DNA, by taking the model variability and dividing by the total variability – r^2=515.14/(515.14+188.23) = .7324. Another value we can get again is an estimate of our total variability *σ,* or the **residual standard error**, by taking the square root of the MSRes = 23.53, much like we did for ANOVA to find the estimate of the pooled standard deviation.

Finally, the ratio of the Mean Squares provides the ***F*** statistic which tests if the slope is significantly different from zero (i.e. if there is a significant non-zero linear relationship between the two variables – H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.) The **Pr(>F)** is the corresponding *p*-value for the *F* test of these hypotheses. In simple linear regression, the 2-sided *t*-test in the **Coefficients** output for the slope is equivalent to the ANOVA *F*-test. Notice that the square of the *t*-statistic for testing about the slope is equal to the *F*-statistic in the ANOVA table, and the corresponding *p*-values are the same.

Interpretation of estimated slope b_1:
According to our regression model, we **estimate** that increasing DNA by one unit has the effect of increasing the predicted plaque by .167 units.

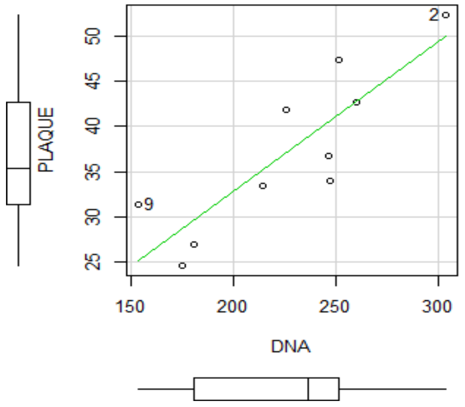
Interpretation of r^2:
According to our model, **73% of variation** in plaque levels can be accounted for by its **linear relationship** with DNA.

Decision for test of a significant linear relationship:
Since the *p*-value = .002 is less than the significance level α = .05, we can reject the null hypothesis that the population slope, β_1, equals 0.

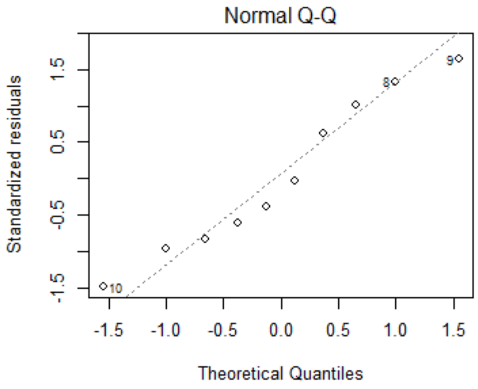
Conclusion: There is sufficient evidence to conclude that in the linear model for plaque based on DNA the population slope, β_1, does not equal zero. Hence, it appears that DNA is a significant linear predictor of plaque.

Checking the Simple Linear Regression Assumptions

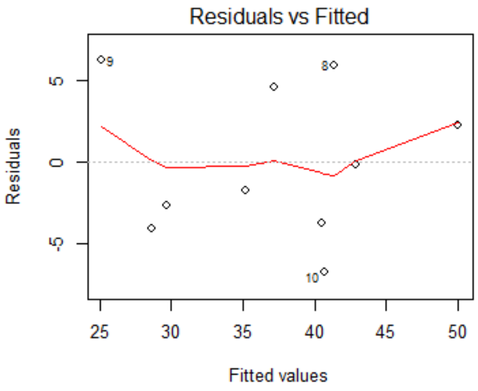
Here is a summary of some graphical procedures that are useful in detecting departures from the assumptions underlying the simple linear regression model.



1. **LINEARITY:** Do a scatter plot of *y* versus *x*.
   The plot should appear to be roughly linear.



2. **NORMALITY:** Examine a QQ plot of the residuals to check on the assumption of normality for the population (true) error terms. An example QQ plot is shown below.



3. **CONSTANT VARIANCE (or STANDARD DEVIATION)** of the population (true) error terms: Make a plot of the residuals versus *the fitted y values* ($\hat{y}$). This plot is called a **residuals vs fitted plot**. The residuals represent what is left over after the linear model has been fit. The residuals vs fitted plot should be a random scatter of points in roughly a horizontal band, with no apparent pattern. An example residuals vs fitted plot is shown at the right. Sometimes this plot can also reveal departures from linearity (i.e. that the regression analysis is not appropriate due to lack of a linear relationship).