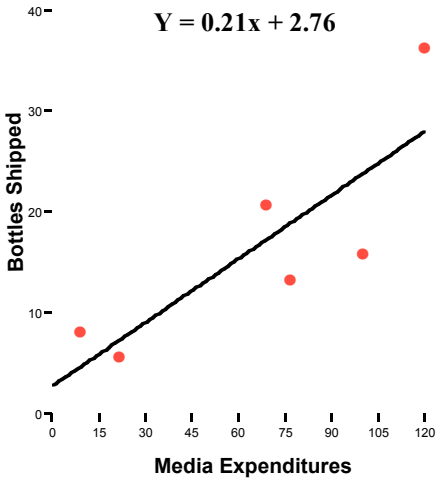
Activity #14: More Simple Linear Regression Section x.x

Resources: beer.sav

Recall the last activity in which we calculated a least-squares regression line and some indices of accuracy. The following table and graph display the relationship between a beer brand's media expenditures and shipment volumes.

Brand	Media Expenditure	Shipment
Busch	8.7	8.1
MGD	21.5	5.6
Bud Light	68.7	20.7
Coors Light	76.6	13.2
Miller Lite	100.1	15.9
Budweiser	120.0	36.3
Mean	**65.933**	**16.633**
Std. Deviation	**43.5017**	**11.0471**
Correlation	**0.829**	
Source: *Superbrands 1998*; 10/20/97		



1) When we conduct a linear regression analysis, we are often interested in finding the most parsimonious model (the simplest model that can explain the variance in the dependent variable. Even though our regression equation:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (Y = 0.21x + 2.76) appears to be relatively simple, it can be broken down into even simpler models.

(1) The simplest model would be one in which we have no model: $\hat{Y}_i = 0$

(2) If that model were found to be inadequate, we could try making the model a bit more complex. We could create a model in which each value of Y is predicted by a single number, b: $\hat{Y}_i = \beta_0$. We would then determine whether or not this model provided a significantly better prediction of Y than the first model.

(3) We could then try an even more complex model in which each value of Y is predicted by a constant (the y-intercept) and a slope: $\hat{Y} = \beta_0 + \beta_1 x$. If this model provided us a more accurate prediction, we may decide to throw out the simpler models and use this full model.

(4) Finally, we could find another independent variable that may help explain variance in the dependent variable (in the beer example, the variable "cost per bottle" might also help predict the number of bottle shipped). If the addition of this second variable provided us with a better prediction, we may decide to use this two-variable model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. This process of adding another predictive variable and testing its impact on the accuracy of prediction could continue.

2) At each stage in building models, the value of adding complexity to the regression model predictive accuracy may be assessed through hypothesis testing procedures. This allows the researcher to decide which terms and independent variables belong in the final model.

We will later learn that this process may be reversed. We may begin with a full regression model and try to simplify that model by eliminating independent variables that add little to the accuracy of predicting the dependent variable.

Using the beer data, let's discover how to test the significance of each term in the regression model.

3) Let's use the simplest model, $\hat{Y}_i = 0$ to predict shipment volumes for our beer brands. How do we measure the accuracy of our prediction?

Least Squares Regression Line: **Y = 0**				
Observed		Predicted	Prediction Error	Squared Error
X *Media Expenditures*	**Y** *Bottles Shipped*	$\hat{Y}$	$Y - \hat{Y}$	$\left(Y - \hat{Y}\right)^2$
8.7	8.1	0	8.1	65.61
21.5	5.6	0	5.6	31.36
68.7	20.7	0	20.7	428.49
76.6	13.2	0	13.2	174.24
100.1	15.9	0	15.9	252.81
120.0	36.3	0	36.3	1317.69
			Sum	**2270.20**

4) Our next model: $\hat{Y}_i = \beta_0$ attempts to predict each value of Y by a single number. If this is our model, what single number will minimize the sum of squared deviations from Y? The following table evaluates the accuracy of this predictive model.

