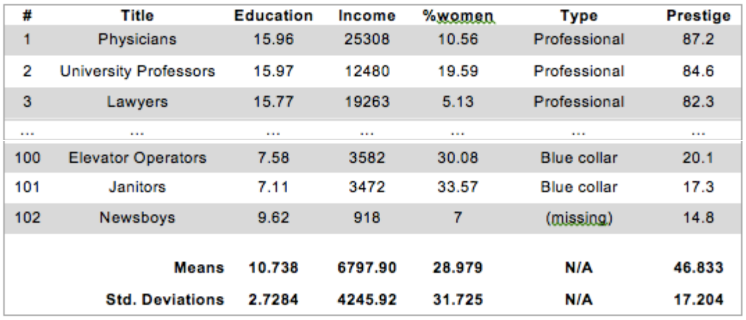
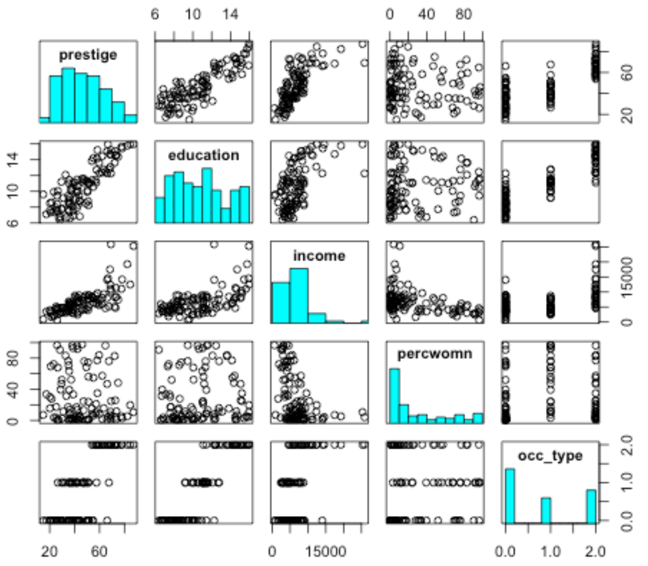
Lesson 11: Multiple regression & Model Selection

Packages used: mosaic, ggvis, dplyr, leaps, car, DAAG, MASS, genridge

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



#	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. In the previous lesson, we modeled prestige as a function of income. Suppose we wanted to know whether education, income, or %women best predicts prestige. We might decide to evaluate three models:

Model: prestige = b_0 + b_1(income)

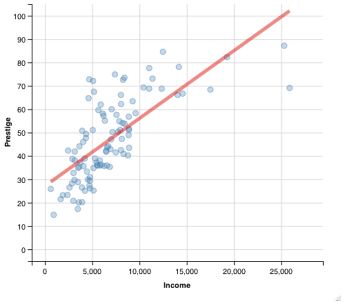
Least-squares line: y = 27.14 + 0.003x

$R^2 = 0.5111$

AIC = 801.88

RMSE = 12.09

F = 104.54 ($p < 0.00001$)



Model: prestige = b_0 + b_1(education)

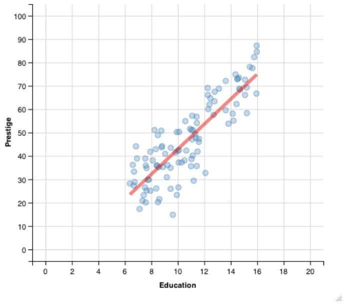
Least-squares line: y = -10.7 + 5.36x

$R^2 = 0.7228$

AIC = 744.01

RMSE = 9.10

F = 260.75 ($p < 0.00001$)



Model: prestige = b_0 + b_1(% women)

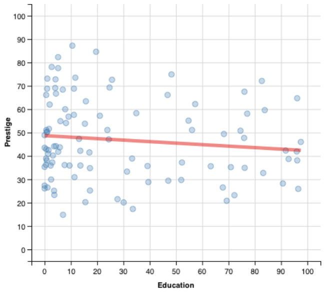
Least-squares line: y = 48.7 – 0.06x

$R^2 = 0.014$

AIC = 873.43

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. If you had to select a single predictor of prestige, which predictor (income, education, %women) would you choose?

4. In the previous lesson, we conduct an F-test to compare a model with no predictors to a model with the income predictor. That F-test indicated the full model with income as a predictor was better than the reduced model.

   Can we extend this process by adding a second predictor? With a single independent variable, we visualize our regression model as a line through a 2-dimensional scatterplot of data. With 2 predictors, we're fitting a 2-dimensional plane to a 3-dimensional scatterplot.

   Write out the models to determine if the combination of income and education provides a better prediction of prestige than a model with no predictors:

   Full model: ______________________________ Reduced model: ______________________________

5. In the previous lesson, we derived simple formulas to calculate the slope and y-intercept of the least-squares line. How do we estimate the coefficients for the best-fitting plane (with 2 predictors) or hyperplane (with 3+ predictors)?

   We can use some simple matrix algebra to find the coefficients that minimize the sum of squared errors. Suppose we have *n* observations in our dataset, with *p* predictors in our full model. Our full model, then, in matrix notation is:

$$Y = Xb + e$$

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

   If linear algebra were a prerequisite for this course, we could show the least-squares solution is: $b = \left(X^T X\right)^{-1} X^T Y$

   Let's have R estimate the parameters for our full and reduced models:

```
reducedmodel <- lm(prestige ~ 1, data = prestige)                # Prestige is a function of a constant
fullmodel <- lm(prestige ~ income + education, data = prestige)  # Prestige = f(income, education)
coef(reducedmodel)                                               # Get coefficients
coef(fullmodel)
```

   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. To compare these models with our omnibus F-test, we need to know R^2 for each model.

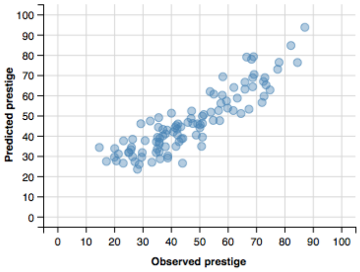
   We know $R^2 = 0$ for the reduced model (since it has no predictors), but how do we calculate R^2 for the full model? What does it mean to have a correlation among more than two variables?

