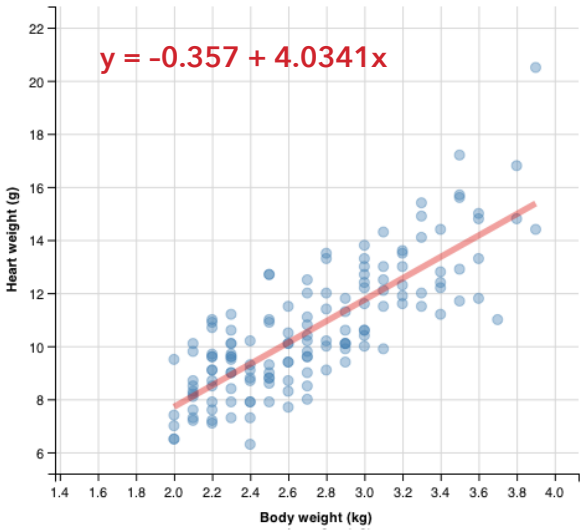
Activity #10: Simple linear regression



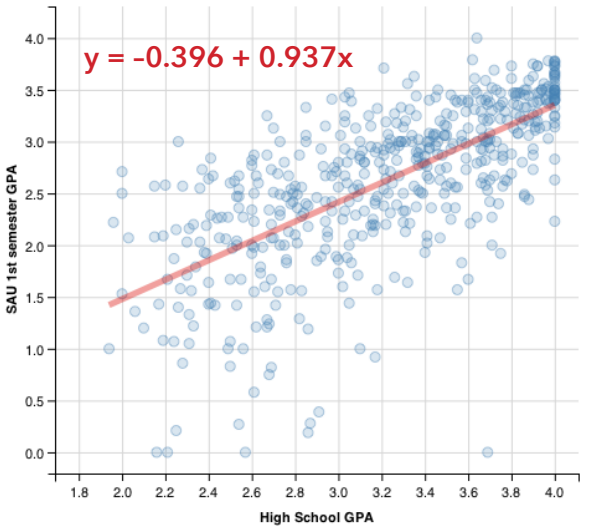
Relationship between body weight (in kg) and heart weight (in g) for 144 house cats

$\bar{X} = 2.723611$ $\bar{Y} = 10.63056$

$s_x = 0.485307$ $s_y = 2.43464$

$n = 144$ $r = 0.804$

R. A. Fisher (1947) The analysis of covariance method for the relation between a part and the whole, *Biometrics* **3**, 65–68.



Relationship between high school GPA and first-semester St. Ambrose GPAs in 2013

$\bar{X} = 3.235433$ $\bar{Y} = 2.63624$

$s_x = 0.543623$ $s_y = 0.75097$

$n = 508$ $r = 0.678$

2013-14 MAP-Works data

1. Above, you can see scatterplots from two datasets. On top of each scatterplot, I've plotted the line that best-fits the data. In this activity, we'll learn what it means to be the "best" fitting line, how to find the formula for that line, how to interpret the slope and y-intercept, and how to evaluate whether the line fits "good enough."

   Looking at the scatterplots and *regression lines* displayed above, do you think the datasets have linear relationships? In other words, if you wanted to sketch a function through the data points that *best* describes the relationship between X and Y, would you choose a line? How well do lines fit these datasets?

2. Interpret the slope and y-intercept for the scatterplot on the top-left. What do they mean with regards to the data in the scatterplot?

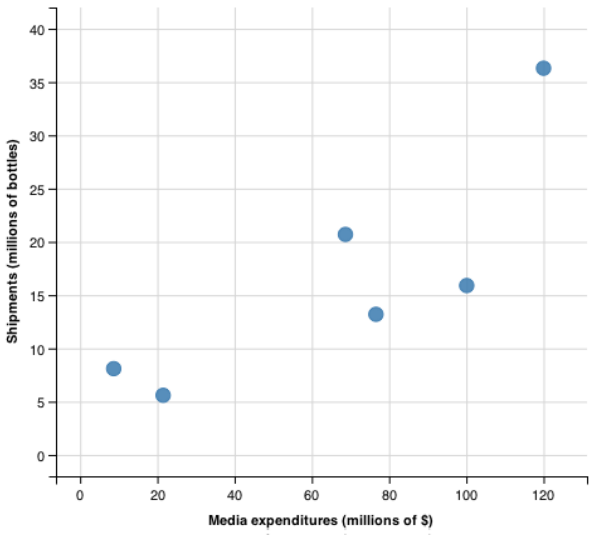
3. Predict the brain weight of a cat that weighs 3 kg. Predict the brain weight of a cat that weighs 20 kg. In which prediction do you have more confidence? How confident are you with your predictions?

4. The top-right scatterplot shows the degree to which high school GPAs might predict first-semester GPAs at St. Ambrose. If we wanted to improve the prediction of first-semester GPAs, what other predictors should we add to our model?

5. To build-up our intuition and derive some important formulas, let's turn to an extremely small dataset. The following table and scatterplot display the relationship between media expenditures and number of bottles shipped for 6 brands of beer:

Brand	Media Expenditures (millions of $)	Bottles Shipped (in millions)
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
mean	65.9333	16.6333
std. dev	43.5017	11.0471
correlation	correlation: r = 0.8288	

Source: Superbrands, 1998; 10/20/1997



Based on the correlation coefficient, one could argue that the variables have a linear relationship. If this is true, we can construct the following model: **shipments = f(media) + error**. Substituting y=shipments and x=media, we get:

$$y_i = f(x) + e_i$$

If we're going to use a **linear** function to model this relationship, we know we're going to need to find the slope (b_1) and y-intercept (b_0) of this line. We can, therefore, write out our linear model as:

$$y_i = f(x \mid b_0, b_1) + e_i$$

This model we created implies that the number of bottles shipped by a beer company is a function of the amount they spend on advertising plus other stuff. Our goal, then, will be to estimate the parameters of this model (the slope and y-intercept) and determine how well the model fits the data.

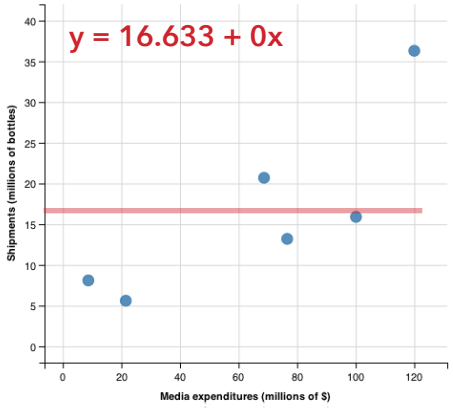
On the scatterplot displayed above, sketch the line that you think best fits the data. Below, estimate the slope and y-intercept of the line you sketched.

Y = _____ (x) + _____

(slope) (y-intercept)

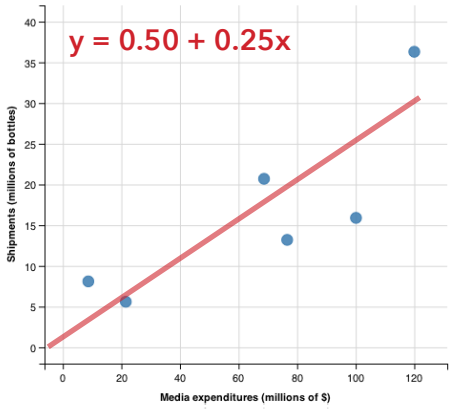
6. I imagine everyone has different values for the slope and y-intercept. How could we determine which line is **best**?

Suppose 3 students sketched the following lines. Which line is best? What do the numbers in the tables represent?

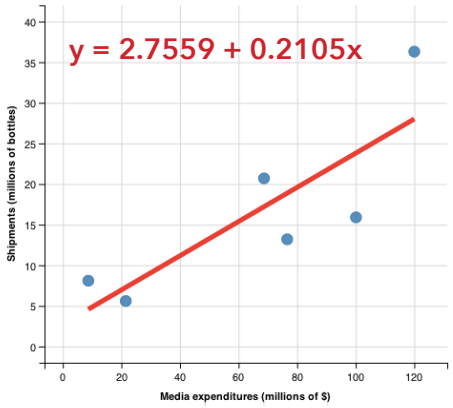


Note: This would be our best prediction if we didn't know anything about media expenditures

