The intercept is approximately 0.5.

The slope is less than one and is approximately 0.67 ($1/1.5 = 2/3 = 0.67$).

b) Compute the variance and covariance for unemployment and minimum wage.

The means of variables:

$$\overline{unemp} = \frac{5+8+7+4+4+7}{6} = \mathbf{5.833}$$

$$\overline{wage} = \frac{5+10+8+5+6+7}{6} = \mathbf{6.833}$$

The variances of variables:

$$\begin{aligned} var(unemp) &= \frac{1}{6-1}\sum_{i=1}^{6}(unemp_i - \overline{unemp})^2 \\ &= \frac{1}{5}\left\{(5-5.83)^2 + (8-5.83)^2 + (7-5.83)^2 + (4-5.83)^2 + (4-5.83)^2 + (7-5.83)^2\right\} \\ &= \frac{14.833}{5} = \mathbf{2.967} \end{aligned}$$

$$\begin{aligned} var(wage) &= \frac{1}{6-1}\sum_{i=1}^{6}(wage_i - \overline{wage})^2 \\ &= \frac{1}{5}\left\{(5-6.83)^2 + (10-6.83)^2 + (8-6.83)^2 + (5-6.83)^2 + (6-6.83)^2 + (7-6.83)^2\right\} \\ &= \frac{18.833}{5} = \mathbf{3.767} \end{aligned}$$

The covariance of variables:

$$\begin{aligned} cov(unemp, wage) &= \frac{1}{6-1}\sum_{i=1}^{6}(unemp_i - \overline{unemp})(wage_i - \overline{wage}) \\ &= \frac{1}{5}\{(5-5.83)(5-6.83) + (8-5.83)(10-6.83) + (7-5.83)(8-6.83) \\ &\quad +(4-5.83)(5-6.83) + (4-5.83)(6-6.83) + (7-5.83)(7-6.83)\} \\ &= \frac{14.833}{5} = \mathbf{2.967} \end{aligned}$$

c) Estimate the regression of a state's unemployment rate on its minimum wage and write the resulting regression equation. Interpret the coefficient on minimum wage.

The regression equation we would like to estimate is

$$unemp = \beta_0 + \beta_1 \, wage + u.$$

The estimated equation is

$$\widehat{unemp} = \hat{\beta}_0 + \hat{\beta}_1 \, \widehat{wage}$$

where

$$\hat{\beta}_1 = \frac{cov(wage, unemp)}{var(wage)}$$

and

$$\hat{\beta}_0 = unemp - \hat{\beta}_1 wage.$$

Using the results from part **b)**, we have

$$\hat{\beta}_1 = \frac{cov(wage, unemp)}{var(wage)} = \frac{2.967}{3.767} = \mathbf{0.7876}$$

$$\hat{\beta}_0 = 5\cdot833 - 0.7876 * 6.833 = \mathbf{0.4513}$$

Therefore the resulting regression equation is

$$\widehat{unemp} = 0\cdot4513 + 0\cdot7876 * wage,$$

and the interpretation would be as following:
Given all other variables constant, if the state's minimum wage increases by 1 dollar, we predict the unemployment rate increases by 0.7876 percentage points.

d) How does your estimated answer in part b) compare to your best guess in part a)?

Very close. The guess for the intercept was 0.5, and the estimated intercept is 0.4513. The guess for the slope was 0.67, and the estimated slope is 0.7876.

e) What is the predicted unemployment rate for a state with a minimum wage of 9 dollars?

If we apply 9 dollars to the estimated regression line, we have

$$\widehat{unemp}_{wage=9} = 0.4513 + 0.7876 * 9 = 7.5398$$

and therefore the predicted unemployment rate for a state with a minimum wage of 9 dollars is 7.5398%.

f) If a state were to reduce its minimum wage from 10 dollars to 7 dollars, what is the predicted change in the unemployment rate?

