Activity #11: Multiple Linear Regression

Scenario: Recall our *prestige* dataset: http://www.bradthiessen.com/html5/data/prestige.csv

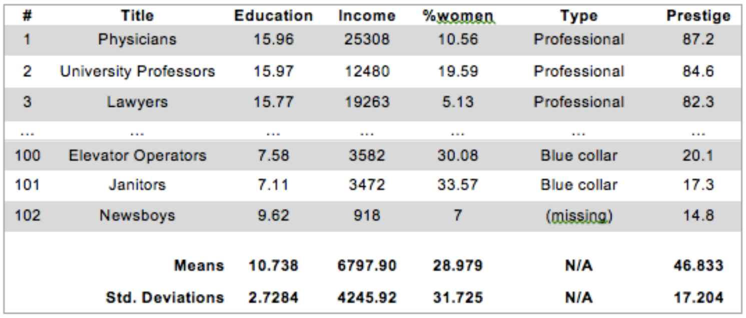
prestige: Pineo-Porter Prestige score (a survey)

education: average years of education for people in that occupation

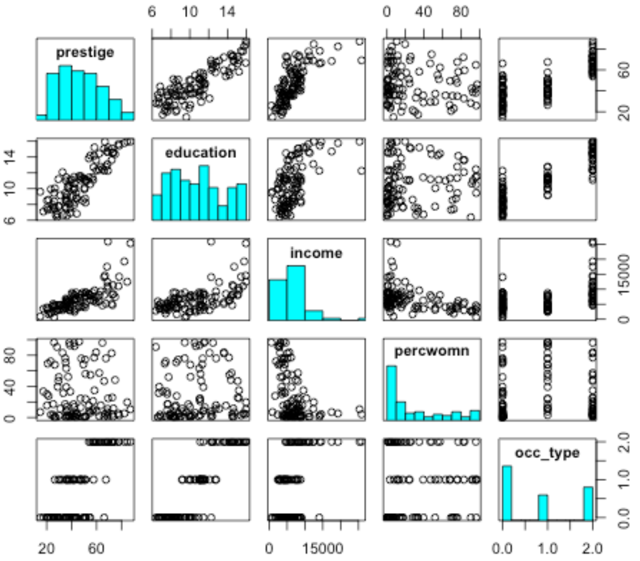
income: average income (1971 Canadian dollars) for people in that occupation

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
...	...	...	...	...	...	...
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
```

$$R^2_{\text{prestige, income}} = 0.5111$$

Source: Canada (1971). Census of Canada. Vol. 3, Part 6. Statistics Canada, 19-21.

1. Last time, we constructed a model in which prestige was predicted by income. We could have also chosen to model prestige as a function of education or % women. Some key results from 3 simple linear regression models are displayed below. Interpret these results.

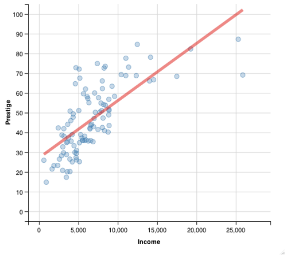
Model: prestige = b_0 + b_1(income)

Least-squares line: y = 27.14 + 0.003x

R^2 = 0.5111

RMSE = 12.09

F = 104.54 (p < 0.00001)



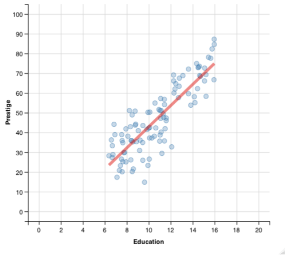
Model: prestige = b_0 + b_1(education)

Least-squares line: y = -10.7 + 5.36x

R^2 = 0.7228

RMSE = 9.10

F = 260.75 (p < 0.00001)



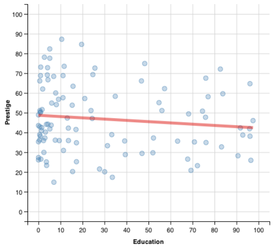
Model: prestige = b_0 + b_1(% women)

Least-squares line: y = 48.7 – 0.06x

R^2 = 0.014

RMSE = 17.17

F = 1.42 (p = 0.2362)



2. The R^2 values from those 3 models sum to 1.2479. How is that possible?

3. The F-test for our first model (in which prestige is a linear function of income) indicates income is a *significant* predictor of prestige (at least compared to a reduced model with no predictors).

   All our models thus far have employed a single independent variable (predictor). We visualized this as a straight line through a 2-dimensional scatterplot of data. Could we improve our prediction by adding another predictor variable? Instead of fitting a 1-dimensional line to a 2-dimensional scatterplot, we'd try to fit a 2-dimensional plane to a 3-dimensional scatterplot.

   Let's try this. Let's compare a reduced model with no predictors to a full model that attempts to predict prestige from both income and education. Write out these models:

   Full model: ______________________________ Reduced model: ______________________________

4. We know simple formulas to calculate the slope and y-intercept of the least-squares regression line, but how do we find the coefficients for the best-fitting plane? Suppose we add another predictor. How would we find the coefficients for the best-fitting hyperplane?

   Linear algebra isn't a prerequisite for this course, so I won't go into much detail, but we can use matrix algebra to find the coefficients for the best-fitting function. Suppose we have *n* observations in our dataset, with *p* predictors in our full model. Our full model, then, in matrix notation is:

$$Y = Xb + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,\mathrm{p}} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,\mathrm{p}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,\mathrm{p}} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

   We could then show the least-squares solution to estimating the coefficients as: $b = \left(X^T X\right)^{-1} X^T Y.$

   To learn more about this, see http://www.stat.purdue.edu/~jennings/stat514/stat512notes/topic3.pdf.

5. I used R to compute the least-squares solutions for our full and reduced models. Interpret the coefficients.

   Reduced: $\hat{y} = \overline{Y} = 46.833$

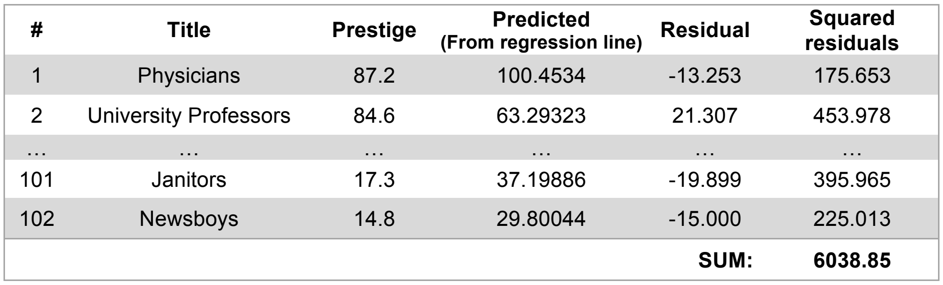
   Full: $\hat{y} = -6.8478 + 0.0014x_1 + 4.1374x_2$

   Full: $\hat{y} = -6.8478 + 0.0014(\text{income}) + 4.1374(\text{education})$

   From the coefficients, can we determine which variable (income or education) is the better predictor of prestige?

6. In the last activity, we used the omnibus F-test to compare the full and reduced models. This test relied on knowing the R^2 values for each model. We know $R^2 = 0$ for the reduced model (with no predictors), but how do we calculate R^2 for the full model? What does it mean to have a correlation among more than two variables?

   When we have two variables, X and Y, the correlation coefficient (R) can be interpreted as the correlation between the observed and predicted Y values. With this definition, we can calculate R with multiple predictors -– we just need to calculate the correlation between our observed Y values and those predicted by the X variables. The following table displays the predicted prestige scores based on our income and education predictors:



#	Title	Prestige	Predicted (From regression line)	Residual	Squared residuals
1	Physicians	87.2	100.4534	-13.253	175.653
2	University Professors	84.6	63.29323	21.307	453.978
…	…	…	…	…	…
101	Janitors	17.3	37.19886	-19.899	395.965
102	Newsboys	14.8	29.80044	-15.000	225.013
				SUM:	**6038.85**

   A computer can calculate the multiple correlation between the observed and expected prestige scores to be 0.893. If we square this value, we get: $R_{y,x_1,x_2} = 0.798$. Interpret this value.

7. Now let's compare our models to see if the full model provides a significantly better prediction than the reduced model. To do this, we can use the omnibus F-test or fill-in an ANOVA summary table. Calculate the F-test:

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

   Now, let's fill-in our ANOVA summary table:

Source	**SS**	**df**	**MS**	**MSR (F)**
Regression (b_1 , b_2 | b_0)		k =		195.55
Error		n - k -1 =		$p < 0.0001$
Total		n -1 =	MS_{total}	$R^2 = 0.798$

   You can verify your calculations using the output pasted on the top of the next page. What conclusion(s) can we make from this?

