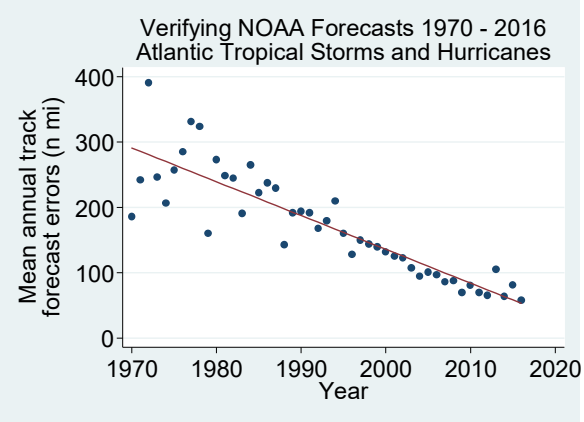
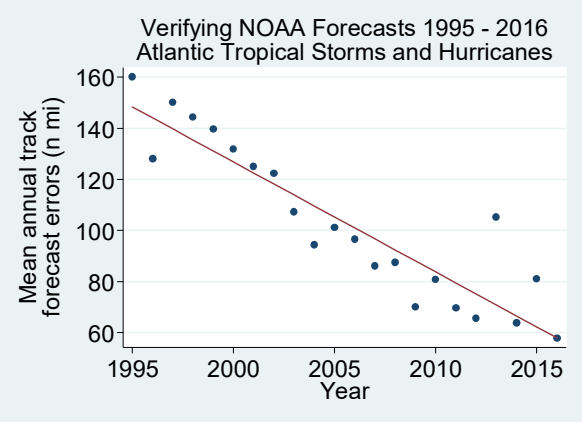
Homework 19: ECO220Y

Required Exercises: Chapter 18: 6, 8, 15, 17

Required Problems:

(1) Exercise 29 in Chapter 19 considers the accuracy of the predictions of the U.S. National Hurricane Center. The National Oceanic and Atmospheric Administration (NOAA) released updated data in April 2017: https://www.nhc.noaa.gov/verification/pdfs/1970-present_OFCL_ATL_annual_trk_errors_noTDs.pdf. For each year from 1970 through 2016 it reports the average 48-hour tracking errors measured in nautical miles (n mi) for all tropical storms and hurricanes each year. Below are scatter diagrams and regression results for the entire period of available data and for only the more recent data (1995 – 2016).






```
. regress mn_err_48hrs year; /* 1970 - 2016 */

      Source |       SS       df       MS              Number of obs =      47
-------------+------------------------------           F(  1,    45) =  160.02
       Model |   231099.21     1   231099.21           Prob > F      =  0.0000
    Residual |  64987.5034    45  1444.16674           R-squared     =  0.7805
-------------+------------------------------           Adj R-squared =  0.7756
       Total |  296086.713    46  6436.66768           Root MSE      =  38.002

------------------------------------------------------------------------------
mn_err_48hrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -5.169415   .4086494   -12.65   0.000    -5.992477   -4.346353
       _cons |    10474.7    814.457    12.86   0.000     8834.303    12115.1
------------------------------------------------------------------------------

.  regress mn_err_48hrs year if year > 1994; /* 1995 - 2016 */

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =  102.43
       Model |  16453.8348     1  16453.8348           Prob > F      =  0.0000
    Residual |  3212.74021    20  160.637011           R-squared     =  0.8366
-------------+------------------------------           Adj R-squared =  0.8285
       Total |   19666.575    21  936.503571           Root MSE      =  12.674

------------------------------------------------------------------------------
mn_err_48hrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -4.310615   .4259205   -10.12   0.000     -5.19907   -3.422161
       _cons |   8748.089   854.1878    10.24   0.000     6966.285    10529.89
------------------------------------------------------------------------------
```

(a) For each periods (i.e. 1970 – 2016 and 1995 – 2016), assess each of the six assumptions (noting any situations where you have incomplete information to fully assess). (Review Lecture 18 if you forgot the six assumptions.)

