Activity #13: Simple Linear Regression

Resources: actgpa.sav; beer.sav; http://mathworld.wolfram.com/LeastFitting.html

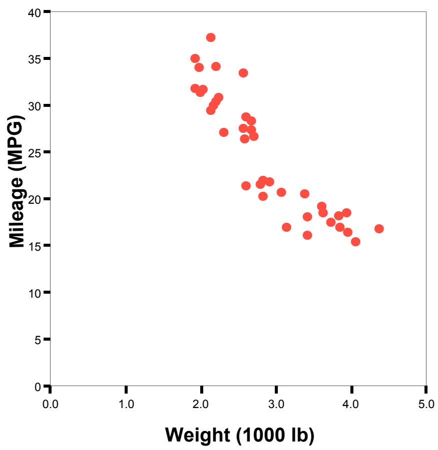
In the last activity, we learned how to quantify the strength of the linear relationship between two variables through correlation coefficients. This is important information, but we're generally more interested in modeling the relationship between two variables. In other words, we want to come up with a way to predict values of a dependent (outcome) variable once we're given a set of independent (predictor) variables.

Let's look at a couple simple examples before we begin.

1) After graduating from SAU, you're quickly hired by Toyota to help them design new cars. Toyota wants you to determine the gas mileage of cars based on their specifications (before they are actually built). Your job, then, is to complete two tasks:
    (1) Explain what factors influence mileage (and determine how much each factor contributes to mileage)
    (2) Use values of those factors to predict the mileage of pre-production automobiles.

   List several factors that influence the mileage of a car. Which factor do you believe has the biggest impact on mileage?

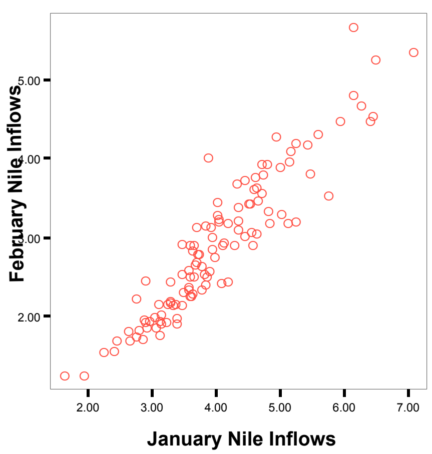
2) Now suppose you're only interested in modeling the relationship between the weight of a car and its mileage. You record the weight and mileage of 38 cars and create the following scatterplot:



What is the general relationship between the variables?

Could you use this scatterplot to make future mileage predictions?

3) Your textbook gives an example of predicting the amount of water that flows through the Nile River based on previous measurements:



Situation: Over 400 high school seniors apply to St. Ambrose each summer. The admissions office looks at each student's high school transcripts as well as their ACT scores in order to make an admissions decision. Students who are not predicted to have success at SAU (based on low ACT scores or poor transcripts) are not admitted to the University. We will attempt to predict an individual's first-year GPA based on that individual's ACT scores.

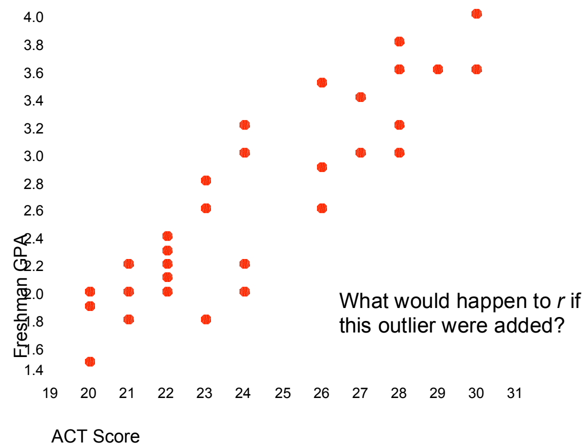
4) Suppose you only know that the average GPA for SAU freshmen is 2.9. What would be your best prediction for the GPA of an applicant who earned a score of 22 on the ACT? How about for an applicant with an ACT score of 29?

5) Without any knowledge of the relationship between ACT scores and GPA, our best guess is the overall mean GPA. Suppose we were given the GPAs of 30 former SAU students along with their ACT scores. The following table lists the *conditional mean GPAs* for these 30 students. Interpret what these conditional means represent. Based on this knowledge, what GPA would you predict for the applicant with an ACT score of 22? Do you believe this prediction will be more accurate than your previous prediction?