```
      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  2,    99) =  195.55
       Model |  23856.5752     2  11928.2876           Prob > F      =  0.0000
    Residual |  6038.85086    99  60.9984935           R-squared     =  0.7980
-------------+------------------------------           Adj R-squared =  0.7939
       Total |  29895.4261   101  295.994318           Root MSE      =  7.8102

------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0013612   .0002242     6.07   0.000     .0009163    .0018061
   education |   4.137444    .348912    11.86   0.000     3.445127    4.829762
       _cons |  -6.847778   3.218977    -2.13   0.036    -13.23493   -.4606292
------------------------------------------------------------------------------
```

8. Let's add one more predictor to our model. Let's see if the combination of income, education, and % women predict prestige better than a model with no predictors. To do this, we would compare:

Reduced: $\hat{y} = b_0 = \overline{Y} = 46.833$

Full: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

Full: $\hat{y} = -6.794 + 0.0013(\text{income}) + 4.1866(\text{education}) - 0.0089(\%\text{women})$

$R^2_{y,\,x_1,\,x_2,\,x_3} = 0.7982$

Once again, make sure you can interpret those coefficients and the squared multiple correlation.

Conduct a test to compare the full and reduced models.

9. If we added another predictor to our model (**any** predictor), what would happen to the value of R^2? Because R^2 will always increase when extra predictors are added (even if those predictors are almost completely unrelated to the dependent variable), it might be better to use another statistic to evaluate the fit of our model.

We can define an *adjusted R-squared* that will only increase only if the additional predictor improves the prediction more than would be expected by chance:

$$R^2_{\text{adjusted}} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k-1} = R^2 - \left(1 - R^2\right)\frac{k}{n-k-1} = 1 - \frac{MS_E}{MS_{\text{Total}}}$$

For this most recent example,

$$R^2_{\text{adjusted}} = R^2 - \left(1 - R^2\right)\frac{k}{n-k-1} = 0.7982 - \left(1 - 0.7982\right)\frac{3}{102-3-1} = 0.798$$

Based on this adjusted R-squared value, what can we conclude about the predictor *% women*?

10. Let’s look one last time at our full model with 3 predictor variables:

$$\text{Full: } \hat{y} = -6.794 + 0.0013(\text{income}) + 4.1866(\text{education}) - 0.0089(\%\text{women})$$

Explain, again, why we can’t simply compare the magnitude of the coefficients to determine which is the most potent predictor of prestige?

If we really want to compare coefficients in our model, we could calculate *standardized beta coefficients*. One way to do this would be to convert all our variables (prestige, income, education, %women) to z-scores before conducting the regression. We could also run a regression with the (untransformed) variables and then convert the coefficients using the following transformation:

$$\beta_k = b_k \frac{s_{x_k}}{s_y}$$

As an example, suppose we want to convert the coefficient of education to a standardized beta coefficient:

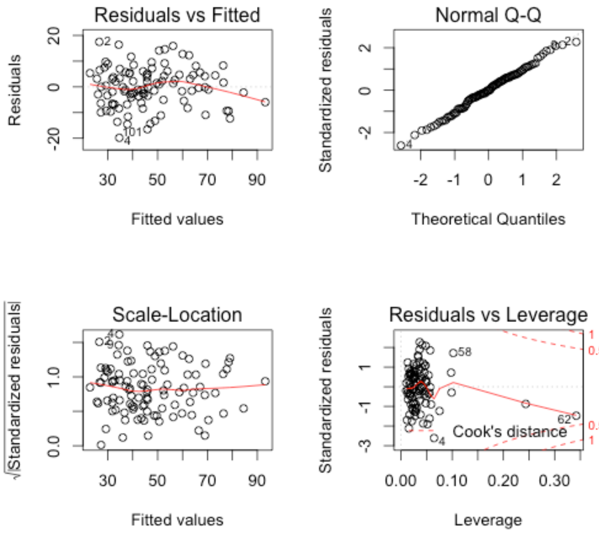
$$\beta_2 = b_2 \frac{s_{x_2}}{s_2} = 4.1866 \frac{2.7284}{17.204} = 0.66396$$

Converting all the coefficients yields the following. Interpret these coefficients.

$$\hat{y} = 0.32418(z_{\text{income}}) + 0.66396(z_{\text{education}}) - 0.01642(z_{\%\text{women}})$$

Explain why we still must be extremely cautious in comparing these standardized coefficients.

11. To the right, I’ve plotted some graphs of the residuals. From these graphs, what can we conclude about the assumptions necessary to conduct a linear regression analysis?



12. Look at the coefficients for our models with two and three predictors:

$$\hat{y} = -6.848 + 0.0014(\text{income}) + 4.1374(\text{education})$$
$$\hat{y} = -6.794 + 0.0013(\text{income}) + 4.1866(\text{education}) - 0.0089(\%\text{women})$$

Notice that the coefficients didn't change significantly when we added a new predictor. That's a good sign that we don't have a *multicollinearity* problem.

Multicollinearity is when two or more predictors in our model are highly correlated with each other (meaning that one can be linearly predicted from the others). To read more about multicollinearity, including how to detect and deal with it, visit: http://en.wikipedia.org/wiki/Multicollinearity

13. Thus far, we've only analyzed the **total contribution** of several independent variables. In other words, we've always compared our full models to a reduced model with no predictors. Suppose, instead, we are interested in finding a model that adequately predicts prestige using the fewest variables possible.