(b) Consider using the results for 1970 – 2016 to make an interval prediction about the forecast error in 2017. Which formula should you use? Would heteroscedasticity present a serious problem for that interval prediction? If so, what would it do to your interval? What about if we used the results from 1995 – 2016? Explain, referencing the relevant given regression results and scatter diagrams.

(c) Find the value of the Root MSE for each of the two regressions: 38.002 and 12.674, respectively. What are the units of measurement of these values? *Why* are they so different? Does it have to do with the smaller number of observations in the second regression?

(d) Find the value of the standard error of the slope coefficient for each of the two regressions: 0.4086494 and 0.425920, respectively. Given the huge difference in the value of the s_e in the two regressions and given that $s_b = \frac{s_e}{s_x\sqrt{n-1}}$, how can you explain the standard errors of the slope coefficients coming out to be very similar values in the two regressions? (Note: You should offer two reasons, making sure to explain.)

(e) After reviewing Section 19.4 (and Lecture 5), are these data an example of summary values? Explain. How does this affect predictions? (You may also find it useful to click the link given with the background for this question to see the data.)

(f) Using the 1995 – 2016 regression, compute the 90% prediction interval for 2017. (Note that $\bar{X}$, which is the average year from 1995 – 2016 inclusive, is $\bar{X} = \frac{1995+1996+\cdots+2016}{22} = 2005.5$ and s_x^2, which is the variance of year from 1995 – 2016 inclusive, is 42.16667.)

(g) Using the 1995 – 2016 regression, compute the 99% confidence interval estimate of the expected forecast error for 2017.

(h) Compute the 99.9% CI estimate of the OLS slope coefficient and fully interpret it.

(2) In a simple linear regression where all of the underlying assumptions hold, which parameter determines the amount of scatter about the line? Which statistic do we use to estimate this parameter?

(3) The OLS slope estimate b_1 is a sample statistic and hence is affected by sampling error. We use it as a point estimate of the unknown population parameter β_1 (i.e. the true slope if we observed x and y for every element of the population). How much is the OLS slope estimate affected by sampling error? In other words, how do we quantify the size of sampling error on the OLS slope estimate? What factors affect it and in which ways?

(4) Each part below is an actual student question (from a previous year) about Lecture 19. (These questions and answers were cut-and-paste from Piazza, which is why they have clunky equation formatting here.) Write up your best answer/explanation to each.

(a) Can you ever observe epsilon (i.e. y_i = alpha + beta*x_i + epsilon_i)? How about in Slide 4 of Lecture 19?

(b) On Slide 10 of Lecture 19, I still don't understand why the variability of the x-variable affects the standard error of the OLS slope.

(c) I still don't understand Assumption 6 and I'm sensing that it is important. Can you explain it? Can you illustrate it with the cGPA and salary example discussed in Lecture 19?

(5) In Lecture 19 we considered regressing starting salary (in $1,000's) on cGPA. Below is a summary of these two variables, a correlation matrix, and the regression results for Sample #1.

```
. summarize salary cGPA;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |        50    52.35282    11.47955   26.94069   75.77021
        cGPA |        50        2.74    .3659988       1.61       3.84

. correlate salary cGPA;

(obs=50)
             |   salary     cGPA
-------------+------------------
      salary |   1.0000
        cGPA |   0.1382   1.0000

. regress salary cGPA;

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    0.93
       Model |  123.340729     1  123.340729           Prob > F      =  0.3385
    Residual |   6333.8788    48  131.955808           R-squared     =  0.0191
-------------+------------------------------           Adj R-squared = -0.0013
       Total |  6457.21953    49   131.77999           Root MSE      =  11.487

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cGPA |   4.334865     4.4837     0.97   0.338    -4.680219    13.34995
       _cons |   40.47529   12.39228     3.27   0.002     15.55894    65.39165
------------------------------------------------------------------------------
```

(a) Find the 95% prediction interval for a person with a 3.2 cGPA.

(b) How should you interpret the answer to part (a): (30.7, 78.0). We are 95% confident that ___ a salary between $30,700 and $78,000.