Since

$$\frac{d\ \widehat{unemp}}{d\ wage} = 0.7876,$$

and the minimum wage is reduced by 3 dollars, the predicted change in the unemployment rate is

$$\frac{d\ \widehat{unemp}}{d\ wage} * (-3) = 0.7876 * (-3) = \mathbf{-2.3628}$$

Another approach:

The predicted unemployment rate when the minimum wage is 10:

$$\widehat{unemp}_{wage\ =10} = 0.4513 + 0.7876 * 10 = 8.3274$$

and the predicted unemployment rate when the minimum wage is 7:

$$\widehat{unemp}_{wage\ =7} = 0.4513 + 0.7876 * 7 = 5.9646,$$

and the predicted change is -2.3628 dollars $(= 5.9646 - 8.3274)$.

g) What percent of the variation in unemployment is explained by the minimum wage (i.e. what is the R-squared)?

We can find the R-squared with the following formula

$$R^2 = 1 - \frac{SSR}{SST}$$

where

$$SSR = \sum_{i=1}^{6} (unemp_i - \widehat{unemp}_i)^2$$

$$SST = \sum_{i=1}^{6} (unemp_i - unemp)^2.$$

The predicted unemployment rate values using the estimated regression line are

State	Unemp	MinWage	Regression	Predicted Unemp
PA	5	5	0.4513 + 0.7876 * 5	4.389
CA	8	10	0.4513 + 0.7876 * 10	8.327
OR	7	8	0.4513 + 0.7876 * 8	6.752
OH	4	5	0.4513 + 0.7876 * 5	4.389
NY	4	6	0.4513 + 0.7876 * 6	5.177
WA	7	7	0.4513 + 0.7876 * 7	5.965

Then

$$\begin{aligned} SSR &= (5 - 4.389)^2 + (8 - 8.327)^2 + (7 - 6.752)^2 + (4 - 4389)^2 + (4 - 5.177)^2 + (7 - 5.965)^2 \\ &= \mathbf{3.150} \end{aligned}$$

$$\begin{aligned} SST &= (5 - 5.833)^2 + (8 - 5.833)^2 + (7 - 5.833)^2 + (4 - 5833)^2 + (4 - 5.833)^2 + (7 - 5.833)^2 \\ &= \mathbf{14.833} \end{aligned}$$

Therefore

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{3.150}{14.833} = \mathbf{0.7876},$$

and the minimum wage explains 78.76 percent of the variation in unemployment.

a) What assumption is required in order use regression through the origin when regressing unemployment on minimum wage? Do you think this is a reasonable assumption in this case? Explain.

We have to assume that the unemployment rate is zero when the minimum wage is zero. It is not a reasonable assumption because cyclical, frictional, and structural motivations for unemployment will still exist even with a zero minimum wage. There will be some people who are simply looking for jobs, some people who have the wrong skills for the jobs available, and some that lose or gain jobs due to cyclicality regardless of whether there is a minimum wage or not.

b) Compute the regression through the origin coefficients and write the resulting equation.

The regression equation we would like to estimate is now

$$unemp = \beta_1 \; wage + u,$$

and the coefficient estimate of the regression through the origin can be found by using the following formula:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{6} unemp_i \; wage_i}{\sum_{i=1}^{6} wage_i^2}.$$

Then,

$$\begin{aligned}\hat{\beta}_1 &= \frac{5 * 5 + 8 * 10 + 7 * 8 + 4 * 5 + 4 * 6 + 7 * 7}{5^2 + 10^2 + 8^2 + 5^2 + 6^2 + 7^2} \\ &= \frac{254}{299} = \mathbf{0.8495},\end{aligned}$$

and therefore the resulting equation is

$$\widehat{unemp} = 0.8495 \; wage.$$

c) What is the predicted unemployment rate for a state with a minimum wage of 9 dollars? Compare your answer to part 1 d).

- 1. d) – Regression with constant
  The predicted unemployment rate with a minimum wage of 9 dollars is

$$\widehat{unemp}_{wage=9} = 0{\cdot}4513 + 0{\cdot}7876 * 9 = \mathbf{7.5398}$$

- Regression without constant
  The predicted unemployment rate with a minimum wage of 9 dollars is

$$\widehat{unemp}_{wage=9} = 0{\cdot}8495 * 9 = \mathbf{7.6455}$$

The predicted value is higher.