   Suppose we calculate the correlation between two variables: X and Y. We already know we find the least squares regression line that linearly transforms the X values into predicted Y values. Since linear transformations have no impact on correlation coefficients, the correlation between X and Y can be interpreted as the correlation between the observed and predicted Y values.

   With this logic, we can calculate R with multiple predictors – we simply need to calculate the correlation between our observed Y values and the Y values predicted by the predictors.

   The following table displays the predicted prestige scores based on our income and education predictors:

Title	Prestige	Prediction (from full model)
Physicians	87.2	100.4534
Professors	84.6	63.2932
…	…	…
Newsboys	14.8	29.8004



   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: $R^2_{y,x_1,x_2} = 0.798$. Interpret this value.

7. To compare our models, we can use the omnibus F-test (or an ANOVA summary table). Let's do both:

```
summary(fullmodel)     # Summarize model (get omnibus F-statistic and p-value)
anova(fullmodel)       # ANOVA table for full model (with each predictor as a separate source of variation)
anova(reducedmodel, fullmodel)       # ANOVA table comparing full with reduced model
```

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

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}	$R^2 = 0.798$

8. Here's some output from R. Verify our calculation and state any conclusions we can make.

```
> summary(fullmodel)

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept) -6.8477787  3.2189771  -2.127   0.0359 *
 income       0.0013612  0.0002242   6.071 2.36e-08 ***
 education    4.1374444  0.3489120  11.858  < 2e-16 ***

 Residual standard error: 7.81 on 99 degrees of freedom
 Multiple R-squared:  0.798,   Adjusted R-squared:  0.7939
 F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16


> anova(fullmodel)
 Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
 income     1 15279.3 15279.3  250.49 < 2.2e-16 ***
 education  1  8577.3  8577.3  140.62 < 2.2e-16 ***
 Residuals 99  6038.9    61.0


> anova(reducedmodel, fullmodel)
 Model 1: prestige ~ 1
 Model 2: prestige ~ income + education
   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
 1    101 29895.4
 2     99  6038.9  2     23857 195.55 < 2.2e-16 ***

> AIC(fullmodel, reducedmodel)
              df      AIC
 fullmodel     4 713.7251
 reducedmodel  2 872.8732
```

9. Let's add another predictor – %women – to our model. Does the combination of income, education, and % women predict prestige better than a model with no predictors? To do this, we would compare:

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

$$\text{Full: } \hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

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

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

Make sure you can interpret those coefficients and the squared multiple correlation. Interpret the following output:

```
fullmodel <- lm(prestige ~ income + education + percwomn, data=prestige)
summary(fullmodel)
```

```
 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
 income       0.0013136  0.0002778   4.729 7.58e-06 ***
 education    4.1866373  0.3887013  10.771  < 2e-16 ***
 percwomn    -0.0089052  0.0304071  -0.293   0.7702

 Residual standard error: 7.846 on 98 degrees of freedom
 Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
 F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

10. Suppose we add another predictor (**any** predictor) to our model. What would happen to the value of R^2?

Because R^2 will increase monotonically when we add predictors (even if those predictors are virtually unrelated to the dependent variable), we may want to use another statistic to evaluate the fit of our model.

We've already encountered Akaike's AIC: $AIC = -2\ln\left(L_{\text{model}}\right) + 2\left(\#\text{ of coefficients in model} + 1\right)$

Remember, lower values of AIC indicate better fit. That means AIC penalizes models that have more parameters. Here are the values of AIC for the 3 models we've fit thus far:

```
             df      AIC
fullermodel   5 715.6358    (predictors = income, education, %women)
fullmodel     4 713.7251    (predictors = income, education)
reducedmodel  2 872.8732    (predictors = none)
```

Adjusted R-squared is an alternative to AIC that evaluates the fit of a model while penalizing models with more parameters. As we add predictors, adjusted R-squared will 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 our model with 3 predictors, $R^2_{\text{adjusted}} = R^2 - \left(1 - R^2\right)\frac{k}{n-k-1} = 0.7982 - \left(1 - 0.7982\right)\frac{3}{102-3-1} = 0.792$

For all 3 of our models:

```
             adjusted R-squared
fullermodel            0.792     (predictors = income, education, %women)
fullmodel              0.794     (predictors = income, education)
reducedmodel           0.000     (predictors = none)
```

From the AIC or adjusted R-squared values, which model might we want to choose?

What do these values indicate about the %women predictor?

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

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

Explain why we can't simply compare the magnitude of the coefficients to determine which predictor is best?

12. If we want to compare coefficients in our model, we could calculate *standardized beta coefficients*.

One way to do this would be to convert all our predictors to z-scores before estimating the regression model. We could also run the regression with our (untransformed) predictors and then convert the coefficients with:

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

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 one of these coefficients.

```
coef(lm(scale(prestige) ~ scale(income) + scale(education) + scale(percwomn), data=prestige)) # scale = z
```

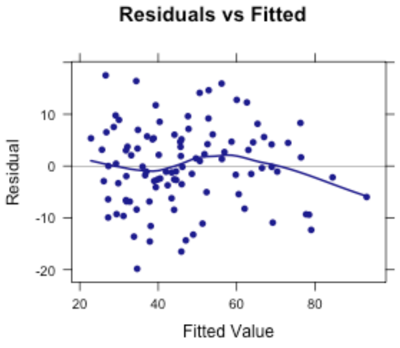
$$\hat{y} = 0.32418\left(z_{\text{income}}\right) + 0.66396\left(z_{\text{education}}\right) - 0.01642\left(z_{\text{\%women}}\right)$$

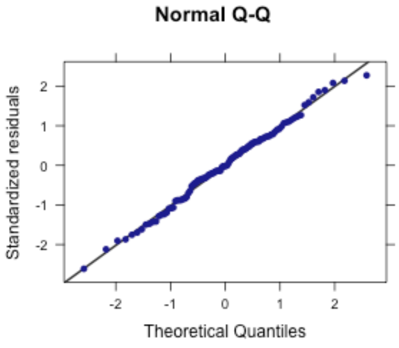
Why is there no intercept?

Explain why we must still be cautious when comparing these beta coefficients.

13. Before we begin the model selection process, let's evaluate the conditions necessary for linear regression. We've already discussed several of these conditions: validity, linearity, independent errors, equal variance of errors, and normality of errors.

Based on the following model plots (from our full model with 3 predictors), evaluate the linearity, equal variances, and normality assumptions.





14. Look at the parameters 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 remained fairly stable when we added a new predictor. That's a good sign we don't have a *multicollinearity* problem.

Multicollinearity is when two or more predictors in our model are highly correlated (meaning that one can be linearly predicted from the others). One effect of multicollinearity is that the coefficients change wildly when we add or subtract predictors.

To detect multicollinearity, we can use the VIF (variance inflation factor): $VIF_k = \dfrac{1}{1-R_k^2}$

where R_k^2 is the R-squared value obtained by regressing predictor k on the remaining predictors.

If you want to learn the details of VIF, I'd suggest: https://onlinecourses.science.psu.edu/stat501/node/347

For now, I'll just note the following rules of thumb:
- If the largest VIF is greater than 10 then there is cause for concern (Bowerman & O'Connell, 1990; Myers, 1990).
- If the average VIF is substantially greater than 1 then the regression may be biased

