Lesson 21: Multiple Linear Regression Analysis

Motivation and Objective: *We've spent a lot of time discussing simple linear regression, but simple linear regression is, well, "simple" in the sense that there is usually more than one variable that helps "explain" the variation in the response variable. Multiple Linear Regression (MLR) is an analysis procedure to use with more than one explanatory variable. Many of the steps in performing a Multiple Linear Regression analysis are the same as a Simple Linear Regression analysis, but there are some differences. In this lesson, we'll start by assuming all conditions of the Multiple Linear Regression model are met (we'll talk more about these conditions in Lesson 22) and learn how to interpret the output. By the end of this lesson, you should understand 1) what multiple regression is, and 2) how to use and interpret the output from a multiple regression analysis.*

What is Multiple Linear Regression?

Multiple Linear Regression is an analysis procedure to use when more than one explanatory variable is included in a "model". That is, when we believe there is more than one explanatory variable that might help "explain" or "predict" the response variable, we'll put all of these explanatory variables into the "model" and perform a multiple linear regression analysis.

Multiple Linear Regression Model

The multiple linear regression model is just an extension of the simple linear regression model. In simple linear regression, we used an "x" to represent the explanatory variable. In multiple linear regression, we'll have more than one explanatory variable, so we'll have more than one "x" in the equation. We'll distinguish between the explanatory variables by putting subscripts next to the "x's" in the equation.

Multiple Linear Regression Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_v x_v + \varepsilon$

where **y = an observed value of the response variable for a particular observation in the population**

β_0 = the constant term (equivalent to the "y-intercept" in SLR)

β_j = the coefficient for the j^{th} explanatory variable ($j = 1, 2, \ldots, v$)

x_j = a value of the j^{th} explanatory variable for a particular observation ($j = 1, 2, \ldots, v$)

ε = the residual for the particular observation in the population

In Simple Linear Regression, it was easy to picture the model two-dimensionally with a scatterplot because there was only one explanatory variable. If we had two explanatory variables, we could still picture the model: the x-axis would represent the first explanatory variable, the y-axis the second explanatory variable, and the z-axis would represent the response variable. The model would actually be an equation of a plane. However, when there are three or more explanatory variables, it becomes impossible to picture the model. That is, we can't visualize what the equation represents. Because of this, β_0 is not called a "y-intercept" anymore but is just called a "constant" term. It is the value in the equation without any "x" next to it. (It is often called a *constant term* in simple linear regression as well, but we can visualize what this constant term is in simple linear regression – it's the y-intercept!)

Question: If all the explanatory variables had a value of 0 and the residual of an observation is 0, what is the value of the response variable?

Answer: ?

Likewise, the numbers in front of the “x’s” are no longer slopes in multiple regression since the equation is not an equation of a line anymore. We’ll call these numbers *coefficients*, which means “numbers in front of”. As we will see, the interpretation of the coefficients (β_1, β_2, etc.) will be very similar to the interpretation of the slope in simple linear regression.

As with Simple Linear Regression, there are certain conditions that must exist in Multiple Linear Regression for conclusions from the analysis to be valid to a particular population of interest. Many of these conditions will be the same or similar as in Simple Linear Regression. We will talk about these conditions and checks of these conditions in Lesson 22. Even though it is important to make sure all of the conditions are met before doing an analysis, we’ll concentrate only on the analysis in this lesson under the *assumption* that all conditions are met. *(Note: this is backwards and is the ONLY time we’ll ever do an analysis without checking the conditions first, but it might be more interesting for all of us to see what the analysis is all about first.)*

Performing the Multiple Linear Regression Analysis

The following ActivStats tutorials discuss how to read the *Minitab* output from a Multiple Linear Regression Analysis. We’ll go through another example in detail explaining and expanding on certain aspects of the output. It is recommended to view the tutorials now and again after the completion of the example to follow.

