Activity #15: More Simple Linear Regression

Resources: string.sav, professor.sav,

1) Studies over the past two decades have shown that certain activities can effect the reorganization of the human central nervous system. For example, it is known that the part of the brain associated with activity of a limb is taken over for other purposes in individuals who have lost a limb. In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of 9 violin players and 6 controls (those who have never played a stringed musical instrument) when the fingers on their left hands were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers on their left hand extensively, might show an increased amount of neuron activity. Shown below is a neuron activity index from the MSI along with the number of years each individual had been playing a stringed instrument.

Subject	Years Played	Neuron Activity
1	0	5.0
2	0	6.0
3	0	7.5
4	0	9.0
5	0	9.5
6	0	11.0
7	5	16.0
8	6	16.5
9	8	11.5
10	10	16.0
11	12	25.0
12	13	25.5
13	17	25.5
14	18	23.0
15	19	26.5
Mean	**7.2**	**15.567**
Std. Dev.	**7.243**	**7.7825**
Source: Elbert, T., "Increased cortical representation of the fingers of the left hand in string players," Science, 270, 13 October, 305-307		

a) Enter this data into SPSS. Make sure you define your variables.

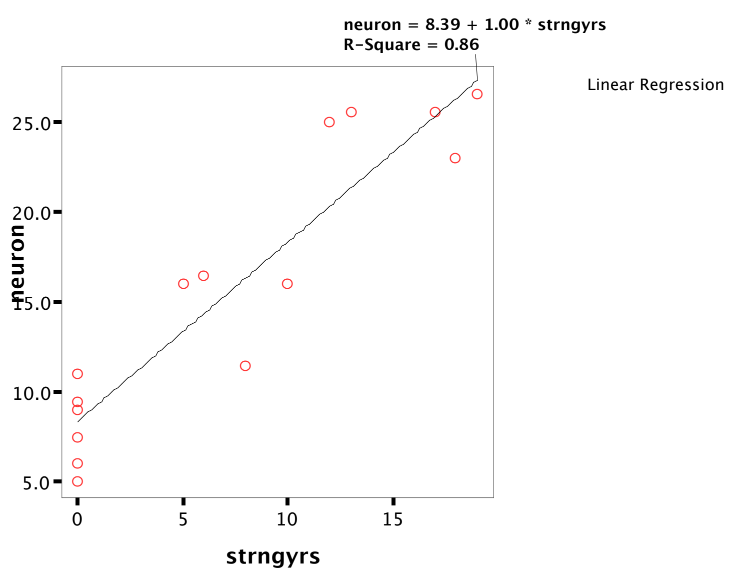
b) Could we run an ANOVA on this data? What type of analysis is most appropriate?

c) State the null and alternative hypotheses.

d) Create a scatterplot of the data. Which variable is the dependent variable?

e) Calculate and interpret the correlation between the variables.

f) Does it appear as though the variables have a linear relationship?



2) In the last activity, we learned how to compare a "full" regression model with a "reduced" regression model. Remember, we will always assume that the constant (y-intercept) in significant. Formally state the full and reduced models in this situation.

Full Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ (each observation is due to a constant and a treatment effect)

Reduced Model: $\hat{Y} = \hat{\beta}_0$ (the treatment effect has no significant impact on the outcome)

3) We also learned how to summarize our calculations into an ANOVA summary table. Fill in the following summary table. Interpret each cell in the table (graphically, if it helps). Then, compute all the required information by hand.

ANOVA (Summary of Calculations)					
Source	Sum of Squares	df	Mean Square	F	Sig.
Regression or $\beta_{1\|0}$	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ 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 $(1-R^2)(SSY)$	$n-k-1$	$\frac{SS_E}{df_E}$		
Total	$\sum_{i=1}^{n}(Y - \bar{Y})^2$ or $(n-1)S_Y^2$	$n-1$	$\frac{SS_{TOT}}{df_{TOT}}$	$\eta^2 = \frac{SS_{reg}}{SS_{TOT}}$	

ANOVA (Calculated from our example data)					
Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	730.2266	1	730.2266	80.6491	<.0001
Error	117.7067	13	9.0543		
Total	847.9333	14			

4) What conclusions can you draw from this analysis?

5) We also learned that we can skip the ANOVA summary table and go straight to the hypothesis test using the following test statistic:

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

Use the above formula to calculate the value of the test statistic. Find the critical F-value for a significance test at alpha = 0.01. Does our calculated test statistic fall in this critical region?