```
vif(fullmodel)                # VIF for each predictor in the full model
```

```
   income education  percwomn
 2.282038  1.845165  1.526593
```

```
mean(vif(fullmodel))          # mean VIF value
```

```
[1] 1.500598
```

Evaluate the multicollinearity assumption based on these VIF calculations.

15. To test the condition of independence-of-errors, we can use the Durbin-Watson statistic: $D = \dfrac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$

where *e* represents the residual (prediction error) for observation *t*.

The D statistic ranges from 0 to 4, with independent errors yielding a value near 2. Values of D larger or smaller than 2 suggesting errors are not independent. R can calculate the D statistic and estimate its p-value:

```
durbinWatsonTest(fullmodel)          # Durbin-Watson statistic for full model
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.4032531      1.170379       0
 Alternative hypothesis: rho != 0
```

From this, evaluate the assumption of independent errors.

Model Selection

16. So far, we've only compared the **total contribution** of 1, 2, and 3 predictors to reduced models with no predictors. Suppose we're interested in finding a model that adequately predicts prestige using relatively few predictors.

    Recall that we found income was a significant predictor of prestige. The R-squared value of $R^2_{\text{Y},1} = 0.5111$ yielded an omnibus F-statistic of 104.54.

    We then found the combination of income and education – $R^2_{\text{Y},12} = 0.7980$ and F = 195.55 – were better than a reduced model with no predictors.

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

    Write out the full and reduced models we want to compare to answer this question:

    Full model: ________________________________ Reduced model: ________________________________

17. We can use our omnibus F-test or ANOVA summary table to compare these models. 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}	

```
anova(fullmodel)              # ANOVA for full model with 2 predictors
```

```
          Df  Sum Sq Mean Sq F value    Pr(>F)    
income     1 15279.3 15279.3  250.49 < 2.2e-16 ***
education  1  8577.3  8577.3  140.62 < 2.2e-16 ***
Residuals 99  6038.9    61.0                      
```

Calculate the omnibus F-test to verify the value of F = 140.62.

Calculate and interpret: $R^2_{\text{Prestige, education | income}}$ =

18. Let’s continue this **forward-selection** process by adding another predictor to our model. Let’s add %women.

Our 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: $R^2_{Y1} = 0.511$ $R^2_{Y12} = 0.798$ $R^2_{Y123} = 0.7982$
$R^2_{Y2} = 0.723$ $R^2_{Y13} = 0.559$
$R^2_{Y3} = 0.014$ $R^2_{Y23} = 0.752$

Use the omnibus F-test to answer our question:

```
anova(fullmodel)                # ANOVA for full model with 3 predictors
```

```
Analysis of Variance Table
         Df  Sum Sq Mean Sq  F value Pr(>F)
income     1 15279.3 15279.3 248.1727 <2e-16 ***
education  1  8577.3  8577.3 139.3167 <2e-16 ***
percwomn   1     5.3     5.3   0.0858 0.7702
Residuals 98  6033.6    61.6
```

Should we include %women as predictor of prestige? Explain.

19. Suppose we decided to use all 3 predictors. We could then construct confidence and prediction intervals for the predicted prestige of a job with income = 5000, education = 10, and %women = 40:

```
predict(fullmodel, list(income=5000,education=10,percwomn=40), interval=“conf") # Confidence Interval
predict(fullmodel, list(income=5000,education=10,percwomn=40), interval=“pred") # Prediction Interval
```

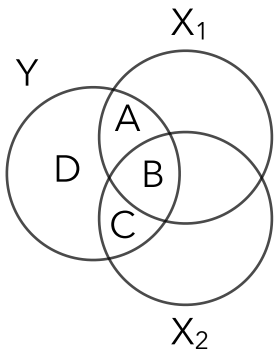
Interpret:

Predicted prestige = 42.74

Confidence interval: (39.78, 45.70)

Prediction interval: (27.27, 58.21)

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

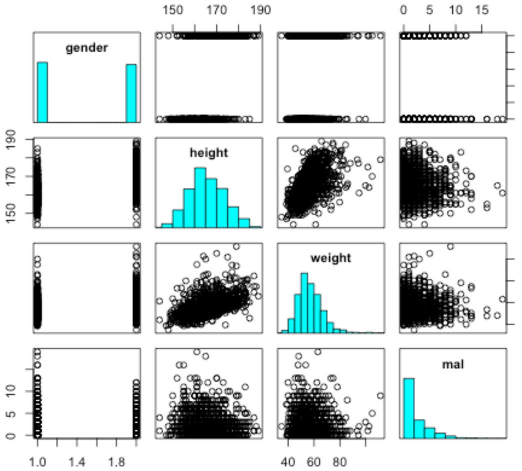


Effect	SSREG	R^2 Values
X₁ and X₂ together	$SS_{X_1 X_2} = A + B + C$	$R^2_{Y12} = \dfrac{A+B+C}{A+B+C+D}$
X₁ alone	$SS_{X_1} = A + B$	$R^2_{Y1} = \dfrac{A+B}{A+B+C+D}$
X₂ alone	$SS_{X_2} = B + C$	$R^2_{Y2} = \dfrac{B+C}{A+B+C+D}$
$X_1 \mid X_2$ = "X₁ unique"	$SS_{X_1 \mid X_2} = (A+B+C) - (B+C) = A$	$R^2_{Y1\mid 2} = \dfrac{A}{A+B+C+D}$
$X_2 \mid X_1$ = "X₁ unique"	$SS_{X_2 \mid X_1} = (A+B+C) - (A+B) = C$	$R^2_{Y2\mid 1} = \dfrac{C}{A+B+C+D}$

21. Let's turn to a simpler dataset to investigate **interaction** within the framework of regression. 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 = male = (1 = male, 0 = female)
x3 = mal = malaise score for each subject at age 22