ActivStats: Go to page 26-1 in the Lesson Book

View: Watch the Nambe Hills Story on Metalware Pieces

View: Learn to Read the Multiple Regression Table in MINITAB
(Note: you will learn HOW to use Minitab to do a MLR analysis in a Lab Activity)

View: Learn More About the Multiple Regression Table in MINITAB

ActivStats: Go to page 26-3 in the Lesson Book

View: Understand How the Values in the Table are Interrelated

Example 21.1: The Literacy Rate Example

Literacy rate is a reflection of the educational facilities and quality of education available in a country, and mass communication plays a large part in the educational process. In an effort to relate the literacy rate of a country to various mass communication outlets, a demographer has proposed to relate literacy rate to the following variables: number of daily newspaper copies (per 1000 population), number of radios (per 1000 population), and number of TV sets (per 1000 population). Here are the data for a sample of 10 countries:

Country	newspapers	radios	tv sets	literacy rate
Czech Republic / Slovakia	280	266	228	0.98
Italy	142	230	201	0.93
Kenya	10	114	2	0.25
Norway	391	313	227	0.99
Panama	86	329	82	0.79
Philippines	17	42	11	0.72
Tunisia	21	49	16	0.32
USA	314	1695	472	0.99
Russia	333	430	185	0.99
Venezuela	91	182	89	0.82

Question: What is the response variable? What are the explanatory variables?

Answer:

Below is the *Minitab* output from a Multiple Linear Regression analysis.

```
Predictor                  Coef       SE Coef        T        P
Constant               0.51486      0.09368       5.50    0.002
newspaper copies       0.0005421    0.0008653     0.63    0.554
radios                -0.0003535    0.0003285    -1.08    0.323
television sets        0.001988     0.001550      1.28    0.247

S = 0.186455   R-Sq = 69.9%   R-Sq(adj) = 54.8%

                   Analysis of Variance
Source          DF       SS          MS         F        P
Regression       3     0.48397     0.16132    4.64    0.053
Residual Error   6     0.20859     0.03477
Total            9     0.69256
```

The multiple linear regression equation

The multiple linear regression equation is just an extension of the simple linear regression equation – it has an "x" for each explanatory variable and a coefficient for each "x".

Question: Write the least-squares regression equation for this problem. Explain what each term in the regression equation represents in terms of the problem.

Answer:

Interpretation of the coefficients in the multiple linear regression equation

As mentioned earlier in the lesson, the coefficients in the equation are the numbers in front of the x's. For example, the coefficient for x_1 (the number of daily newspapers) is 0.00054. Each "x" has a coefficient. How these numbers are determined is beyond the scope of this course. We'll trust the output to give us these values. But, we should understand what these values mean in the context of the problem. The interpretation of each coefficient will be very similar to the interpretation of the slope in simple linear regression, *with some subtle but important differences*.

Let's start with the interpretation of the coefficient for *newspaper copies* (x_1). Like the slope in simple linear regression, it tells us that we predict the literacy rate to increase by 0.00054 for every additional daily newspaper copy in that country (per 1000 people in the population). But, there is more. *To properly interpret the coefficient of daily newspaper copies, the other two variables can't be changing – only the number of daily newspaper copies increases by 1.* So, a way to interpret the coefficient of *number of daily newspaper copies* is as follows:

> ***For every additional daily newspaper copy per 1000 people in a population, literacy rate is predicted to increase by 0.00054, keeping the number of radios and TV sets the same.***

Although the above interpretation is technically correct, a better interpretation is as follows:

> ***For countries with the same number of radios and same number of TV sets per 1000 people in the population, literacy rate is predicted to be 0.00054 higher for every additional daily newspaper copy per 1000 people in the population.***

The idea with the second interpretation is that the number of radios and TV sets has to stay the same. So, if we had two countries that had the same number of radios and TV sets per 1000 people in the population but one of the countries had one more daily newspaper copy than the other country (per 1000 people in the population), we'd predict the literacy rate for that country with one additional newspaper copy to be 0.00054 more than the other country.

Let's try interpreting the coefficient of radios.

Question: Here is an interpretation of the coefficient of radios: *For countries with the same number of daily newspaper copies and same number of TV sets (per 1000 people in the population), literacy rate is predicted to be .00035 higher for every additional radio per 1000 people in the population.* Which of the following is true regarding this interpretation?

A) This is a correct interpretation of the coefficient of *radios*.
B) This is not a correct interpretation of the coefficient of *radios*. "higher" should be replaced with "lower".
C) This is not a correct interpretation of the coefficient of *radios*. You do not need to compare only countries with the same number of daily newspaper copies and TV sets.
D) This is not a correct interpretation of the coefficient of *radios*. You do not need the word "predicted" in the interpretation.
E) B, C, and D above.