6) Does playing a stringed musical instrument increase neural activity? What proportion of variance in the dependent variable is due to the independent variable?

7) Have SPSS run a linear regression analysis on the data. Does the output match your calculations? Does it match the output from Stata pasted below?

Source	SS	df	MS		
Model	730.206005	1	730.206005	Number of obs =	15
Residual	117.727328	13	9.05594834	F(1, 13) =	80.63
				Prob > F =	0.0000
				R-squared =	0.8612
				Adj R-squared =	0.8505
Total	847.933333	14	60.5666667	Root MSE =	3.0093

neuron	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
strngyrs	.9971405	.1110454	8.98	0.000	.7572415	1.23704
_cons	8.387255	1.114887	7.52	0.000	5.978688	10.79582

Note: Each year, I give myself 20 minutes to search for data that might be interesting. This is the best I could do. Sorry.

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

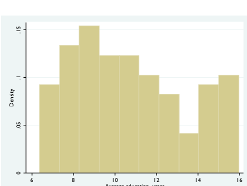
Source: Canada (1971). Census of Canada. Vol. 3, Part 6. Statistics Canada, 19-21.

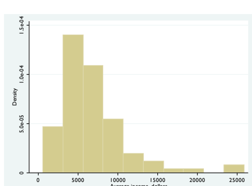
Download the data at: http://web.me.com/bradthiessen/data/prestige.sav or http://web.me.com/bradthiessen/data/prestige.dta

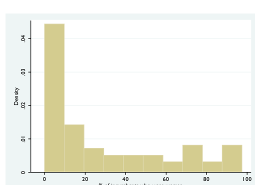
Variables:

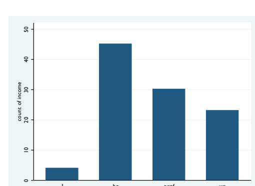
Title: Name of occupation
Education: Average years of education for occupational incumbents (in 1971)
Income: Average income, in dollars, of incumbents (in 1971)
%women: Percentage of incumbents who are women (in 1971)
Type: Type of occupation (blue collar, white collar, professional/managerial/technical)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

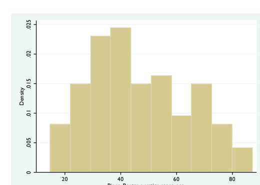
#	Title	Education	Income	%women	Type	Prestige
1	Physicians	15.96	25308	10.56	Professional	87.2
2	University Professors	15.97	12480	19.59	Professional	84.6
3	Lawyers	15.77	19263	5.13	Professional	82.3
4	Architects	15.44	14163	2.69	Professional	78.1
5	Physicists	15.64	11030	5.13	Professional	77.6
6	Psychologists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


Education


Income

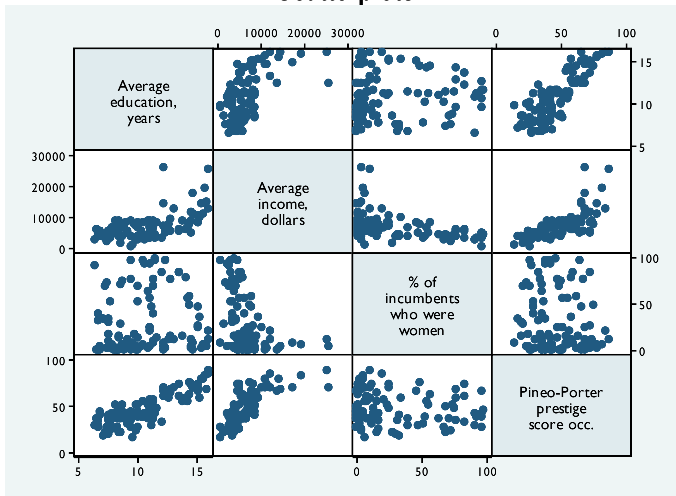

% women


Type


Prestige

Download this data into SPSS.
Which variables will have the strongest relationship with prestige?

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

8) Look at the scatterplots and correlations. What conclusions can you make? Do you think we have linear relationships?

