Activity #16: Multiple Linear Regression

So far, we have used linear regression to make predictions from a single independent variable. This was displayed as a straight line through a two-dimenstional scatterplot of observations.

We can expand this concept of least-squares regression to situations in which we want to make predictions from multiple dependent variables. Instead of a line shooting through a 2-dimensional plot, we will have a plane shooting through a 3-dimensional plot. We will begin by looking once again at the occupational prestige data.

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

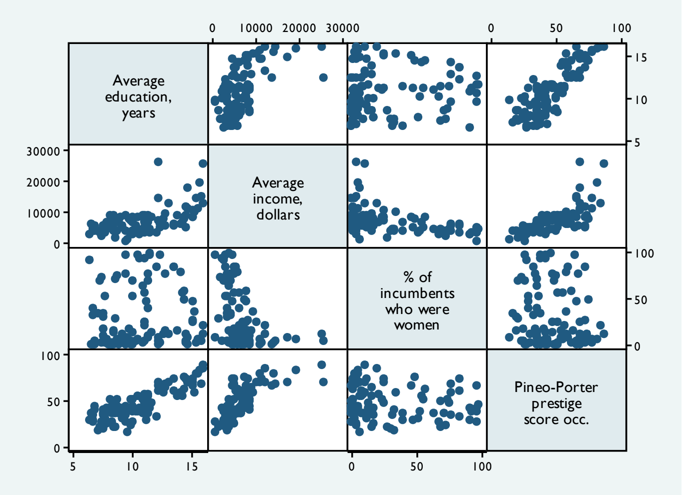
Download the data at: http://web.me.com/bradthiessen/data/prestige.sav or http://web.me.com/bradthiessen/data/prestige.dta

Variables:

Title:	Name of occupation
Education:	Average years of education for occupational incumbents (in 1971)
Income:	Average income, in dollars, of incumbents (in 1971)
%women:	Percentage of incumbents who are women (in 1971)
Type:	Type of occupation (blue collar, white collar, professional/managerial/technical)
Prestige:	*Pineo-Porter Prestige* score (from a survey conducted in the mid-1960s)

#	Title	Education (X_2)	Income (X_1)	%women (X_3)	Type (X_4)	Prestige (Y)
1	Physicians	15.96	25308	10.56	Professional	87.2
2	University Professors	15.97	12480	19.59	Professional	84.6
…	…	…	…	…	…	…
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**

Scatterplots



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
```

$$R^2_{\text{prestige, income}} = 0.511$$

1) Recall that we regressed prestige on income and concluded that income is a significant predictor of prestige. The Stata output is shown below. Make sure you understand all the numbers in the output!

```
      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  1,   100) =  104.54
       Model |  15279.2563     1  15279.2563           Prob > F      =  0.0000
    Residual |  14616.1698   100  146.161698           R-squared     =  0.5111
-------------+------------------------------
       Total |  29895.4261   101  295.994318           Root MSE      =   12.09

------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0028968   .0002833    10.22   0.000     .0023347    .0034589
       _cons |   27.14118   2.267704    11.97   0.000     22.64212    31.64024
------------------------------------------------------------------------------
```

2) Let's try to improve our prediction by adding another independent (predictor) variable. Let's see how well the combination of income and education predict prestige. To do this, let's write out our full and reduced models.

Reduced Model: $\hat{Y} = \hat{\beta}_0$ or Prestige is predicted by its mean

Full Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ or Prestige is predicted by income and education

3) I had Stata compute the regression lines for the full and reduced models. To do this, we would need to use some matrix algebra. Since linear algebra is not a prerequisite for this course (and since I believe it is a waste of time to do this by hand), we'll rely on technology to compute our regression coefficients.

Reduced Model: $\hat{Y} = 46.8333$ (Without any predictor, our best guess is to use the mean prestige score)

Full Model:

$$\hat{Y} = -6.8478 + 0.0014X_1 + 4.1374X_2$$

$$or$$

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

Interpret these coefficients.

4) When we conducted a simple linear regression analysis, we calculated R, the correlation between Y and X values. I didn't mention it at the time, but the value of R can also be interpreted as the correlation between observed and predicted Y values. With this definition, we can calculate R with multiple predictors – we just need to calculate the correlation between the predicted and observed Y values. The following table lists the observed and predict Y values (based on our regression equation).

#	**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 calculates the correlation between the observed and predicted prestige scores to be: $R_{Y,X_1X_2} = 0.8933$.

5) If we square this correlation, we get R-squared = 0.798. Interpret this value.

6) We still don't know if our full model is significantly better than our reduced model. To determine this, we can once again create a summary table or use our omnibus F test. Let's start with the omnibus F test. Calculate it and write your conclusion.

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

7) We already know our conclusion, but let's create the summary table. The following table displays the formulas you will need.

ANOVA (Summary of Calculations)					
Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ or $R^2(SSY)$	k	$\frac{SS_{reg}}{df_{reg}}$	$\frac{SS_{reg}}{SS_E}$	${}_{\alpha}F^{k}_{n-k-1}$
Error or Residual	$\sum_{i=1}^{n} (Y - \hat{Y}_i)^2$ or $\left(1-R^2\right)\left(SSY\right)$	$n-k-1$	$\frac{SS_E}{df_E}$		
Total	$\sum_{i=1}^{n} (Y - \bar{Y})^2$ or $(n-1)S_Y^2$	$n-1$	$\frac{SS_{TOT}}{df_{TOT}}$	$\eta^2 = \frac{SS_{reg}}{SS_{TOT}}$	

8) Complete the summary table. Check to ensure your F statistic is the same as what you got from the omnibus F test. Also, check your SSE against the table at the top of the previous page. Finally, check your values against the Stata output pasted below.

ANOVA (Calculated from our example data)					
Source	Sum of Squares	df	Mean Square	F	Sig.
Regression					
Error					
Total					