In this example, we found income was a significant predictor of prestige. In fact, we found $R^2_{Y,1} = 0.5111$ led to an omnibus F-statistic of 104.54.

We also found that the combination of income and education provided a significantly better prediction of prestige than a model with no predictors. For this model, we found $R^2_{Y,12} = 0.7980$.

Our question, now, is: *Did adding education as a predictor significantly improve our prediction?*

To answer this, let's write out the full and reduced models we'd like to compare:

Full model: ______________________________ Reduced model: ______________________________

14. To compare these models, we can use an omnibus F-test or fill-in the ANOVA summary table. Verify these calculations.

Source	**SS**	**df**	**MS**	**MSR (F)**
income & education	23856.55	2	11928.3	195.55
income	15276.56	1	15276.6	104.54
education | income	8579.99	1	8580	**140.7**
Error	6038.876	99	61	
Total	29895.426	101	MS_{total}	

15. Use the omnibus F-test to verify that F-statistic of 140.7.

16. Calculate and interpret $R^2_{\text{Y}2|1}$

17. Before we move on to another example, let's attempt to answer one final question. So far, we've shown:
    - Income is a significant predictor of prestige (compared to a model with no predictors)
    - The combination of income and education are significant predictors of prestige (compared to a null model)
    - The combination of income, education, and %women are significant predictors of prestige (vs. null model)
    - Education significantly improves the prediction of prestige over a model with only income as the predictor

    Our final question is: *Does %women significantly improve our prediction over a model with income and education?*
    or
    Should we add %women to predict prestige if we're already using income and education?

    Write out the full and reduced models of interest.

    Full model: ______________________________ Reduced model: ______________________________

    Using R, I calculated the following multiple correlation coefficients:

    $R^2_{\text{Y}1} = 0.511$ $R^2_{\text{Y}12} = 0.798$ $R^2_{\text{Y}123} = 0.7982$
    $R^2_{\text{Y}2} = 0.723$ $R^2_{\text{Y}13} = 0.559$
    $R^2_{\text{Y}3} = 0.014$ $R^2_{\text{Y}23} = 0.752$

    Use the omnibus F-test to answer our question:

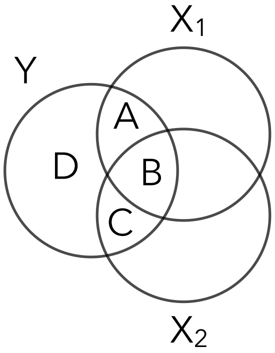
18. Just to review a concept we learned last time, we could use our full model (with all three predictors) to predict the prestige of a job with: income = 5000, education = 10, %women = 40. Using R, I came up with a 95% confidence interval and a 95% prediction interval for this prediction. Interpret both intervals:

    Predicted prestige = 42.74

    Confidence interval: (39.78, 45.70)

    Prediction interval: (27.27, 58.21)

19. The following figure and table attempt to visualize the contribution of two predictors on a dependent variable.

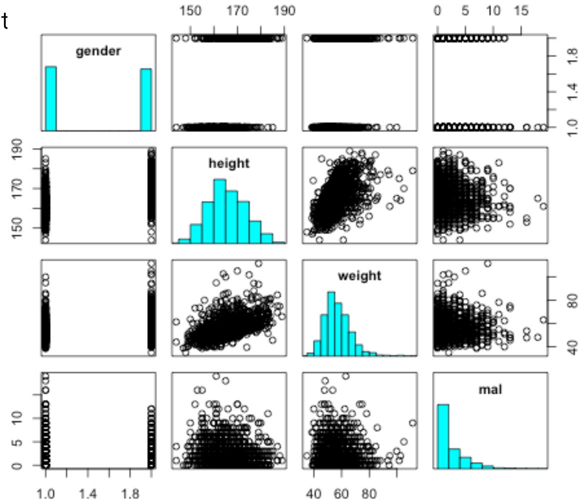


Effect	SSSS_{REG}	R^2 Values
X_1 and X_2 together	$SS_{X_1X_2} = A+B+C$	$R^2_{Y12} = \frac{A+B+C}{A+B+C+D}$
X_1 alone	$SS_{X_1} = A+B$	$R^2_{Y1} = \frac{A+B}{A+B+C+D}$
X_2 alone	$SS_{X_2} = B+C$	$R^2_{Y2} = \frac{B+C}{A+B+C+D}$
$X_1 \mid X_2$ = "X_1 unique"	$SS_{X_1\mid X_2} = (A+B+C)-(B+C) = A$	$R^2_{Y1\mid 2} = \frac{A}{A+B+C+D}$
$X_2 \mid X_1$ = "X_1 unique"	$SS_{X_2\mid X_1} = (A+B+C)-(A+B) = C$	$R^2_{Y2\mid 1} = \frac{C}{A+B+C+D}$

20. Let's turn to a simple dataset to practice our multiple regression tests and to investigate the concept of interaction. The **htwt** dataset lists 4 measurements for 1000 subjects:

    y = weight = weight of each subject at age 16 (in kg)
    x1 = height = height of each subject at age 16 (in cm)
    x2 = gender = female or male
    x3 = mal = malaise score for each subject at age 22

variable	mean	std. dev
weight	57.17209	9.656277
height	166.163	8.025138
gender	(50.9% female, 49.1% male)	
mal	2.591	2.842851



21. Suppose we're interested in modeling an individual's weight as a function of their height. A computer would find:

$$\hat{y} = -46.764 + 0.62551(\text{height})$$

We could use the omnibus F-test to determine if height is a significant predictor of weight, but I'm interested in a different question: *Is this prediction the same for males and females?*

Before we address that question, let's see how well the combination of height and gender predict weight. To do this, we'd compare: Full: $\hat{y} = b_0 + b_1(\text{height}) + b_2(\text{female})$

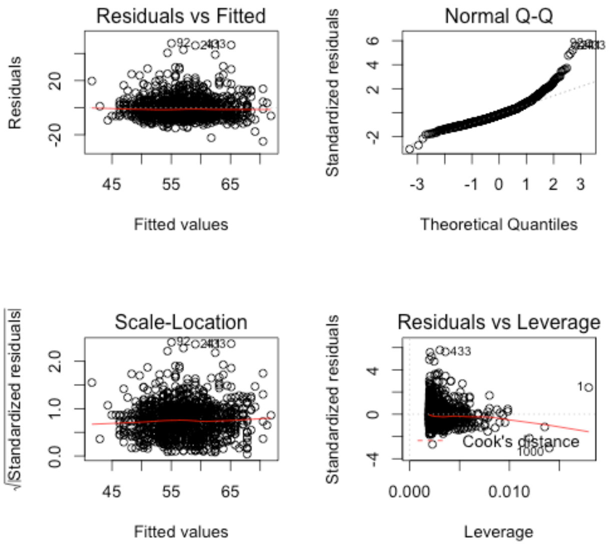
Reduced: $\hat{y} = b_0$

A computer found the following least-squares coefficients: Full: $\hat{y} = -53.788 + 0.67175(\text{height}) - 1.3439(\text{male})$

Reduced: $\hat{y} = 57.17209$

Interpret the coefficient for the gender variable (female = 0, male = 1). What does -1.3439 represent?

22. Interpret the rest of the output from running this regression analysis:



