## Part A: Problems

1. A firm that manufactures metal chairs has been concerned about the number and cost of machine breakdowns. The problem is that much of the equipment is old and is not reliable. The cost of replacing the machines is high and the owner of the firm is not sure that metal chairs are going to be popular next season. To help with the decision the owner collects information on the age of the twenty machines (in years) and the cost of the repairs on the machines last month. The estimated regression, along with the standard errors and the t-statistics are below.

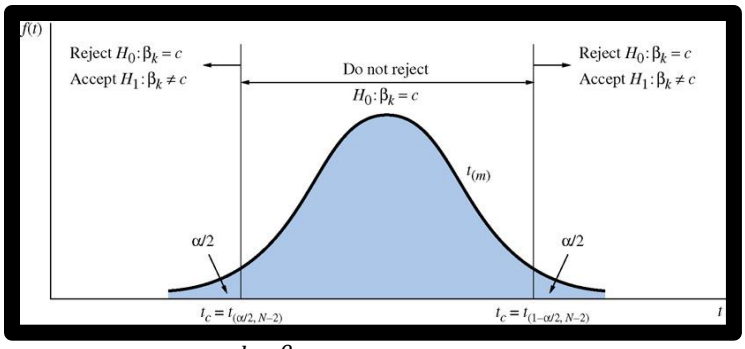
$$\widehat{Cost} = (???) + 29.68Age,$$
(se) (58.69) (????)
(t) (1.957) (4.844)

a. What is the estimate for the intercept? What is the standard error for the slope?
Since $t = \frac{b_i - \beta_i}{se(b_i)}, 1.957 = \frac{b_1 - 0}{58.69}, b_1 = 1.957 * 58.69 = 114.856.$
The estimate for the intercept is \$114.86.
Since $t = \frac{b_i - \beta_i}{se(b_i)} = 4.844 = \frac{29.68 - 0}{se(b_2)}, \ se(b_2) = 6.127.$
The standard error for the slope is 6.127.

b. Interpret the estimated coefficients.
The cost of monthly repairs on a machine that is brand new is on average \$114.86. Since we have no information on the range of ages in the machines, we cannot use this interpretation.
As the machines age by one year, cost of repairs increases by \$29.68 on average.

c. Construct a 95% confidence interval estimate for the slope.
$$\alpha = 0.05, \quad {}^{\alpha}/_{2} = 0.025, \quad df = n - 2 = 20 - 2 = 18$$
$$b_i \pm t_{\frac{\alpha}{2},df} se(b_i) = 29.68 \pm (2.101 * 6.127) = [16.807, 42.553]$$
At the 95% confidence level the marginal effect on monthly repair cost of aging a year is an increase of between \$16.81 and \$42.55.

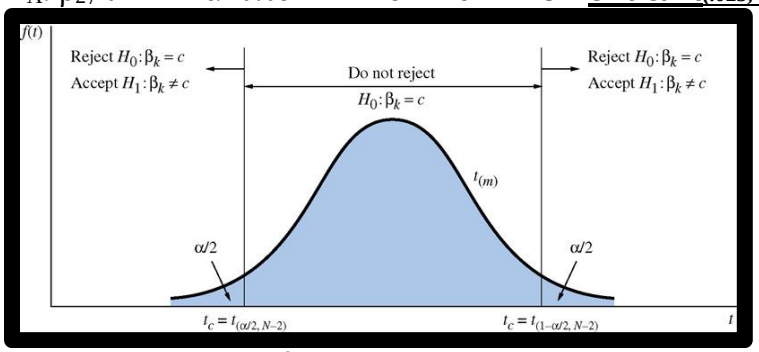
d. Test if the change in the cost of repairs is actually \$20 for each year the machine ages.
$H_0$: $\beta_2 = 20$
$H_A$: $\beta_2 \neq 20$ $\alpha = 0.05$ df= 20-2= 18 **critical $t_{(.025, 18)}$= -2.101 or 2.101**




Test statistic: $t = \frac{b_i - \beta_i}{se(b_i)}$ ,t = (29.68-20)/6.127 = 1.58

Reject $H_O$ if t < -2.101 or t > 2.101. -2.101 < 1.58 < 2.101, *Do not reject $H_O$*. There is insufficient evidence to reject the hypothesis that the change in monthly repair cost is $20 for each year the machine ages at the 5% significance level. **NOTE: Since $20 was in the confidence interval created in part c you could have used this to come to the same conclusion.**

e. Test the significance of the model.
$H_0$: $\beta_2=0$
$H_A$: $\beta_2\neq0$ α=0.05 df= 20-2= 18 **critical $t_{(.025, 18)}$= -2.101 or 2.101**




Test statistic: $t = \frac{b_i - \beta_i}{se(b_i)}$ ,t = (29.68-0)/6.127 = 4.844

Reject $H_O$ if t < -2.101 or t > 2.101. 2.101 < 4.844, *Reject $H_O$*.
There is evidence to reject the hypothesis that there is no linear relationship between age and the monthly cost of repairs at the 5% significance level. The model is statistically significant.

2. A real estate agent who specializes in commercial real estate wanted a better method, a more precise method, of judging the likely selling price (in $1,000s) of apartment buildings. He started by recording the price of a number of apartment buildings that had sold recently and the size in square feet (in 1,000s) of the building. The data is in a file entitled A3Q1Apart. **Submit your Stata log file in your assignment or in the D2L dropbox.**

a. Estimate the following simple linear regression model:

$$Price = \beta_1 + \beta_2 size + e$$

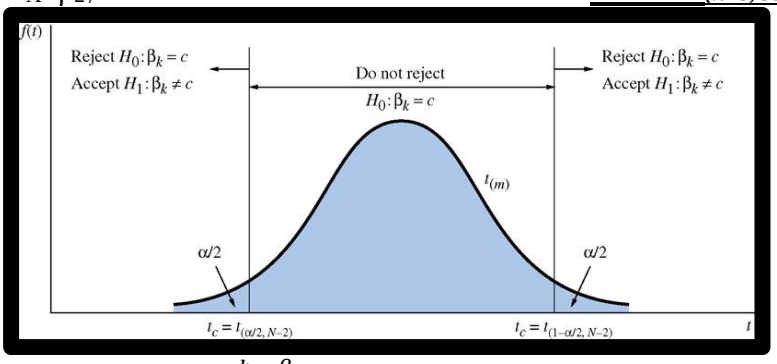
$$\widehat{Price} = 4082.11 + 44.86size \qquad R^2 = 0.1211$$
s.e. 1172.75 19.60

```
regress price size
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    5.24
       Model |  54013997.4     1  54013997.4            Prob > F      =  0.0277
    Residual |  391849896     38  10311839.4            R-squared     =  0.1211
-------------+------------------------------           Adj R-squared =  0.0980
       Total |  445863894     39  11432407.5            Root MSE      =  3211.2

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        size |   44.86097   19.60124     2.29   0.028     5.180332    84.5416
       _cons |   4082.111   1172.753     3.48   0.001     1707.997    6456.225
------------------------------------------------------------------------------
```

b. Interpret the estimated coefficients.
For each additional thousand square feet in an apartment the price increases on average by $44.86 thousand or $44,860.
The average selling price for an apartment building with no square feet is $4082.111 thousand or $4,082,111. This would then be the average value of the commercial property, without a building. Since the smallest building in the sample is 1500 square feet this interpretation may be valid.

c. Test the significance of the model.
$H_0$: $\beta_2=0$
$H_A$: $\beta_2\neq0$ $\alpha=0.05$ df= 40-2= 38 **critical $t_{(.025,\ 38)}$= ±2.024**




Test statistic: $t = \frac{b_i - \beta_i}{se(b_i)}$ ,t = (44.86097-0)/ 19.60124 = 2.29
Reject $H_O$ if t < -2.024 or t > 2.024. 2.29 > 2.024, *Reject $H_O$.*
There is evidence to reject the hypothesis that there is no linear relationship between size and price at the 5% significance level. The model is statistically significant.

d. What is the coefficient of determination, $R^2$? Interpret the coefficient of determination.