9) Before we begin our regression analysis, run an analysis to determine if the three occupation types differ in prestige. What type of analysis would you need to conduct? Do our data meet the assumptions necessary to conduct this analysis? Interpret the results.

type	Summary of Pineo-Porter prestige Mean	Std. Dev.	Freq.
0	36.085714	11.34732	49
1	42.243478	9.5158157	23
2	67.906666	8.8192554	30
Total	46.833333	17.204485	102

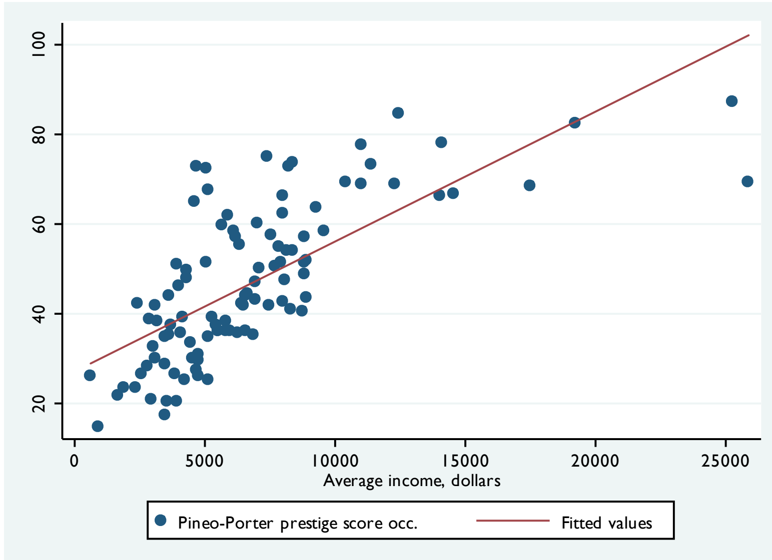
Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	19467.1511	2	9733.57554	92.40	0.0000
Within groups	10428.275	99	105.336111		
Total	29895.4261	101	295.994318		

Bartlett's test for equal variances: chi2(2) = 2.4469 Prob>chi2 = 0.294

Row Mean-Col Mean	(Bonferroni Tests) 0	1
1	6.15776 0.059	
2	31.821 0.000	25.6632 0.000

10) Let's find the OLS regression line to interpret the relationship between occupational prestige & income. Using the correlation coefficient, sample means, and standard deviations, you can verify the results obtained from Stata:




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

From this output, write out the estimated regression line and interpret the coefficients.

11) The correlation between income and prestige was found to be 0.7149. How much of the variance in occupational prestige is accounted for by average occupational income?

12) Given the full and reduced models listed below, complete the summary table.

Full Model: $Y_i = \beta_0 + \beta_1 (X_1) + \varepsilon_i$ or $\text{prestige}_i = \beta_0 + \beta_1 (\text{income}_i) + \varepsilon_i$

Reduced Model: $Y_i = \beta_0 + \varepsilon_i$ or $\text{prestige} = \beta_0 + \varepsilon_i$

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression					
Error					
Total					