```
      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
------------------------------------------------------------------------------
```

9) Let's go one step further. Let's see if the combination of income, education, and %women significantly predict prestige better than a model with no predictors. To do this, we would compare the following models:

Reduced Model: $\hat{Y} = \beta_0$ vs. **Full Model:** $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Stata computes the following coefficients: $\hat{Y} = -6.794 + 0.001X_1 + 4.187X_2 - 0.009X_3$.

Interpret the coefficients.

10) Stata also computes the correlation between the observed prestige scores and the prestige predicted by this equation. This correlation, when squared, was found to be: $R^2_{Y123} = 0.7982$. Interpret this value and then conduct a test to see if the full model is significantly better than the reduced model.

11) Thus far, we've only analyzed the ***total contribution*** of several independent variables. Suppose we are interested in something else. Suppose we want to predict prestige using the most parsimonious model possible. In other words, we want to maximize the prediction accuracy using as few predictors as possible.

In this example, we know income was a significant predictor of prestige. In fact, we found $R_{Y1}^2 = 0.5111$ led to a significant omnibus F test.

We also found the combination of income and education was a significant predictor of prestige. For this model, we found $R_{Y12}^2 = 0.7980$.

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

To answer this, we need to write out the full and reduced models we would like to compare.

Reduced Model:

Full Model:

12) To answer our question, we can once again use our summary table or omnibus F test. We can calculate SS values through the usual formulas and/or use the same omnibus F test. Fill-in the missing row of values. How did you calculate these values? Should we add SAT scores to our prediction model?

Source	SS	df	Mean Square	F	Sig.
X_1 and X_2					
X_1: HS GPA					
$X_2 \mid X_1$ "Education given Income" or "What education adds to our prediction"					
Error					
Total	29895.426	101			

13) Verify the value of our F statistic using the omnibus F-test:

14) Calculate and interpret $R_{Y2\mid1}^2$:

15) Let’s finish this prestige example by attempting to answer one final question. We have shown:

a) Income is a significant predictor of prestige
b) The combination of income and education are significant predictors of prestige
c) The combination of income, education, and %women are significant predictors of prestige.
d) Education significantly improves the prediction of prestige over just income alone

The final question we will attempt to answer is:

e) Does %women significantly improve our prediction over income and education? Another way of asking this is: *Should we add %women to predict prestige if we’re already using income and education?*