(A) all people with a 3.2 cGPA have
(B) the benefit of an extra 1 point of cGPA is
(C) people with a cGPA of 3.2 on average have
(D) the mean benefit of an extra 1 point of cGPA is
(E) a randomly selected person with a cGPA of 3.2 has

(c) Find the 95% confidence interval for a 3.2 cGPA.

(d) How should you interpret the answer to part (c): (49.1, 59.6). We are 95% confident that ___ a salary between $49,100 and $59,600.

(A) all people with a 3.2 cGPA have
(B) the benefit of an extra 1 point of cGPA is
(C) people with a cGPA of 3.2 on average have
(D) the mean benefit of an extra 1 point of cGPA is
(E) a randomly selected person with a cGPA of 3.2 has

(e) How should you interpret the interval (-4.680219, 13.34995) from the STATA regression results?

(6) Returning to the starting salary cGPA example, answer these multiple-choice questions.

(a) Consider these regression results for a large sample:

Salary-hat = 42.000 + 4.000*cGPA
(6.000) (3.000)

Is the association between cGPA and salary "statistically significant"?

(A) No because the P-value will be large
(B) No because the P-value will be small
(C) Yes because the P-value will be large
(D) Yes because the P-value will be small
(E) Yes if the significance level is the conventional α = 0.05

(b) Consider these regression results for a large sample:

Salary-hat = 42.000 + 4.000*cGPA
(6.000) (0.500)

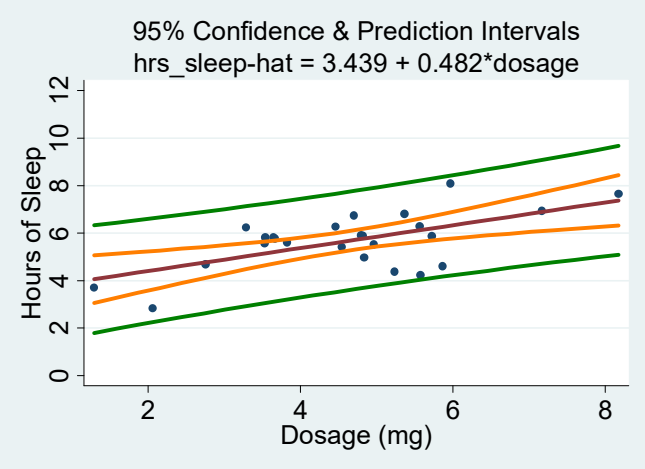
Is the association between cGPA and salary "statistically significant"?

(A) No because the P-value will be very large
(B) No because the P-value will be very small
(C) Yes because the P-value will be very large
(D) Yes because the P-value will be very small
(E) Yes if the significance level is α = 0.10 but not at the conventional α = 0.05 level

(7) Recall the experimental drug trial example (sleeping aid) discussed in many times (e.g. Lectures 4 & 5 and HW 18). Here are the STATA regression results for that example as well a prediction and confidence intervals.

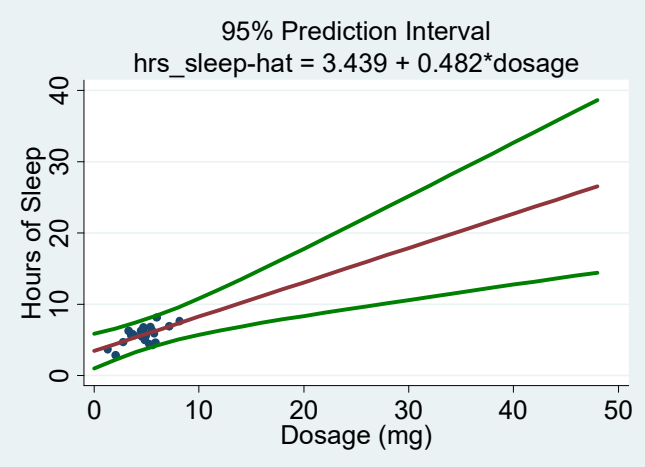
```
. regress hrs_sleep dosage;

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =   13.89
       Model |  12.6255781     1  12.6255781           Prob > F      =  0.0011
    Residual |  20.9040126    23  .908870111           R-squared     =  0.3766
-------------+------------------------------           Adj R-squared =  0.3494
       Total |  33.5295906    24  1.39706628           Root MSE      =  .95335
------------------------------------------------------------------------------
   hrs_sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dosage |   .4816382   .1292249     3.73   0.001     .2143161    .7489602
       _cons |   3.439461   .6260549     5.49   0.000     2.144368    4.734555
------------------------------------------------------------------------------
```