```
Analysis of Variance Table

           Df  Sum Sq  Mean Sq  F value  Pr(>F)
hght+gndr   2  25487  12743.5  187.7700  < .0001
height      1  25173  25172.9  370.9122  < .0001
gender      1    314    313.8    4.6231  0.03178
Residuals 997  67664     67.9

RMSE = 8.24

R-squared = 0.2736
Adjusted R-squared = 0.2721
```

23. I also had the computer calculate all possible multiple correlations. Interpret these values.

$R^2_{Y1} = 0.2702$ $\quad R^2_{Y12} = 0.2736$ $\quad R^2_{Y123} = 0.2741$

$R^2_{Y2} = 0.0569$ $\quad R^2_{Y13} = 0.2712$

$R^2_{Y3} = 0.0029$ $\quad R^2_{Y23} = 0.0570$

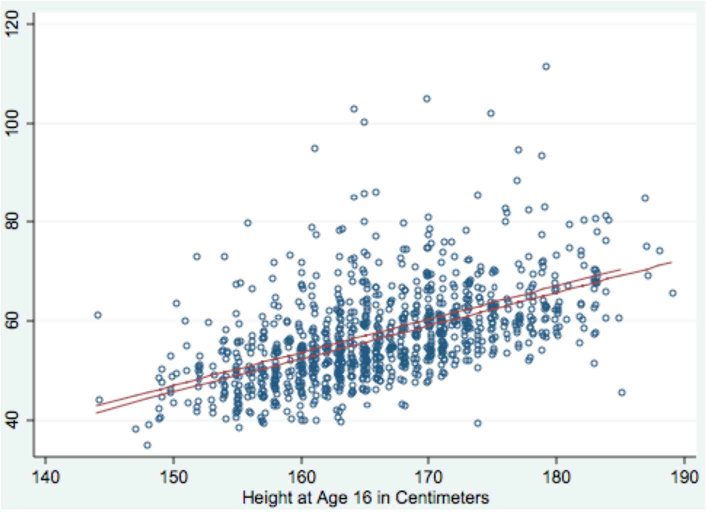
24. Suppose we want to know if gender explains variation in weight beyond what height already explains. In other words, suppose we want to know if gender significantly improves our prediction of weight, after controlling for height. Write out the full and reduced models, run an omnibus F-test, and write out your conclusions.

25. Now let's go back to our question: *Does height predict weight the same way for males and females?*

When we choose a model such as

$$\hat{y} = -53.788 + 0.67175(\text{height}) - 1.3439(\text{male})$$

we're indicating that weight differs by a constant amount for males and females. No matter what height we substitute into this model, males and females with that same height will differ by 1.3439 kg (see the parallel regression lines to the right)



If we want to model an **interaction** between height and gender, we need to put that into our model. We could do this in one of two ways:

a) Split our data into two sets (one dataset for males and another for females). We could then run a separate regression analysis for each dataset.

b) Incorporate an interaction (product) term into our model and run a single regression analysis.

26. Using option (a), I split the data into two groups and conducted two regression analyses. I found the following coefficients:

$$\text{Males: } \hat{y} = -72.01376 + 0.77066(\text{height})$$

$$\text{Females: } \hat{y} = -33.75055 + 0.54792(\text{height})$$

Using option (b), I input the following model into R:

$$\text{Full model: } \hat{y} = b_0 + b_1(\text{height}) + b_2(\text{female}) + b_{12}(\text{height x female})$$

and obtained these coefficients:

$$\hat{y} = -33.75 + 0.5479(\text{height}) - 38.2632(\text{male}) + 0.2227(\text{height x male})$$

$$R^2 = 0.2795,\ R^2_{\text{adjusted}} = 0.2773$$

27. The plot below displays how this interaction term allows the slopes to differ for males and females. To interpret this interaction term (and its coefficient), we can do some manual arithmetic:

For males: $\hat{y} = -33.75 + 0.5479(\text{height}) - 38.2632(\text{male}) + 0.2227(\text{height x male})$

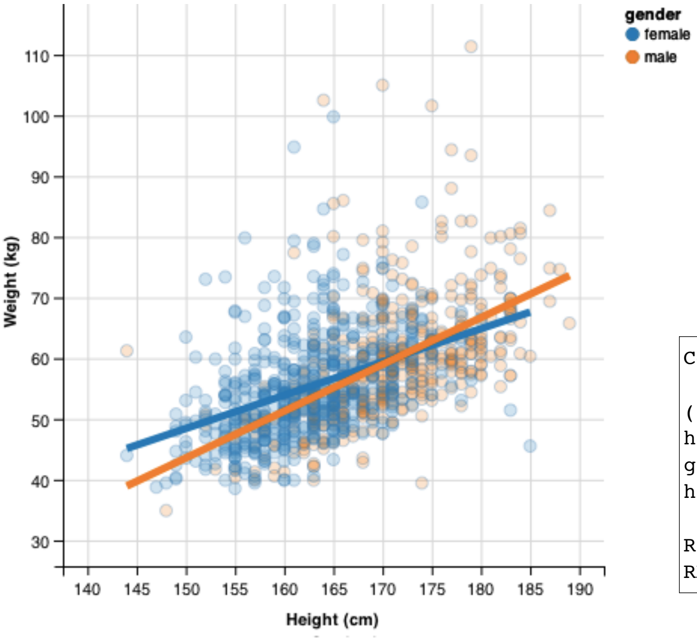
$$\hat{y} = -33.75 + 0.5479(\text{height}) - 38.2632(1) + 0.2227(\text{height x } 1)$$

$$\hat{y} = -72.0132 + 0.7706(\text{height})$$

For females: $\hat{y} = -33.75 + 0.5479(\text{height}) - 38.2632(\text{male}) + 0.2227(\text{height x male})$

$$\hat{y} = -33.75 + 0.5479(\text{height}) - 38.2632(0) + 0.2227(\text{height x } 0)$$

$$\hat{y} = -33.75 + 0.5479(\text{height})$$



Notice these coefficients (from the model with the interaction effect) are the same as what we got from the two separate regression analyses:

$$\text{Males: } \hat{y} = -72.01376 + 0.77066(\text{height})$$

$$\text{Females: } \hat{y} = -33.75055 + 0.54792(\text{height})$$

With this full model (including the interaction term), we could test the significance of each coefficient:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -33.75055    9.43230  -3.578 0.000363
height               0.54792    0.05825   9.407  < 2e-16
gendermale         -38.26321   12.96338  -2.952 0.003235
height:gendermale    0.22274    0.07812   2.851 0.004445

R-squared = 0.2795     Adjusted R-squared = 0.2773
RMSE = 8.209
```