3. The following regression equation examines the relationship between house prices and the number of parks in a city.

$$\widehat{houseprice} = 260{,}000 + 8{,}000 * numberparks$$

a) Interpret the coefficient on the number of parks.

Given all other conditions constant, the predicted house price is 8,000 dollars higher on average with one more park in a city.

b) What is the predicted house price in a city with 30 parks? In a city with 50 parks?

The predicted house price in a city with 30 parks is

$$\widehat{houseprice}_{30\,parks} = 260{,}000 + 8{,}000 * 30 = \mathbf{500{,}000},$$

and the predicted house price in a city with 50 parks is

$$\widehat{houseprice}_{50\,parks} = 260{,}000 + 8{,}000 * 50 = \mathbf{660{,}000}.$$

c) How many parks would there need to be in a city in order for the predicted house price to be 436,000 dollars?

If the predicted house price is 436,000 dollars,

$$\begin{aligned} & 436,000 = 260,000 + 8,000 * numberparks \\ \Rightarrow \quad & 8,000 * numberparks = 176,000 \\ \Rightarrow \quad & numberparks = \frac{176,000}{8,000} = \mathbf{22}, \end{aligned}$$

and therefore there would need to be 22 parks in a city in order for the predicted house price to be 436,000 dollars, according to the regression equation.

d) Suppose that a house in a town with 41 parks sells for 510,000 dollars. What is the error in the predicted house price?

The predicted house price is

$$\widehat{houseprice}_{41\ parks} = 260,000 + 8,000 * 41 = \mathbf{588{,}000}.$$

The error is

$$\begin{aligned} u_{41parks} &= houseprice_{41\ parks} - \widehat{houseprice}_{41\ parks} \\ &= 510,000 - 588,000 = \mathbf{-78{,}000} \end{aligned}$$

e) New York City has 1,700 parks. What is the predicted house price in NY? Do you think this is reasonable? If no, explain the shortcoming of the regression.

e) New York City has 1,700 parks. What is the predicted house price in NY? Do you think this is reasonable? If no, explain the shortcoming of the regression.

The predicted house price in NY is

$$\widehat{houseprice}_{1700parks} = 260\,000 + 8\,000 * 1,700 = \mathbf{13,860,000}\ .$$

The predicted value is not reasonable. The regression is estimated with the house price and the number of parks across the cities with different sizes of population. Since the number of parks is positively correlated with the population size on average, the predicted house price in the city with a large population (like New York City) should be biased upwardly.

4. You decide that a more reasonable regression would have the number of parks per 1,000 households (HHs). You get the following result.

$$\widehat{houseprice} = 220,000 + 480,000 * parksperthousandHHs$$

a) Interpret the coefficient on parks per thousand HHs.

Given all other conditions constant, it is predicted that the house price is 480,000 dollars higher in the city which has one more park per 1,000 households.

b) What is the predicted house price in a city with 30 parks and 20,000 households?

The number of parks per 1,000 households is

$$parksperthousandHHs = \frac{30}{\frac{20,000}{1,000}} = \frac{30}{20} = \mathbf{1.5}.$$

and according to the regression equation we have a predicted house price of

$$\widehat{houseprice}_{1.5pptH} = 220,000 + 480,000 * 1.5 = \mathbf{940,000}.$$

The predicted house price is 940,000 dollars.

c) What is the predicted house price in New York if there are 1,700 parks and 8 million people? How does this compare to your answer in part 3. e)?

The number of parks per 1,000 households in New York City is

$$parksperthousandHHs_{NYC} = \frac{1,700}{\frac{8,000,000}{1,000}} = \frac{1,700}{8,000} = \mathbf{0.2125},$$

and according to the regression equation we have a new predicted house price;

$$\widehat{houseprice}_{NYC} = 220,000 + 480,000 * 0.2125 = \mathbf{322,000}.$$

The predicted house price with the regression specified in **4** is 322,000 dollars, which is more reasonable average value for house price than the predicted value in **3. e)**.