```
variable     mean      std. dev
weight      57.172    9.656277
height     166.163    8.025138
gender     (50.9% female, 49.1% male)
mal          2.591    2.842851
```



Suppose we're interested in modeling an individual's weight as a function of their height. We could compute:

$$\hat{y} = -46.764 + 0.62551(\text{height}) \qquad R^2 = .2702 \qquad R^2_{adj} = .2695 \qquad s_{y|x} = 8.253 \qquad AIC = 7063$$

We might then decide to see how well the combination of height and gender predict weight by comparing:

$$\text{Full: } \hat{y} = b_0 + b_1(\text{height}) + b_2(\text{female}) \qquad \text{to get} \qquad \text{Full: } \hat{y} = -53.788 + 0.67175(\text{height}) - 1.3439(\text{male})$$

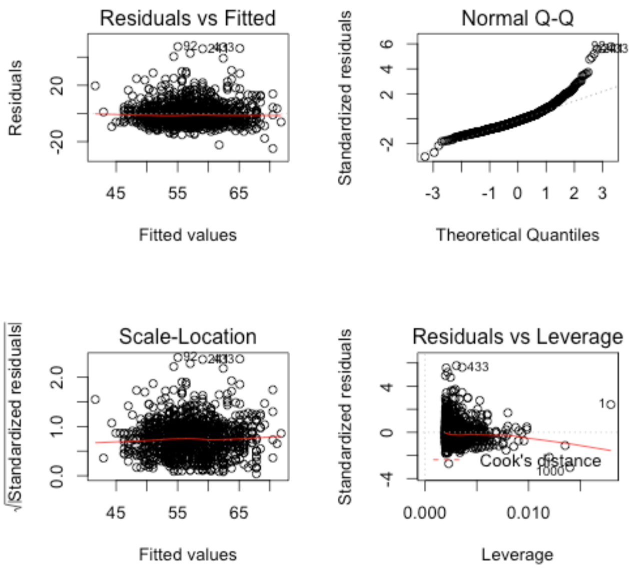
$$\text{Reduced: } \hat{y} = b_0 \qquad\qquad \text{Reduced: } \hat{y} = 57.17209$$

with $R^2 = .2736 \qquad R^2_{adj} = .2721 \qquad s_{y|x} = 8.238 \qquad AIC = 7060$

Interpret that coefficient (–1.3439) for the male predictor variable.

22. Interpret the following output and plots:

```
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq  F value  Pr(>F)
height      1  25173 25172.9 370.9122 < 2e-16
gender      1    314   313.8   4.6231 0.03178
Residuals 997  67664    67.9
```



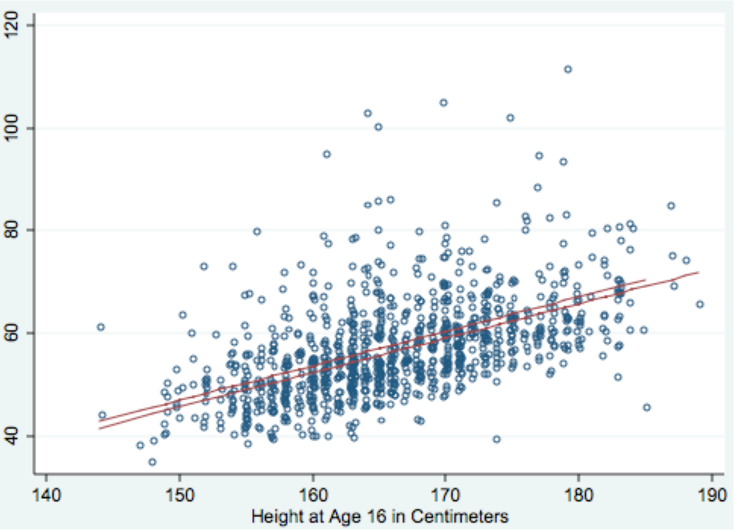
23. From the output displayed above, calculate: $R^2_{\text{weight, height, gender}}$ =

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

When we construct 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.

Let's try both options:

Option (a): Split our data into two sets (one for males and another for females); run separate regression for each.

```
## Split data to create data.frame for males and another for females
males <- htwt %>%
  filter(gender=="male")

females <- htwt %>%
  filter(gender=="female")

## Run regression for each data.frame

male.model <- lm(weight ~ height, data=males)
coef(male.model)

female.model <- lm(weight ~ height, data=females)
coef(female.model)
```

The computer found the following parameter estimates: Males: $\hat{y} = -72.01376 + 0.77066(\text{height})$

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

Explain how the effect of height on weight differs by gender.

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

We could add an interaction term into our model: $\hat{y} = b_0 + b_1(\text{height}) + b_2(\text{female}) + b_{12}(\text{height x female})$