Least Squares Regression Line: **Y = (Mean of Y)**					
Observed		Predicted	Prediction Error	Squared Error	*SSregression*
X *Media Expenditures*	**Y** *Bottles Shipped*	$\hat{Y}$	$Y - \hat{Y}$	$\left(Y - \hat{Y}\right)^2$	$\overline{Y}^2$
8.7	8.1	16.633	-8.533	72.812	*276.657*
21.5	5.6	16.633	-11.033	121.727	*276.657*
68.7	20.7	16.633	4.067	16.540	*276.657*
76.6	13.2	16.633	-3.433	11.785	*276.657*
100.1	15.9	16.633	-0.733	0.537	*276.657*
120.0	36.3	16.633	19.667	386.791	*276.657*
			Sum	**610.192**	***1659.95***

5) The last column represents the sum of squared deviations that is not due to error (or the amount of variation that is due to the regression model). You should also notice that SSE + Ssreg = SSY (or the sum of squares for the "no model" model). This will always be the case. The variation in the dependent variable can always be partitioned into:

(1) the amount (or proportion) of variance due to the regression model
(2) the amount (or proportion) of variance due to error (not accounted for by the model)

This partitioning of SS is often summarized in an ANOVA table:

ANOVA					
Source	Sum of Squares	df	Mean Square	F	Sig.
β_0	1660	1	1660	13.61	.02
Error	610	4	122		
Total	2270	5			

6) Let's examine that ANOVA summary table. We already know what the Sums of Squares represent. SStotal represents the total amount of variation in Y (or the sum of the squared vertical distance between the Y values and the x-axis). SSE represents the amount of variance in Y not explained by the model. The degrees of freedom for each row is the number of values from which the SS is calculated. For example, SStotal is based on n=6 values. Likewise, SSregression is based on one value, the mean. SSE is based on N-1 df because one df is lost for the estimation of β_0.

Mean squares are calculated by dividing the SS by its respective df. The F-value (MSR) is then compared to a critical value with 1 df in the numerator and 4 df in the denominator. You can see this MSR is significant at a 0.02 level.

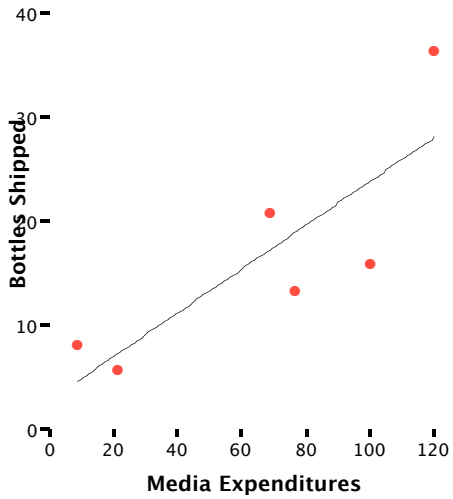
7) In a little bit, we will see why an F-test is used to check the significance of regression coefficients. Before we go on, I must state that we very rarely test for significance of β_0. The vast majority of measurements in a variety of fields are different from zero, so we just assume β_0 is significant. We will not test for the significance of β_0 in this class.

Another model with a single value: $\hat{Y} = \hat{\beta}_1 x_1$ could be compared to the "no model" case to see if it added significant predictive accuracy. Since we will always assume that β_0 is significant, we will not conduct this test either.