28. The p-value for the interaction term (p = 0.004445) indicates the interaction term improves our prediction. We could also use the omnibus F-test to determine if this interaction term significantly improves our prediction (beyond a model without the interaction term). Below, I've conducted the omnibus F-test and then pasted output from R that compared the full model (with interaction) to a reduced model without interaction. Interpret.

$$F_{996}^{1} = \frac{(.2795 - .2736)/(3-2)}{(1 - .2795)/(1000 - 3 - 1)} = 8.13 \quad (p = 0.00445)$$

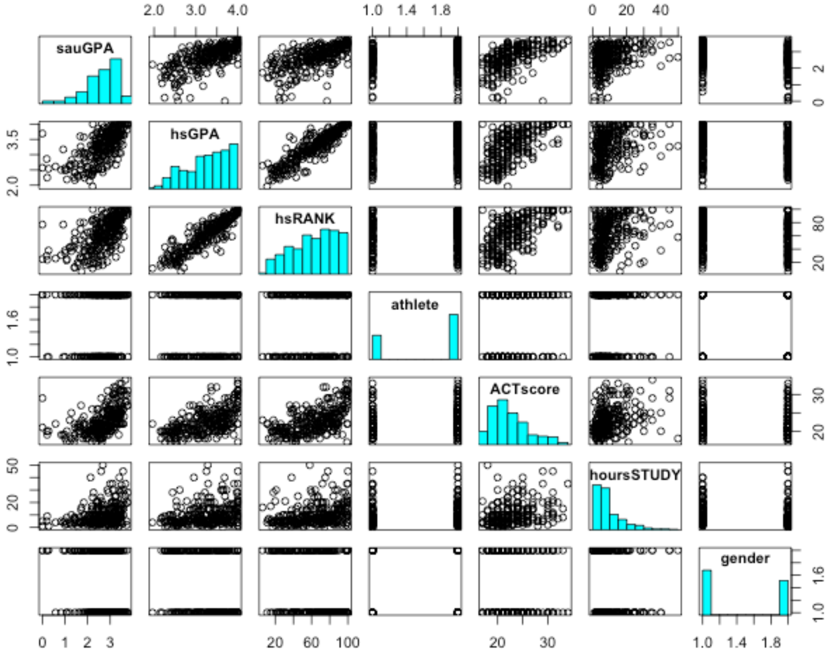
```
Analysis of Variance Table

Model 1: weight ~ height + gender
Model 2: weight ~ height * gender
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1    997 67664
2    996 67116  1    547.83 8.1297 0.004445
```

Notice the value of our F-statistic is equal to the t-statistic (from question #27) squared

Scenario: Let's see how well we can predict the fall semester GPAs of St. Ambrose freshmen based on:

- HS GPA = high school GPA
- Athlete = student athlete?
- Hours studying = hours studying per week
- HS %ile rank = high school percentile rank
- ACT score = ACT Composite score
- Gender = male or female

	y	x_1	x_2	x_3	x_4	x_5	x_6
Student	1st sem. GPA	HS GPA	HS %ile rank	Athlete	ACT score	Hours studying	Gender
1	2.87	2.82	43	no	24	5	male
2	3.16	3.49	76	no	32	7	male
…	…	…	…	…	…	…	…
255	1.69	3.26	70	yes	21	4	male
Mean	2.65	3.275	63.27	34.9%	22.96	10.62	56.5%
Std. Dev	0.75	0.52	24.48	athletes	3.66	8.90	female



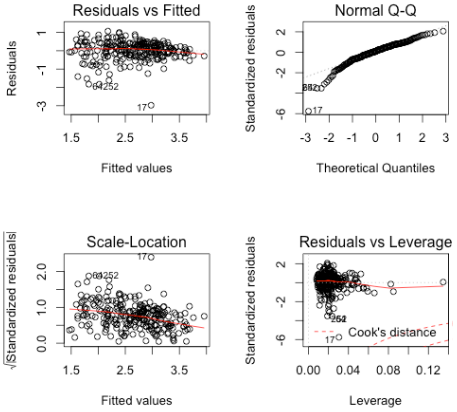
Data: http://www.bradthiessen.com/html5/data/gpadata.csv
Note: I only kept records with no missing data. How could we handle missing data?