Answer: ?

We'll leave the interpretation of the coefficient of TV sets for you to do on your own. There are a couple of ActivStats tutorials that summarize and illustrate what we've been discussing:

ActivStats: Go to page 26-2 in the Lesson Book

View: Learn How Regression Coefficients Change with New Predictor Variables

View: Understanding Removing the Linear Effects of a Variable

Confidence intervals for the coefficients in the multiple linear regression equation

As in simple linear regression, the coefficients in the regression equation are based on a *sample* of countries. Had we collected data on *all* countries, the coefficients may have been different. The hope is that the sample of countries is representative of all countries so that the coefficients in the equation are close to what they would be had we had data on all countries. If we wanted to estimate what the true coefficients are had we collected data on all countries, we could construct confidence intervals for each coefficient in the same fashion as was done in simple linear regression. Each confidence interval would give us a range of possible values (with a certain level of confidence) for the coefficient.

Let's see how it's done – we'll see how similar this is to simple linear regression. Let's construct a 95% confidence interval for β_3, the coefficient for TV sets.

As usual, a confidence interval is of the form of **best estimate** ± **margin of error**
The "best estimate" is the coefficient from the sample data (b_3). The "margin of error" = (t*)(SE(b_3)). In general:

Formula for confidence interval for a coefficient (β_i):

$$b_i \pm (t^*_{n-v-1})(SE(b_i))$$

Note 1: the degrees of freedom for the t critical value is the DFE in the Analysis of Variance table. (Recall, DFE = n – v – 1 where v = the number of explanatory variables)*
Note 2: the subscript “i” in the formula are for the specific explanatory variable. So, if we’re finding the confidence interval for TV sets, we’ll use the coefficient for TV sets (b_3) and the standard error for TV sets (SE(b_3)).

As usual again, we need three pieces of information to construct the bounds of the confidence interval, two of which can be found in the output: b_3 and SE(b_3) – both are highlighted in red in the output below:

```
Predictor                 Coef       SE Coef        T        P
Constant               0.51486     0.09368       5.50    0.002
newspaper copies       0.0005421   0.0008653     0.63    0.554
radios                -0.0003535   0.0003285    -1.08    0.323
television sets        0.001988    0.001550      1.28    0.247
```

The other piece of information is the t* critical value for a 95% confidence interval. To find this value, we need the degrees of freedom.

Question: What are the degrees of freedom for the t* value in this problem?

Answer:

Question: Determine the lower and upper bounds for the 95% confidence interval for β_3.

Answer:

Question: Which of the following is the best interpretation of the 95% confidence interval for β_3?

A] We’re 95% sure that literacy rate will either go down by .0018 or go up by .00578.
B] We’re 95% sure that a country’s literacy rate will change by -0.0018 to +0.00578.
C] For countries with the same number of daily newspapers and same number of radios (per 1000 people in the population), we’re 95% sure that a country’s literacy rate will change by -0.0018 to +0.00578.
D] For countries with the same number of daily newspapers and same number of radios (per 1000 people in the population), we’re 95% sure that the literacy rate will be between 0.0018 lower to 0.00578 higher for a country with 1 more TV set per 1000 people in the population.
E] We’re 95% sure that the literacy rate will be between 0.0018 lower to 0.00578 higher for a country with 1 more TV set per 1000 people in the population.

Answer:

The confidence intervals for the other two coefficients will be left for you to do. Remember to use the proper point estimate and standard error. For example, to find the bounds for a confidence interval for β_2, use b_2 = -0.00035 and SE(b_2) = 0.00033.

Using the multiple linear regression equation for prediction

One of the uses of a regression analysis is for prediction. Predicting using a multiple linear regression equation is just an extension of predicting with a simple linear regression equation. We just have to make sure to put the right values in for the right x's.

Question: Predict literacy rate for a country that has 200 daily newspaper copies (per 1000 in the population), 800 radios (per 1000 in the population), and 250 TV sets (per 1000 in the population).

Answer: ?

Determining a final model – how to choose "significant" predictors of the response variable

Another reason for performing a multiple linear regression analysis is to determine which (if any) of the explanatory variables are significant predictors of the response variable. Typically, researchers may include explanatory variables they *think* are useful predictors of the response variable (i.e. help to "explain" the response variable). But, are they? For example, in Example 21.1, researchers included three variables associated with mass communication as possible predictors of literacy rate. But, are all needed? That is, does each explanatory variable help explain some of the variation in the response variable *after accounting for the effects of the other explanatory variables in the model*? A two-step process will be used to answer this question.