$R^2 = 0.1211$.

12.11% of the variation in the price of commercial apartment buildings was explained by the variation in the size, or number of square feet.

e. Use the regression results to predict the selling price ($\widehat{Price}$), the standard error of the forecast (sef), and the residuals ($e_i$).

f. Estimate with 95% confidence the mean price of a 55,000 square foot building.

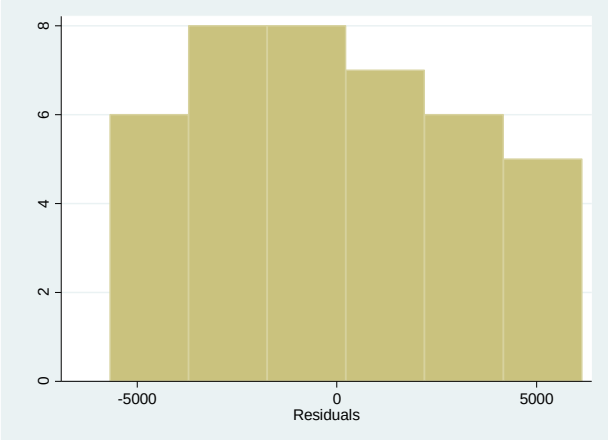
```
-------------------------+
     | size       yhat       sef |
     |-------------------------------|
 41. |   55   6549.464   3251.165 |
     +-------------------------------+
```

$\hat{y} \pm t_{0.025,38} se(f) = 6549.464 \pm 2.024 * 3251.165 =$ [-30.894, 13129.822]

The mean price of a 55,000 square foot building is between 0 and $13,129,822 with 95% confidence.

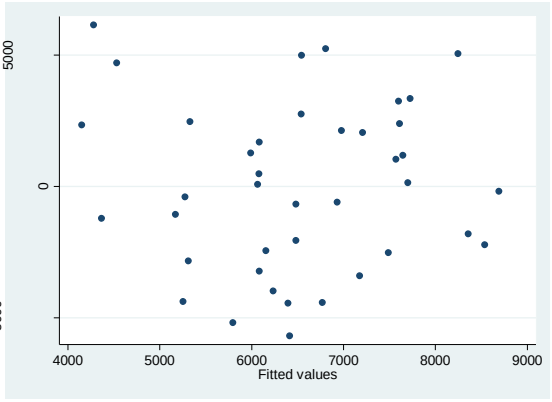
g. Do a histogram of the residual. Do the residuals appear to be approximately normally distributed?

The errors seem to be approximately normally distributed.



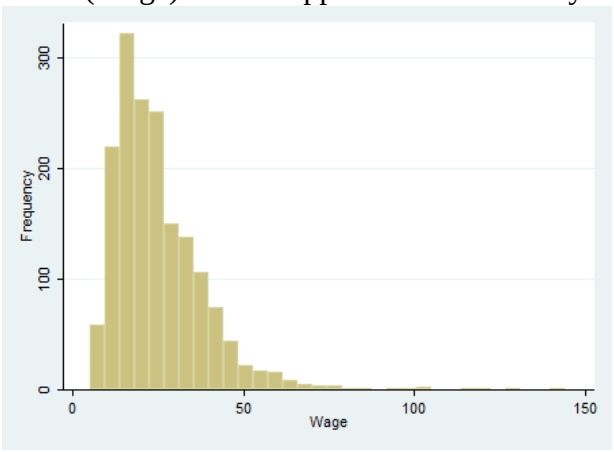

h. Plot the residuals against the predicted selling price. Does it appear that heteroscedasticity is problem?

The variance seems to be constant, with no recognizable funnel shape.

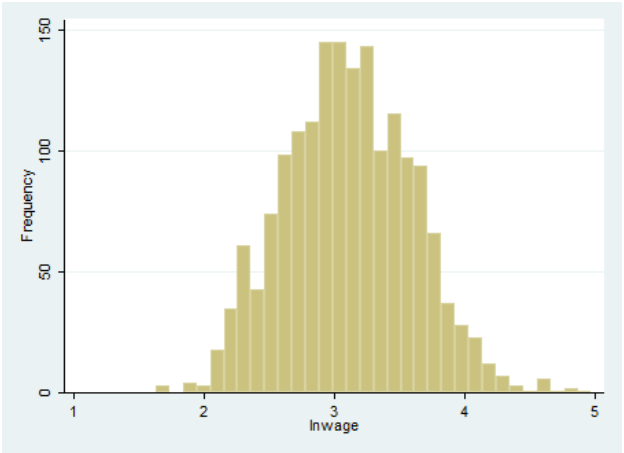



3. How much does education affect wage rates? Use the data file and data definitions from Assignment #2, question 4, *LandruAB2007.dat.* **Submit your Stata log file in your assignment or in the D2L dropbox.**

   a. Create a new variable, ln of HourlyWageRate, *ln(Wage).*
   b. Construct histograms of the variable HourlyWageRate, and the new variable, ln(Wage). Which appears more normally distributed?




Graph A




Graph B

The histogram of the original hourly wage rates, Graph A above, is skewed right, while the graph of the natural log (ln) of the hourly wage rate, Graph B above, is mound shaped, appearing to be very roughly bell shaped and symmetrical. Graph B would be closer to being the shape of a normal distribution.

c. Estimate the linear regression $Wage = \beta_1 + \beta_2 Educ + e$, then predict wages and the residual.

$Wage = 2.976 + 1.619 Educ$ $R^2 = 0.0904$
(se) 1.753 0.124

```
regress HourlyWageRate YrsEducation
      Source |       SS           df       MS                 Number of obs =    1719
-------------+----------------------------------               F(  1,  1717) =  170.65
       Model |  30572.6835         1  30572.6835               Prob > F      =  0.0000
    Residual |  307601.694      1717  179.150666               R-squared     =  0.0904
-------------+----------------------------------               Adj R-squared =  0.0899
       Total |  338174.377      1718  196.841896               Root MSE      =  13.385

------------------------------------------------------------------------------
HourlyWage~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 YrsEducation |   1.619408   .1239649    13.06   0.000      1.37627    1.862546
       _cons |   2.976186    1.75334     1.70   0.090     -.4627222    6.415093
------------------------------------------------------------------------------
```

```
. predict yhat
(option xb assumed; fitted values)

. predict ehat, residuals
```

d. Estimate the log-linear regression $\ln(Wage) = \beta_1 + \beta_2 Educ + e$, then predict wages and the residual. (***Do not forget to give your yhat and residuals a different name here).