Now you regress the natural log of house price on parks per thousand households.

$$\ln(\widehat{houseprice}) = 12{\cdot}9 + 06 * parksperthousandHHs$$

d) Interpret the coefficient on parks per thousand HHs.

Given all other conditions constant, it is predicted that the house price is 60 percent higher in the city which has one more park per 1,000 households.

e) What is the expected change in house price if parks per thousand increases from 1 to 1.5?

Since we have

$$\frac{d\ \ln(\widehat{houseprice})}{d\ parksperthousandHHs} = 0{\cdot}6$$

and the parks per 1,000 households increased by 0.5,

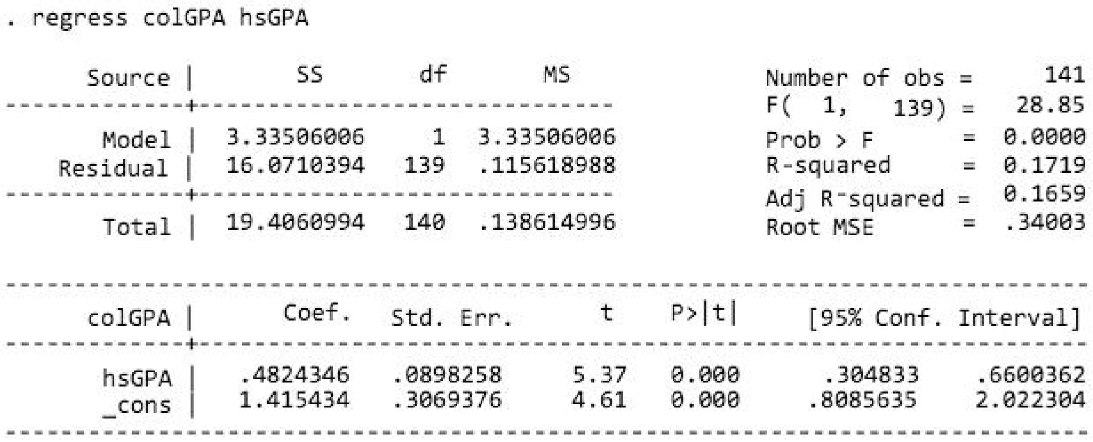
$$\frac{d\ \ln(\widehat{houseprice})}{d\ parksperthousandHHs} * 0.5 = 0{\cdot}6 * 0.5 = 0{\cdot}3$$

and therefore the house price is expected to increase by 30 percent.

5. **The Stata data set "college_gpa" has data on students college gpa, high school gpa, lectures skipped during the 30 week academic year, and other characteristics. We wish to estimate how predictive high school gpa predicts college gpa. You must submit your do-file (commands).**

   a) **Regress college GPA on high school GPA and write the estimated regression line.**

   After importing the data set, when we run a regression we have the following information on Stata:



```
. regress colGPA hsGPA

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  1,   139) =   28.85
       Model |  3.33506006     1  3.33506006           Prob > F      =  0.0000
    Residual |  16.0710394   139  .115618988           R-squared     =  0.1719
-------------+------------------------------           Adj R-squared =  0.1659
       Total |  19.4060994   140  .138614996           Root MSE      =  .34003

------------------------------------------------------------------------------
      colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsGPA |   .4824346   .0898258     5.37   0.000      .304833    .6600362
       _cons |   1.415434   .3069376     4.61   0.000     .8085635    2.022304
------------------------------------------------------------------------------
```

The slope estimate is 0.4824 and the intercept estimate is 1.415. Therefore the estimated regression line is

$$\widehat{colGPA} = 1{\cdot}415 + 0{\cdot}4824 * hsGPA.$$

b) Interpret the intercept of your regression. Interpret the coefficient on high school GPA in your regression line. Does this coefficient make sense? Explain.