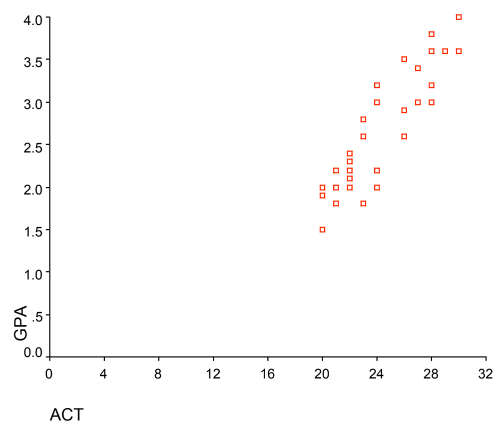
ACT Score	Average GPA
20 (n = 3)	1.8
21 (n = 3)	2.0
22 (n = 5)	2.2
23 (n = 3)	2.4
24 (n = 4)	2.6
25 (n = 0)	???
26 (n = 3)	3.0
27 (n = 2)	3.2
28 (n = 4)	3.4
29 (n = 1)	3.6
30 (n = 2)	3.8

6) Using the table of conditional means, predict the GPA of a student whose ACT score was 25. Make another prediction for a student whose ACT score was 17. Do you see any problems with this *conditional means* approach?

7) Let's use all the available information. The following scatterplot displays the relationship between a student's ACT scores and GPA for 30 former SAU students. Before we make our prediction, let's examine the relationship between these variables. Does it appear as though GPA and ACT scores have a linear relationship? Estimate the value of the correlation coefficient.



8) When two variables appear to have a linear relationship, we can describe the overall relationship by fitting a straight line through the points. You remember that y = mx + b, where m = slope and b = y-intercept, is the equation for a straight line. Carefully draw a straight line that best fits the data in the above scatterplot. By estimating the slope and y-intercept, make a guess as to the equation of your "best-fit" line. (The same scatterplot graphed in a different "window" has been provided to help you).



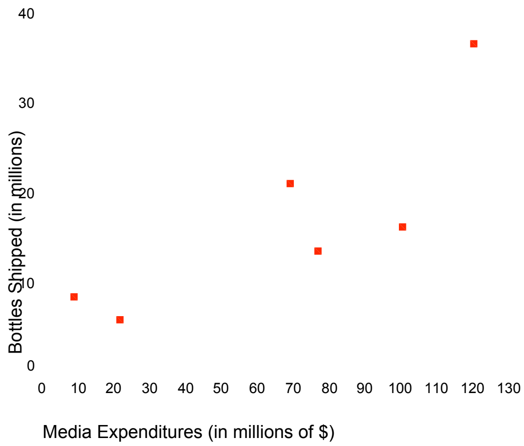
9) We would expect that no two people in this class drew exactly the same line through the data. How can we tell which line really does give us the "best fit?" Can you think of a way to measure how well a line fits a scatterplot of data?

:

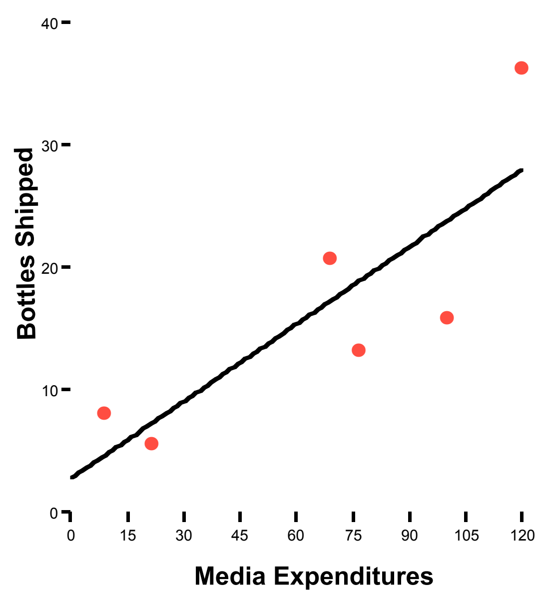
10) To illustrate how we determine the line that best fits the data, let's move on to a smaller data set. The following table displays the media expenditures (in millions of dollars) and the number of bottles shipped (in millions) for 6 major brands of beer. Would you expect a relationship between these two variables? If so, what kind of relationship would you expect?