$\ln(Wage) = 2.296 + 0.059 Educ$ $R^2 = 0.0964$
(se) 0.061 0.004

```
regress lnWage YrsEducation
      Source |       SS           df       MS                 Number of obs =    1719
-------------+----------------------------------               F(  1,  1717) =  183.22
       Model |  40.2937149         1  40.2937149               Prob > F      =  0.0000
    Residual |  377.593128      1717  .21991446               R-squared     =  0.0964
-------------+----------------------------------               Adj R-squared =  0.0959
       Total |  417.886843      1718  .243240304              Root MSE      =  .46895

------------------------------------------------------------------------------
      lnWage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

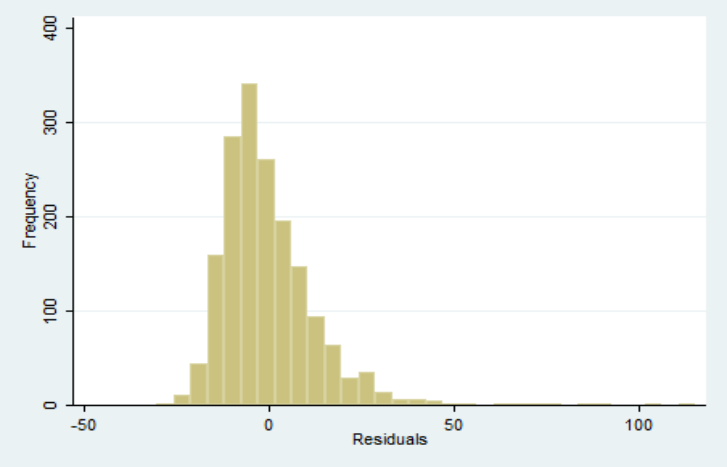
```
------------------------------------------------------------------------------
 YrsEducation |   .0587906   .0043433    13.54   0.000     .050272    .0673093
        _cons |   2.296204   .0614305    37.38   0.000     2.175717     2.41669
------------------------------------------------------------------------------
. predict yhat1
(option xb assumed; fitted values)

. predict ehat1, residuals
```

e. What is the estimated return to education in each model? That is, for an additional year of education, what percentage of increase in wages can an average worker expect?
For each extra year of education the linear model predicts a marginal effect of an average increase of $1.62 per hour. At the mean of the hourly wage rate, $25.49, that would be an increase of $1.62/$25.49 = 0.0636, or 6.36% for an extra year of formal education. This percentage is not constant at all wage rates.
For each extra year of education the log-linear model predicts an elasticity of an average increase of 0.0588, or 5.88% increase in wages for an extra year of formal education. This percentage is constant at all wage rates.

f. Construct histograms of the residuals from each model. Does one set of residuals appear more compatible with normality than the other?

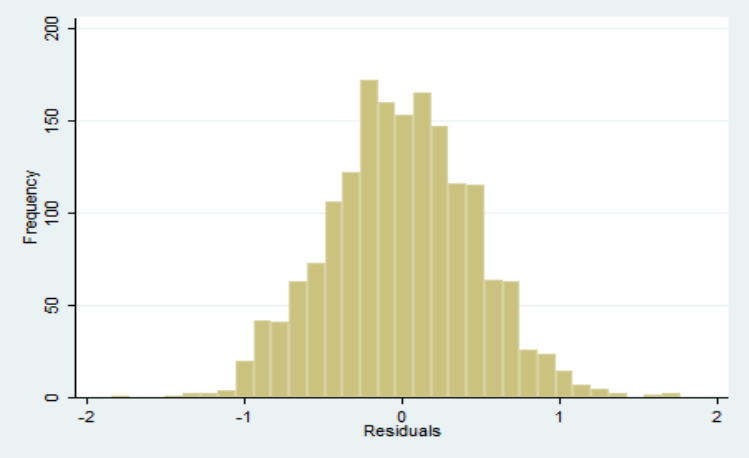
Graph C – Residuals from linear model




The residuals from the linear model appear to have a right tail (or positive tail), see Graph C. The residuals do not appear to be normally distributed, so would need to be tested for normality. The residuals look much like the original histogram of Hourly Wage Rates in part b, Graph A.

The residuals from the log-linear model, see Graph D, appear more mound shaped. The log - linear model residuals also appear to follow the general shape of the histogram of the ln(Wage), in part b, Graph B, and are closer to being the shape of a normal distribution than the linear model residual histogram.

Graph D – residuals from log-linear model




g. Calculate the "generalized" $R^2$ for the log-linear model (see slide 4-37) and compare it to the $R^2$ of the linear model. Which model fits the data better?

gen predictedy1=exp( yhat1+.10995)

correlate HourlyWageRate predictedy1
(obs=1719)

| | Hourly~e | predic~1 |
|---|---|---|
| HourlyWage~e | 1.0000 | |
| predictedy1 | 0.3059 | 1.0000 |

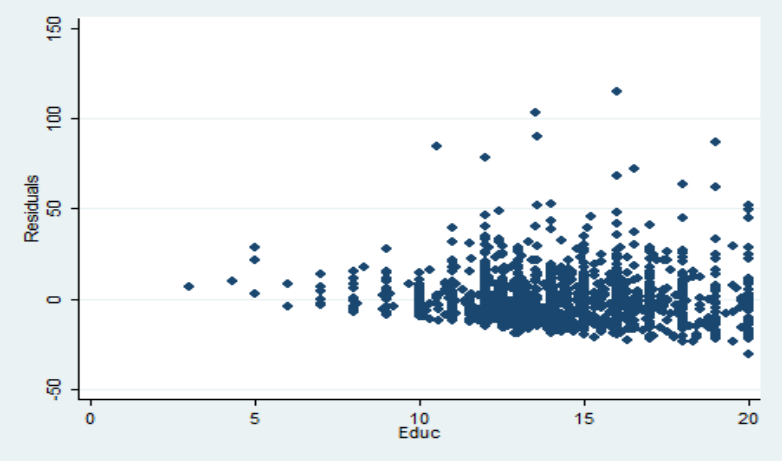
$$R_g^2 = r_{y,\hat{y}}^2 = 0.3059^2 = 0.0936$$

The "generalized" $R^2$ for the log-linear model is 0.0936 while the $R^2$ for the linear model is 0.0904. The log-linear model does a slightly better job of explaining the variation in the hourly wage rate than the linear model. Neither model explains very much.

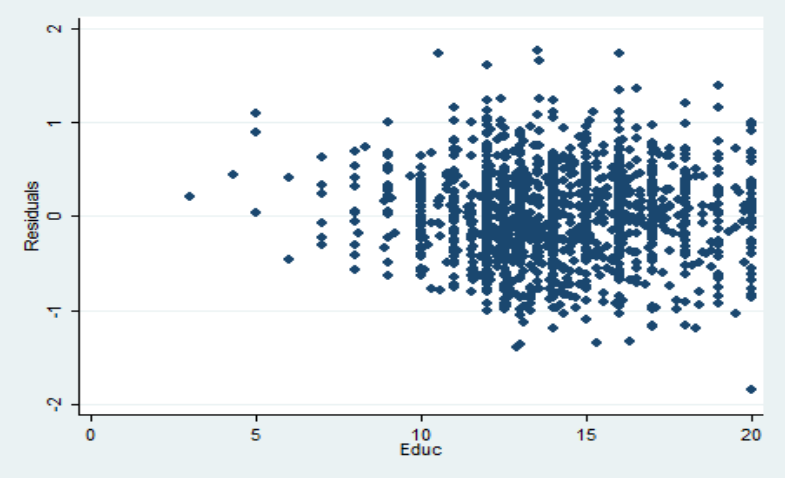
h. Plot the residuals from each model against *Educ*. Do you observe any patterns?

The graph of the residuals from the linear model scattered against the years of formal education, Graph E, has a pattern. The variance in the residuals is not constant, as there is a "fan" shape visible. As the number of years of formal education increases the residuals are spreading out more and more from the expected value of zero. The spread is also much larger in the positive residuals than in the negative residuals, as can be seen with a scale of 150 to negative 50. In contrast the graph of the residuals from the log-linear model scattered against the years of formal education, Graph F, does not have a pattern. The residuals appear to be more randomly scattered with negative and positive values of equal magnitude.

Graph E – Residuals from linear model




Graph F – Residuals from log-linear model




i. Using each model, predict the hourly wage rate, *HourlyWageRate,* for a worker with 14 years of education from both models and compare the prediction to the average wage of all workers with 14 years of education. Which prediction is more accurate?