Interpretation of the intercept: If a student's high school GPA is zero, her predicted college GPA is 1.415 points.

Interpretation of the slope: If one student's high school GPA is 1 point higher than another student, her GPA is expected to be 0.4824 points higher in college.

Yes, it makes sense. If we assume that high school GPA reflects the person's ability or study habits, we can expect that the person who has a higher ability or better study habits would have a higher college GPA.

c) If student A has a high school GPA that is 1.5 points higher than the high school GPA of student B, then what is the predicted difference in their college GPAs?

Since we have

$$\frac{d\ \widehat{colGPA}}{d\ hsGPA} = 0.4824$$

and the difference in high school GPAs is 1.5 points,

$$\frac{d\ \widehat{colGPA}}{d\ hsGPA} * 1.5 = 0.4824 * 1.5 = \mathbf{0.7236}$$

and therefore student A is expected to have 0.7236 points higher GPA than student B in college.

d) What percent of the variation in college GPA is predicted by high school GPA?

As reported in the regression result in part a), R^2 is 0.1719. Therefore 17.19 percent of the variation in college GPA is predicted by high school GPA.

Commands in a Do file for question 5:

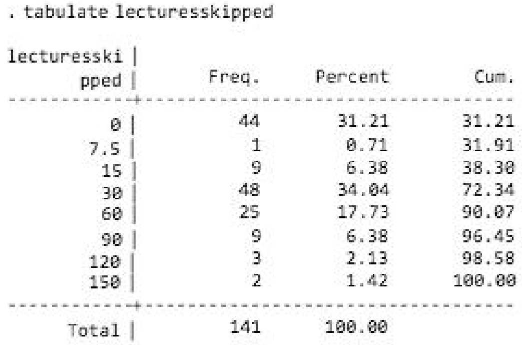
```
clear
use college_gpa.dta


* Run a regression of college GPA on high school GPA
regress colGPA hsGPA
```

6. **Using the same data, we will examine the relationship between skipping lectures and college GPA. The variable "lecturesskipped" is total lectures skipped during the academic year.**

a) **What is the modal number of lectures skipped?**

To find out the mode, we can use `tabulate` command:

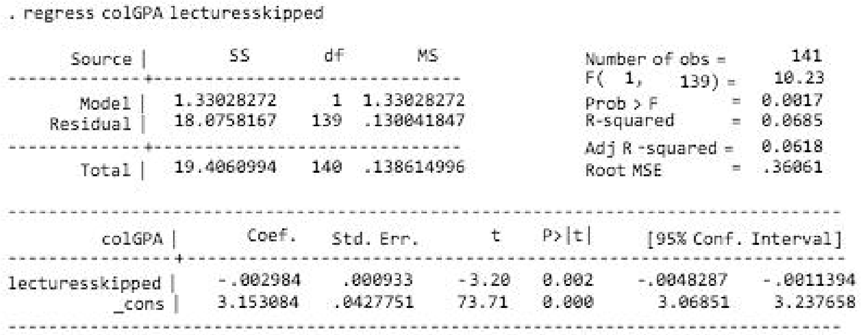


```
. tabulate lecturesskipped

lecturesski |
       pped |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         44       31.21       31.21
        7.5 |          1        0.71       31.91
         15 |          9        6.38       38.30
         30 |         48       34.04       72.34
         60 |         25       17.73       90.07
         90 |          9        6.38       96.45
        120 |          3        2.13       98.58
        150 |          2        1.42      100.00
------------+-----------------------------------
      Total |        141      100.00
```

According to the results above, the modal number of lectured skipped is 30.

b) **Regress college GPA on lectures skipped and write the regression line. Interpret the coefficient on skipped in a sentence.**



```
. regress colGPA lecturesskipped

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  1,   139) =   10.23
       Model |  1.33028272     1  1.33028272           Prob > F      =  0.0017
    Residual |  18.0758167   139  .130041847           R-squared     =  0.0685
-------------+------------------------------           Adj R-squared =  0.0618
       Total |  19.4060994   140  .138614996           Root MSE      =  .36061

------------------------------------------------------------------------------
         colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
lecturesskipped |   -.002984    .000933   -3.20   0.002    -.0048287   -.0011394
          _cons |   3.153084   .0427751   73.71   0.000      3.06851    3.237658
------------------------------------------------------------------------------
```