Brand	Media Expenditure	Shipment
Busch	8.7	8.1
Miller Genuine Draft	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
Source: *Superbrands 1998*; 10/20/97		

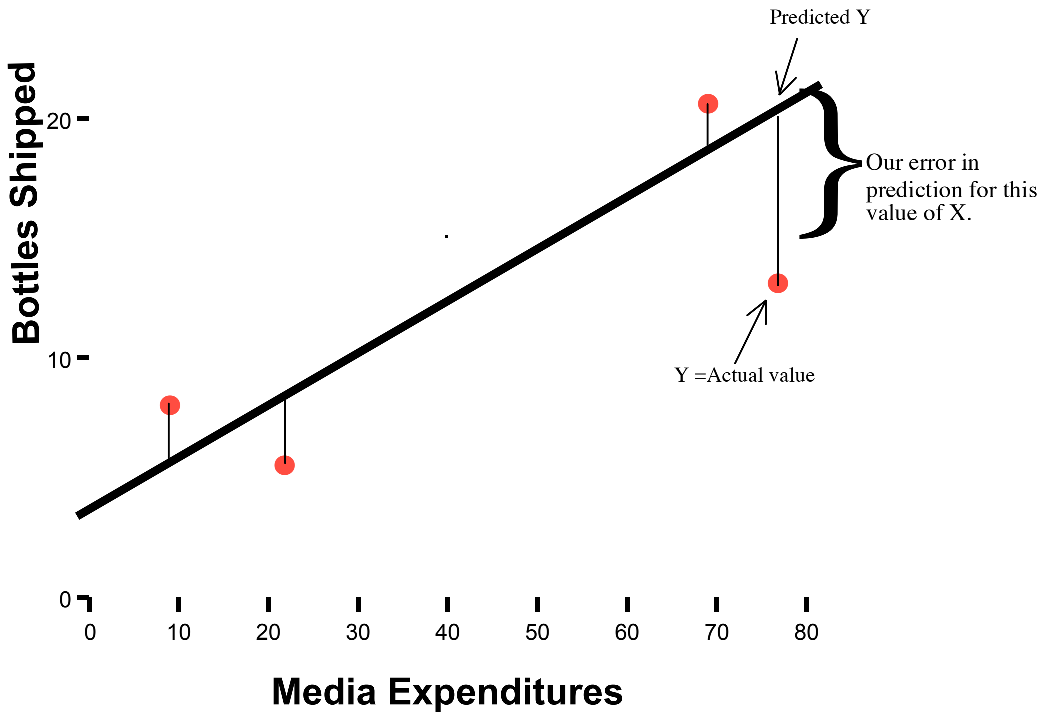
11) The following scatterplot displays the two variables for the six brands of beer. Does it appear as though the variables have a linear relationship? Estimate the correlation coefficient and draw a straight line through the data. Estimate the equation for that best-fit line.



12) I had SPSS calculate the regression line for this set of data. It turns out that $y = 0.21x + 2.76$ is the line of best fit. You can see this *least-squares regression line* on the following scatterplot. Does this line provide us with 100% accuracy in our predictions? Use this equation and/or the regression line to predict the shipment for a brand that spent $45 million on advertising. You should also use the equation to find the predicted shipment for a brand that spent $76.6 on advertising (which is exactly what Coors spent).



13) No regression line will provide us with perfect accuracy in prediction (there's always some amount of measurement error). In calculating the equation of the best-fit line, we try to minimize the total amount of distance between each data point and the prediction line (this distance will always be greater than zero, due to the fact that our prediction is never perfect). Since we use X (media expenditures) to predict Y (bottle shipped), we want a line that is as close as possible to the points in the vertical direction. The following graph displays the errors we wish to minimize:



14) Because some of the distances will be negative and some will be positive, we find the square of each distance (absolute values aren't used, since they are difficult to work with mathematically). **The *least-squares regression line* minimizes the sum of the squared errors**. Let's calculate the squared errors and see if the regression line really does minimize the vertical distances between the observed values and the predicted values. How do we calculate the predicted values of Y?