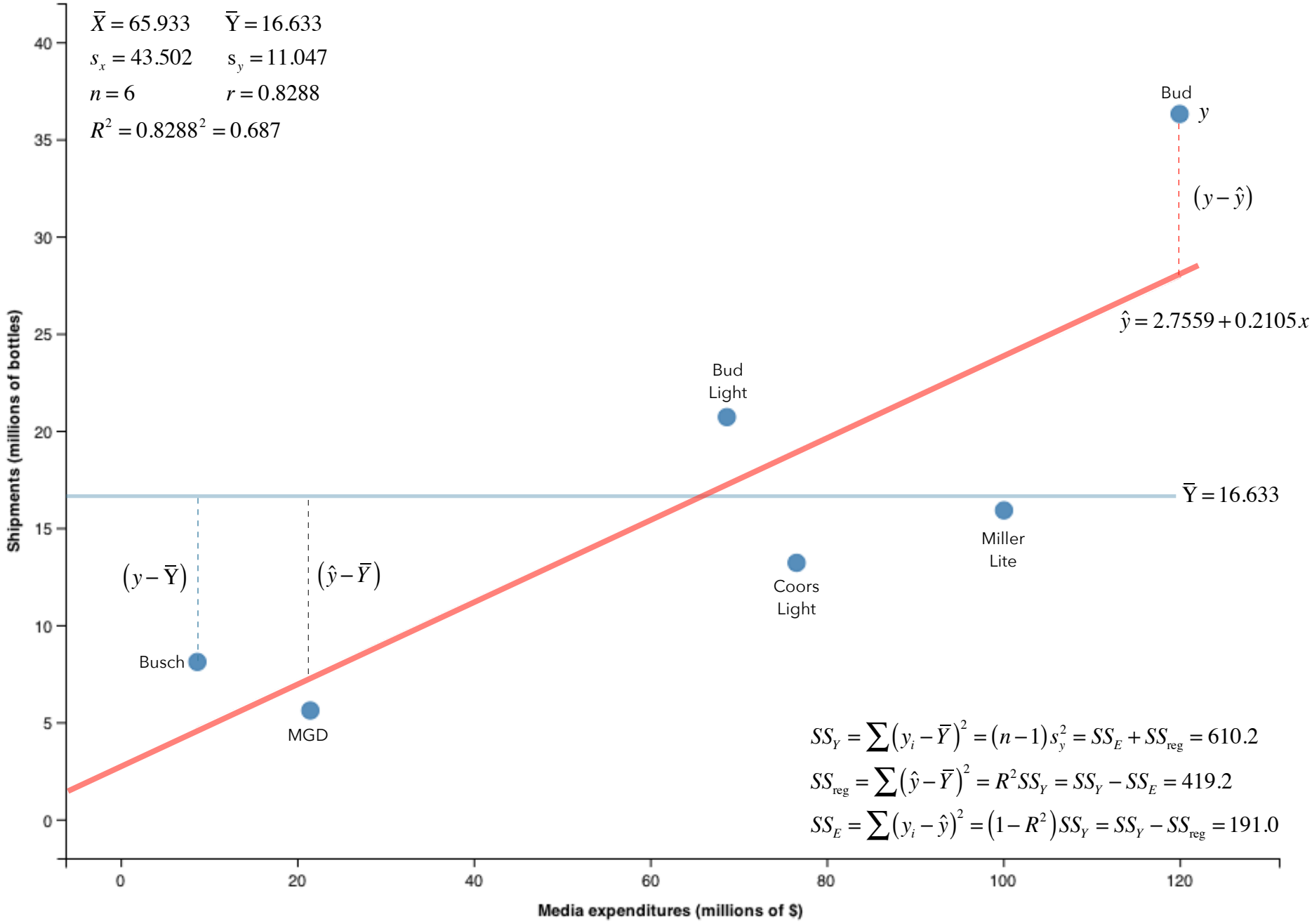
Observed		Predicted	error	error^2
Media (x)	Shipped (y)	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
8.7	8.1	16.633	-8.533	72.812
21.5	5.6	16.633	-11.033	121.727
68.7	20.7	16.633	4.067	16.5405
76.6	13.2	16.633	-3.433	11.785
100.1	15.9	16.633	-0.733	0.537
120.0	36.3	16.633	19.667	386.791
			Sum = 0.002	**610.193**



Observed		Predicted	error	error^2
Media (x)	Shipped (y)	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
8.7	8.1	2.675	5.425	29.431
21.5	5.6	5.875	-0.275	0.076
68.7	20.7	17.675	3.025	9.151
76.6	13.2	19.65	-6.45	41.602
100.1	15.9	25.525	-9.625	92.641
120.0	36.3	30.5	5.8	33.64
			Sum = -2.1	**206.541**



Observed		Predicted	error	error^2
Media (x)	Shipped (y)	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
8.7	8.1	4.58725	3.51275	12.339
21.5	5.6	7.28165	-1.68165	2.828
68.7	20.7	17.21725	3.48275	12.13
76.6	13.2	18.8802	-5.6802	32.265
100.1	15.9	23.82695	-7.92695	62.837
120.0	36.3	28.0	8.2841	68.626
			Sum = -0.009	**191.025**


$\bar{X} = 65.933$ $\bar{Y} = 16.633$
$s_x = 43.502$ $s_y = 11.047$
$n = 6$ $r = 0.8288$
$R^2 = 0.8288^2 = 0.687$
Bud
y
$(y - \hat{y})$
$\hat{y} = 2.7559 + 0.2105x$
Bud Light
$\bar{Y} = 16.633$
Miller Lite
Coors Light
$(y - \bar{Y})$
$(\hat{y} - \bar{Y})$
Busch
MGD
$SS_Y = \sum (y_i - \bar{Y})^2 = (n-1)s_y^2 = SS_E + SS_{reg} = 610.2$
$SS_{reg} = \sum (\hat{y} - \bar{Y})^2 = R^2 SS_Y = SS_Y - SS_E = 419.2$
$SS_E = \sum (y_i - \hat{y})^2 = (1 - R^2) SS_Y = SS_Y - SS_{reg} = 191.0$
Shipments (millions of bottles)
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100
120
Media expenditures (millions of $)

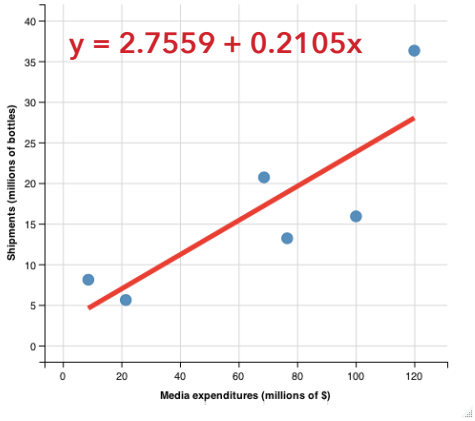
7. It turns out the best-fitting line for the data is y = 2.7559 + 0.2105x. Use that line to "predict" the number of bottles shipped by a company that spent $76.6 million on advertising.

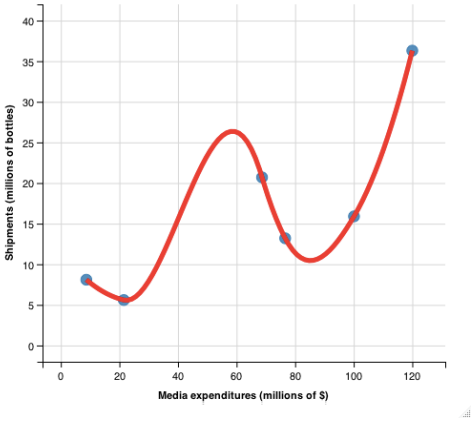
   Predicted number of bottles shipped = f(76.6) = _____

   Coors Light spent $76.6 million on advertising and shipped 13.2 million bottles. How far off was your prediction?

8. No line – not even the best line – will fit all the data with perfect accuracy. There's always some amount of random error (and, probably, measurement error).

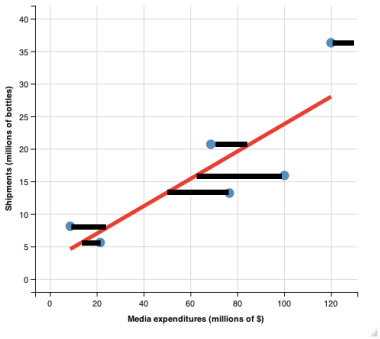
   Why, then, wouldn't we simply connect-the-dots to create our prediction model? Why might we prefer an imperfect line to a perfect connect-the-dots model?



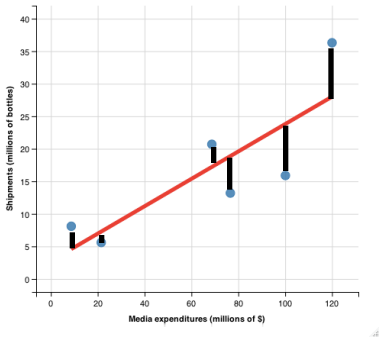


9. We're always going to have some amount of unexplained, random, or measurement error in the data, so the line will never fit perfectly. The best-fitting line, however, will minimize the total amount of error (the sum of the distances between the points and the line).