Step 1: perform an F-test

The first step is to determine if *any* of the explanatory variables are significant predictors of the response variable. If none are, there is no need to continue the analysis. However, if at least one is, then we can continue with the analysis.

To determine if *any* of the explanatory variables are significant predictors of the response variable, an F-test is performed. The F-test in multiple regression tests a different hypothesis than in simple linear regression.

Hypotheses for the F-test in multiple linear regression.

Null hypothesis;

H_0: all the coefficients = 0 **or** **H_0:** $\beta_1 = \beta_2 = \ldots = \beta_v = 0$

This implies that none of the explanatory variables are significant predictors of the response variable.

Alternative hypothesis:

H_A: at least one coefficient is not 0 **or** **H_A: at least one** $\beta_i \neq 0$

This implies that at least one of the explanatory variables is a significant predictor of the response variable.

It is important to note that the alternative hypothesis is that *at least* one of the explanatory variables is a significant predictor of the response variable. So, if there is evidence to reject the null hypothesis from the F-test, it does NOT say that *all* of the explanatory variables are significant predictors. It just says that there is *at least* one that is – it won't tell us how many or which one(s), just that there is *at least* one that is a significant predictor of the response variable. To determine which one or ones, t-tests on each explanatory variable need to be performed. More on that later. Let's continue with the F-test.

Recall from Lesson 20, the F-statistic = MSM / MSE. The numerator degrees of freedom for the F-statistic = DFM (which equals v, the number of explanatory variables), while the denominator degrees of freedom = DFE (n – v – 1). The F-statistic for the F-test testing the null hypothesis given in the yellow box above is given in the multiple linear regression output and is highlighted in red below for the Literacy Rate example:

```
                        Analysis of Variance
Source              DF        SS          MS        F       P
Regression           3     0.48397     0.16132    4.64    0.053
Residual Error       6     0.20859     0.03477
Total                9     0.69256
```

Question: Verify that the F-statistic in the output above equals MSM / MSE.

Answer: ?

Question: What are the degrees of freedom for this F-statistic?

Answer: ?

The p-value for the F-test testing the null hypothesis in the yellow box above is also given in the output – it is highlighted in red in the output below:

```
                        Analysis of Variance
Source              DF        SS          MS        F       P
Regression           3     0.48397     0.16132    4.64    0.053
Residual Error       6     0.20859     0.03477
Total                9     0.69256
```

There are several other resources that can be used to determine the p-value for this F-test:

1) The F-distribution calculator – click on the link to get to this applet. See Lesson 20 on how to use the applet.
2) An ActivStats tutorial contains instructions on how to use their F-table to determine the p-value:

   ActivStats: Go to page 26-3 in the Lesson Book

   View and Do: Apply the F-table to Regression

   Note: the important part of this tutorial is towards the end where the narrator explains how to scroll through the table to get to the observed F-statistic for a given numerator and denominator degrees of freedom.
3) F-tables, which can be found in Lesson 20 on MyStatLab. See Lesson 20 on how to use the F-tables to approximate the p-value.

Since the p-value is given in the *Minitab* output, we'll use that to answer the question of interest. But, be able to use any one of the other methods listed above to find the p-value just in case it is not given in the Analysis of Variance table.

Question: State a conclusion in the context of the problem.

Answer: ? 🔊

As mentioned in the answer above (but is worth repeating once more), if the p-value from the F-test is less than 0.10, we should continue the analysis. But, if the p-value is greater than 0.10, then there is no evidence to indicate that *any* of the explanatory variables are significant predictors of the response variable and, therefore, there would be no need to continue to the next step. In the Literacy Rate example, we had suggestive (but weak) evidence that there could be at least one explanatory variable that is a significant predictor of the response variable. Therefore, we need to move to step 2.

Step 2: perform a t-test on each explanatory variable

One last time: only do this step if there was even the slightest evidence to reject the null hypothesis from the F-test. Rejecting the null hypothesis from the F-test is an indication that at least one of the explanatory variables helps to explain the response variable. To determine which one or ones are significant predictors, t-tests on each explanatory variable will be performed.