29. We're going to investigate several questions regarding St. Ambrose first semester GPAs:
    a) How well do ACT scores predict first-semester GPAs at St. Ambrose?
    b) Do ACT add to our prediction of SAU GPAs beyond what high school GPAs predict?
    c) Do student athletes have higher or lower SAU GPAs?
    d) Do the self-reported hours studying per week predict SAU GPAs beyond ACT and high school data?
    e) Is there an interaction between gender and athletics?

Before we address these questions, though, let's see how good of a prediction we can get if we use all our predictor variables. One thing to notice is the strong correlation (r = 0.903) between HS GPA and HS Rank. If we include both predictors, we'll have a multicollinearity problem. I'm only going to include the HS GPA predictor.

On the top of the next page, I fit a full model with all our predictor variables (no interaction terms) and compared it to a reduced model with no predictors. Interpret this output.

```
Full model:          Coefficient  95% Confidence interval
                     (intercept)       -1.043, -0.054
R² = 0.5184          hsGPA             +0.525, +0.852
adj. R² = 0.5088     Not athlete       -0.178, +0.120
RMSE = 0.5248        ACTscore          +0.020, +0.066
F = 53.612           hoursSTUDY        +0.002, +0.017
p < 2.2e-16          Male              -0.422, -0.128
```



Which predictors have significant coefficients?

30. How much better of a prediction could we get if we include **all** the interaction terms, too? To check this, I fit a model with all the interaction terms and obtained an R-squared value of 0.55. What does this tell us about those interaction terms? Below, I've pasted output from an F-test comparing the model with interaction terms to a model with no interaction terms. Interpret.

```
Analysis of Variance Table
Model 1: sauGPA ~ hsGPA + athlete + ACTscore + hoursSTUDY + gender
Model 2: sauGPA ~ hsGPA * athlete * ACTscore * hoursSTUDY * gender
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    249 68.583
2    223 64.091 26     4.492 0.6011 0.9384
```

31. Looking, once again, at the output at the top of this page, it looks like the **athlete** variable does not help us predict first-semester GPAs. Let's eliminate that variable, estimate the coefficients, and compare it to the full model that did have the athlete variable.

```
No athlete model:    Coefficient  95% Confidence interval
R² = 0.5181          (intercept)       -1.052, -0.080
adj. R² = 0.5104     hsGPA             +0.526, +0.852
RMSE = 0.5239        ACTscore          +0.020, +0.065
F = 67.21            hoursSTUDY        +0.002, +0.017
p < 2.2e-16          Male              -0.399, -0.129
```

```
Analysis of Variance Table
Model 1: sauGPA ~ hsGPA + ACTscore + hoursSTUDY + gender
Model 2: sauGPA ~ hsGPA + athlete + ACTscore + hoursSTUDY + gender
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    250 68.624
2    249 68.583  1  0.041239 0.1497 0.6991
```

Eliminating the athlete variable didn't have much of an impact on our R-squared and RMSE values. Also, notice that now all our predictors are now significant. The F-test at the bottom says the model without the athlete variable does not differ significantly from the model with the athlete variable. All of this indicates it's safe to eliminate the athlete variable from our prediction.

32. We could continue this **backwards-selection process** by eliminating another variable and seeing how much of an impact it has on the predictive potency of our model. Let's do this by eliminating the **hours studying** variable:

```
No hours model:     Coefficient  95% Confidence interval
R^2 = 0.5072        (intercept)       -1.128, -0.156
adj. R^2 = 0.5014   hsGPA             +0.541, +0.869
RMSE = 0.5288       ACTscore          +0.026, +0.070
F = 86.13           Male              -0.411, -0.139
p < 2.2e-16

      Model 1: sauGPA ~ hsGPA + ACTscore + gender
      Model 2: sauGPA ~ hsGPA + ACTscore + hoursSTUDY + gender
        Res.Df    RSS Df Sum of Sq      F  Pr(>F)
      1    251 70.175
      2    250 68.624  1    1.5516 5.6525 0.01818 *
```

Eliminating the hours studying variable did have an impact on our predictive potency. While our R-squared dropped just slightly (from 0.5181 to 0.5072), that was a significant drop. If we don't care so much about a 1% drop, we could decide to keep the variable in the model. Otherwise, we could choose to eliminate it.

Let's say that we decide this model is our "best" model. We could then use our model to make predictions:

$$\hat{y} = -0.6416 + 0.7051(\text{hsGPA}) + 0.0480(\text{ACTscore}) - 0.2753(\text{male})$$

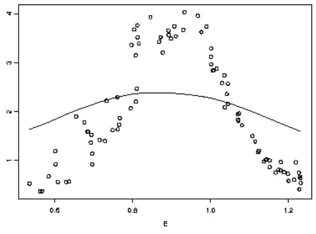
We know how well this model fits the data we used to estimate the coefficients, but how accurate would this model be for new data?

The data we used were from first-year students in 2013. Suppose I gathered high school GPAs, ACT scores, and gender for this year's first-year students. I could then predict the Fall GPAs of these students using our model.

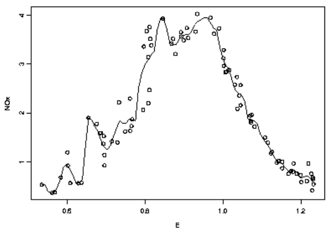
On the 2013 data, our model had an R-squared value of 0.5072. If we fit our model to this year's data, would you expect the R-squared value to be greater than, less than, or equal to 0.5072? Explain.

33. What could we do to ensure we don't *overfit* the data we use to estimate our model?

Bias-Variance Tradeoff



High Bias - Low Variance



Low Bias - High Variance

"overfitting" - modeling the random component

34. Let’s quickly investigate some other questions we can address with our data. Determine what models we could fit to address each question. Then, we can use our data in-class to attempt to answer each question.

a) How well do ACT scores predict first-semester GPAs at St. Ambrose?

Full model: ______________________ Reduced model: ______________________

How could we attempt to answer the question?

b) Do ACT add to our prediction of SAU GPAs beyond what high school GPAs predict?