Write out the full and reduced models we would like to compare.

Reduced Model:

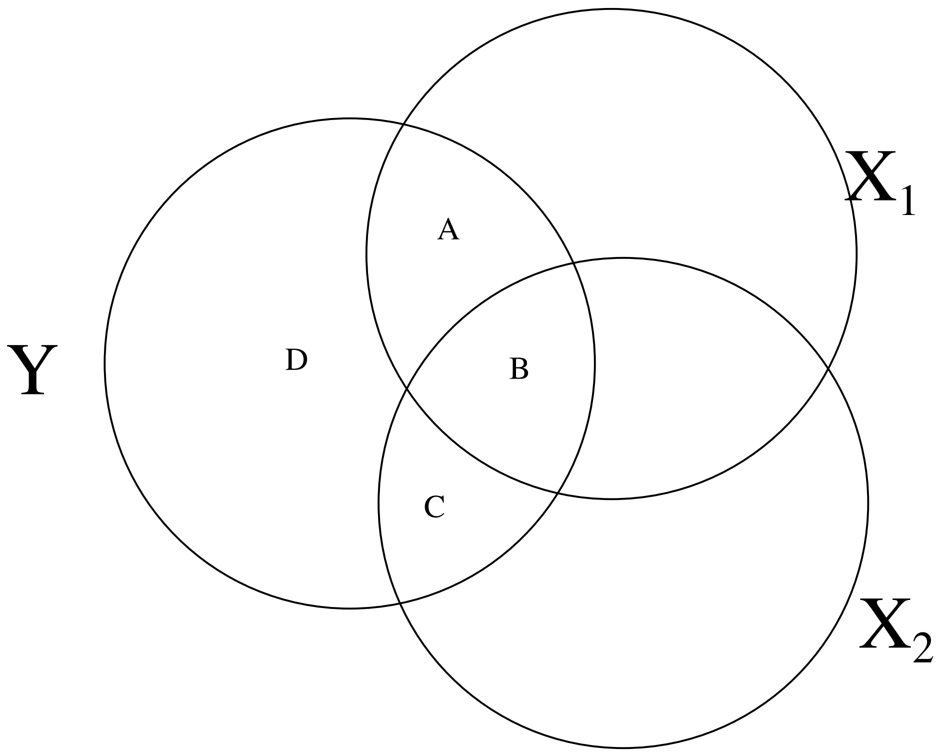
Full Model:

16) Using a computer, I calculated the following multiple correlations:

$$\begin{array}{lll} R_{Y1}^2 = 0.511 & R_{Y12}^2 = 0.798 & \\ R_{Y2}^2 = 0.723 & R_{Y13}^2 = 0.559 & \\ R_{Y2}^2 = 0.014 & R_{Y23}^2 = 0.752 & R_{Y123}^2 = 0.7982 \end{array}$$

Use the omnibus F test to answer question e).

The following display attempts to visualize the contribution of two independent variables to the prediction of one dependent variable.



Effect	SS_{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"	$SS_{X_2 \mid X_1} = (A+B+C) - (A+B) = C$	$R^2_{Y2\mid 1} = \frac{C}{A+B+C+D}$

Situation: Suppose you work in the admissions office at SAU. Your goal is to predict which students will be successful at SAU.

Download the data at: http://web.me.com/bradthiessen/data/gpa.sav or http://web.me.com/bradthiessen/data/gpa.dta

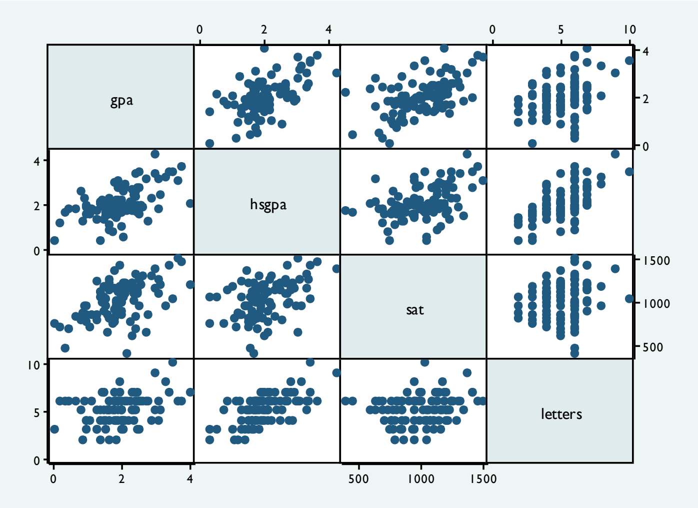
17) What would you use to determine if students are successful at SAU? What will you use to predict student success?

Dependent variable – measure of success:

Independent variables – predictors of success:

18) Suppose you have access to the following data for recent SAU graduates.

Subject	Y College GPA	X_1 High School GPA	X_2 SAT Total Score	X_3 Letters of Recommendation (Rating on 1-10 scale)
1	2.04	2.01	1070.00	5
2	2.56	3.40	1254.00	6
3	3.75	3.68	1466.00	6
…	…	…	…	…
98	2.08	2.53	1212.00	4
99	.70	1.78	818.00	6
100	.89	1.20	864.00	2
Mean	**1.9805**	**2.0486**	**1014.76**	**5.19**
Std. Dev.	**0.74923**	**0.72234**	**217.34705**	**1.495**



From this data, a computer calculated the following regression coefficients:

$$\hat{Y} = -0.153 + 0.376X_1 + 0.001X_2 + 0.023X_3$$

Interpret the coefficients.

19) A computer also calculated the following correlations:

$$R^2_{Y1} = 0.2972 \qquad R^2_{Y12} = 0.3985$$
$$R^2_{Y2} = 0.2733 \qquad R^2_{Y13} = 0.2974$$
$$R^2_{Y3} = 0.1226 \qquad R^2_{Y23} = 0.3319 \qquad R^2_{Y123} = 0.3997$$

Interpret 0.2972, 0.3985, and 0.3997. Why does high school GPA only account for 29.7% of the variance in SAU GPA?

20) Write out full and reduced models and run an omnibus F test to answer the following questions:

1) Is high school GPA a good predictor or college GPA? Write out the full and reduced models.

2) Is the combination of high school GPA and SAT scores a good predictor of college GPA?

3) Does adding X_2 to our model improve our prediction accuracy? Complete the summary table.

Source	SS	df	Mean Square	F	Sig.
X_1 and X_2 and X_3	22.214	3	7.4048	21.31	p<.0001
X_1 and X_2	22.146	2	11.073	32.13	p<.0001
$X_3 \mid X_1, X_2$ “Letters given HS GPA and SAT”					
Error	33.358	96	0.3475		
Total	55.573	99	0.5613		