Hypotheses for the t-tests in multiple linear regression:

Null hypothesis

H_0: coefficient for a particular explanatory variable is 0

OR

H_0: $\beta_i = 0$, where i = 1, 2, …, v

This implies that the particular explanatory variable being tested does NOT help to explain the response variable after accounting for the effects of the other explanatory variables in the model.

Alternative hypothesis

H_A: coefficient for a particular explanatory variable is NOT 0

OR

H_0: $\beta_i \neq 0$, where i = 1, 2, …, v

This implies that the particular explanatory variable being tested does help to explain the response variable after accounting for the effects of the other explanatory variables in the model.

One comment about the hypotheses before continuing: notice how the hypotheses are written in words – both include a part that is underlined stating, "after accounting for the effects of the other explanatory variables in the model." In multiple regression, the coefficients and standard errors of the coefficients for each of the variables are determined based on the other explanatory variables being in the model. For example, the coefficient and standard error of the coefficient for *newspaper copies* is 0.00054 and 0.000865, respectively (see output). But, these values were determined based on having both *radios* and *TV sets* in the model. If one or both of those other explanatory variables was not in the model, the coefficient and standard error of the coefficient for *newspaper copies* may change. If the coefficient and standard error of the coefficient changes, then the t-statistic and p-value would also change, which could lead to a different conclusion about *newspaper copies*!! Therefore, when writing the hypotheses in words (and the conclusion) for the t-tests, it is critical to make sure we include the part, "after accounting for the effects of the other explanatory variables in the model."

Let's start with the first explanatory variable: *newspaper copies*.

H_0: $\beta_1 = 0$ which implies the number of daily newspaper copies in a country does not help to explain that country's literacy rate after accounting for the effects of the number of radios and the number of TV sets in the country.

H_A: $\beta_1 \neq 0$ which implies the number of daily newspaper copies in a country does help to explain that country's literacy rate after accounting for the effects of the number of radios and the number of TV sets in the country.

As in simple linear regression, $\text{t-statistic}_{DFE} = \dfrac{b_i - 0}{SE(b_i)}$.

Question: Calculate the t-statistic (with degrees of freedom) for *newspaper copies*. For your convenience, the output is given below:

```
Predictor                   Coef       SE Coef        T       P
Constant                 0.51486       0.09368     5.50   0.002
newspaper copies       0.0005421     0.0008653     0.63   0.554
radios                -0.0003535     0.0003285    -1.08   0.323
television sets         0.001988      0.001550     1.28   0.247
```

Answer: ?

The two-sided p-value for the t-tests testing the null hypothesis given in the yellow box above are given in the column titled “P” in the *Minitab* output. For example, the p-value for *daily newspaper copies* is 0.554. This p-value can also be found using the t-distribution calculator (remember to multiply the given p-value by two to get the two-sided p-value), or the t-table on page A-62 in the text. If the p-value is given in the output, use that. But, be ready to determine the p-value using the t-distribution calculator or the t-table if the p-value is not given in the output.

Question: Based on the p-value, which of the following is a correct conclusion about *daily newspaper copies*?

A] There is not enough evidence to indicate that the number of daily newspaper copies in a country is a significant predictor of that country’s literacy rate.

B] There is not enough evidence to indicate that the number of daily newspaper copies in a country is a significant predictor of that country’s literacy rate, keeping the number of radios and TV sets the same.

C] There is not enough evidence to indicate that the number of daily newspaper copies in a country is a significant predictor of that country’s literacy rate, after accounting for the effects of the number of radios and number of TV sets in the country.

D] There is some evidence to indicate that the number of daily newspaper copies in a country is a significant predictor of that country’s literacy rate, after accounting for the effects of the number of radios and number of TV sets in the country.

E] There is strong evidence to indicate that the number of daily newspaper copies in a country is a significant predictor of that country’s literacy rate, keeping the number of radios and TV sets the same.

Answer: ?