Full model: ______________________ Reduced model: ______________________

How could we attempt to answer the question?

c) Do the self-reported hours studying per week predict SAU GPAs beyond ACT and high school GPA?

Full model: ______________________ Reduced model: ______________________

How could we attempt to answer the question?

35. The final question is: *Do student athletes have higher or lower SAU GPAs?*

To address this question, we could conduct a t-test (or randomization-based test of the two groups):

```
Two Sample t-test
data:  sauGPA by athlete

sample estimates:    athlete mean = 2.501573       not athlete mean = 2.729

alternative hypothesis: true difference in means is not equal to 0
t = -2.3323, df = 253, p-value = 0.02047

95 percent confidence interval:  -0.41952830 -0.03539793
```

From this, what would we conclude?

36. We could also address this question by comparing:

$$\text{Full: } \hat{y} = b_0 + b_1(\text{athlete})$$

$$\text{Reduced: } \hat{y} = b_0$$

$$F = \frac{(0.02105 - 0)/(1-0)}{(1-0.02105)/(255-1-1)} = 5.44 \quad (p = 0.02047)$$

How does this compare to the t-test?

37. As we'll soon see, the t-test (and ANOVA) are simply special cases of linear regression. Regression allows us, though, to develop and test more complex models. For example, we have already concluded that athletes have lower GPAs than non-athletes. Would this difference hold if we controlled for ACT scores? In other words, if we have two students with the same ACT score, does being an athlete have an association with a lower GPA. To test this, we could compare:

$$\text{Full: } \hat{y} = b_0 + b_1(\text{ACTscore}) + b_2(\text{athlete})$$

$$\text{Reduced: } \hat{y} = b_0 + b_1(\text{ACTscore})$$

$$F = \frac{(0.2955 - 0.2875)/(2-1)}{(1-0.2955)/(255-2-1)} = 2.8587 \quad (p = 0.09212)$$

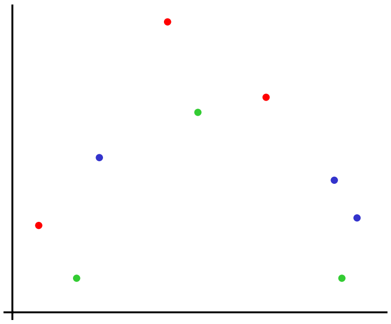
What conclusions can we make? Do athletes have lower first-semester GPAs?

38. Let's revisit the questions about choosing the best model and ensuring our model doesn't overfit our data.

These are extremely important questions in predictive modeling. Here's a quick overview of some approaches:

Cross-validation: Read more about this method at http://www.autonlab.org/tutorials/overfit10.pdf

- Randomly divide the data into k pieces (let's say k = 10)
- Use k-1 of those pieces (90% of the data; called the *training set*) to estimate the model coefficients
- Compute prediction error on the remaining piece (10% of the data; called the *test set*)
- Do this for each piece (10% of the data)
- Average the k (10) prediction error estimates. This gives us the predictive accuracy of the model.
- Repeat this process for other competing models. Whichever gives the smallest mean error is the "best"
- Estimate coefficients for that "best" model using all of the data



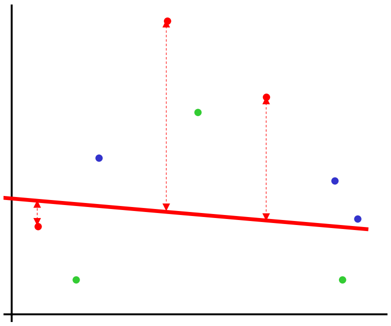
Let's see this process work on the small dataset pictured to the left.
We randomly split the data into 3 pieces (red, blue, and green dots)

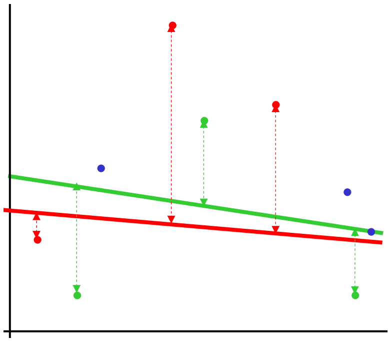
Below (left): We fit a model to the green and blue dots
and measure error using red dots

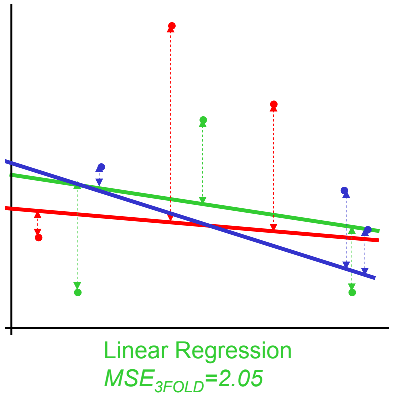
Below (middle): We fit a model to the red and blue dots
and measure error using green dots

Below (right): We fit a model to the red and green dots
and measure error using blue dots

We then take the average of those mean square errors.
We'd repeat this process with different models (other predictor variables) and choose the model that produces the smallest average mean square error.