The estimated regression line is

$$\widehat{colGPA} = 3.153 - 0.00298 * lecturesskipped.$$

If a student skips one additional lecture, the college GPA is expected to fall by 0.00298.

c) **What is the predicted college GPA for a student who misses 20 lectures during the academic year?**

When we plug in 20 for $lecturesskipped$, we have

$$\widehat{colGPA}_{20skipped} = 3.153 - 0.00298 * 20 = \mathbf{3.0934},$$

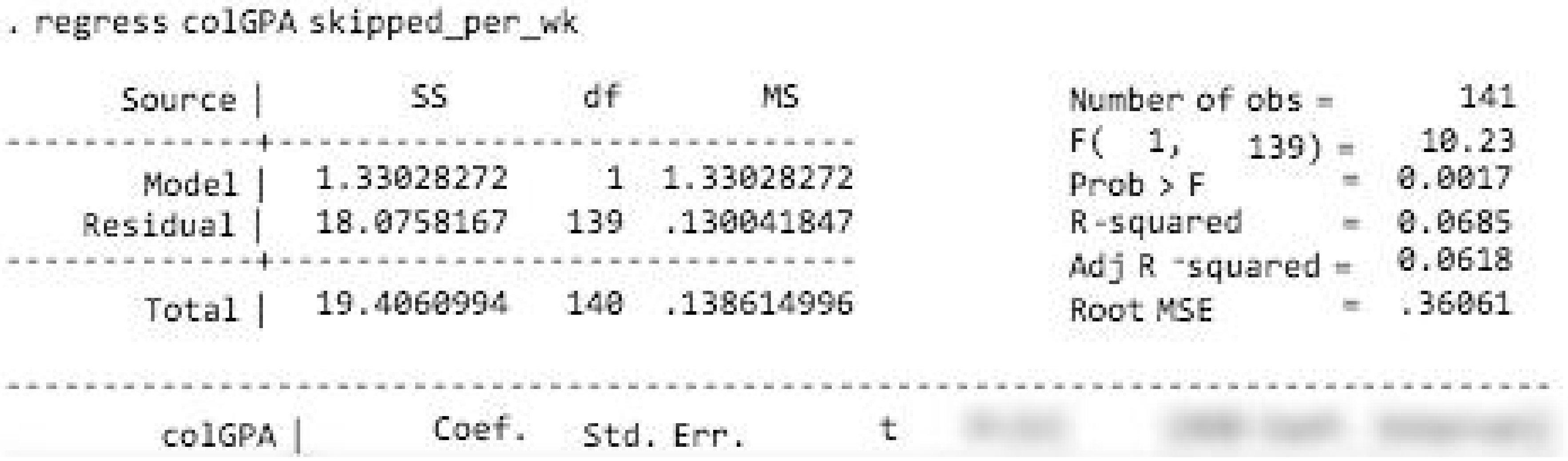
and the predicted college GPA for a student who missed 20 lectures is 3.09.

d) Make a new variable that is lectures skipped per week during the academic year, rerun the regression, and write the regression line. Interpret the coefficient on skipped in a sentence. What is the relationship between the slope computed in parts b) and d)?