It will be left as an exercise for you to do the t-tests for the other two variables. But, in looking at the output below, the t-statistic for *radios* is -1.08 with a two-sided p-value of 0.323 and the t-statistic for *TV sets* is 1.28 with a two-sided p-value of 0.247. (All t-statistics have 6 degrees of freedom.)

```
Predictor                   Coef       SE Coef        T       P
Constant                 0.51486       0.09368     5.50   0.002
newspaper copies       0.0005421     0.0008653     0.63   0.554
radios                -0.0003535     0.0003285    -1.08   0.323
television sets         0.001988      0.001550     1.28   0.247
```

Based on these p-values, it doesn’t appear that any of the explanatory variables are significant predictors of literacy rate! So, even though the p-value from the F-test indicated suggestive (but weak) evidence that there was at least one explanatory variable that was a significant predictor of literacy rate, the t-tests indicated that none of them were! But (and here’s the important part), that’s *after accounting for the effects of the other explanatory variables!!* If one of the explanatory variables was not in the model, we might have different conclusions about the other two variables! But, how do we know without removing one of the variables? And, if we're going to remove a variable, which one do we remove? That's all part of the next topic in determining a final model.

Step 3: Backwards selection process

An objective in multiple regression is to determine the predictors (i.e. explanatory variables) that accurately describe what happens with the response variable. That statement is pretty vague, and is meant to be. The idea is to determine a "best-fitting" model. But, what's "best-fitting" may depend on the problem. One part of determining a "best-fitting" model (but not necessarily the only part) is to determine which variables are significant predictors of the response variable. Researchers may include a number of explanatory variables in a model because they *think* all of them are significant predictors. But, they may not be. There are a number of "selection" methods (called *stepwise* methods) that are used to include only significant predictors in a final model. The one that we'll concentrate on here is called the **backwards selection process**.

In a backwards selection process, all explanatory variables are included in the initial model. Then the "least significant" explanatory variable is removed from the model and the model is re-fit with the remaining explanatory variables. Again, the least-significant explanatory variable is removed and the model is re-fit with the remaining explanatory variables. This process of removing one explanatory variable at a time is continued until all remaining explanatory variables are "significant" predictors of the response variable.

What does "least significant" and "significant" mean? That is somewhat subjective, but the guidelines we'll use here is that the least significant explanatory variable is the one with the highest p-value from the t-test. A variable is "significant" if its p-value from the t-test is less than 0.05 (or so). Putting it altogether, the backwards selection process will remove the explanatory variable with the highest p-value from the t-test *as long as its p-value is greater than 0.05 (or so)*. Then we'll refit the model with the remaining explanatory variables. (Remember, when one explanatory variable is removed, the coefficients and standard error of the coefficients may change, which will change the t-statistic and the p-value.) We'll continue to remove the least significant explanatory variable (one at a time) until all remaining explanatory variables have p-values less than 0.05 (or so). *Note: the "or so" is included because 0.05 is an arbitrary set cut-off point. We may decide to keep an explanatory variable with a p-value of 0.06, for example.*

Let's see how this works in practice with the Literacy Rate example. Here is the relevant output:

```
Predictor                  Coef      SE Coef       T       P
Constant                0.51486      0.09368    5.50   0.002
newspaper copies        0.0005421    0.0008653  0.63   0.554
radios                 -0.0003535    0.0003285 -1.08   0.323
television sets         0.001988     0.001550   1.28   0.247
```

Question: Which of the following is true regarding what to do first using the backwards selection process?

A) Remove *newspaper copies* since it has the t-statistic closest to 0.
B) Remove *radios* since it has a negative coefficient.
C) Remove *newspaper copies* since it has the highest p-value from the t-test.
D) Remove *TV sets* since it has the lowest p-value from the t-test.
E) Remove all of them since none of them are significant predictors of literacy rate.

Answer: ? 🔊

Question: True or false? Suppose all p-values from the t-test are much less than 0.05. We would still remove the explanatory variable with the highest p-value.

Answer: ?

After removing a variable, run the analysis again with the remaining variables and do the t-tests on the explanatory variables. Here is the relevant output after removing *newspaper copies* from the analysis. Note how the coefficients, standard error of the coefficients, t-statistics, and p-values have changed.

```
Predictor                Coef      SE Coef       T       P
Constant               0.53008    0.08646      6.13   0.000
radios               -0.0004736   0.0002548  -1.86   0.105
television sets        0.0027812  0.0008551   3.25   0.014
```

Question: What are the degrees of freedom for the t-tests in the above output?

Answer: ?

Question: Which variable would get removed in the backwards selection process?

A] *radios* since its p-value is the highest and it's greater than 0.05.
B] *TV sets* since its p-value is less than 0.05.
C] Both variables since at least one has a p-value greater than 0.05.
D] neither since at least one p-value is less than 0.05.