```
summarize HourlyWageRate if YrsEducation==14

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
HourlyWage~e |       155    26.00516    11.41325         6.9      78.32

list in 98
     +-------------------------------------------------------------------------------------------------+
 98. | Prov | Age | Gender | Marita~s | Labour~s | YrsExp~e | Hourly~e | Uniono~e | Indust~e | YrsEdu~n |
     |   48 |  40 |      1 |        1 |        1 |       20 |    25.51 |        1 |        7 |       14 |
     |-------------------------------------------------------------------------------------------------|
     |   lnWage    |       yhat        |      yhat1      |      ehat1      |       ehat        |   predic~1    |
```

| 3.239071 | 25.64789 | 3.119272 | .1197981 | -.1378928 | 25.26001 |
+---+

The predicted hourly wage rate for a worker with 14 years of formal education is $25.65 using the linear model, $25.26 using the log-linear model (with the correction factor), and the average wage of all workers with 14 years of education is $26.01. The prediction of the linear model is closer to the average wage in the sample.

j. Based on the results from parts b - i, which functional form would you use?
The residuals in the linear model do not appear in the histogram to be normally distributed, nor do the residuals in the scatter plot against years of education look to have constant variance. The linear model has a coefficient of determination of 0.0904.
In contrast, the residuals in the log-linear model are approximately bell shaped in the histogram, and have no pattern in the scatter plot against years of education, so could have constant variances. The generalized coefficient of determination is 0.936 in the log-linear model. The log-linear model is a better fit to the data and should be used instead of the linear model.

## Part B: Data Project **LIBRARY DATA**

**NOTE: You may be asked to sign into the Library Resources. Make sure you have your UCID card before you start.**

- To get to data through the Library's Website start at the Home page and select "Search Collections" from the tabs near the top.
- On the new webpage select "Data" from near the bottom right
- Now select the link to Spatial and Numeric Data Services, and then "Data & Statistics".
- Now you will have various tabs for Aggregate Data, Microdata, Time Series, and Data Citation & Software.
- Select Microdata, LANDRU, Survey of Labour and Income Dynamics,

You will see a large number of data sets, (good idea just to look around a bit) scroll down until you find **Survey of Labour and Income Dynamics (SLID)** and click on the tab. Here you will see data sets from **1993-2010**, I want you to extract the following variables from the '**Person file**' for **2010**. Click the boxes for the variables listed below.

**ECAGE26** "Person's age , refyear, external cross-sec file"
**ECSEX99** "Sex of respondent on external cross-sectional files"
**MARST26** "Marital status of person as of December 31 of refyear"
**PVREG25** "Province of residence group, household, December 31, refyear"
**ALFST28** "Annual labour force status"
**YRXFTE11** "Number of years of work experience, full-year full-time"

**NOCG2E6** "NOC-S 2006; NOC-S 2001 (End of reference year)"
**IMPHWE1** "Hourly wage at end of job or end of refyear"
**UNCOLL1** "Flag - Union member or covered by collective agreement"
**N07C3G10** "Grouping 3, industry code of employer based on NAICS 2007"
**HLEVEG18** "Highest level of education of person, 1st grouping"
**YRSCHL18** "Number of years of schooling completed by person (elem, high school, post secondary)"

Follow the directions done in the Lab to get four files by email and save all the original files (that you get a link to in your email), and then save your data through SPSS into a format you can import into Stata.

## Create a log file and a do file of all the commands used in this exercise. You will need to submit these into D2L.

a) Open the file that is "**.sps" in Word. This file shows you the variables and what the various codes mean. You will need this to edit your dataset.

b) Read your data into Stata. Save the raw data file as Stata data. **You should have 49,787 records when you start.**

**c)** Check every variable and drop observations that do not answer directly the survey question. The survey allows individuals to answer 'don't know' or 'unknown' or 'interim processing code' or 'refusal' or 'not applicable".

**Save your data with a new file name, so you have an original file with the raw data, and your edited data file.**

```
drop if pvreg25>60
(447 observations deleted)

. drop if ecsex99>2
(0 observations deleted)

. drop if marst26>6
(77 observations deleted)

. drop if alfst28>7
(6906 observations deleted)

. drop if yrxfte11>96
(7787 observations deleted)

. drop if uncoll1>3
(11128 observations deleted)

. drop if nocg2e6>25
(732 observations deleted)
```

```
. drop if yrschl18==999.7
(0 observations deleted)

. drop if imphwe1==999.99
(19 observations deleted)

. drop if n07c3g10>16
(130 observations deleted)

. drop if hleveg18>12
(101 observations deleted)