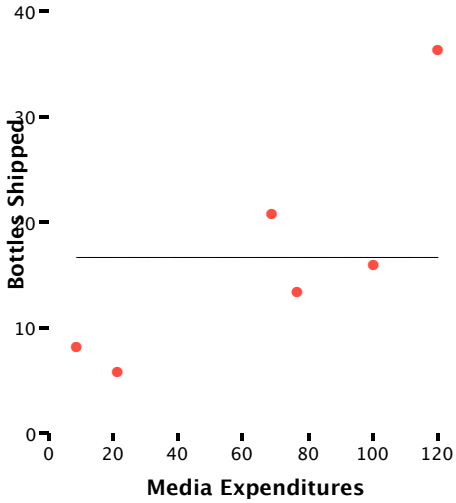
8) A much more common and interesting approach to testing hypotheses in simple linear regression is to examine the effect of adding the β_1 term to the model after the β_0 term has been entered. The general procedure is similar to the previous case. The predicted values of the full model are compared to the predicted values of the reduced model to find the increase in predictive power. The increase in predictive power is divided by error variance to find a ratio to test for additional predictive power.

In our example, we wish to compare the following two models:

Full model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$



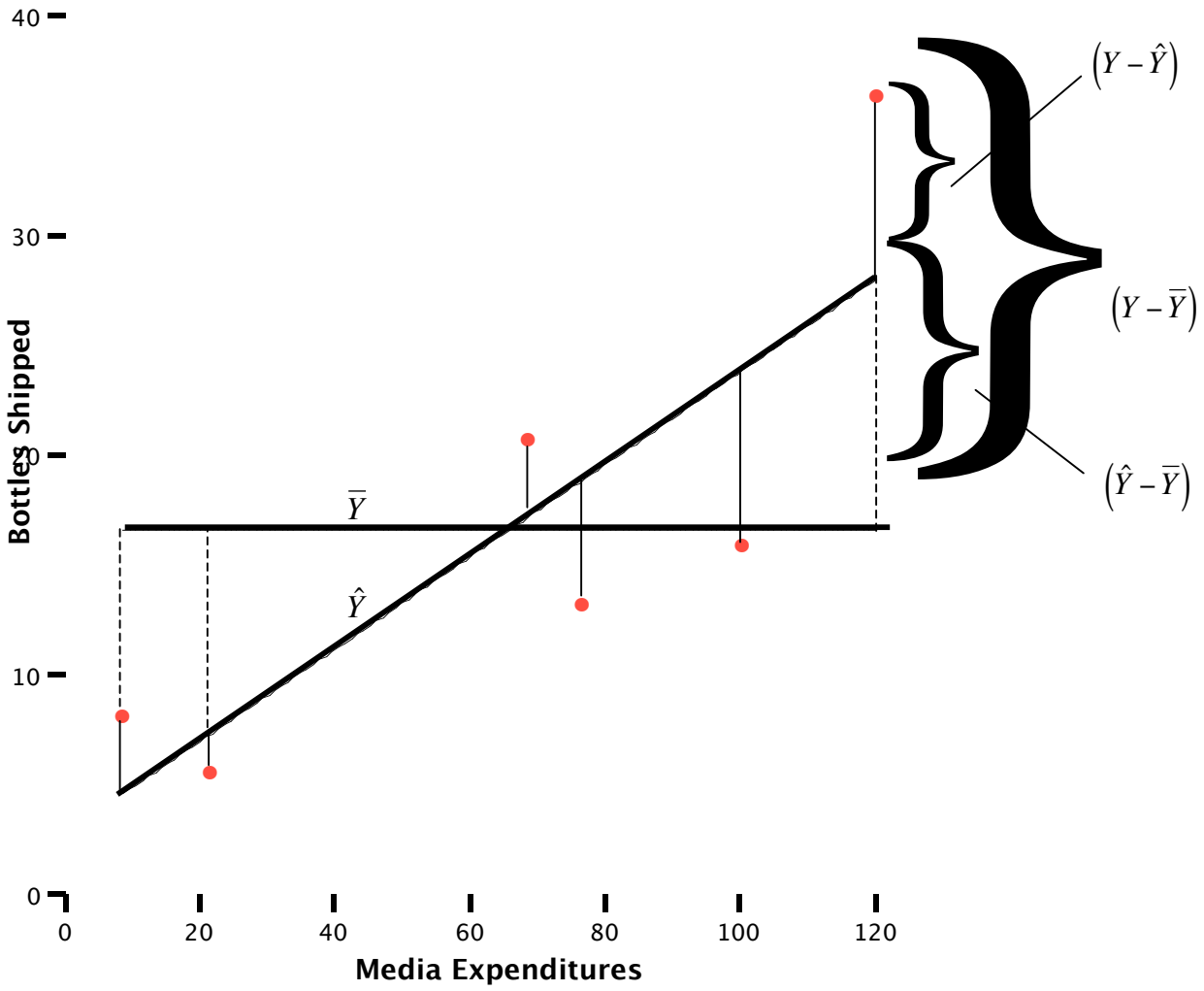
Reduced Model: $\hat{Y}_i = \beta_0$



It appears as though the full mode significantly reduces the total error in prediction. If we quantify the sums of squares, we can conduct a hypothesis test to see if this decrease in error is statistically significant. The table on the top of the next page compares the predictive accuracy of the full and reduced models:

	Full Model			Gain in accuracy
Observed	Predicted	Error	Squared Error	Over reduced model

To help visualize the columns of this table, take a look at the following display. Identify each column of the table in the display and express what each column represents.



9) The first five columns should be familiar to you, with the fifth column representing SSE (the sum of squared residuals). Column 7 represents the increase in predictive power of the full model over the partial model. This is what we call SSreg (the increased amount of variability predicted by the full model). This SSreg is often written mathematically as: $SS_{\beta_{1|0}}$ ("the sum of squares of β_1 given β_0").

You should note that SSreg and SSE for the full model sum to SSE for the partial model; ignoring errors due to rounding. In our example, $417 + 191 \approx 610$. This means that the variability in Y that cannot be predicted by the reduced model is partitioned into two parts: (1) that which can be predicted by the addition of the X term to the model and (2) that which cannot be explained by X.