```
interaction.model <- lm(weight ~ height * gender, data=htwt)           # Notice * gives us interaction
int.model <- lm(weight ~ height + gender + height:gender, data=htwt)     # Another way to specify model
```

and estimate these coefficients: $\hat{y} = -33.75 + 0.5479(\text{height}) - 38.2632(\text{male}) + 0.2227(\text{height x male})$

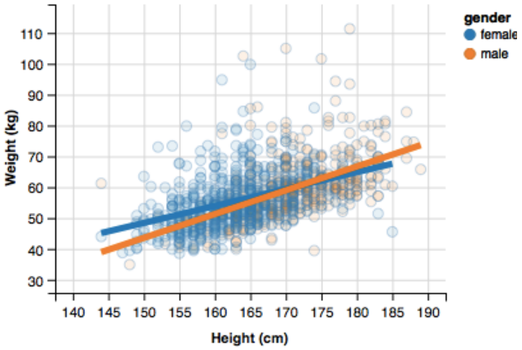
To interpret this interaction term (and its coefficient), we can do some manual arithmetic

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 the coefficients (from the model with the interaction term) are the same as those from the two separate regression analyses.

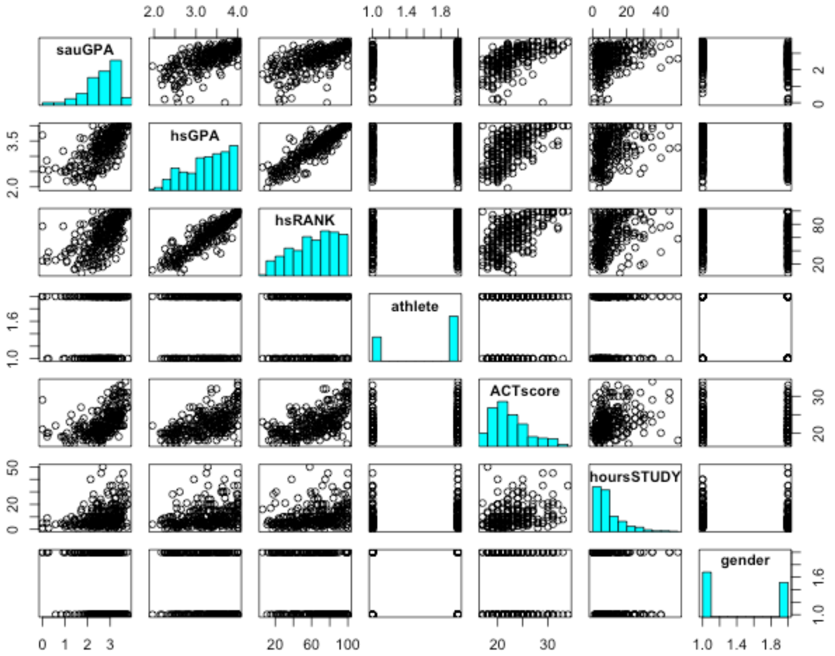
We could test our interaction effect with the omnibus F-test:

```
               Df  Sum Sq Mean Sq  F value    Pr(>F)
height          1   25173 25172.9 373.5646 < 2.2e-16
gender          1     314   313.8   4.6562  0.031181
height:gender   1     548   547.8   8.1297  0.004445
Residuals     996   67116    67.4
```

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?

25. In the previous example, we used a **forward-selection** process to evaluate different prediction models. This time, let's try a **backwards-selection** process. Let's start with a full model containing all our predictors, including the interaction terms. Then, we'll remove predictors from our model to see if the fit significantly worsens.

But first, let's investigate the multicollinearity condition:

```
interact.model <- lm(sauGPA ~ hsGPA*athlete*ACTscore*hoursSTUDY*gender, data=gpa) # All interactions
mean(vif(interact.model))      # VIF to check for multicollinearity
```

```
[1] 46531.9
```

Based on the mean VIF of this model, we have a serious multicollinearity problem. That's to be expected, since we expect these predictors to be correlated. For example, the correlation between HS GPA and HS Rank is r = 0.903. We shouldn't include both those predictors in a model.

26. When I summarize that full model with all the interaction terms, here's the output from R:

```
Coefficients:                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                               -2.858533   8.896733  -0.321    0.748
hsGPA                                                      1.426448   2.616383   0.545    0.586
athletenot athlete                                        -3.421158  10.001927  -0.342    0.733
ACTscore                                                   0.187316   0.413151   0.453    0.651
hoursSTUDY                                                -0.189176   0.984171  -0.192    0.848
gendermale                                                -2.068469  10.249080  -0.202    0.840
hsGPA:athletenot athlete                                   0.782668   2.923433   0.268    0.789
hsGPA:ACTscore                                            -0.043748   0.117887  -0.371    0.711
athletenot athlete:ACTscore                                0.114351   0.467675   0.245    0.807
hsGPA:hoursSTUDY                                           0.059479   0.280278   0.212    0.832
athletenot athlete:hoursSTUDY                              0.322663   1.043032   0.309    0.757
ACTscore:hoursSTUDY                                        0.004147   0.041278   0.100    0.920
hsGPA:gendermale                                           0.371719   3.018346   0.123    0.902
athletenot athlete:gendermale                             12.750745  12.524280   1.018    0.310
ACTscore:gendermale                                        0.057158   0.479190   0.119    0.905
hoursSTUDY:gendermale                                      0.189588   1.092736   0.173    0.862
hsGPA:athletenot athlete:ACTscore                         -0.024624   0.132866  -0.185    0.853
hsGPA:athletenot athlete:hoursSTUDY                       -0.098374   0.295006  -0.333    0.739
hsGPA:ACTscore:hoursSTUDY                                 -0.001345   0.011449  -0.117    0.907
athletenot athlete:ACTscore:hoursSTUDY                    -0.007738   0.044410  -0.174    0.862
hsGPA:athletenot athlete:gendermale                       -3.125614   3.679553  -0.849    0.397
hsGPA:ACTscore:gendermale                                 -0.010055   0.136975  -0.073    0.942
athletenot athlete:ACTscore:gendermale                    -0.633812   0.584537  -1.084    0.279
hsGPA:hoursSTUDY:gendermale                               -0.042746   0.316015  -0.135    0.893
athletenot athlete:hoursSTUDY:gendermale                  -0.507255   1.302175  -0.390    0.697
ACTscore:hoursSTUDY:gendermale                            -0.005725   0.045755  -0.125    0.901
hsGPA:athletenot athlete:ACTscore:hoursSTUDY               0.002488   0.012228   0.203    0.839
hsGPA:athletenot athlete:ACTscore:gendermale               0.155965   0.165924   0.940    0.348
hsGPA:athletenot athlete:hoursSTUDY:gendermale             0.112163   0.372868   0.301    0.764
hsGPA:ACTscore:hoursSTUDY:gendermale                       0.001181   0.012903   0.092    0.927
athletenot athlete:ACTscore:hoursSTUDY:gendermale          0.028069   0.055376   0.507    0.613
hsGPA:athletenot athlete:ACTscore:hoursSTUDY:gendermale -0.006461   0.015409  -0.419    0.675

Residual standard error: 0.5361 on 223 degrees of freedom
Multiple R-squared:   0.55,   Adjusted R-squared:  0.4874
F-statistic: 8.791 on 31 and 223 DF,  p-value: < 2.2e-16
AIC = 437.512
```

Forget about trying to interpret those coefficients. What does that F-statistic (F = 8.791) tell us?

What do the p-values for each parameter estimate tell us?

27. Let's try a second model that has all the predictors with <u>no</u> interaction terms.