Lest Squares Regression Line: **Y = 0.21x + 2.76**				
Observed		Predicted	Prediction Error	Squared Error
X *Media Expenditures*	**Y** *Bottles Shipped*	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
8.7	8.1	4.587	3.513	12.341
21.5	5.6	7.275	-1.675	2.806
68.7	20.7	17.187	3.519	12.383
76.6	13.2	18.846	-5.646	31.877
100.1	15.9	23.781	-7.881	62.110
120.0	36.3	27.96	8.34	69.556
		Sum	0.17	**191.073**
		Expected value	0.00	← *Why?*

15) To demonstrate the fact that Y = 0.21x + 2.76 is the least-squares regression line, suppose we thought the best fitting line was Y = 0.3x + 2.0 (an arbitrary guess). The following table displays the predictions based on this line as well as the sum of squared errors of prediction. Is this prediction line more or less accurate than the least-squares regression line we found earlier?

Another Possible Prediction Line: **Y = 0.3x + 2.0**				
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	4.61	3.49	21.2521
21.5	5.6	8.45	-2.85	8.1225
68.7	20.7	22.61	-1.91	3.6481
76.6	13.2	24.98	-11.78	138.7684
100.1	15.9	32.03	-16.13	260.1769
120.0	36.3	38.00	-1.7	2.89
		Sum	-30.88	**434.858**

By definition, the least-squares regression line will minimize the sum of squared *residuals* $\left(Y - \hat{Y}\right)^2$

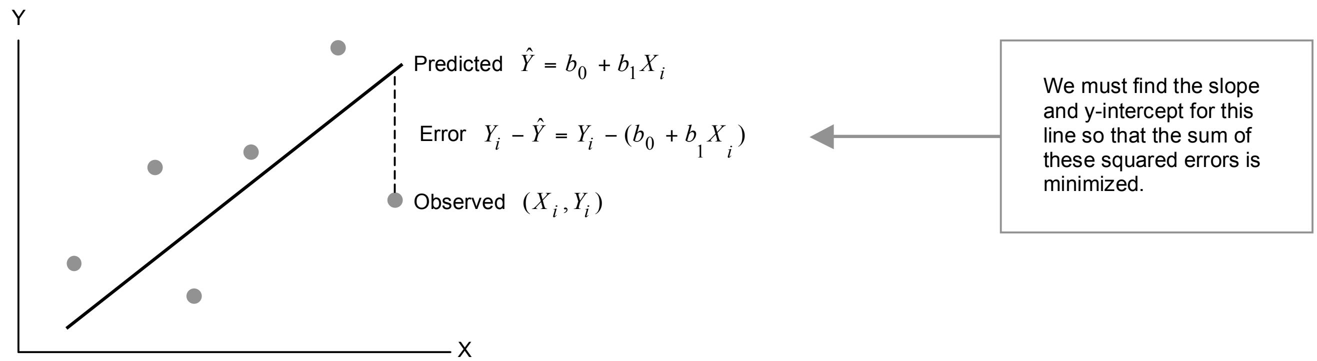
These *residuals* (or errors) represent the vertical distance between each observed Y value and the Y value predicted by our regression line.

16) As you know, the formula for a straight line is: Y = mx + b. In linear regression, the equation for the least-squares regression line is often written as: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Interpret the coefficients of this equation. The least-squares regression line for the beer data was found to be: Y = 0.21 +2.76. Interpret the coefficients for this specific regression line.

17) When we studied ANOVA, we wrote out formal models. We can do the same with linear regression. Each individual score, Y, is modeled by: $Y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{E}$ which reminds us that error (E) is equal to the distance between an observed value and the predicted value: $\hat{E} = Y - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right) = \left(Y - \hat{Y}\right)$

Now that we have a basic understanding of the geometric and algebraic properties of the least-squares regression line, you might wonder how we go about calculating the slope and y-intercept of the best-fitting line.

Derivation of the Parameters of the Least Squares Regression Line



Let Q represent the sum of squared errors: $Q = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$

We need to find values for b_0 and b_1 that will minimize Q. We know that to minimize a function, we must set its first derivative equal to zero and solve. Because we have two variables in this function, we'll need to take partial derivatives of Q with respect to b_0 and b_1.