Everything we've done can be summarized in an ANOVA table:

ANOVA					
Source	Sum of Squares	df	Mean Square	F	Sig.
β_0	1660	1	1660	13.61	.02
Error β_0	610	4	122		
$\beta_{1\|0}$	417	1	417	8.73	.04
Error (Full)	191	4	47.75		
Total	2270	5			

The significant F-ratio for the full model (at an alpha-level of .05) indicates that the independent variable does significantly improve our prediction accuracy.

10) I realize these concepts are difficult to understand all at once. Right now, it is my hope that you are able to follow the logic behind significance testing in a linear regression analysis. We will go over the specific steps of these hypothesis tests in a bit. First, let's see if we can figure out why we use F-ratios to test the significance of factors in our predictive models.

The following page lists the summary statistics we can calculate from our dataset. I recommend that you verify these statistics on your own time. For now, just explain what each statistic represents.

$\overline{X} = \frac{1}{n}\sum x_i = 65.933$	$\overline{Y} = \frac{1}{n}\sum y_i = 16.633$
$S_x^2 = \frac{\sum\left(X - \overline{X}\right)}{n-1} = 1892.397$	$S_y^2 = \frac{\sum\left(Y - \overline{Y}\right)}{n-1} = 122.038$
$S_x = \sqrt{\frac{\sum\left(X - \overline{X}\right)}{n-1}} = 43.5017$	$S_y = \sqrt{\frac{\sum\left(Y - \overline{Y}\right)}{n-1}} = 11.0471$
$r_{xy} = \frac{1}{n-1}\sum \frac{(x_i - \overline{X})(y_i - \overline{Y})}{s_x s_y} = 0.829$	$R_{xy}^2 = \frac{SS_{reg}}{SSY} = \frac{SSY - SSE}{SSY} = 0.687$
$S_{Y\|X}^2 = \frac{\sum\left(Y - \hat{Y}\right)^2}{n-2} = \sigma_Y^2\left(1 - R^2\right)\left(\frac{n-1}{n-2}\right) = 47.756$	$S_{Y\|X} = \sqrt{S_{Y\|X}^2} = S_Y\sqrt{1-R^2}\sqrt{\frac{n-1}{n-2}} = 6.9106$
$1 - R_{xy}^2 = \frac{SSE}{SSY} = 0.313$	N = 6
$\hat{\beta}_1 = \frac{S_y}{S_x} r_{xy} = 0.21$	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 2.67$
$SSY = \sum\left(Y - \overline{Y}\right)^2 = (n-1)S_Y^2 = 610$	$SSE = \sum\left(Y - \hat{Y}\right)^2 = (n-2)S_{Y\|X}^2 = \left(1 - r^2\right)SSY = 191$
$SS_{reg} = \frac{SSY - SSE}{SSY} = \sum\left(\hat{Y} - \overline{Y}\right)^2 = \left(R^2\right)SSY = 417$	You can see most of these statistics are related (just rearrange terms)