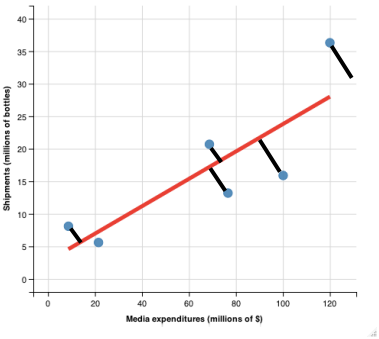
   If distances between points and the prediction line represent error, which distances (errors) are we interested in minimizing? Do we want to minimize the horizontal, vertical, or perpendicular distances? Why?



horizontal errors

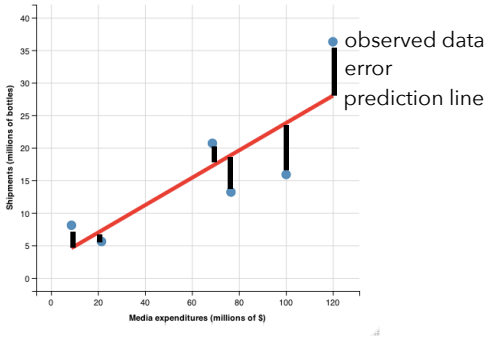


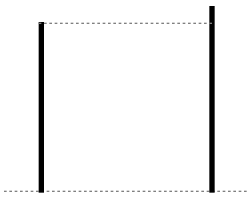
vertical errors
least squares
assumes x values are "good" measures
or that we chose the x values



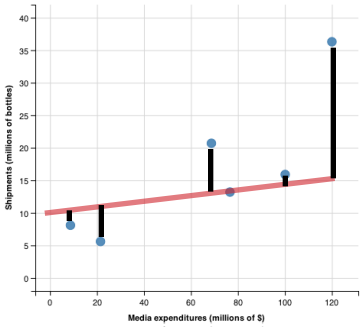
perpendicular errors
error-in-both-variables regression
orthogonal (Deming) regression
"perpendicular" changes as units change

10. Assuming we're interested in the error in predicting y values from given x values, we want to minimize the vertical distances. Below, I've drawn in these vertical errors for two potential regression lines. I then calculated the total length of the lines to find the sum of these errors:





total error (sum of all line lengths)



Since the sum of the errors for the line on the left is less than that for the line on the right, the line on the left better fits our data. Now all we have to do is find the sum of the errors for every possible line we could draw for our data.

That could take forever, so let's use some math to find the formula for the line that best fits a given dataset.

To do this, let's establish some notation: $(x_i, y_i) \leftarrow$ the coordinate of a data point

$$\hat{y}_i = b_0 + b_1 x_i + e \leftarrow \text{our linear model}$$

$$e = y_i - (\hat{y}_i) = y_i - (b_0 + b_1 x_i) = y_i - b_0 - b_1 x_i \leftarrow \text{error}$$

We want to find the line that minimizes the sum of those errors. We want to minimize: $\sum_{i=1}^{n} y_i - b_0 - b_1 x_i$

The problem is that some of the errors will be positive (when the observed data is above the prediction line) and some errors will be negative (when the observed data is below the prediction line). If we find the sum of these values, the positive and negative errors will cancel each other out. More importantly, the sum of errors would actually be minimized with a regression line that is drawn way up above all the data points (and, therefore, is not a good fit at all).

How can we deal with this issue? How can we ensure all the errors are positive?

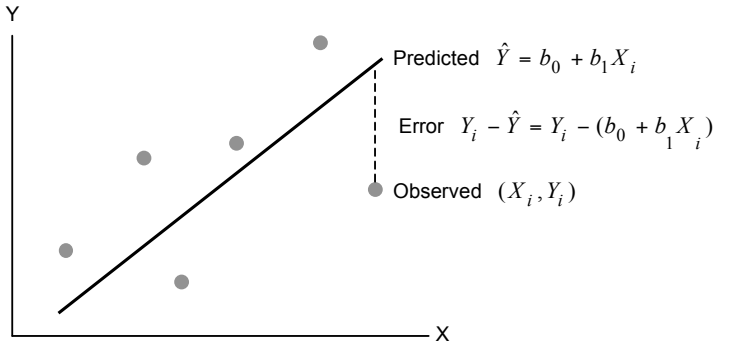
We could take the absolute value of our errors (like I did in the example at the top of this page). We could minimize: $\sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|$

This is the approach used in *quantile regression* (which we'll learn about later in the semester). One of the problems with absolute values is that they're difficult to work with algebraically.

Another way to ensure all the errors are positive would be to square each error. We'd then want to minimize:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

This gives us a nice (mathematically speaking) function we can minimize using Calculus. It also has the feature/bug of magnifying outliers. When we square large errors (outliers), those squared errors get huge.



We must find the slope and y-intercept for this line so that the sum of these squared errors is minimized.

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: $-2\sum (Y_i - b_0 - b_1 X_i) = 0$ $\quad \sum (Y_i - b_0 - b_1 X_i) = 0$

$$\sum Y_i = n b_0 + b_1 \sum X_i$$

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: $-2\sum X_i (Y_i - b_0 - b_1 X_i) = 0$ $\quad \sum X_i (Y_i - b_0 - b_1 X_i) = 0$

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

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

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: $b_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$

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

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

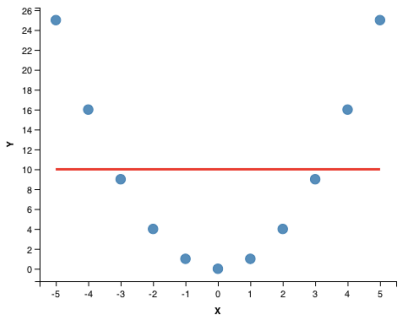
$$b_0 = \overline{Y} - b_1 \overline{X} \quad \text{and} \quad b_1 = r\frac{S_y}{S_x}$$

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:

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

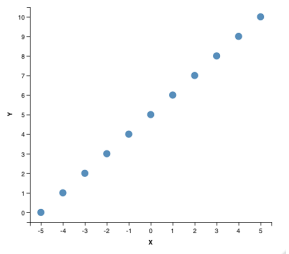
Now that we have formulas to calculate the slope and y-intercept of our least-squares regression line (the line of "best" fit), we need to find some way to determine if that line is any good. We can find the least-squares regression line for any dataset, but it doesn't mean the line is a good fit (or meaningful).

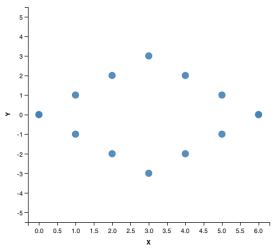
See the example to the right:



11. We're going to try out several measures of how well our regression line fits the data. Let's see if we can figure out the value of each measure under two situations: (a) a model that fits perfectly, and (b) a model that doesn't fit at all.

Eventually, we'll want to fill-in this table:





Measure / Index	Value for perfect fit	Value for no fit
$SS_E = \sum (y_i - \hat{y})^2$	0	$\sum (y_i - \bar{Y})^2 = SS_y = (n-1)s_y^2$
$s_{y\|x}^2 = \frac{SS_E}{df_E} = \frac{\sum (y_i - \hat{y})^2}{n-2}$	_____	$\left(\frac{n-1}{n-2}\right) s_y^2 = \frac{SS_Y}{n-2}$
$s_{y\|x} = \sqrt{\frac{SS_E}{df_E}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$	0	$s_y \sqrt{\frac{n-1}{n-2}} = \sqrt{\frac{SS_Y}{n-2}}$
$1 - R^2 = \frac{SS_E}{SS_Y} = \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{Y})^2}$		
$R^2 = \frac{SS_{\text{reg}}}{SS_Y} = \frac{\sum (\hat{y} - \bar{Y})^2}{\sum (y_i - \bar{Y})^2}$		

12. The first potential measure of how well a regression line fits the data would be to simply look at SSE (the sum of the squared errors). What would SSE equal if the line fit the data perfectly?

Now suppose we have uncorrelated variables – knowing the value of X would not tell us anything about the value of Y. What would the least-squares regression line look like in this case? We'd need to find the value of *m* that would minimize the following:

$$\sum (y_i - M)^2$$

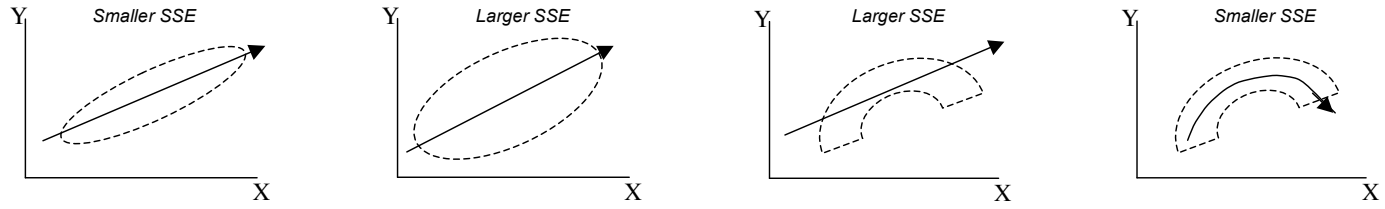
If you remember from a previous statistics class, this value is minimized when M equals the sample mean. Therefore, our best prediction for uncorrelated variables would be:

$$\sum (y_i - \bar{Y})^2$$

That formula should look familiar. That's our good friend SStotal from ANOVA (or SSy, as we'll refer to it in regression). What's the largest value we could possible get for SSE?