Partial derivative of Q with respect to b_0: (we treat b_0 as a variable and all other terms as constants)

(Chain Rule)

$$\frac{\partial Q}{\partial b_0} = \frac{\partial \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2}{\partial b_0} = 2\sum (Y_i - b_0 - b_1 X_i) \frac{\partial (Y_i - b_0 - b_1 X_i)}{\partial b_0} = -2\sum (Y_i - b_0 - b_1 X_i)$$

We set this partial derivative equal to zero: $-2\sum (Y_i - b_0 - b_1 X_i) = 0$

$$\sum (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

Partial derivative of Q with respect to b_1: (we treat b_1 as a variable and all other terms as constants)

(Chain Rule)

$$\frac{\partial Q}{\partial b_1} = \frac{\partial \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2}{\partial b_1} = 2\sum (Y_i - b_0 - b_1 X_i) \frac{\partial (Y_i - b_0 - b_1 X_i)}{\partial b_1} = -2\sum X_i (Y_i - b_0 - b_1 X_i)$$

Set the partial derivative equal to zero: $-2\sum X_i (Y_i - b_0 - b_1 X_i) = 0$

$$\sum X_i (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

Now we must solve this system of two *normal* equations…

System of normal equations:
$$\sum Y_i = nb_0 + b_1 \sum X_i$$
$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

This system can be solved to get: $$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

and

$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \overline{Y} - b_1 \overline{X}$$

We can rewrite b_1 given the following information: $$S_{xy} = \sum (x_i - \overline{X})(y_i - \overline{Y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i$$

$$S_{xx} = \sum (x_i - \overline{X})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

$$S_{yy} = \sum (y_i - \overline{Y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

Therefore, $$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{S_{xy}}{S_{xx}} = r\frac{S_y}{S_x}$$

So, the line that minimizes the sum of squared errors has the following slope and y-intercept parameters:

$$b_0 = \overline{Y} - b_1 \overline{X} \quad \text{and} \quad b_1 = r\frac{S_y}{S_x}$$

In our example, r = 0.829; s_y = 43.5017; s_x = 11.0471. Using the mean values of X and Y, we can compute:

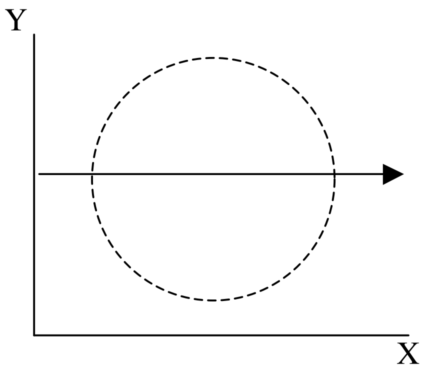
$$\hat{\beta}_1 = r\frac{s_y}{s_x} = (0.829)\left(\frac{11.0471}{43.5017}\right) = 0.21 \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 16.633 - (0.21)(65.933) = 2.76$$

Now that we can compute a regression line and interpret its coefficients, we need to find some way of measuring the accuracy of our prediction line. We already know that the least-squares regression line is the line of best fit (it minimizes the sum of squared residuals (oftentimes called $SS_{residual}$ or SSE)). What we don't know is whether or not that best-fitting line actually does a good job of fitting the data. (For example, imagine a scatterplot of two uncorrelated variables. The shape of the scatterplot would be a circle. We could fit the least-squares line to the data, but it still wouldn't fit the data very well.

18) The first index of accuracy we may want to evaluate is SSE, the sum of squared residuals $\left(Y-\hat{Y}\right)^2$. To evaluate how well SSE serves as an index of accuracy, let's calculate the maximum and minimum values of SSE.

The minimum value of SSE would occur when every observed value of Y falls upon the prediction line. If this is the case, there would be zero distance between each point and its predicted value. Therefore, when we have a perfect prediction, SSE = 0.

The maximum value of SSE would occur when we have uncorrelated variables (knowing the value of X would not tell us anything about the value of Y). The scatterplot of uncorrelated variables would look like a circle:



What would the least-squares regression line look like in this case? Well, we always want to minimize the sum of squared residuals.

Minimize: $\sum\left(Y-a\right)^2$ where *a* represents the predicted value of Y.

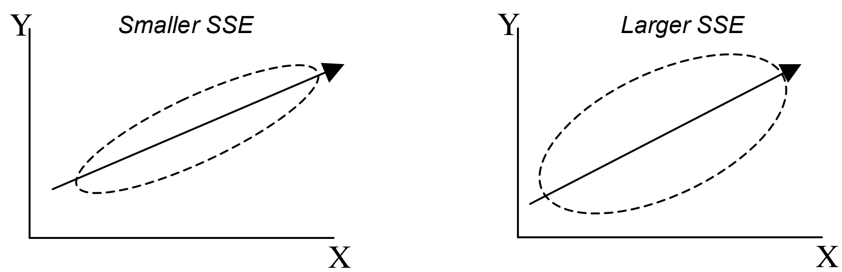
We know that this is minimized when a = the mean of Y (by definition, the mean is the value that minimizes the sum of squared deviations).