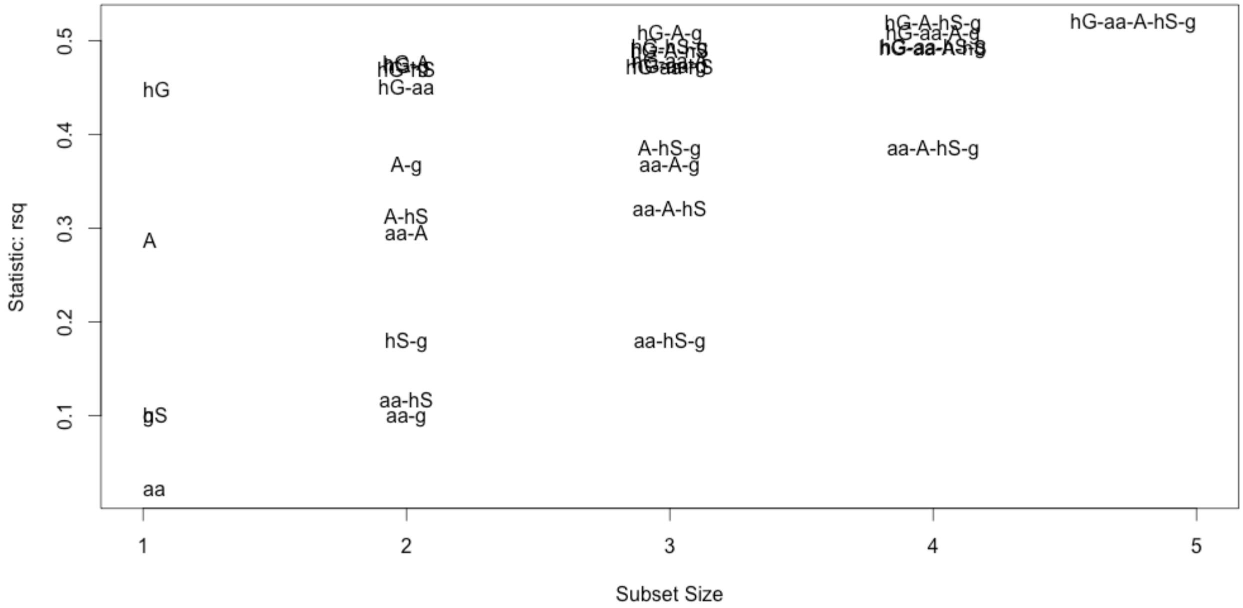


Average cross-validated mean square error:	Model
0.281	GPA = f(hsGPA + athlete + ACTscore + hoursSTUDY + gender)
0.280	GPA = f(hsGPA + + ACTscore + hoursSTUDY + gender)
0.283	GPA = f(hsGPA + + ACTscore + + gender)
0.297	GPA = f(hsGPA + + ACTscore)
0.311	GPA = f(hsGPA +)
0.337	GPA = f(full model including all possible interaction terms)

From this process (and the 6 models displayed above), what model would we choose as "best"?

Best subsets regression: Read more at https://onlinecourses.science.psu.edu/stat501/node/89

- With 5 predictor variables, how many different models could we fit from our data (not counting interactions)?
- Why not try to fit all of these models?
- Fit all the models with 1 predictor. Determine which model is best (possibly using R-squared values)
- Identify the best 2-, 3-, 4-, and 5-predictor models.
- Compare these best-fitting models and choose one



Above, I've printed a display of the R-squared values I obtained from all possible subsets.
The best one-predictor model includes high school GPA (hG)
The best two-predictor model includes high school GPA and ACT scores (hg + A)
It looks like there's not much difference between the 3- and 4-predictor models, so I might choose the simpler model.

Stepwise regression - Possibly using AIC (Akaike info. criterion), F-tests, or BIC (Bayesian information criterion)

Read more, including criticisms of this approach, at http://en.wikipedia.org/wiki/Stepwise_regression

- Add or subtract a predictor from your model,
- Test to see if that addition (or deletion) of a predictor made the prediction significantly better (or worse)

Below, I've pasted the final result from a stepwise regression using AIC.

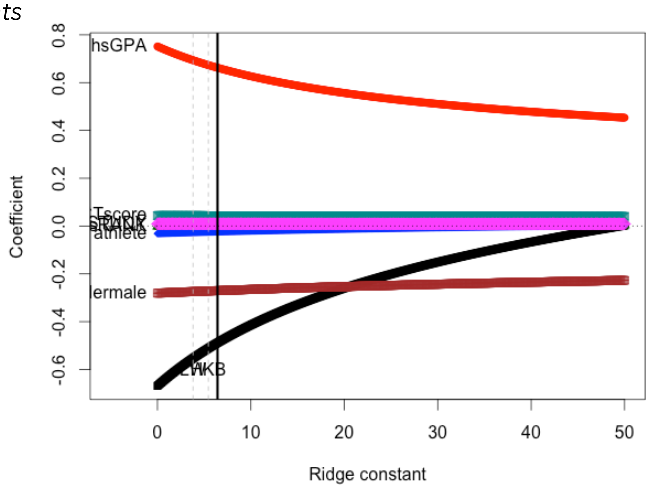
```
Call:
lm(formula = sauGPA ~ hsGPA + ACTscore + hoursSTUDY + gender,
    data = gpa)

Coefficients:
(Intercept)        hsGPA     ACTscore   hoursSTUDY   gendermale
    -0.5662       0.6889       0.0425       0.0093      -0.2640
```

Ridge regression: Read more at: http://en.wikipedia.org/wiki/Tikhonov_regularization

- With cross-validation, best-subsets, and stepwise methods, our model selection is a discrete process: individual predictors are either in or out.
- These methods can have high variance. A different set of data could lead to a completely different model.
- Ridge regression allows a predictor to be partly included in a model. It shrinks the size of the coefficient.
- The benefits of ridge regression are most apparent when we have a multicollinearity problem.

Below, I've pasted output using ridge regression. I used a model that included both HS GPA and HS Rank to introduce collinearity. Notice the smaller magnitude of the coefficients under the ridge regression method. Lambda was selected to be 6.43 to estimate the coefficients

```
                Linear         Ridge
                Regression     Regression
   Predictor    Coefficient    Coefficient
 (intercept)    -0.6688        -0.4884
       hsGPA    +0.7514        +0.6622
      hsRANK    -0.0016        +0.0002
 not athlete    -0.0316        -0.0243
    ACTscore    +0.0436        +0.0432
  hoursSTUDY    +0.0093        +0.0093
        male    -0.2808        -0.2702
```



The plot shows the shrinkage of the coefficients as we increased lambda. Note that we wouldn't want to <u>*use*</u> *the ridge regression coefficients (because they have bias). We use ridge regression to determine if our coefficient estimates are stable as we increase bias. If the estimates remain stable (like most in the plot displayed above), we have evidence that multicollinearity is not a problem.*

Lasso (least absolute shrinkage and selection operator): http://statweb.stanford.edu/~tibs/lasso/simple.html

- Ridge regression shrinks the magnitude of the coefficients, but it does not allow coefficients to become zero.
- With the Lasso, coefficients can drop to zero (so we can use this process to select variables in our model).

Note: Our course website contains other examples of multiple regression we can review together. The datasets associated with these examples can be found at:

- **csat:** http://www.bradthiessen.com/html5/data/csat.csv
- **teenage gambling:** http://www.bradthiessen.com/html5/data/gambling.csv
- **race/income:** http://www.bradthiessen.com/html5/data/income.csv
- **state cancer rates:** http://www.bradthiessen.com/html5/data/cancer.csv
- **smoking/birthweight:** http://www.bradthiessen.com/html5/data/birthweight.csv