If we add one more observation to our data, what has to happen to the value of SSE?

13. The size of SSE depends on a few factors, such as the amount of variation in our data, the number of observations we have in our data, and the degree to which a line fits the data.



It seems problematic that adding data would automatically increase the size of SSE. Perhaps it would be better to calculate the average squared error (or mean squared error). This would give us the *variance of the estimate*:

$$s^2_{y|x} = \frac{SS_E}{df_E} = \frac{\sum (y_i - \hat{y})^2}{n-2}$$

This variance of the estimate represents the average squared distance from each observation to the prediction line. In a situation with perfect fit, what would this measure equal? Write that in the table on the previous page.

With uncorrelated variables, what would be the maximum value of the variance of the estimate?

$$\max\left\{s^2_{y|x}\right\} = \frac{\sum (y_i - \bar{Y})^2}{n-2} = \left(\frac{n-1}{n-2}\right)\left(\frac{\sum (y_i - \bar{Y})^2}{n-1}\right) = \left(\frac{n-1}{n-2}\right)s^2_y = \frac{SS_Y}{n-2}$$

14. I don't know about you, but I'd prefer our measure of good-fit to <u>not</u> be in squared units. We can fix that easily enough:

$$s_{y|x} = \sqrt{s^2_{y|x}} = \sqrt{\frac{SS_E}{df_E}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

This measure is called the **standard error of the estimate**. What does it represent? Sketch a scatterplot and show what the standard error of the estimate would be visually.

Since this is simply the square root of the variance of the estimate, it's easy to find the values under situations with perfect and no fit. These values have been filled-in the table on the previous page.

15. The maximum value of each of our measures is still unbounded. Ideally, our measure would have a fixed minimum and maximum.

    Suppose we took our total sums of squares (the variation in Y) and partitioned it. We know some of that variation is unexplained by our regression line, so we could calculate the following:

$$\frac{SS_E}{SS_Y} = \frac{\sum\left(y_i - \bar{Y}\right)^2}{\sum\left(y_i - \hat{y}\right)^2} = 1 - r^2$$

    This happens to equal one minus our correlation coefficient squared. Under the perfect-fit and no-fit scenarios, what would be the value of this ratio of error variance to total variance? Write those values in our table.

16. That measure seems backwards – it equals zero when we have perfect fit and it equals 1 when we have no fit. Let's invert that by taking:

$$r^2 = R^2 = \frac{SS_Y - SS_E}{SS_Y} = \frac{SS_{reg}}{SS_Y} = \frac{\sum\left(\hat{y} - \bar{Y}\right)^2}{\sum\left(y_i - \hat{y}\right)^2}$$

    This is the **coefficient of determination** and it has the same interpretation as eta-squared in an ANOVA. Fill-in the table to show the values of this measure under perfect and no-fit situations.

17. In most cases, the best measures of how well a line models a dataset are the coefficient of determination and the standard error of the estimate (or the RMSE, the *root mean squared error)*. Identify the advantage of each measure to determine how well a line fits a dataset.

18. We've derived the least squares criterion and formulas to calculate the slope and y-intercept for that line of best fit. We've also derived some measures we can use to indicate how well that best-fitting line actually fits the data. There are only a few things left to do to fully understand simple linear regression:

    a) Practice using technology to estimate these regression lines
    b) Figure out how to determine if a regression line fits the data "good enough"
    c) Investigate the assumptions we're making when we estimate these least-squares regression lines.

19. Let's take another look at the beer data. To estimate the least squares regression line in Stata, I would simply use the command:

```
regress shipment media
```

To regress shipments on media expenditures in R, I would use the command:

```
lm(shipment ~ media, data=beer)
```

This tells R to find a linear model (lm) where shipment is a function of media. With just this command, we would get the following output:

Brand	Media Expenditures (millions of $)	Bottles Shipped (in millions)
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
mean	65.9333	16.6333
std. dev	43.5017	11.0471
correlation	correlation: r = 0.8288	

Source: Superbrands, 1998; 10/20/1997

```
Coefficients:
(Intercept)        media
     2.7559       0.2105
```

If we want to know more than the slope and intercept of the best-fitting line, we need to use some additional commands. First, we can store our linear model (under the name "model" in this example). We can then get a summary of the model using the summary() command:

```
model = lm(ship~media, data=beer)
summary(model)
```

This produces the following output:

```
Residuals:
     1      2      3      4      5      6
 3.513 -1.681  3.484 -5.678 -7.925  8.287

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.75591    5.46812   0.504   0.6408
media        0.21048    0.07104   2.963   0.0414 *
---
Residual standard error: 6.911 on 4 degrees of freedom
Multiple R-squared:  0.6869, Adjusted R-squared:  0.6087
F-statistic: 8.777 on 1 and 4 DF,  p-value: 0.04145
```

What, if anything, can we interpret from that output? We can also use other R commands to get more information:

Command	What the command does…
`plot(model)`	Creates diagnostic plots to check assumptions (output on the next page)
`coef(model)`	Returns the coefficients of the model
`confint(model)`	Returns confidence intervals for the coefficients of the model
`vcov(model)`	Returns the variance/covariance matrix
`residuals(model)`	Returns the residuals (errors) for each observation in the dataset
`predict(model)`	Returns the predicted values for each observation in the dataset
`anova(model)`	Summarizes the model in an ANOVA summary table (output on the next page)

Let's take a look at the diagnostic plots and ANOVA summary table on the next page:

20. We'll investigate this more later, but here are the assumptions of the linear regression model (in decreasing order of importance:

- Validity: The data you are analyzing maps to the research question you are trying to answer.
  Diagnosis: Take a careful look at the purpose of your study and the data you've collected
  How to fix: Get better data

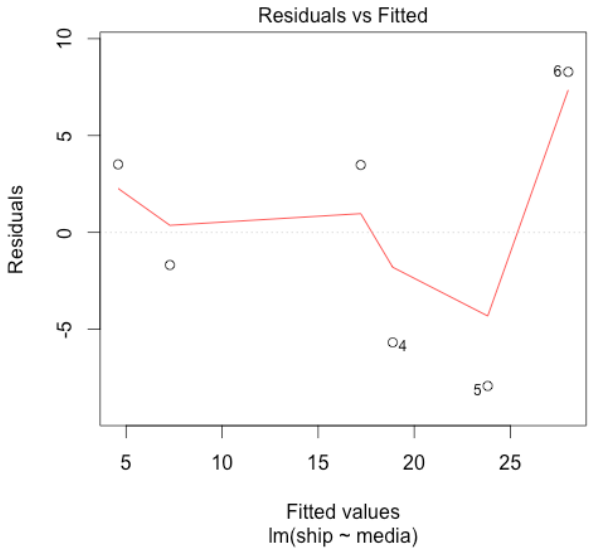
- Additivity and linearity: The deterministic component of the model is a linear function of the predictors.
  Diagnosis: Look at plots of observed vs predicted or residuals vs predicted values. The points should be symmetrically distributed around a diagonal line in the former plot or around horizontal line in the latter plot, with a roughly constant variance.
  How to fix: You could transform your data (if it seems appropriate) or add a nonlinear component

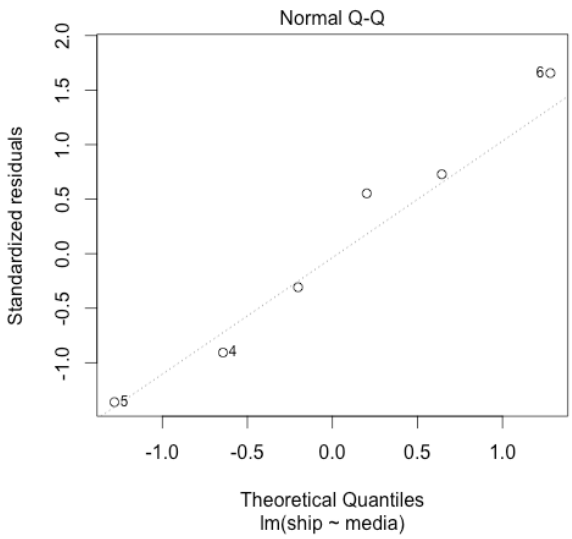
- Independent errors: No correlation among errors
  Diagnosis: If you have time series data, be careful that consecutive errors are not related.

- Equal variance of errors (*homoskedasticity*): The variance in the errors is the same across all levels of X.
  Diagnosis: Look at the plot of residuals vs predicted values. If the residuals grow larger as a function of X, you have a problem.