```
no.interact <- lm(sauGPA ~ hsGPA + athlete + ACTscore + hoursSTUDY + gender, data=gpa)
```

Based on the following output, should we keep the interaction terms in our model?

```
Full model:          Coefficient   95% Confidence interval
AIC = 402.7862       (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
```

28. We could use our omnibus F-test to compare the model with all predictors (and no interaction terms) to the model with all the predictors and interaction terms.

```
anova(interact.model, no.interact)
```

```
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    223 64.091
2    249 68.583 -26    -4.492 0.6011 0.9384
```

Based on that, what conclusion do you make regarding the interaction terms?

29. If you look at the output pasted in question #27, you'd notice the **athlete** variable doesn't seem to help our prediction. Let's eliminate it and run the omnibus F-test:

```
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
AIC = 400.94

      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
```

Based on that, what conclusion do you make regarding the athlete predictor?

30. We could continue this **backwards-selection process** by eliminating the **hours studying** variable:

```
No hours model:     Coefficient  95% Confidence interval
R² = 0.5072         (intercept)      -1.128, -0.156
adj. R² = 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
AIC = 404.64

      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 *
```

Based on the F-statistic, its p-value, or AIC, what conclusion do you make regarding the hours studying predictor?

31. Adding or removing a single predictor at a time can be tedious. We can automate this process to fit every combination of our predictors using best subsets regression.

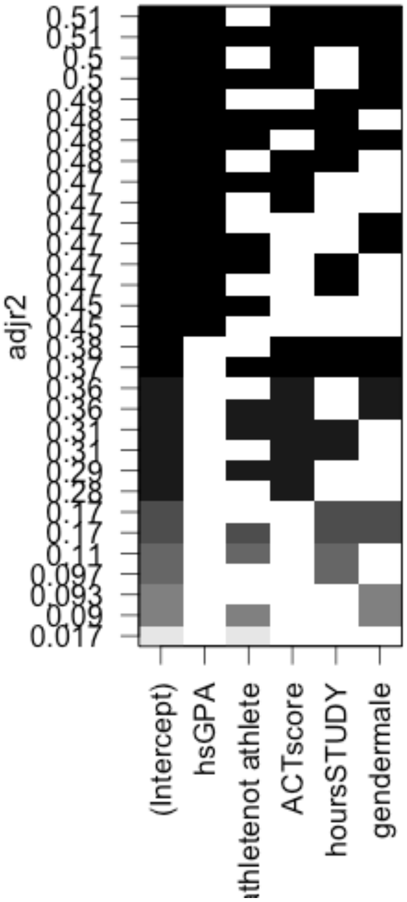
With 5 predictors to choose from, we could fit:
- 1 model with no predictor
- 5 models each having a single predictor
- 10 models each having 2 predictors
- 10 models each having 3 predictors
- 5 models each having 4 predictors
- 1 model with all 5 predictors

That gives us a total of $2^p = 2^5 = 32$ possible regression models to compare. You can see how this method becomes computationally complex when we have a larger number of predictors.

Best subsets regression fits all these models and then compares them using a criterion (such as R-squared or AIC). We'll use the **leaps** package in R to use best subsets regression:

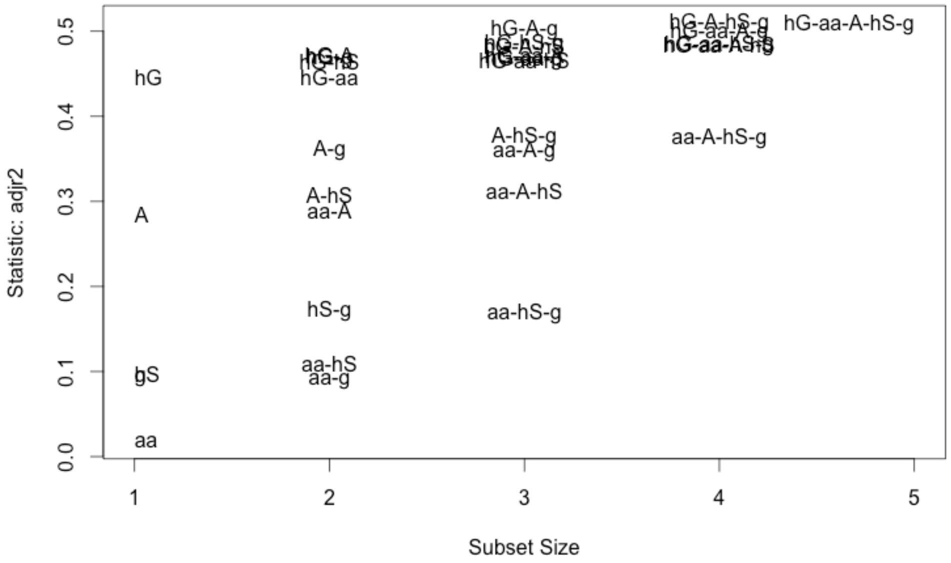
```
library(leaps)         # Load leaps package

# Run best subsets regression and keep the nbest models
leaps<-regsubsets(sauGPA ~ hsGPA + athlete + ACTscore + hoursSTUDY + gender, data=gpa, nbest=10, nvmax=32)

summary(leaps)         # View results
plot(leaps,scale="r2") # Creates the black-box plot displayed below