13) You should have gotten the following results. What conclusions can we make?

```
      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
-------------+------------------------------           Adj R-squared =  0.5062
       Total |  29895.4261   101  295.994318           Root MSE      =   12.09 = Sy|x

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

14) Use the omnibus F-test to verify the F statistics comparing our full and reduced models.

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

15) Using the Stata output at the top of this page, we could predict the prestige of an occupation with an average income of $7,000:

$$\hat{Y} = 27.14118 + 0.0028968(7000) = 47.42$$

How confident are we that a $20,000 job will have a prestige score of exactly 47.42?

16) The Stata output also shows confidence intervals for the regression coefficients. For example, a 95% confidence interval for the income coefficient was found to be (0.0023, 0.0035). Interpret this interval. Does this mean we are 95% confident that increasing an occupation's income by $1000 will be associated with a 2.3347 – 3.4589 increase in prestige?

17) Let's go back to predicting the prestige of a $7,000 per year occupation. We don't really believe the prestige will be exactly 47.42. Thankfully, we can use the following formulas to calculate a confidence interval:

$$\hat{Y} \pm t_{\frac{\alpha}{2},n-2} S_{Y|X}\sqrt{\frac{1}{N} + \frac{\left(X_0 - \bar{X}\right)^2}{(N-1)S_x^2}}$$

where

$$S_{Y|X} = \sqrt{\frac{\left(Y-\hat{Y}\right)^2}{n-2}} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{(1-R^2)SS_T}{n-2}} = \sqrt{\frac{(1-R^2)s_Y^2(n-1)}{n-2}} = s_y\sqrt{(1-R^2)}\sqrt{\frac{(n-1)}{(n-2)}}$$

A 95% confidence interval for the *average prestige of all $7,000 occupations* is then calculated to be:

$$S_{Y|X} = \sqrt{146.16} = \sqrt{\frac{14616.1698}{102-2}} = 17.204\sqrt{(1-.7149^2)}\sqrt{\frac{(102-1)}{(102-2)}} = RootMSE = 12.089$$

$$\hat{Y} \pm t_{\frac{\alpha}{2},n-2} S_{Y|X}\sqrt{\frac{1}{N}+\frac{\left(X_0-\bar{X}\right)^2}{(N-1)S_x^2}} = \left(47.42\right) \pm (1.984)(12.09)\sqrt{\frac{1}{102}+\frac{\left(7000-6797.90\right)^2}{(102-1)(4245.92^2)}} = 47.42 \pm 2.38$$

We are 95% confident the average prestige for all occupations with $7,000 per year incomes is between 45.04 and 49.80.

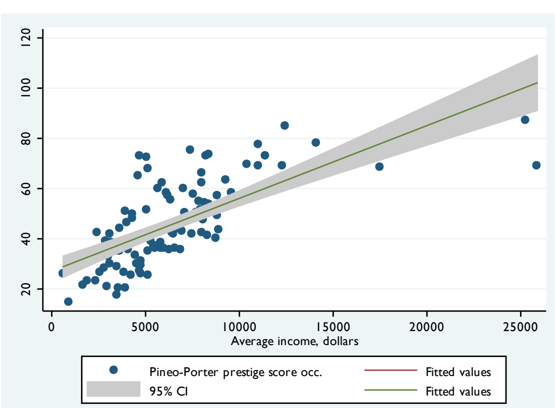
18) Will this confidence interval have the same width for all values of income?

19) If you look closely at our interpretation, you'll see that a confidence didn't give us exactly what we wanted. We wanted an interval to predict the prestige of a single $7,000 per year occupation. To do this, we would need to calculate a prediction interval. If we want an interval about one future observation, will the interval be wider or more narrow than our confidence interval?

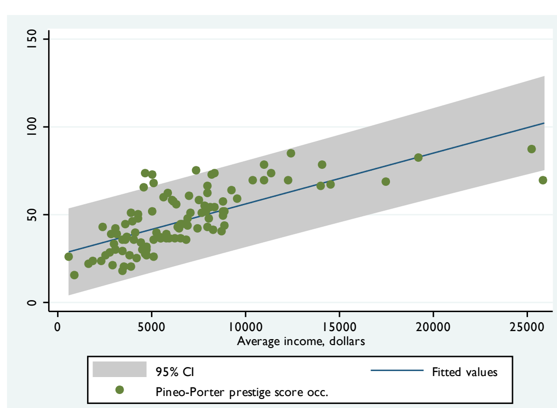
20) The formula for a prediction interval is:

$$\hat{Y} \pm t_{\frac{\alpha}{2},n-2} S_{Y|X}\sqrt{1+\frac{1}{N}+\frac{\left(X_0-\bar{X}\right)^2}{(N-1)S_x^2}} = \left(47.42\right) \pm (1.984)(12.09)\sqrt{1+\frac{1}{102}+\frac{\left(7000-6797.90\right)^2}{(102-1)(4245.92^2)}} = 47.42 \pm 24.10$$

We predict with 95% confidence the prestige of an occupation with $7,000 income will be between 23.31 and 71.52.