. drop if yrschl18==99.7
(0 observations deleted)
```

d) Summarize the data. Comment on which of the summary statistics are valid to interpret, and which are not.

summarize

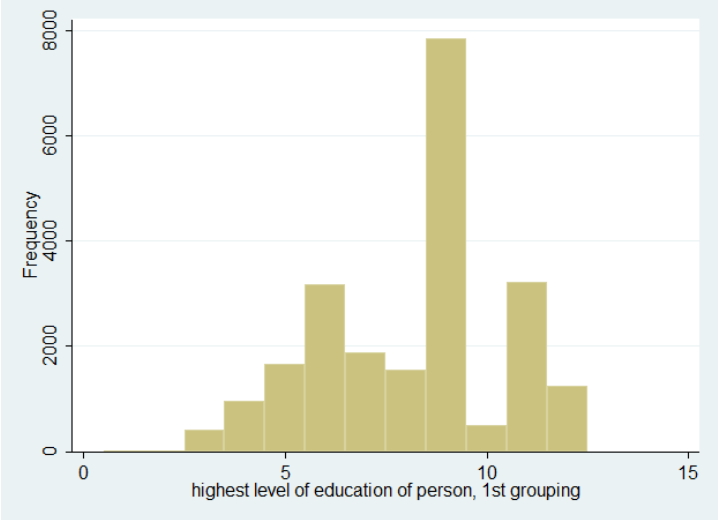
| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| year99 | 22460 | 2010 | 0 | 2010 | 2010 |
| ecage26 | 22460 | 39.38348 | 13.96567 | 16 | 69 |
| ecsex99 | 22460 | 1.513446 | .4998303 | 1 | 2 |
| marst26 | 22460 | 2.966563 | 2.221834 | 1 | 6 |
| pvreg25 | 22460 | 33.36897 | 14.30438 | 10 | 59 |
| alfst28 | 22460 | 1.987578 | 1.853458 | 1 | 7 |
| yrxfte11 | 22460 | 15.76794 | 12.79287 | 0 | 50 |
| nocg2e6 | 22460 | 11.85703 | 6.76568 | 1 | 25 |
| imphwe1 | 22460 | 22.41416 | 12.94953 | 6 | 156.61 |
| uncoll1 | 22460 | 2.353028 | .9241178 | 1 | 3 |
| n07c3g10 | 22460 | 9.20886 | 4.067965 | 1 | 16 |
| hleveg18 | 22460 | 8.182769 | 2.257705 | 1 | 12 |
| yrschl18 | 22460 | 13.58898 | 2.618269 | 0 | 20 |

It would be valid to interpret the mean and standard deviation for age, years of fulltime experience, hourly wage rate, and number of years of education.

It is not valid to interpret the mean and standard deviation on anything that uses a number to represent a characteristic: gender, marital status, province, annual labour force status, type of job or occupation, union membership, industry code, or highest level of education.

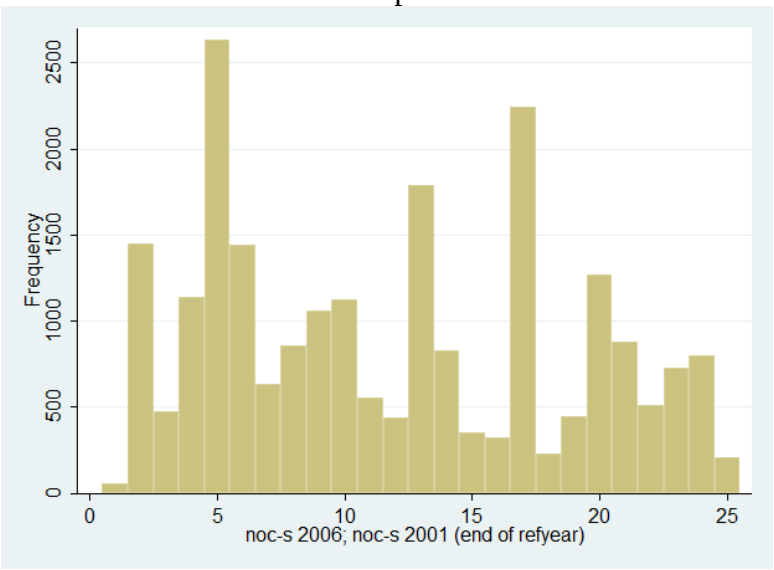
e) Create a histogram of the education levels using **HLEVEG18.** What level is the most

common? What is the second most common? What proportion of the sample graduated high school or more?



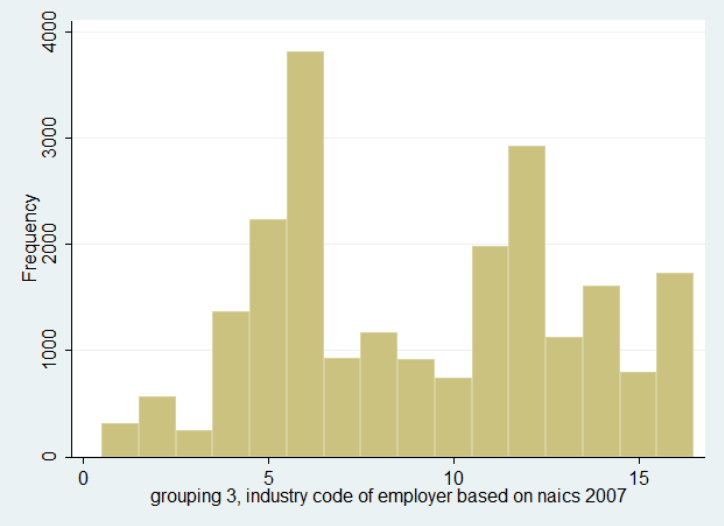

The most common level of education attainment is non-university postsecondary certificate. The second highest would be graduated high school and bachelor's degree (seem to be about the same number). There are 19,386 people in the survey who graduated high school or achieved a higher educational attainment out of the 22,460 records, so **86.3%**.

f) Create a histogram of the occupations using **NOCG2E6.** What occupations are the three most common in the sample?



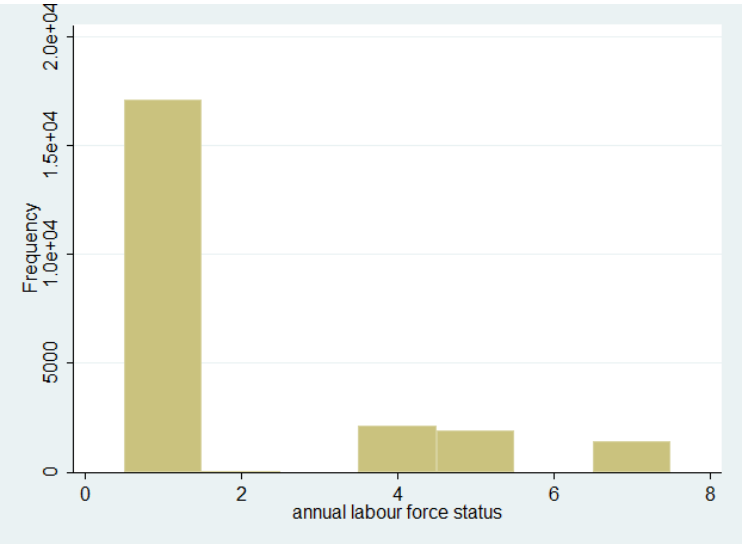

The three most common occupations in the sample are clerical occupations/including supervisors, sales/service occupations n.e.c., and retail salespersons/sales clerks/cashiers.

g) Create a histogram of the industries using **N07C3G10.** What are the three industries that are the most common in this sample?




The three most common industries in the sample are manufacturing, trade, and health care/social assistance.

h) Create a histogram of the **ALFST28.** What are the four most common labour force survey statuses? Does your graph make sense with what you learned in Economics 203?




Although there are seven codes in the Labor Force status, only four are visible on the histogram. These are individual employed all year, individuals employed part of the year and unemployed part of the year, individuals employed part of the year and not in the labour force part of the year, and individuals employed, unemployed and not in labour force during year. These individuals were all employed for some or all of the year.
In Economics 203 you should have seen that roughly 2/3 of the population was in the labour force, and that about 8% of the labour force is unemployed at any one time in the year. This histogram shows most of the labour force employed all year, with a fraction of the people considered not in the labour force, even if they were employed or unemployed during the year.

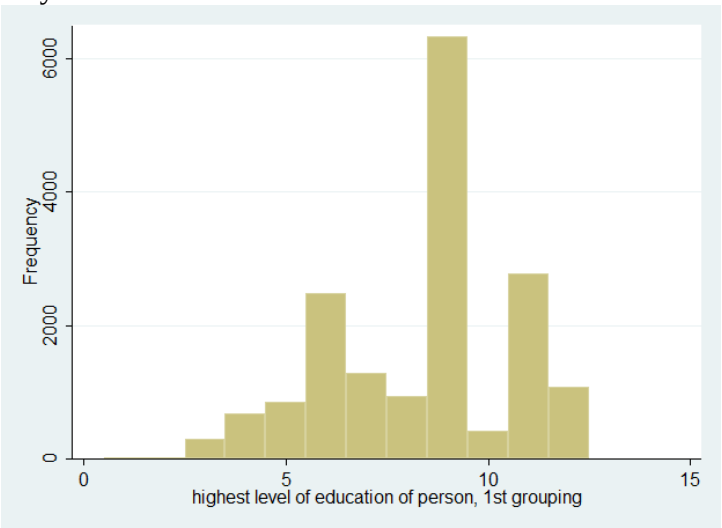
i) Keep only individuals that are **employed fulltime and all year**. Save this new dataset with a new file name.

drop if alfst28>1
(5378 observations deleted)

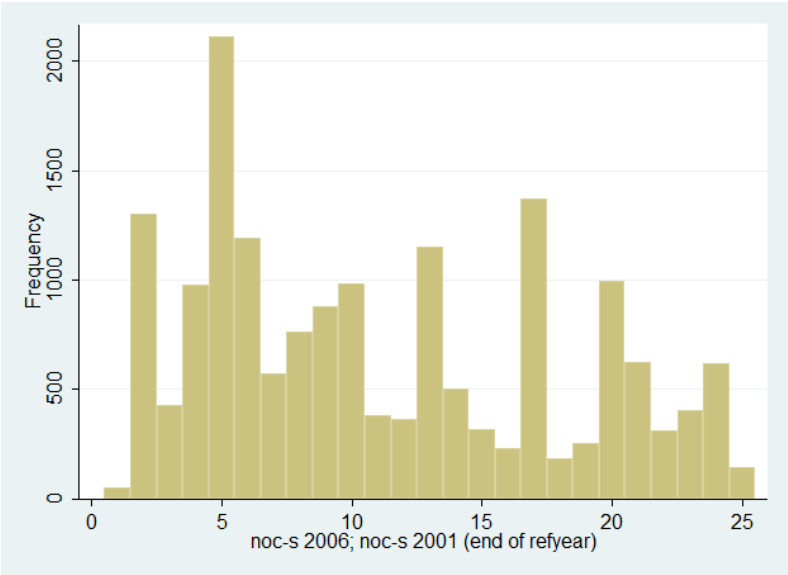
*For information only:*
*summarize*

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| year99 | 17082 | 2010 | 0 | 2010 | 2010 |
| ecage26 | 17082 | 41.35739 | 12.87487 | 16 | 69 |
| ecsex99 | 17082 | 1.526109 | .4993325 | 1 | 2 |