# Creates the plot at the top of the next page
library(car)
#subsets(leaps, statistic="adjr2")
```

```
Selection Algorithm: exhaustive
          hsGPA athlete ACTscore hoursSTUDY gendermale
1  ( 1 )  "*"   " "     " "      " "        " "       
1  ( 2 )  " "   " "     "*"      " "        " "       
1  ( 3 )  " "   " "     " "      "*"        " "       
1  ( 4 )  " "   " "     " "      " "        "*"       
1  ( 5 )  " "   "*"     " "      " "        " "       
2  ( 1 )  "*"   " "     "*"      " "        " "       
2  ( 2 )  "*"   " "     " "      " "        "*"       
2  ( 3 )  "*"   " "     " "      "*"        " "       
2  ( 4 )  "*"   "*"     " "      " "        " "       
2  ( 5 )  " "   " "     "*"      " "        "*"       
2  ( 6 )  " "   " "     "*"      "*"        " "       
2  ( 7 )  " "   "*"     "*"      " "        " "       
2  ( 8 )  " "   " "     " "      "*"        "*"       
2  ( 9 )  " "   "*"     " "      "*"        " "       
2  ( 10 ) " "   "*"     " "      " "        "*"       
3  ( 1 )  "*"   " "     "*"      " "        "*"       
3  ( 2 )  "*"   " "     " "      "*"        "*"       
3  ( 3 )  "*"   " "     "*"      "*"        " "       
3  ( 4 )  "*"   "*"     "*"      " "        " "       
3  ( 5 )  "*"   "*"     " "      " "        "*"       
3  ( 6 )  "*"   "*"     " "      "*"        " "       
3  ( 7 )  " "   " "     "*"      "*"        "*"       
3  ( 8 )  " "   "*"     "*"      " "        "*"       
3  ( 9 )  " "   "*"     "*"      "*"        " "       
3  ( 10 ) " "   "*"     " "      "*"        "*"       
4  ( 1 )  "*"   " "     "*"      "*"        "*"       
4  ( 2 )  "*"   "*"     "*"      " "        "*"       
4  ( 3 )  "*"   "*"     "*"      "*"        " "       
4  ( 4 )  "*"   "*"     " "      "*"        "*"       
4  ( 5 )  " "   "*"     "*"      "*"        "*"       
5  ( 1 )  "*"   "*"     "*"      "*"        "*"       
```



Notice that I chose adjusted R-squared as my criterion to compare models.

From all of this, which model would you choose?



32. Suppose we ultimately decide to make predictions with the model that includes HSGPA, ACT scores, and gender:

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

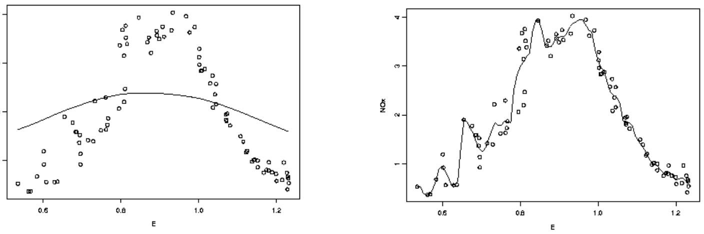
The R-squared and AIC values tell us how well this model fits the data we used to estimate the coefficients, but how accurate would this model be for new data?

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. Using the following visualization as a guide, explain the bias-variance trade-off:

Bias-Variance Tradeoff



High Bias - Low Variance

Low Bias - High Variance

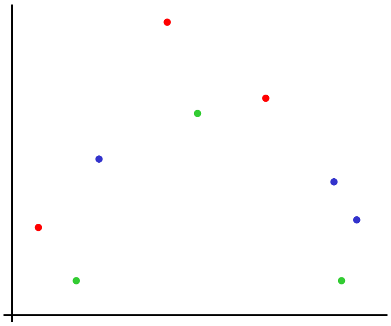
“overfitting” - modeling the random component

34. Because our models will have the tendency to overfit our sample data (and not generalize to other datasets), we might want to split our data before fitting our models.

One way to do this is to split our data into **training** and **testing** subsets. The training dataset would be a random sample of most of the observations in our data. We'd use this train (fit) our models and select the best model. We could then test our model on the testing data. Since the testing data is "new" to our model, it would give us a sense of how well our model generalizes beyond our sample data.

If splitting our data once is a good idea, why don't we split our data multiple times? Rather than taking multiple random samples (which could use the same observation multiple times), we could use k-fold cross validation. To use this method, we would:

- 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



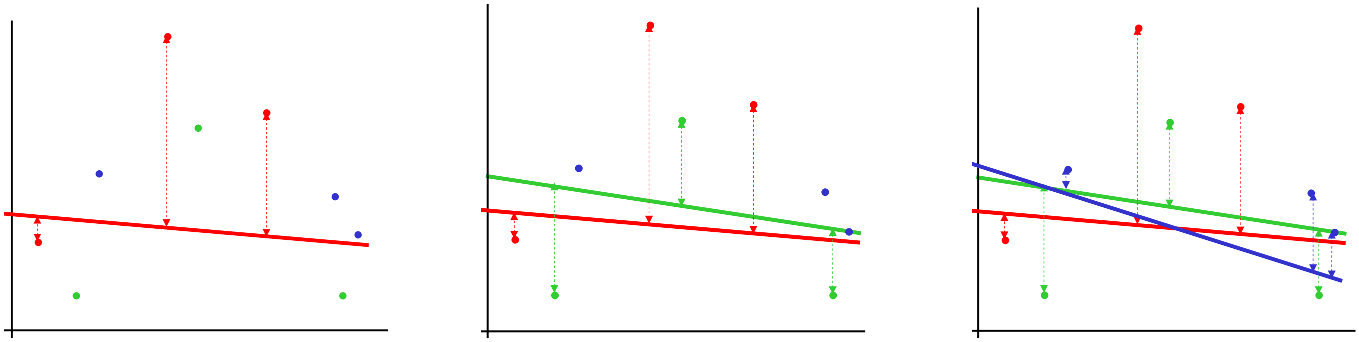
Let's see this process work on the small dataset pictured to the left.
We randomly split the data into 3 pieces (red, blue, and green dots)

Below (left): Model fit to the green and blue dots; error measured with red dots

Below (middle): Model fit to red and blue dots; error measured with green dots

Below (right): Model fit to red and green dots; error measured with 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.