Confidence Interval



Prediction Interval

21) Recall the main assumptions necessary to conduct a simple linear regression are linearity, independence, normality, and homoscedasticity. Up until this point, we have stated these assumptions with respect to our observed data. In other words, we checked to see if our data had an approximately normal distribution and equal variances. We're going to restate these assumptions now:

Normality assumption: The dependent variable follows a normal distribution across values of the independent variable
or
The residuals follow a normal distribution across values of the independent variable
or

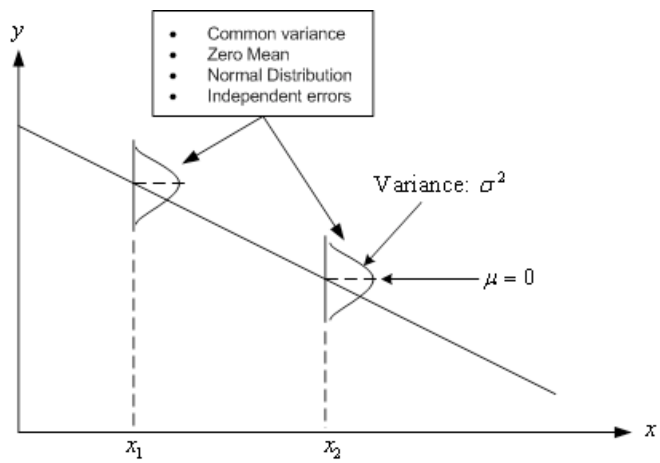
$$\varepsilon_i \mid x \sim N\left(0,\sigma^2\right)$$

Homoscedasticity assumption: The variance of the dependent variable is constant across values of the independent variable.
or
The variance of the residuals are constant across values of the independent variable
or

$$\text{var}\left(y_i \mid x_i\right) = \text{var}\left(\varepsilon_i \mid x_i\right) = \sigma^2$$

We can display these assumptions graphically:

.



We can check these assumptions **after** we conduct a regression analysis by performing residual diagostics.

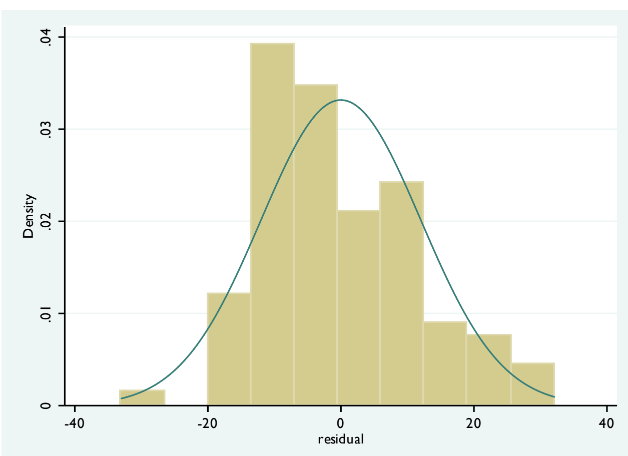
Recall that a residual is our prediction error. It's how far our predictions are from what we actually observe: $\text{Residual} = \text{Y} - \hat{\text{Y}}$

Once we find our least squares regression line, we can graph the residuals to see if the assumptions seem reasonable.

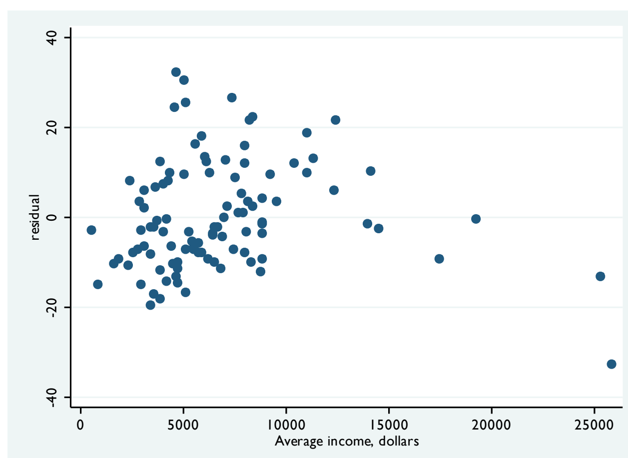
The residuals for some data in this example are displayed on the next page. Make sure you understand how these residuals were calculated.

The next page also shows two graphs. See what you can conclude from these graphs.

#	Title	Income	Prestige	Predicted (From regression line)	Residual	Squared residuals
1	Physicians	25308	87.2	100.4534	-13.253	175.653
2	University Professors	12480	84.6	63.29323	21.307	453.978
3	Lawyers	19263	82.3	82.94222	-0.642	0.412
4	Architects	14163	78.1	68.16854	9.931	98.634
5	Physicists	11030	77.6	59.09287	18.507	342.514
6	Psychologists	7405	74.9	48.59198	26.308	692.112
7	Chemists	8403	73.5	51.48298	22.017	484.749
8	Civil Engineer	11377	73.1	60.09806	13.002	169.050
…	…	…	…	…	…	…
18	Medical Technicians	5180	67.5	42.1466	25.353	642.795
19	Secondary Teachers	8034	66.1	50.41406	15.686	246.049
…	…	…	…	…	…	…
26	Elementary Teachers	5648	59.6	43.5023	16.098	259.136
…	…	…	…	…	…	…
98	Launderers	3000	20.8	35.83157	-15.032	225.948
99	Bartenders	3930	20.2	38.5256	-18.326	335.828
100	Elevator Operators	3582	20.1	37.51751	-17.418	303.370
101	Janitors	3472	17.3	37.19886	-19.899	395.965
102	Newsboys	918	14.8	29.80044	-15.000	225.013
	Means	**6797.90**	**46.8333**	**46.8333**	**0.00**	**SUM = 14616.17 (SSE)**
	Std. Deviations	**4245.92**	**17.2045**	**12.2996**	**12.03**	



Histogram of residuals to check for normality



Scatterplot of residuals by income to check homoscedasticity

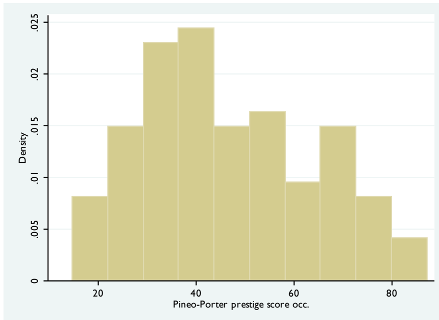
I also had Stata run a test for heteroskedasticity. What can we conclude from this test?

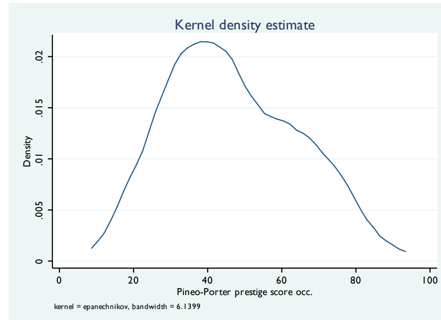
```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of prestige