We use `generate` command to make a new variable. Since there are 30 weeks in a academic year, the new variable, the number of lectures skipped per week (named `skipped_per_wk`), can be made with the following command:

```
generate skipped_per_wk = lecturesskipped / 30
```

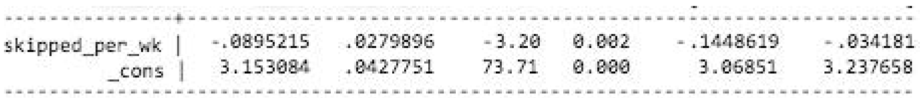
Then running a regression, we have



```
. regress colGPA skipped_per_wk

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  1,   139) =   10.23
       Model |  1.33028272     1  1.33028272           Prob > F      =  0.0017
    Residual |  18.0758167   139  .130041847           R-squared     =  0.0685
-------------+------------------------------           Adj R-squared =  0.0618
       Total |  19.4060994   140  .138614996           Root MSE      =  .36061

------------------------------------------------------------------------------
      colGPA |      Coef.   Std. Err.      t    [illegible]     [illegible] [illegible]
```



```
----------------+----------------------------------------------------------------
skipped_per_wk |  -.0895215   .0279896    -3.20   0.002    -.1448619    -.034181
         _cons |   3.153084   .0427751    73.71   0.000     3.06851     3.237658
---------------------------------------------------------------------------------
```

The estimated regression equation is

$$\widehat{colGPA} = 3.153 - 0.0895 * skipped\_per\_wk$$

The slope coefficient is -0.0895, and if a student skips one additional lecture per week, the college GPA is expected to fall by 0.0895.

Since we ran a regression on a rescaled variable (divided by 30) compared to the previous regression, the coefficient estimate is also simply rescaled:

$$\frac{1}{30}\hat{\beta}_1^* = \hat{\beta}_1$$

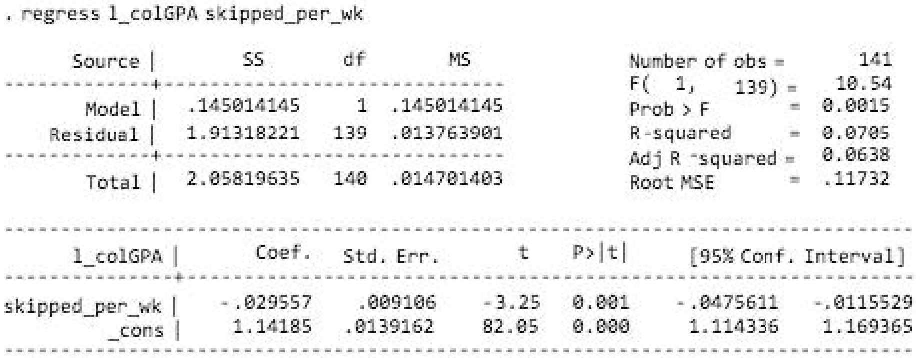
$$\Rightarrow \frac{1}{30} * (-0.0895) = -0.002984 \qquad \text{(same as the slope found in } \mathbf{b)}\text{)} \quad \square$$

e) Make a new variable that is the natural log of GPA. Regress the natural ln of GPA on lectures skipped per week and write the regression. Interpret the coefficient on skipped in a sentence.

By running the following command, we can make a new variable (named `l_colGPA`) that is the natural log of college GPA;

```
generate l_colGPA = ln(colGPA)
```

Running a regression of the new variable on lectures skipped per week variable, we have



```
. regress l_colGPA skipped_per_wk

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  1,   139) =   10.54
       Model |  .145014145     1  .145014145           Prob > F      =  0.0015
    Residual |  1.91318221   139  .013763901           R-squared     =  0.0705
-------------+------------------------------           Adj R-squared =  0.0638
       Total |  2.05819635   140  .014701403           Root MSE      =  .11732

---------------------------------------------------------------------------------
      l_colGPA |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
---------------+-----------------------------------------------------------------
skipped_per_wk |   -.029557    .009106    -3.25   0.001    -.0475611   -.0115529
         _cons |    1.14185   .0139162    82.05   0.000     1.114336    1.169365
---------------------------------------------------------------------------------
```

The estimated regression equation is

$$\ln(\widehat{colGPA}) = 1.142 - 0.0296 * skipped\_per\_wk$$

The slope coefficient is -0.0296, and if a student skips one additional lecture per week, the college GPA is expected to fall by 2.96 percent.