- Normality of errors
  Diagnosis: Look at a P-P or Q-Q plot of the residuals. The residuals should fall near the diagonal line. You could also run a test for normality, like the Shapiro-Wilk or Kologorov-Smirnov tests. Note that the dependent and independent variables in a regression model do not need to be normally distributed by themselves--only the prediction errors need to be normally distributed

Here are the diagnostic plots for our beer dataset:





Finally, here's the ANOVA summary table for our linear model:

```
Analysis of Variance Table
Response: ship
          Df Sum Sq Mean Sq F value  Pr(>F)
media      1 419.17  419.17  8.7773 0.04145 *
Residuals  4 191.02   47.76
```

21. Wait a second – ANOVA summary tables? confidence intervals? What is going on? Let's take a step back before we move on.

    I want to make sure you can calculate these least-squares regression lines without much trouble. Let's go ahead and replicate the regression lines from the beginning of this activity.

    Copy the cat data from this website: http://www.bradthiessen.com/html5/data/cats.csv
    and paste it into this applet: http://www.rossmanchance.com/applets/RegShuffle.htm?hideExtras=2

    Click USE DATA and you'll see the 144 observations plotted to the right. Now, simply click SHOW REGRESSION LINE to find the formula.

    You can also click SHOW RESIDUALS to see the errors. Below that, you can get the value of R-SQUARED or a REGRESSION TABLE that seems to have some sort of t-test for the coefficients of our model. You can also get a CONFIDENCE INTERVAL for the slope.

    This time, paste the high school / SAU GPA data from: http://www.bradthiessen.com/html5/data/actgpa2.csv

    Write out the formula for the least-squares regression line, record SSE, and R-squared.

22. As you can see, R-squared for this GPA dataset equals 0.46. 46% of the variation in St. Ambrose first semester GPAs is accounted for by high school GPAs. Does that seem small or large to you?

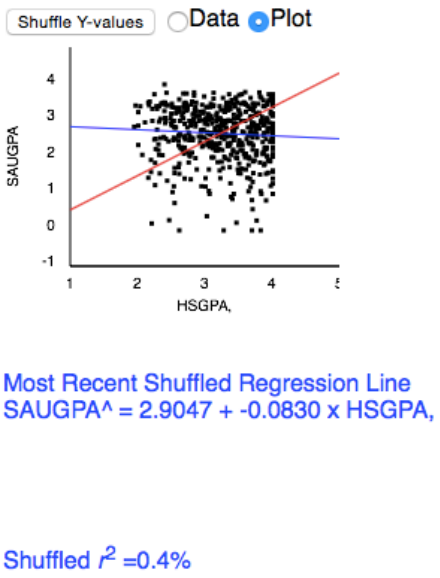
    The slope of our regression line was found to be 0.9372. Even if high school GPAs were completely uncorrelated with St. Ambrose GPAs, we'd expect a sample of data to have some correlation. How unlikely were we to observe a slope of 0.9372 or greater if the data were uncorrelated?

    To investigate this, we can run a randomization-based test for the slope of our regression line. Go back to the applet and check the SHOW SHUFFLE OPTIONS box. Shuffle the data once and show the PLOT (not DATA) for that shuffle. You should get something like the plot displayed to the right.

    The red line represents our observed regression line. The blue line shows the regression line we'd get by randomly shuffling the Y values in our dataset. Why can we do this? If our null hypothesis is that these variables are uncorrelated, then there's no reason to keep them paired together. Any Y value could be paired with any X value.

    Shuffle the data at least 10,000 times and explain what the plot on the top-right of the screen represents. Finally, use the applet to estimate (and interpret) a p-value.

    (The R code for this activity shows how to conduct this test in R)



23. We can also use bootstrap methods to construct a confidence interval for our slope parameter.

    Once again, copy the data from: http://www.bradthiessen.com/html5/data/actgpa2.csv
    Paste it into this applet: http://lock5stat.com/statkey/bootstrap_2_quant/bootstrap_2_quant.html

    Change the top selection to get a bootstrap interval for the SLOPE.
    Generate 10,000 samples and describe what is happening.
    Finally, record and interpret a 95% confidence interval for the slope of our regression line.

24. You'll practice using randomization-based tests and bootstrap confidence intervals for the slope in the assignment aligned with this activity. For now, let's move on to parametric methods used to test how well a regression model fits a given dataset.

    When conducting a linear regression analysis, we're often interested in finding the most parsimonious model that can explain *enough* of the variance in the dependent variable. To find this "best" model, we might compare several competing models, each increasing in complexity (*nested* models). For example:

    a) We might start with the most basic model that predicts the same value for Y regardless of X. All variation in observed Y values would be modeled by random error: $\hat{y}_i = b_0 + e_i$. What value would we choose for b_0?

    b) We could then add one predictor to the model to create: $\hat{y}_i = b_0 + b_1 x_1 + e_i$. We could compare the performance of this model to the previous model to determine if the improvement in prediction justified the (relative) complexity of adding the predictor.

    c) We could then add yet another predictor: $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + e_i$. Once again, we could compare this model to the previous model. If this new model provided a significantly better prediction (explained a significant amount of previously unexplained variance), then we could decide to keep this new model. If the model didn't improve our prediction by very much, we might decide to keep the previous, simpler model.

    At each stage in building our regression model, we can assess the value of adding predictors (complexity) through randomization-based or parametric hypothesis testing methods. These methods can help us determine which predictors to keep in our model.

    We could also work through this process backwards. We could start with a relatively complex model, take away the predictor that explains the lease amount of variance in Y, and determine if the simpler model was significantly worse than the more complex model.

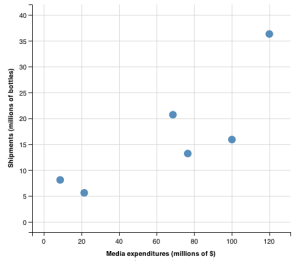
25. When comparing regression models, it's helpful to write out the **full model** (the more complex model) and the **reduced model** (less complex). When you're analyzing your own data, you'll choose these models (based on your experience with whatever data or research question you're working with). For now, I'll force us to choose specific models.

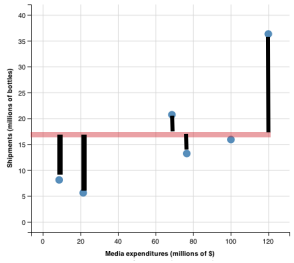
    We'll work one last time with this beer data. I want to know if X (media expenditures) predicts Y (bottles shipped) better than a model with no predictors. Write out our full and reduced models:

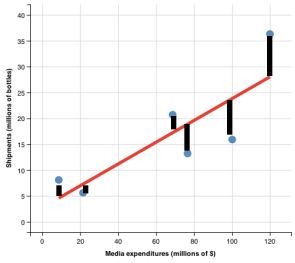
    Full model: _____ Reduced model: _____

26. As we've already seen, the sample mean minimizes the sum of squared errors (if we have no predictor variables). Therefore, what does SSE represent in our reduced model?

    Our full model is the least-squares regression line (using one predictor variable). As you can see below, the full model reduced our error variance by 610.193 – 191.025 = 419.168. What does this value represent?







Observed		Reduced Model			Full Model		
Media (x)	Shipped (y)	predicted	error	error^2	predicted	error	error^2
8.7	8.1	16.633	-8.533	72.812	4.587	3.513	12.339
21.5	5.6	16.633	-11.033	121.727	7.282	-1.682	2.828
68.7	20.7	16.633	4.067	16.541	17.217	3.483	12.130
76.6	13.2	16.633	-3.433	11.785	18.880	-5.680	32.265
100.1	15.9	16.633	-0.733	0.537	23.827	-7.927	62.837
120.0	36.3	16.633	19.667	386.791	28.016	8.284	68.626
			Sum	**610.193**		Sum	**191.025**

27. Fill-in these SSy and SSE values in the ANOVA summary table. How many degrees of freedom will we have?

Source of variation	**SS**	**df**	**MS**	**MSR (F)**
Regression ($b_1 \mid b_0$)	_____	_____	_____	_____
Error	_____	_____	_____	(blank)
Total	_____	_____	MS_{total}	R^2 = _____

28. Complete the ANOVA summary table and estimate a p-value. What conclusion could we make? Remember, you can always use the F-distribution applet at: http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#F

29. The only difference between our full and reduced models is the b_1 coefficient (the slope). If $b_1 = 0$, the full model would be the same as our reduced model. Another way, then, to compare our full and reduced models would be to test the hypothesis: **H_0: $b_1 = 0$**. We can test this hypothesis with a t-test.

$$t_{n-2} = \frac{(\text{observed value}) - (\text{hypothesized value})}{\text{standard error}} = \frac{\hat{b}_1 - 0}{SE_{b_1}} =$$

$$t_{n-2} = \frac{\hat{b}_1 - 0}{SE_{b_1}} = \frac{\hat{b}_1}{\frac{s_{y|x}}{s_x\sqrt{n-1}}} = \frac{r_{xy}\frac{s_y}{s_x}}{\frac{s_y\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}}{s_x\sqrt{n-1}}} = \frac{r_{xy}\frac{s_y}{s_x}s_x\sqrt{n-1}}{s_y\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}} = \frac{r_{xy}\sqrt{n-1}}{\sqrt{1-r^2}\sqrt{\frac{n-1}{n-2}}}$$

$$t_{n-2} = \sqrt{\frac{r_{xy}^2(n-1)}{(1-r^2)\left(\frac{n-1}{n-2}\right)}} = \sqrt{\frac{r_{xy}^2(n-2)}{(1-r^2)}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{r_{xy}-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r_{xy}-0}{SE_{r_{xy}}}$$

What did that derivation just show? Conduct this t-test and state an appropriate conclusion. Compare the value of your t-statistic to the value of the MSE you calculated for the ANOVA.