         chi2(1)      =     3.09
         Prob > chi2  =   0.0788
```

If we are worried about normality and homoskedasticity, we have several options. Two of these options are to:

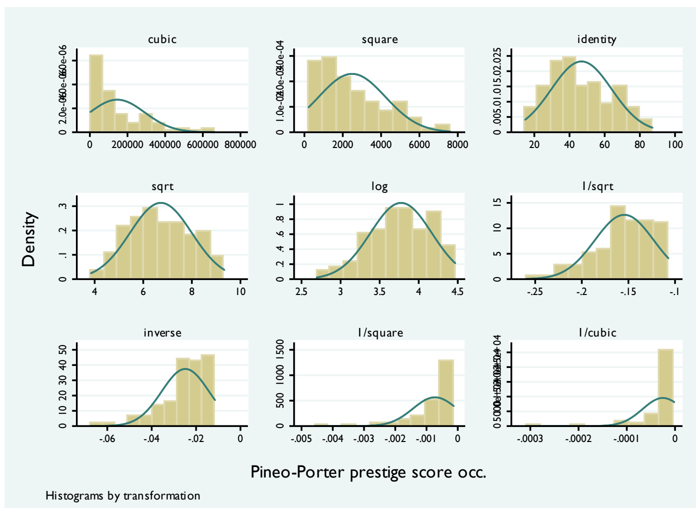
1) Transforming the dependent variable

The histogram below indicates that prestige may have a slight positive skew. The *kernal density* on the right shows a more general picture of the shape of the distribution.





The graphs below show what would happen if, instead of analyzing prestige scores, we were to analyze the logarithm of prestige, prestige-squared, or other transformations of the dependent variable. If any of these graphs appear to be more normally distributed, we may choose to use the transformed data in our analysis.



2) Run a robust regression analysis

Robust linear regression

Number of obs = 102
F(1, 100) = 48.28
Prob > F = 0.0000
R-squared = 0.5111
Root MSE = 12.09

prestige	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0028968	.0004169	6.95	0.000	.0020697	.0037239
_cons	27.14118	2.886142	9.40	0.000	21.41515	32.8672

Quantile (Median) regression

Number of obs = 102

Raw sum of deviations 1447 (about 43.5)

Min sum of deviations 954.6664

Pseudo R2 = 0.3402

prestige	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0030293	.0003073	9.86	0.000	.0024196	.0036391
_cons	23.94584	2.518318	9.51	0.000	18.94957	28.94211