Therefore, the maximum value of SSE (minimum prediction accuracy) is:

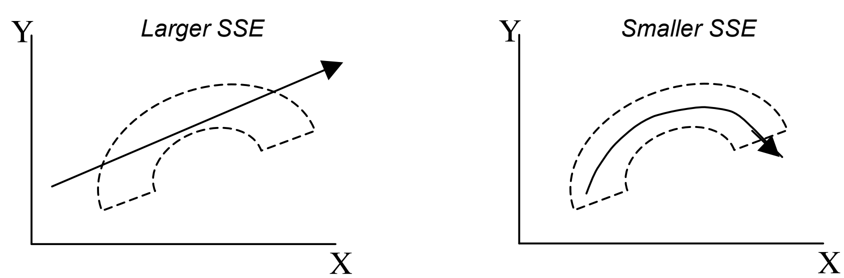
$\sum\left(Y-\overline{Y}\right)^2$ which we remember is called SSY or SS_{TOTAL} in ANOVA.

Some factors that influence the size of SSE are:

(1) Variation around the linear regression line



(2) Nonlinearity (appropriateness of a linear model).



Some problems with using SSE as an index of accuracy:

(1) It varies with *n* (adding observations almost always increases SSE). We'd like a per-observation index…

$$\frac{SSE}{n-2}=\frac{\sum\left(Y-\hat{Y}\right)^2}{n-2}=S_{Y|X}^2$$ *Variance of the estimate (variance of Y given X)*

(2) It's expressed in squared units. $\sqrt{\frac{SSE}{n-2}}=\sqrt{\frac{\sum\left(Y-\hat{Y}\right)^2}{n-2}}=S_{Y|X}$ *Standard error of estimate*

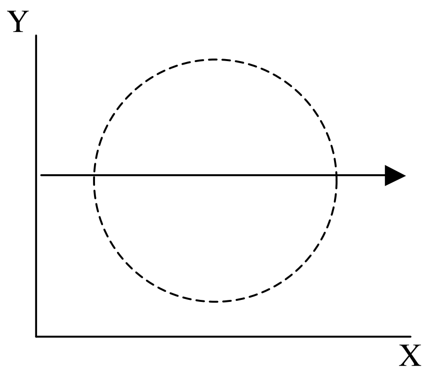
19) Let's now evaluate the merits of using the variance of the estimate as an index of the accuracy of our prediction.

$$S^2_{Y|X} = \frac{\sum\left(Y-\hat{Y}\right)^2}{n-2}$$

This represents the "average" squared vertical distance between predicted and observed values.

The minimum value of $S^2_{Y|X}$ would occur when we have 100% accuracy in prediction. It doesn't take too much thought to realize the value of $S^2_{Y|X}$ would be zero when each observed value lies on the regression line.

To see what the maximum value of $S^2_{Y|X}$ would be, let's once again look at a scatterplot of independent (uncorrelated) variables.



Recall that our least-squares regression line would minimize the sum of squared residuals. The residuals are minimized when:

$$SSE = \sum\left(Y-\overline{Y}\right)^2$$ (which represents the maximum value of SSE)

Therefore, the maximum value of $S^2_{Y|X}$ would be:

$$Max\left\{S^2_{Y|X}\right\} = \frac{\sum\left(Y-\overline{Y}\right)^2}{n-2} = \left(\frac{n-1}{n-2}\right)\left(\frac{\sum\left(Y-\overline{Y}\right)^2}{n-1}\right) = \left(\frac{n-1}{n-2}\right)S^2_Y$$

20) We already know the problem with $S^2_{Y|X}$ -- it's expressed in squared units. We can quickly evaluate the maximum and minimum values of the standard error of estimate, $S_{Y|X}$.