| marst26 | 17082 | 2.644772 | 2.100519 | 1 | 6 |
| pvreg25 | 17082 | 33.74113 | 14.15116 | 10 | 59 |
| alfst28 | 17082 | 1 | 0 | 1 | 1 |
| yrxfte11 | 17082 | 17.43947 | 12.2508 | 0 | 50 |
| nocg2e6 | 17082 | 11.19055 | 6.744819 | 1 | 25 |
| imphwe1 | 17082 | 24.27924 | 13.22505 | 6 | 156.61 |
| uncoll1 | 17082 | 2.277134 | .9492123 | 1 | 3 |
| n07c3g10 | 17082 | 9.31179 | 4.029728 | 1 | 16 |
| hleveg18 | 17082 | 8.397494 | 2.228121 | 1 | 12 |
| yrschl18 | 17082 | 13.76947 | 2.641468 | 0 | 20 |

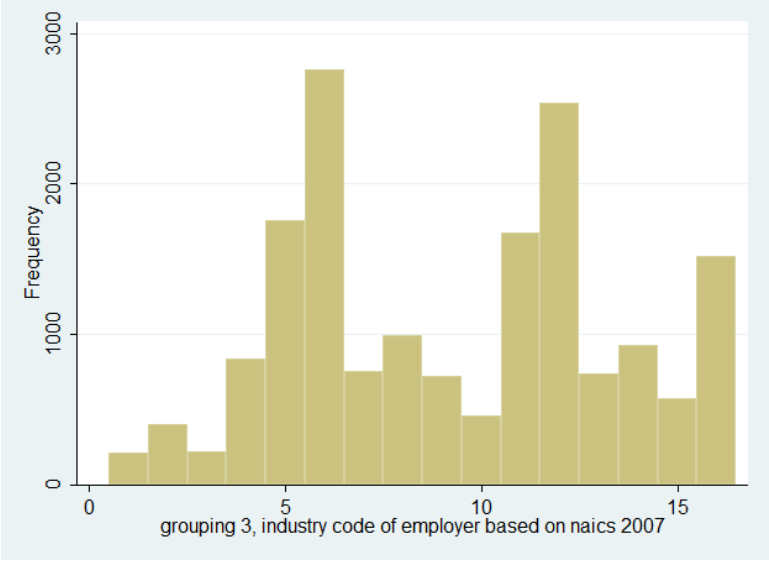
j) Repeat the histograms from parts e) to g). Compare the graphs and comment on what changed between using the whole sample and using only people employed fulltime and all year.



The most common here are identical to the sample with everyone, non-university postsecondary certificate , graduated high school and bachelor's degree although the frequency is not as large for the non-university postsecondary certificate.




In the sample of people who work fulltime and all year, the three most common occupations are other management occupations, clerical occupations/including supervisors, and sales/service occupations n.e.c.. Retail salespersons/ sales clerks/cashiers is the fourth highest in the sample of people who work fulltime and all year, but was third highest in the sample of the people who completed all the questions.




The three most common industries in the sample of people who work fulltime and all year are manufacturing, trade, and health care/social assistance, the same as the whole sample.

k) Using the industry (**N07C3G10**) and province (**PVReg25**), summarize the sample of

individuals that work in "Public Administration" and live in Alberta. Do this for the same industry for individuals that live in Saskatchewan.

**Alberta**

*summarize if pvreg25==48 & n07c3g10==16*

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| year99 | 122 | 2010 | 0 | 2010 | 2010 |
| ecage26 | 122 | 41.7623 | 11.23384 | 18 | 68 |
| ecsex99 | 122 | 1.598361 | .4922513 | 1 | 2 |
| marst26 | 122 | 2.909836 | 2.112552 | 1 | 6 |
| pvreg25 | 122 | 48 | 0 | 48 | 48 |
| alfst28 | 122 | 1 | 0 | 1 | 1 |
| yrxfte11 | 122 | 18.29508 | 11.36523 | 0 | 50 |
| nocg2e6 | 122 | 7.303279 | 4.794779 | 1 | 23 |
| imphwe1 | 122 | 32.56295 | 11.37935 | 6.14 | 66.28 |
| uncoll1 | 122 | 1.655738 | .9160686 | 1 | 3 |
| n07c3g10 | 122 | 16 | 0 | 16 | 16 |
| hleveg18 | 122 | 9.065574 | 1.848554 | 4 | 12 |
| yrschl18 | 122 | 14.38197 | 2.361468 | 10 | 20 |

**Saskatchewan**

*. summarize if pvreg25==47 & n07c3g10==16*

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| year99 | 109 | 2010 | 0 | 2010 | 2010 |
| ecage26 | 109 | 42.80734 | 11.06251 | 17 | 66 |
| ecsex99 | 109 | 1.513761 | .5021192 | 1 | 2 |
| marst26 | 109 | 2.266055 | 1.970346 | 1 | 6 |
| pvreg25 | 109 | 47 | 0 | 47 | 47 |
| alfst28 | 109 | 1 | 0 | 1 | 1 |
| yrxfte11 | 109 | 19.88073 | 11.34046 | 0 | 45 |
| nocg2e6 | 109 | 8.220183 | 5.724149 | 1 | 23 |
| imphwe1 | 109 | 29.34633 | 9.911478 | 10.12 | 72 |
| uncoll1 | 109 | 1.688073 | .949632 | 1 | 3 |
| n07c3g10 | 109 | 16 | 0 | 16 | 16 |
| hleveg18 | 109 | 8.807339 | 2.016033 | 3 | 12 |
| yrschl18 | 109 | 14.11835 | 2.438661 | 6 | 20 |

l) Using the information for the sub-sample in part k) for Alberta, test if the mean hourly wage for individuals who work in "Public Administration" is less than $18.00 per hour.

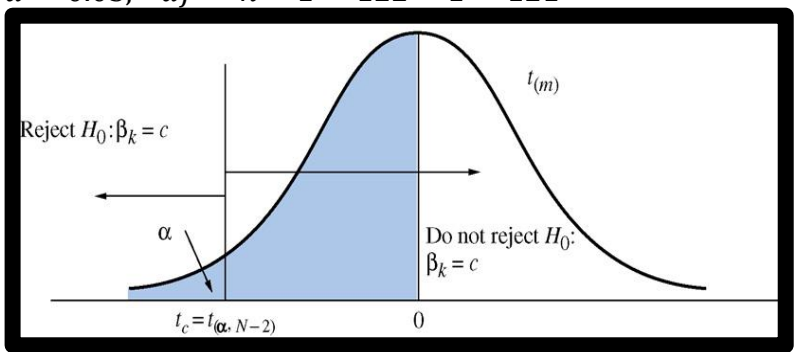
ttest imphwe1==18.00 if pvreg25==48 & n07c3g10==16

One-sample t test

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 imphwe1 |     122    32.56295    1.030238    11.37935    30.52332    34.60258
------------------------------------------------------------------------------
   mean = mean(imphwe1)                                      t =  14.1355
Ho: mean = 18.00                         degrees of freedom =      121

 Ha: mean < 18.00           Ha: mean != 18.00           Ha: mean > 18.00
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000      Pr(T > t) = 0.0000
```