30. This time, conduct a test of the hypothesis: **H_0: $r_{xy} = 0$**.

31. A test for the slope of a regression line is the same as a test for the correlation between x and y. So why do we use an ANOVA to compare our full and reduced models? Follow along:

$$F = MSR = \frac{MS_{reg}}{MS_E} = \frac{SS_{reg} / df_{reg}}{SS_E / df_E} = \frac{SS_{reg}\left(df_E\right)}{SS_E\left(df_{reg}\right)} = \frac{r^2 SS_Y\left(n-2\right)}{\left(1-r^2\right)SS_Y\left(1\right)} = \frac{r^2\left(n-2\right)}{\left(1-r^2\right)} = t^2_{n-2}$$

32. It can also be shown that we can calculate our omnibus F-statistic with the following:

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

Verify this formula gives us the same value for our MSR (as the ANOVA table in question #27).

33. We've already seen that randomization-based methods can test whether a slope coefficient is significant. We can also use randomization-based methods to compare regression models.

Paste the beer data into: http://lock5stat.com/statkey/advanced_2_quant/advanced_2_quant.html

Generate at least 10,000 randomized samples and report the p-value. How does it compare to our p-value from the omnibus F-test?

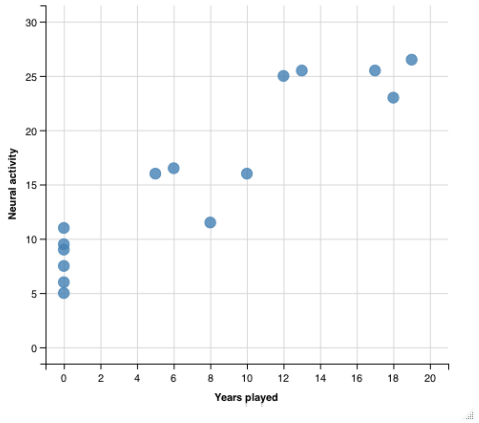
media	beer
8.7	8.1
21.5	5.6
68.7	20.7
76.6	13.2
100.1	15.9
120.0	36.3

Scenario: A number of studies have shown that certain activities can affect the reorganization of the human central nervous system. For example, it's known that the par of the brain associated with activity of a limb is taken over for other purposes in individuals who have lost a limb.

In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of 9 violin players and 6 controls (those who have never played a stringed musical instrument) when the fingers on their left hands were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers on their left hand extensively, might show an increased amount of neuron activity. Shown below is a neuron activity index from the MSI along with the number of years each individual had been playing a stringed instrument:

Subject	Years played	Neural activity
1	0	5.0
2	0	6.0
3	0	7.5
4	0	9.0
5	0	9.5
6	0	11.0
7	5	16.0
8	6	16.5
9	8	11.5
10	10	16.0
11	12	25.0
12	13	25.5
13	17	25.5
14	18	23.0
15	19	26.5
mean	7.2	15.56667
std. dev	7.24273	7.782459
	correlation:	r = 0.928

Elbert, T., "Increased cortical representation of the fingers of the left hand in string players," Science, 270, 13 October, 305-307



You can download this data at:
http://www.bradthiessen.com/html5/data/violin.csv

ANOVA for Regression applet:
http://lock5stat.com/statkey/advanced_2_quant/advanced_2_quant.html

34. Our goal is to determine whether neural activity increases as the number of years playing the violin increases. Suppose we decide to conduct an ANOVA. How would we do this? What conclusions could we draw?

35. Paste the data into the **ANOVA for Regression** applet. On the top-right, you'll see a graph of the data along with the estimated slope and y-intercept for the least-squares regression line. Write out the full and reduced models of interest, along with the formula for the best-fitting line.

Full model: _____ Reduced model: _____

Formula for best-fitting regression line: _____

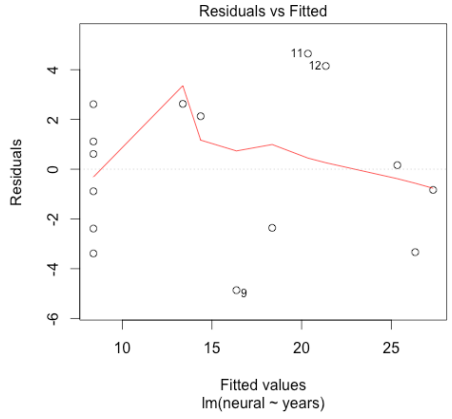
36. Click the **ANOVA TABLE** button on the top-right and fill-in the following table. Then, verify all these calculations.

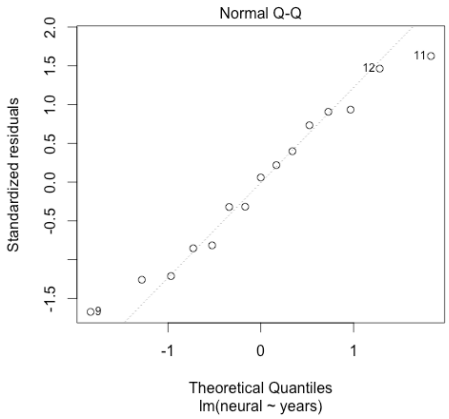
Source of variation	SS	df	MS	MSR (F)
Regression ($b_1 \mid b_0$)	_____	_____	_____	_____
Error	_____	_____	_____	(blank)
Total	_____	_____	MS_{total}	R^2 = _____

37. What conclusions can you draw from this analysis?

38. Replicate that MSR by calculating the omnibus F-statistic.

39. Explain what the following plots indicate with regards to the assumptions underlying linear regression.





40. Finally, use the applet to conduct a randomization-based comparison of our full and reduced models. How do the results compare to our parametric methods?

Scenario: Some occupations are considered to be more prestigious than others (inspiring more respect or admiration). For example, most people would agree that a heart surgeon has a more prestigious occupation than a waitress. We're going to examine some factors that may influence the prestige of various occupations.

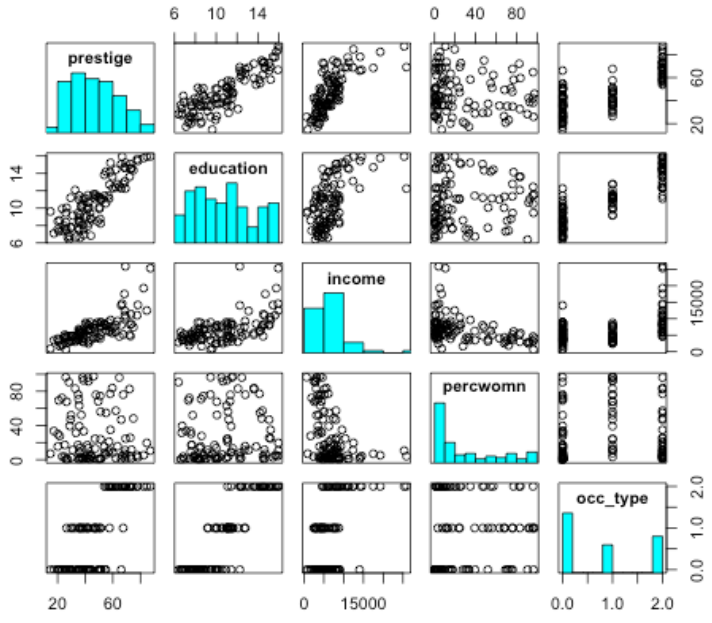
Data: http://www.bradthiessen.com/html5/data/prestige.csv
Source: Canada (1971). Census of Canada. Vol. 3, Part 6. Statistics Canada, 19-21.

prestige: Pineo-Porter Prestige score (a survey)
education: average years of education for people in that occupation
income: average income (1971 Canadian dollars) for people in that occupation
percwomn: % of workers in that occupation who are female
type: 0=blue collar, 1=white collar, 2=professional/technical/managerial