```
library(DAAG)  # Load DAAG library for cross-validation
cross.validated.model <- lm(sauGPA ~ hsGPA + athlete + ACTscore + hoursSTUDY + gender, data=gpa)
CVlm(gpa, cross.validated.model, m=10)
```

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"?

35. With forward-selection, backwards-selection, best-subsets regression, and cross-validation, our model selection is a discrete process: each predictor is either in or out of the model. These discrete processes can have high variance. A different set of data could lead to a completely different model with completely different predictors.

Ridge regression (Tikhonov regularization) is a method that allows each predictor to be partly included in models.

Recall our least squares criterion. We estimate parameters in a regression model to minimize:

$$SS_E = \sum_{i=1}^{N}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2$$

Ridge regression is similar, except coefficients are estimated to minimize:

$$\sum_{i=1}^{N}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} b_j^2 = SS_E + \lambda\sum_{j=1}^{p} b_j^2$$ where λ is a **tuning parameter**.

As you can see, the criterion for ridge regression contains two components. As with ordinary least squares regression, ridge regression seeks coefficients that fit the data well (by minimizing the first component: SSE)

The second term, called a shrinking penalty, is smaller when the coefficient estimates are close to zero, so it has the effect of shrinking the coefficient estimates towards zero.

The tuning parameter λ controls the relative impact of these two components on the coefficient estimates. When λ=0, the penalty term has no effect and ridge regression will yield the least squares estimates. When λ is very large, the coefficient estimates will approach zero. Choosing an appropriate value for λ is very important (and, unfortunately, won't be covered in this course).

While we want to shrink the coefficient estimates, we're typically not interested in shrinking the intercept (which is simply the mean value of our dependent variable when all predictor variables equal zero). For this reason, we typically center our data before performing ridge regression by taking each predictor and subtracting its mean.

Let's fit all our predictors, including both HSGPA and HSrank, on a data.frame that has been centered:

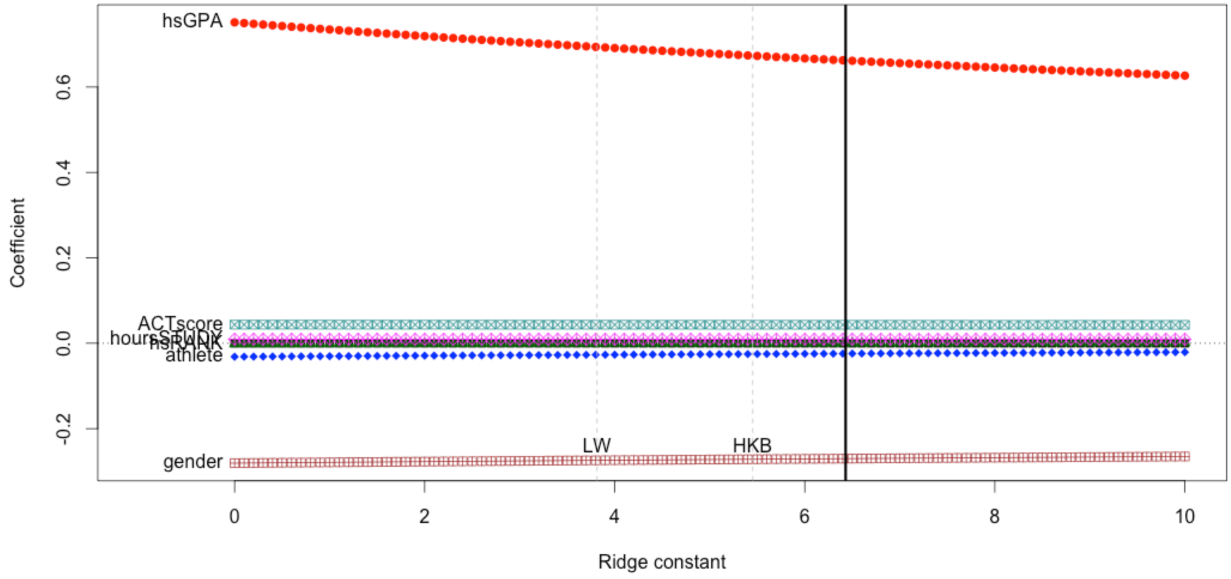
```
# Center our predictors
gpa.centered <- gpa    # Copy our data.frame to gpa.centered
gpa.centered$athlete <- as.numeric(gpa.centered$athlete)   # Convert all to numeric variables
gpa.centered$gender <- as.numeric(gpa.centered$gender)-1   # -1 to make gender a 0/1 variable
gpa.centered <- data.frame(scale(gpa.centered, center = TRUE, scale = FALSE))  # Center but not z-scores

# Find least squares estimates of coefficients
coef(lm(sauGPA ~ hsGPA + hsRANK + athlete + ACTscore + hoursSTUDY + gender, data=gpa.centered))

# Run ridge regression with lambda between 0 and 50
library(MASS)  # Load MASS package
ridge <- lm.ridge(sauGPA ~ hsGPA + hsRANK + athlete + ACTscore + hoursSTUDY + gender, data=gpa.centered,
                  lambda = seq(0, 50, .1))

# Find the “best” value of lambda
select(lm.ridge(sauGPA ~ hsGPA + hsRANK + athlete + ACTscore + hoursSTUDY + gender,
                  data=gpa.centered, lambda = seq(0, 50, .01)))

# Get ridge regression coefficient estimates using lambda = 6.43
lm.ridge(sauGPA ~ hsGPA + hsRANK + athlete + ACTscore + hoursSTUDY + gender, data=gpa.centered,
         lambda = 6.43)

# Plot ridge regression coefficients for various lambda values between 0-10
library(genridge)              # Load genridge package
traceplot(ridge)               # Plot coefficients
abline(v=6.43, lty=1, lw=3)    # Add line at lambda = 6.43
```

The output is displayed on the next page.

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
hsGPA	+0.7510	+0.6622
hsRANK	-0.0016	+0.0002
not athlete	-0.0316	-0.0243
ACTscore	+0.0436	+0.0432
hoursSTUDY	+0.0093	+0.0093
male	-0.2810	-0.2702

The plot shows the shrinkage of the coefficients as we increased lambda. Note that we wouldn't want to use 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.



36. Unlike the discrete processes (e.g., forward-selection, best subsets) ridge regression does not allow us to remove any predictors from a model. The penalty will shrink all coefficient estimates towards zero, but it won't set them equal to zero (which would mean we could remove them from the model).

The **lasso (least absolute shrinkage and selection operator)** not only shrinks estimates towards zero; it actually forces some coefficient estimates to be zero when λ is large.

The lasso coefficients minimize the following criterion:

$$\sum_{i=1}^{N}\left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|b_j\right| = SS_E + \lambda\sum_{j=1}^{p}\left|b_j\right|$$

Notice this is extremely similar to the ridge regression criterion, except penalty term uses absolute values of coefficients (rather than squared values).

Other investigations of our GPA dataset

37. 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?

38. 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?

39. 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?

40. 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?