$H_0: \mu \geq \$18$
$H_A: \mu < \$18$
$\alpha = 0.05, \quad df = n - 1 = 122 - 1 = 121$




$t_c = -1.6575$

$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{32.56295 - 18}{11.379/\sqrt{122}} = 14.1355$ $P(t_{121} < 14.1355) = 1$

Reject $H_o$ if $t < t_{.05,121} = -1.6575$ -1.66575 < 14.1355, so do not reject $H_o$

**There is not sufficient evidence in this sample to reject the hypothesis that the mean hourly wage of individuals working in the Public Adminsitration industry in the province of Alberta in 2010 is greater than or equal to $18.00 at the 5% significance level.**

m) Using the information for the sub-samples in part k), test if the mean hourly wage for individuals in Alberta is larger than the mean hourly wage for individuals in Saskatchewan, who work in "Public Administration", by $3.

sdtesti 122 . 11.37935 109 . 9.911478

Variance ratio test

```
------------------------------------------------------------------------------
       |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |     122           .    1.030238    11.37935           .           .
       y |     109           .    .9493474    9.911478           .           .
---------+--------------------------------------------------------------------
combined |     231           .           .           .           .           .
```

---

ratio = sd(x) / sd(y) f = 1.3181
Ho: ratio = 1 degrees of freedom = 121, 108

Ha: ratio < 1 Ha: ratio != 1 Ha: ratio > 1
Pr(F < f) = 0.9282 2*Pr(F > f) = 0.1436 Pr(F > f) = 0.0718

***Let Alberta be population 1 and Saskatchewan be population 2.***

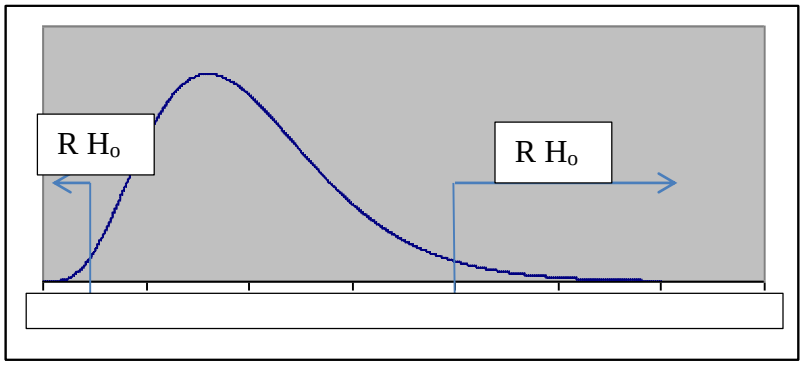
***STEP 1 – test that the variances are equal.***

$H_0 : \sigma_1^2 / \sigma_2^2 = 1$

$H_1 : \sigma_1^2 / \sigma_2^2 \neq 1$

$F = s_1^2 / s_2^2 = 129.4896/98.2374 = 1.32$ $2*P(F_{121,108}>1.32)=2*0.0718=0.1436$

Reject $H_o$ if *p-value* < α, **0.1436 > 0.05 Do not Reject $H_o$**

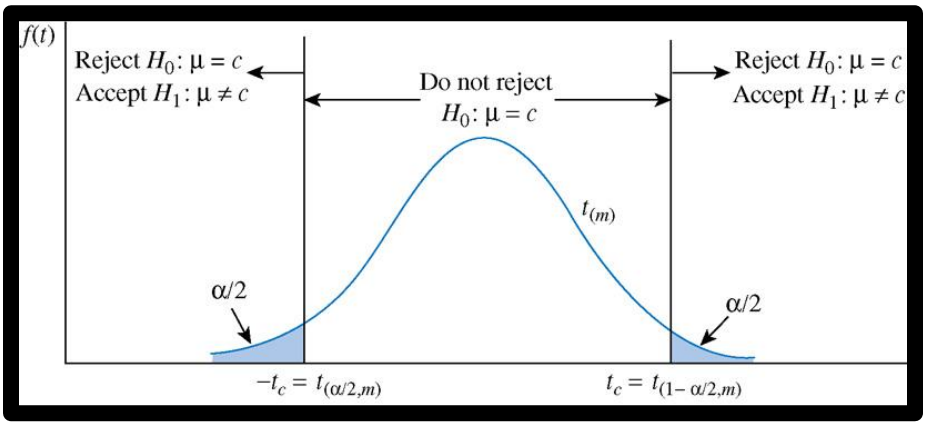



**There is insufficient evidence to reject the hypothesis that the variances are equal on the mean wages in Alberta and Saskatchewan in the Public Administration sector of the economy with 95% confidence.**

***STEP 2 Test the means***

$H_0 : (\mu_1 - \mu_2) = 3$ $H_1 : (\mu_1 - \mu_2) \neq 3$

df = 122 + 109 – 2 = 229 α = 5% $t_{0.25,\ 229} = \pm 1.974$



Reject $H_o$ if t > 1.974 or if t < -1.974

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.1534 \quad S_p^2 = \left(\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}\right) = 114.7506$$

-1.974 < 0.1534 < 1.974 **<u>Do not reject $H_o$</u>**

**<u>There is insufficient evidence to reject the hypothesis that the mean wages in Alberta are $3 more than the mean wages Saskatchewan in the Public Administration sector of the economy, in 2010, at the 5% significance level.</u>**

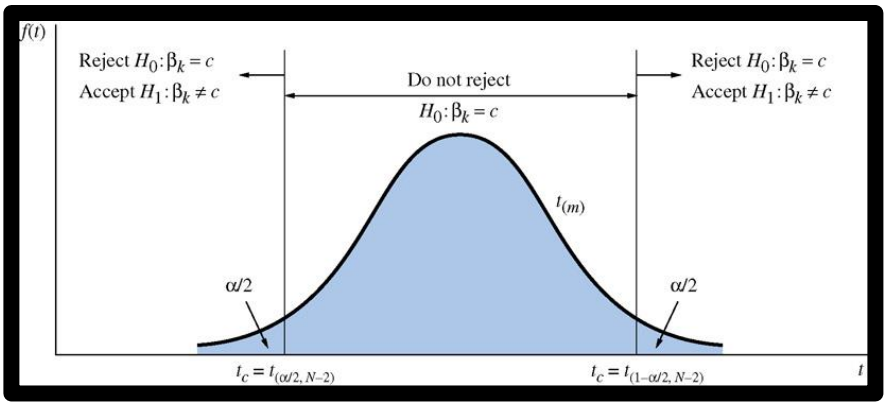
n) Using **NOCG2E6** variable, regress hourly wage (**IMPHWE1)** against years of experience (**YRXFTE11)** for individuals employed as "Other Management Occupations". $Wage = \beta_1 + \beta_2 Experience + e$

$$Wage = 27.32 + 0.35 Experience \quad R^2 = 0.0577$$
(s.e.) (0.96) (0.04)

```
regress imphwe1 yrxfte11 if nocg2e6==2
      Source |       SS           df       MS                 Number of obs =    1299
-------------+----------------------------------               F(  1,  1297) =   79.45
       Model |   17760.024     1       17760.024               Prob > F      =  0.0000
    Residual |  289918.953  1297   223.530419                  R-squared     =  0.0577
-------------+----------------------------------               Adj R-squared =  0.0570
       Total |  307678.977  1298   237.040815                  oot MSE       =  14.951
------------------------------------------------------------------------------
     imphwe1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    yrxfte11 |   .3517152   .0394582     8.91   0.000     .2743062    .4291241
       _cons |   27.31559   .9604597    28.44   0.000     25.43136    29.19981
```

o) Interpret the coefficients from your regression in part (n).
The average wage for someone in "Other Management Occupations" with no experience was $27.32 in Canada in 2010. As there are 3 records in the dataset for people with no experience working in this occupation this interpretation can be used.
Every extra year of experience increased hourly wage by $0.35, on average, in the "Other Management Occupations" in Canada in 2010.

p) Test the significance of the model.
$H_0$: $\beta_2 = 0$
$H_A$: $\beta_2 \neq 0$
$\alpha = 0.05$ df= 1299 - 2= 1297 critical $t_{(.025,\ 1297)} = \pm 1.96$.