Since the standard error of estimate is just the square root of the variance of estimate, the minimum value will be zero. The maximum value, when we try to predict values of Y with an uncorrelated X, is:

$$Max\left\{S_{Y|X}\right\} = \sqrt{\frac{\sum\left(Y-\overline{Y}\right)^2}{n-2}} = \sqrt{\frac{n-1}{n-2}}S_Y$$

(where S_Y is the standard deviation of Y)

21) We can create other indices of accuracy by partitioning SSY. Remember that SSY represents the total sums of squares (or the sum of squared distances from the observed Y values to the mean of Y). First, let's look at:

$$\frac{SSE}{SSY} = \frac{\sum\left(Y-\hat{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2} = \left(1-r^2\right)$$

What does this ratio, which we will call $(1-r^2)$, represent?

22) What is the value of $(1-r^2)$ when we have 100% accuracy in prediction? What's the value of this index of accuracy when we have uncorrelated variables?

23) The index ($1-r^2$) is at a minimum when we have perfect prediction accuracy. Its maximum value occurs when we have no accuracy. This is opposite of what we would intuitively like to see. To "fix" this, we could find the value of r^2. What is the formula for r^2 and what does it represent?

$$r^2 = \frac{SSY - SSE}{SSY} = \frac{\sum\left(Y-\bar{Y}\right)^2 - \sum\left(Y-\hat{Y}\right)^2}{\sum\left(Y-\bar{Y}\right)^2} = \frac{SS_{reg}}{SSY}$$

24) We will see that these indices of accuracy are important statistics in linear regression analyses. The following table summarizes the properties of these indices:

Index	Highest Accuracy	Lowest Accuracy
$SSE = \left(Y-\hat{Y}\right)^2$	0	$SSY = \left(Y-\bar{Y}\right)^2$
$S_{Y\mid X}^2 = \frac{SSE}{n-2} = \frac{\sum\left(Y-\hat{Y}\right)^2}{n-2}$	0	$\left(\frac{n-1}{n-2}\right)S_Y^2$
$S_{Y\mid X} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum\left(Y-\hat{Y}\right)^2}{n-2}}$	0	$\sqrt{\frac{n-1}{n-2}}S_Y$
$\left(1-r^2\right) = \frac{SSE}{SSY} = \frac{\sum\left(Y-\hat{Y}\right)^2}{\sum\left(Y-\bar{Y}\right)^2}$	0	1.0
$r^2 = \frac{SSY-SSE}{SSY} = \frac{SS_{reg}}{SSY}$	1.0	0

25) In this course, we will use a computer to calculate regression lines and summary statistics. We must focus on evaluating the assumptions of linear regression, testing the significance of r and the regression coefficients, selecting regression models, and interpreting regression analyses. The assumptions for linear regression are similar to the assumptions we made in ANOVA:

(1) Existence: We have a Y distribution for each value of X (this assumption is often ignored)
(2) Independence: Y scores are statistically independent (there are no links among subject sampling procedures)
(3) Linearity: The subpopulation means all fall on a straight line (see the illustration in your textbook)
(4) Homoscedasticity: Subpopulations Y (or E) variances are equal for all X values.
(5) Normality: Y (or equivalent E) scores for each X value are normally distributed.

Some general comments about testing these assumptions and robustness:

(1) There are few clear-cut guidelines for determining the seriousness of assumption violations
(2) Seldom do we have equal sample sizes for X values (balanced designs are more robust)
(3) Tests for violations in assumptions often have low power (like the F-max test in ANOVA)
(4) Assumption tests also have assumptions that may or may not be met.

26) We end this activity with a look at some output produced by SPSS. A linear regression analysis was conducted on the “beer” dataset:

Descriptive Statistics			
	Mean	Std. Deviation	N
Bottles Shipped	16.633	11.0471	6
Media Expenditures	65.933	43.5017	6

Correlations			
	Bottles Shipped	Media Expenditures	Sig.
Bottles Shipped	1.000	0.829	.021
Media Expenditures	0.829	1.000	.021

					Change Statistics				
Model	R	R-Square	Adj. R-Square	Std Error of Estimate	R-Square Change	F-Change	df1	df2	Sig. F-Change
1	.829[a]	.687	.609	6.9106	0.687	8.777	1	2	.041
a: Predictors = Media Expenditures. Dependent = Bottles Shipped									

ANOVA					
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	419.169	1	419.169	8.777	.041
Residual	191.024	4	47.756		
Total	610.193	5			

Coefficients							
	Unstandardized		Stndrdzd			95% CI	
Model	B	Std Error	Beta	t	Sig.	Lower	Upper
(Constant)	2.756	5.468		.504	.641	-12.426	17.938
Media Expend.	.210	.071	.829	2.963	.041	.013	.408