Answer: ?

So, we'll remove *radios* from the model since it has the highest p-value AND its p-value is greater than 0.05. We'll refit the model with only *TV sets* and run the analysis. Below is the output:

```
Predictor                Coef      SE Coef      T       P
Constant              0.56790     0.09606    5.91   0.000
television sets       0.0013886   0.0004710  2.95   0.018

S =            R-Sq =

               Analysis of Variance

Source          DF       SS          MS        F       P
Regression       1     0.36065    0.36065    8.69   0.018
Residual Error   8     0.33191    0.04149
Total            9     0.69256
```

Question: Which of the following is true?

A] *TV sets* would remain in the model because we always need to have at least one explanatory variable in the model.
B] *TV sets* would remain in the model since its p-value is less than 0.05.
C] *TV sets* would be eliminated from the model since it has the highest p-value from the remaining explanatory variables.

Answer: ?

At long last, we have a final model. Once a final model is obtained, the output from that model can be used to answer some questions. (Such questions will be saved for the extra practice in the What You Need to Know for this lesson.) Here are a couple of final comments on the backwards selection process:

- There may be situations where none of the variables get eliminated using the backwards selection process. If all variables have p-values less than 0.05, they all stay in the model!
- In the Literacy Rate example, we are left with a model that has only one significant predictor of literacy rate. Thus, this has become a simple linear regression problem. Not all problems will reduce to a simple linear regression problem using the backwards selection process – that just happened to occur in this problem.
- If the p-value from the F-test indicates that at least one explanatory variable is a significant predictor of the response variable, then at least one explanatory variable should end up remaining in the model when doing the backwards selection process. (If the p-value from the F-test was high, indicating no evidence that any of the explanatory variables help predict the response variable, a backwards selection process would eliminate all the variables. Thus, if the p-value from the F-test is high, there is no need to continue the analysis since all variables would be eliminated in a backwards selection process anyway.
- The degrees of freedom for the t-test will change every time an explanatory variable is eliminated from the model. (Make sure you understand why.)
- The backwards selection process is only one of many ways to determine a final model. We will not go into the details of the others and the backwards selection process is the only one you will be responsible to know. However, you should read the author's notes on selecting the best multiple regression model (page 827 and top of page 828 in the text).
    - One comment that should be emphasized here is that we want to try to determine the best model with as few explanatory variables as possible. We do want to include enough so that we are explaining a good proportion of the variation in the response variable (i.e. high R-square), but having too many explanatory variables in a model makes it hard to interpret and understand.
    - On the same note, whenever we add another explanatory variable to the model, the R-square automatically goes up. Well, if the explanatory variable does not explain anything about the response variable, R-square would stay the same. The point is, R-square will not go down when another variable is added to the model. This may seem good, but not necessarily. If the variable added to the model doesn't explain too much more of the variation in the response variable (i.e. R-square doesn't increase too much), it may not be an important variable to keep in the model.

Summary

Multiple Linear Regression is the analysis to use when the response variable is quantitative and there is more than one explanatory variable. Much of the analysis is similar to simple linear regression. The major differences between simple and multiple regression is what is being tested with the F-test and with the t-test. In multiple regression, the F-test is like an initial screening – it will tell us if at least one of the explanatory variables is a significant predictor of the response variable and, therefore, whether we need to continue the analysis or not. If the conclusions from the F-test tell us that there's evidence that at least one explanatory variable helps to explain the response variable, then we do a t-test on each explanatory variable to determine if that explanatory variable helps explain the response variable *after accounting for the effects of the other explanatory variables in the model*. This last part is important – if one of the explanatory variables was not in the model, conclusions about the remaining explanatory variables may change. We use this idea when using the backwards selection process to find a model that includes only significant predictors of the response variable – the backwards selection process removes the explanatory variable with the highest p-value from the t-tests *as long as its p-value is greater than 0.05 (or so)*. The process continues until all remaining explanatory variables have p-values less than 0.05, or so. Such explanatory variables are included in the final model and an analysis is performed on this final model.

All of what is done in a multiple linear regression analysis relies on certain conditions being met. We ignored these conditions in this lesson, but we'll talk about these conditions and doing a complete multiple regression analysis in the next lesson.