Reject $H_O$ if t < -1.96 or t > 1.96.

Test statistic: $t = \frac{b_i - \beta_i}{se(b_i)}$ ,t = (.3517-0)/0.039 = 8.91

1.96 < 8.91, **Reject $H_O$.**

There is evidence to reject the hypothesis that there is no linear relationship between the hourly wage rate and experience. The model is statistically significant.

q) Compute the predicted hourly wage for an individual with 20 years of experience. Compute the standard error for your prediction.

```
predict yhat if nocg2e6==2
(option xb assumed; fitted values)
(15783 missing values generated)

. predict sef if nocg2e6==2, stdf
(15783 missing values generated)

list nocg2e6 yrxfte11 yhat sef in 579

     +-------------------------------------------------------+
     |    nocg2e6      yrxfte11      yhat       sef |
     |-------------------------------------------------------|
579. | other mana         20      34.34989   14.95689 |
     +-------------------------------------------------------+
```

The predicted hourly wage for an individual with 20 years of experience, working in in the "Other Management Occupations" in Canada in 2010 would be an average of $34.35 per hour with a standard error of the forecast of $14.96.

r) Create a new variable, log of hourly wage. For individuals employed as "Other Management Occupations" regress the log of hourly wage on years of experience.

$$\ln(Wage) = \beta_1 + \beta_2 Experience + e$$

$$\ln(Wage) = 3.19 + 0.01 Experience \qquad R^2 = 0.0577$$
(s.e.) (0.029) (0.001)

```
gen lwage=ln(imphwe1)
```

```
. regress lwage yrxfte11 if nocg2e6==2
      Source |       SS           df       MS                 Number of obs =    1299
-------------+----------------------------------               F(  1,  1297) =   95.32
       Model |  20.0486898     1      20.0486898                Prob > F      =  0.0000
    Residual |  272.789706  1297  .210323597                    R-squared     =  0.0685
-------------+----------------------------------               Adj R-squared =  0.0677
       Total |  292.838395   1298  .225607392                   Root MSE      =  .45861
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    yrxfte11 |   .0118171   .0012104     9.76   0.000     .0094427    .0141916
       _cons |   3.192812   .0294615   108.37   0.000     3.135014    3.250609
```

s) Interpret the coefficient on years of experience from the regression in part (r).
A one year increase in experience leads to an average increase in wage of [100* 0.0118% ] or a 1.18% increase in the hourly wage in the "Other Management Occupations" in Canada in 2010.

t) Compute the predicted wage for an individual with 20 years of experience using the regression results from part (r).

```
predict yhat1 if nocg2e6==2
(option xb assumed; fitted values)
(15783 missing values generated)

. predict sef1 if nocg2e6==2, stdf
(15783 missing values generated)

. list nocg2e6 yrxfte11 yhat1 sef1 in 579

     +-------------------------------------------+
     |   nocg2e6   yrxfte11      yhat1      sef1 |
     |-------------------------------------------|
579. | other mana         20   3.429154   .4587931 |
     +-------------------------------------------+

. gen yhatc=exp(yhat1+.1051617985) if nocg2e6==2
(15783 missing values generated)

list nocg2e6 yrxfte11 yhat sef yhat1 sef1 yhatc in 579

     +-------------------------------------------------------------------------
> ----+
     |   nocg2e6   yrxfte11       yhat       sef      yhat1       sef1      yhatc |
     |-------------------------------------------------------------------------
> ----|
579. | other mana         20   34.34989   14.95689   3.429154   .4587931   34.27157 |
     +-------------------------------------------------------------------------
```

The predicted wage for an individual with 20 years of experience in the "Other Management Occupations" in Canada in 2010 is $34.27.

u) Compare the predicted wage from part (q) and part (t).

The predicted hourly wage for an individual with 20 years of experience, working in in the "Other Management Occupations" in Canada in 2010 , using the linear model, would be an average of $34.35 per hour in contrast to $34.27 using the log-linear model.

***For information only:***

summarize if nocg2e6==2 & yrxfte11==20

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| year99 | 36 | 2010 | 0 | 2010 | 2010 |
| ecage26 | 36 | 42.66667 | 3.397478 | 36 | 55 |
| ecsex99 | 36 | 1.388889 | .4944132 | 1 | 2 |
| marst26 | 36 | 1.583333 | .9964222 | 1 | 4 |
| pvreg25 | 36 | 31.22222 | 13.3419 | 10 | 59 |
| alfst28 | 36 | 1 | 0 | 1 | 1 |
| yrxfte11 | 36 | 20 | 0 | 20 | 20 |
| nocg2e6 | 36 | 2 | 0 | 2 | 2 |
| **imphwe1** | **36** | **37.17694** | **15.70046** | **12.22** | **68.13** |
| uncoll1 | 36 | 2.805556 | .5766625 | 1 | 3 |
| n07c3g10 | 36 | 9.333333 | 4.362175 | 1 | 16 |
| hleveg18 | 36 | 9.944444 | 1.755716 | 6 | 12 |
| yrschl18 | 36 | 15.5 | 2.489177 | 11 | 20 |
| yhat | 36 | 34.34989 | 0 | 34.34989 | 34.34989 |
| sef | 36 | 14.95689 | 0 | 14.95689 | 14.95689 |
| lwage | 36 | 3.513269 | .4844639 | 2.503074 | 4.221417 |
| yhat1 | 36 | 3.429154 | 0 | 3.429154 | 3.429154 |
| sef1 | 36 | .4587931 | 0 | .4587931 | .4587931 |
| yhatc | 36 | 34.27157 | 0 | 34.27157 | 34.27157 |

v) Can you comment of the goodness of fit of the model estimated in parts (n) and (r), how do they compare? **HINT** *To work with the log – linear model we need to create the Generalized* $R^2$ (see slide 4-37).

correlate imphwe1 yhatc if nocg2e6==2
(obs=1299)

| | imphwe1 | yhatc |
|---|---|---|
| imphwe1 | 1.0000 | |
| yhatc | 0.2260 | 1.0000 |

$$R_g^2 = (r_{y,\hat{y}})^2 = 0.226^2 = 0.0510$$

In both models the years of experience are significant as is the constant. The coefficient of determination in the original, linear model is 0.0577 while the generalized coefficient of determination for the log-linear model is 0.0510. While neither model predicts the variation in the hourly wage rate very well, the linear model is a slightly better model. This is not a surprising result, as other factors are important in predicting wage rates, such as education and the type of industry.

**w) SAVE YOUR DATA FOR THE NEXT ASSIGNMENT!**