11) Now suppose we wish to conduct a hypothesis test to see if β_1 is significantly different from zero (to see if it should be included in our prediction model). What general form do our test statistics take? Write out the formula for the test statistic of interest.

12) Let's rearrange some of the terms in this test statistic. See if you can follow along:

$$t_{n-2} = \frac{\hat{\beta}_1 - 0}{\frac{S_{Y|X}}{S_x\sqrt{n-1}}} = \frac{\frac{S_y}{S_x} r_{xy}}{\frac{S_y\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}}{S_x\sqrt{n-1}}}$$

So far, we've just substituted other formulas for several of the terms.

$$= \frac{\frac{S_y}{S_x} r_{xy}}{\frac{S_y\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}}{S_x\sqrt{n-1}}} = \frac{\frac{S_y}{S_x} r_{xy}\left(S_x\sqrt{n-1}\right)}{S_y\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}} =$$

$$= \frac{\frac{S_y}{S_x} r_{xy}\left(S_x\sqrt{n-1}\right)}{S_y\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}} = \frac{r_{xy}\sqrt{n-1}}{\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}} =$$

$$= t^2_{n-2} = \frac{r^2(n-1)}{\frac{(1-r^2)(n-1)}{(n-2)}} = \frac{r^2(n-2)}{(1-r^2)} =$$

$$= t_{n-2} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r^2}} =$$

This is the test statistic for testing the significance of a correlation coefficient.

13) Let's stop here to stress that last statement. If you wish to test the significance of a correlation coefficient, you conduct a simple t-test with n-2 degrees of freedom. Let's test the significance of our correlation of 0.829.

14) The value of that test statistic should look familiar. It is the same value as the test statistic for testing the beta coefficient in a linear regression analysis. This is important. Since r and β_1 are so closely related (look at their formulas), we can use the same hypothesis test to test the significance of either statistic. But that still doesn't explain why we used an F-test to test the significance of the regression terms.

Let's look once again at that F-test we conducted. By rearranging terms...

$$F = \frac{SS_{reg} / df_{reg}}{SSE / df_E} = \frac{r^2(SSY)/df}{(1-r^2)(SSY)/df} = \frac{r^2/(n-2)}{(1-r^2)/(1)} = t^2_{n-2}$$

We can see that an F-test is equivalent to the square of a t-test. We know from our answer to #13 that our t-statistic was calculated as 2.96. Therefore, the F-test we calculated way back in #9 should be $(2.96)^2 = 8.76$. You can verify that we did indeed obtain that value (with some rounding error).

Our test statistic for testing the significance of a regression term or a correlation coefficient is:

$$F^{k_{full}-k_{reduced}}_{n-k_{full}-1} = \frac{\left(R^2_{full} - R^2_{reduced}\right)/\left(k_{full} - k_{reduced}\right)}{\left(1-R^2\right)/\left(N - k_{full} - 1\right)}$$

Where: k = # of independent variables
Full = full model (includes the term we want to test)
Reduced = reduced model (simplified model)
N = number of observations
R = correlation coefficient.

In the next activity, we will go through examples of linear regression analysis..