#	Title	Education	Income	%women	Type	Prestige
1	Physicians	15.96	25308	10.56	Professional	87.2
2	University Professors	15.97	12480	19.59	Professional	84.6
3	Lawyers	15.77	19263	5.13	Professional	82.3
4	Architects	15.44	14163	2.69	Professional	78.1
5	Physicists	15.64	11030	5.13	Professional	77.6
6	Psychologists	14.36	7405	48.28	Professional	74.9
7	Chemists	14.62	8403	11.68	Professional	73.5
8	Civil Engineer	14.52	11377	1.03	Professional	73.1
...	...	...	...	...	...	...
18	Medical Technicians	12.79	5180	76.04	White collar	67.5
19	Secondary Teachers	15.08	8034	46.8	Professional	66.1
...	...	...	...	...	...	...
26	Elementary Teachers	13.62	5648	83.78	Professional	59.6
...	...	...	...	...	...	...
98	Launderers	7.33	3000	69.31	Blue collar	20.8
99	Bartenders	8.5	3930	15.51	Blue collar	20.2
100	Elevator Operators	7.58	3582	30.08	Blue collar	20.1
101	Janitors	7.11	3472	33.57	Blue collar	17.3
102	Newsboys	9.62	918	7	(missing)	14.8
	Means	**10.738**	**6797.90**	**28.979**	**N/A**	**46.833**
	Std. Deviations	**2.7284**	**4245.92**	**31.725**	**N/A**	**17.204**

Correlations:

	education	income	%women	prestige
education	1.0000			
income	0.5776	1.0000		
%women	0.0619	-0.4411	1.0000	
prestige	0.8502	0.7149	-0.1183	1.0000



41. Before attempting to model prestige, I wanted to know if the 3 occupation types differed in prestige. Interpret these results:

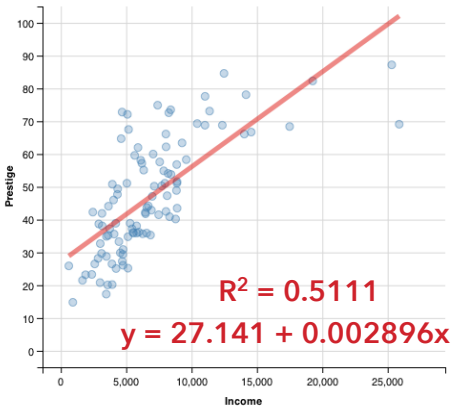
```
  occ_type  n     mean       sd
1        0 49 36.08571 11.347320
2        1 23 42.24348  9.515816
3        2 30 67.90667  8.819255


Pairwise comparisons (Bonferroni)
  0       1
1 0.059   -
2 < 2e-16 4.4e-14
```

Source	SS	df	MS	MSR (F)
Type	17796	2	9733.576	92.40
Error	12100	99	105.336	$p = 2.2 \times 10^{-16}$
Total	29896	101	MS_{total}	$\eta^2 = 0.5953$

```
Bartlett test of homogeneity of variances
data:  prestige by occ_type
Bartlett's K-squared = 2.4469, df = 2, p= 0.2942
```

42. The relationship between prestige and income is displayed below. Interpret the coefficients of our model (which you could verify using the summary statistics on the previous page).



When I conducted this regression analysis in R, it gave me the following output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.141176  2.268e+00   11.97   <2e-16 ***
income       0.0028968  2.833e-04   10.22   <2e-16 ***
```

Interpret those p-values and the R-squared value.

43. Write out the full and reduced models. Complete the ANOVA summary table. How did we already know the MSR?

Full model: _____ Reduced model: _____

Source of variation	SS	df	MS	MSR (F)
Regression ($b_1 \mid b_0$)	15279	_____	15279	104.54
Error	14616	_____	146.16	(blank)
Total	29895	_____	MS_{total}	$R^2 = 0.5111$

44. Calculate and interpret the RMSE (root mean square error). What does it mean in this study?

45. Use the omnibus F-test to verify the F-statistic from the ANOVA summary table on the previous page.

46. With our least-squares regression line, we could predict the prestige of a job with an average income of $7000:

y = 27.141 + 0.002896(7000) = 47.41877

We know that prediction won't be perfectly accurate, so it might make sense to construct a confidence interval for our regression coefficients. Using R, I found the following confidence intervals:

```
                    2.5 %       97.5 %
(Intercept) 22.642116976 31.640235760
income       0.002334692  0.003458907
```

Interpret the 95% confidence interval for the slope of our regression line: (0.00233, 0.00345). Why does this not mean we're 95% confident that increasing an occupation's income by $1000 will be associated with a 2.33 – 3.45 increase in prestige.

We could use bootstrap methods or the following formula to construct a confidence interval for our regression line:

$$\hat{y} \pm \left(t_{n-2}^{\alpha/2}\right) s_{y|x} \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{X}\right)^2}{(n-1)s_x^2}}$$

where

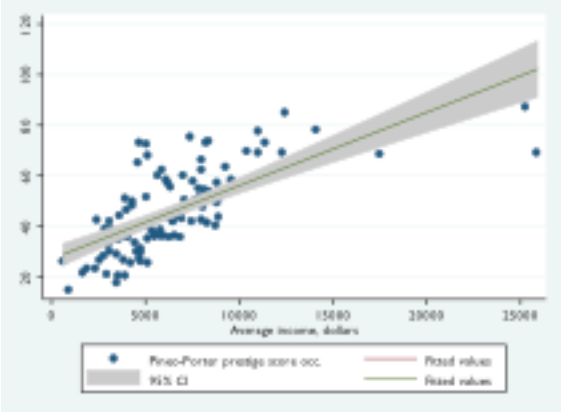
$$s_{y|x} = \sqrt{\frac{\left(y_i - \bar{Y}\right)^2}{(n-2)}} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{\left(1-R^2\right)SS_Y}{n-2}} = \sqrt{\frac{\left(1-R^2\right)(n-1)s_y^2}{n-2}} = s_y\sqrt{1-R^2}\sqrt{\frac{n-1}{n-2}} = \sqrt{MSE}$$

A 95% confidence interval for the **average prestige of all occupations with $7000 incomes** is, then:

$$s_{y|x} = \sqrt{146.16} = 12.089$$

$$47.41877 \pm (1.984)(12.089)\sqrt{\frac{1}{102} + \frac{(7000 - 6797.90)^2}{(102-2)(4245.92)^2}} = 47.41877 \pm 2.38$$

47. Will this confidence interval have the same width (uncertainty) for all values of income? Explain.



The confidence interval is displayed on the plot to the right.

48. Based on our interpretation, this confidence interval didn't give us exactly what we wanted. We wanted an interval to predict the prestige of a single occupation that has a \$7000 income. The interval we calculated predicts the **average** prestige **all** occupations with \$7000 incomes.

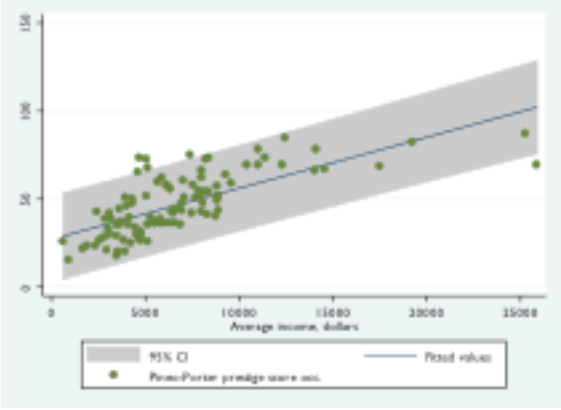
If we construct an interval to predict a single future observation, that interval must be **WIDER** **MORE NARROW** than our confidence interval.

To construct a prediction interval for our regression line, we use:

$$\hat{y}_i \pm \left(t_{n-2}^{\alpha/2}\right) s_{y|x} \sqrt{1+\frac{1}{n}+\frac{\left(x_0-\bar{X}\right)^2}{(n-1)s_x^2}}$$

$$47.41877 \pm (1.984)(12.09)\sqrt{1+\frac{1}{102}+\frac{(7000-6797.90)^2}{(102-1)(4245.92)^2}}$$

$$47.41877 \pm 24.10$$



The prediction interval is displayed to the right.

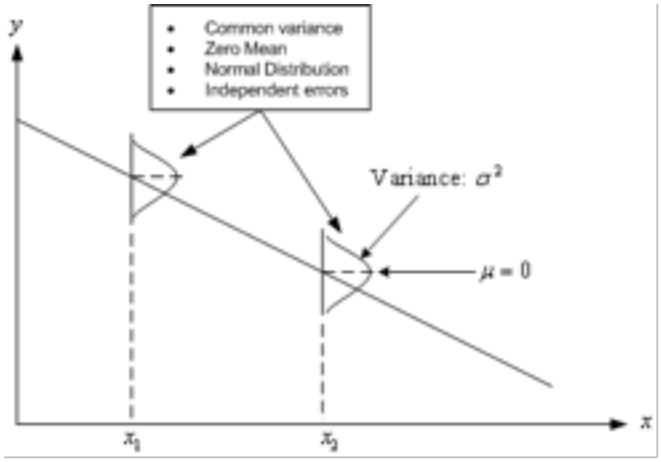
Obtaining confidence or prediction intervals in R is easy. Once you've specified your model, you apply the interval to new data using:

```
predict(model, newdata, interval="confidence")    predict(model, newdata, interval="predict")
```

The output, when our new data is a job with an income of \$7000, is:

```
     fit      lwr      upr                          fit      lwr      upr
47.41877 45.04112 49.79642                    47.41877 23.31552 71.52202
```

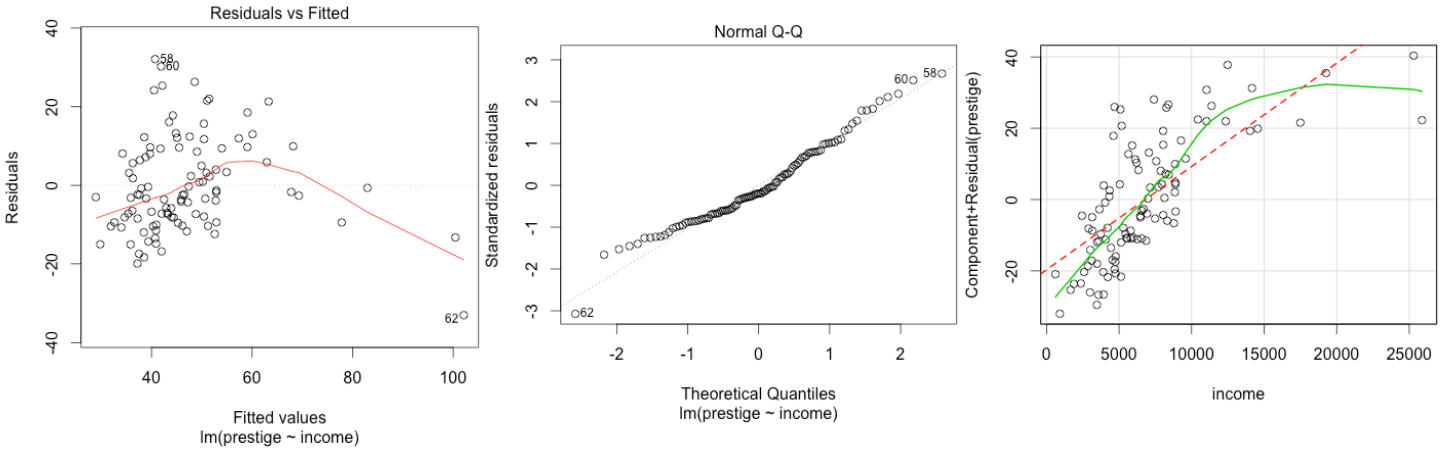
49. Recall the assumptions underlying regression. The diagram to the right attempts to display many of these assumptions.



Below, I've pasted output from the plots(model) command in R. Interpret these graphs and evaluate whether the assumptions appear to be satisfied in this case.

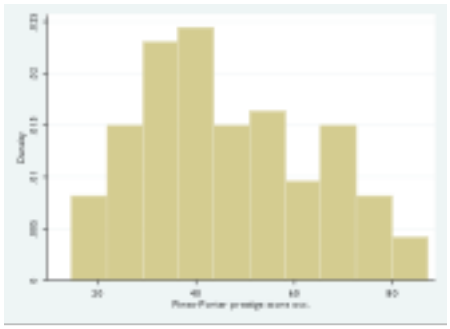
Non-constant error variance test:

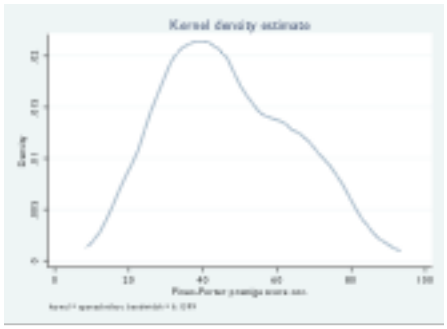
```
Variance formula: ~ fitted.values
Chisquare = 3.088455   Df = 1    p = 0.07885
```



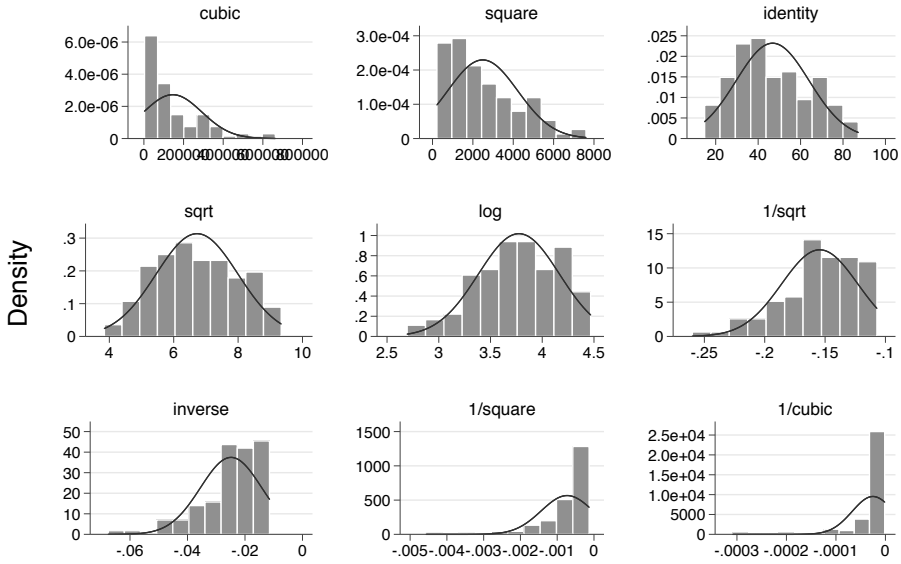
50. If we're worried about the normality and/or heteroscedasticity of our residuals, we have a few options.

a) We could transform our dependent variable to make it better approximate a normal distribution. Here's the distribution of our prestige data:





The figure on the next page displays the distributions we would get if we were to transform the prestige data using logarithms, exponents, or other transformations.



If a transformation makes the data better approximate a normal distribution, it may mean the residuals will better approximate a normal distribution. Be careful with this, though. Once you transform the data, your linear model may become much more difficult to interpret.

To learn more about transformations, check out http://onlinestatbook.com/2/transformations/tukey.html or http://onlinestatbook.com/2/transformations/box-cox.html

b) You could use *robust* regression methods (as we'll learn in a future activity). Interpret the following:

Ordinary least squares regression

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.141176   2.268e+00   11.97   <2e-16 ***
income       0.0028968  2.833e-04   10.22   <2e-16 ***
```

Robust linear regression

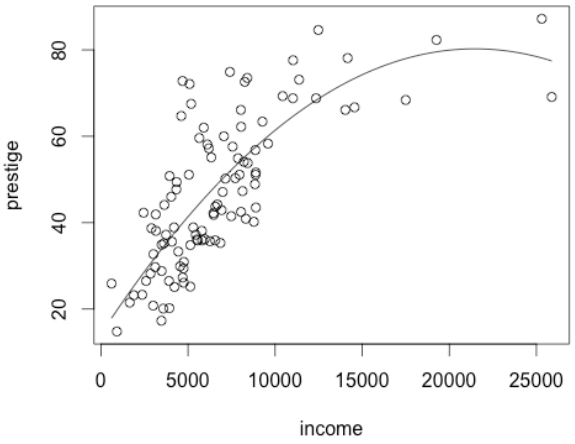
```
Robust linear regression                        Number of obs =       102
                                                F(  1,   100) =     48.28
                                                Prob > F      =    0.0000
                                                R-squared     =    0.5111
                                                Root MSE      =     12.09
------------------------------------------------------------------------------
             |               Robust
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0028968   .0004169     6.95   0.000     .0020697    .0037239
       _cons |   27.14118   2.886142     9.40   0.000     21.41515     32.8672
------------------------------------------------------------------------------
```

Quantile (median) regression

```
Quantile (Median) regression                    Number of obs =       102
  Raw sum of deviations     1447 (about 43.5)
  Min sum of deviations 954.6664                Pseudo R2     =    0.3402
------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0030293   .0003073     9.86   0.000     .0024196    .0036391
       _cons |   23.94584   2.518318     9.51   0.000     18.94957    28.94211
------------------------------------------------------------------------------
```

c) You might also want to try to fit a model that isn't linear. We'll also learn some of these methods in the future.

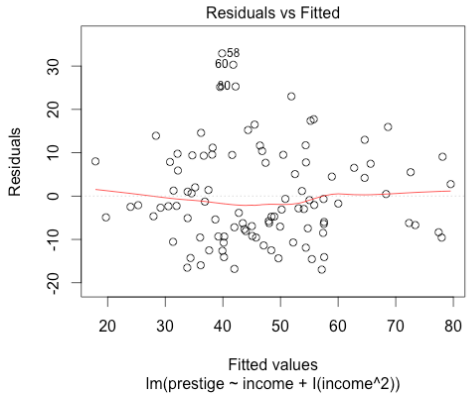
Model: $y = b_0 + b_1 x_1 + b_2 x_1^2 + e$



Best-fitting quadratic function:

$$y = 14.183 + 0.00615x - 0.000000143x^2$$

Below: Residuals vs. fitted plot:



Lowess (locally locally weighted scatterplot smoothing):