dosage (mg)	95% Prediction Interval (hours)	95% Confidence Interval (hours)
2	(2.27, 6.53)	(3.60, 5.21)
4	(3.35, 7.38)	(4.94, 5.79)
6	(4.28, 8.37)	(5.79, 6.87)
8	(5.09, 9.50)	(6.31, 8.28)



(a) Interpret the results from the 6 mg dosage row such that doctors and patients can understand you.

(b) Consider this graph of the prediction interval. Do the wider bounds from the formula fully account for the uncertainties of an out-of-sample prediction?



(c) Answer this multiple-choice question.

Of the 25 observations, only 48% (12/25) are inside 95% CI. Why?

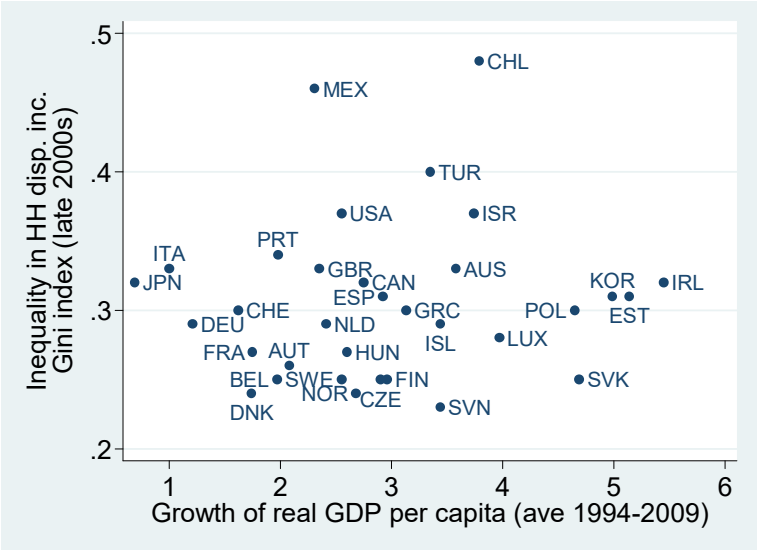
(A) Because the CI refers to the mean at each dosage not to an individual person
(B) A CI is seldom narrower than the scattering of the data: it has to reflect sampling error
(C) On average about 95% would be in that interval but because of sampling error fewer happen to be

(d) Answer this multiple-choice question.

Of the 25 observations, 100% (25/25) are inside 95% PI. Why?

(A) Because the PI refers to an individual person at each dosage not to the mean
(B) On average 95% would be in that interval but because of sampling error more happen to be
(C) Because the PI must reflect the variability of the data and hence the observations in the data cannot lie outside the PI

(8) Recall the six assumptions discussed in Lecture 18 and recall the growth and inequality example from HW 18. Below the scatter diagram is reproduced.



(a) Answer this multiple-choice question.

In estimating $inequality_j = \alpha + \beta GDP_{growth_j} + \varepsilon_j$ does Assumption #6 hold?

(A) Yes

(B) No because $inequality_j$ is correlated with ε_j

(C) No because GDP_{growth_j} is correlated with $inequality_j$

(D) No because GDP_{growth_j} is correlated with other factors that also affect $inequality_j$

(E) No because $inequality_j$ is determined by many other factors that have nothing to do with GDP_{growth_j}

(b) How should you interpret the OLS results:

inequality-hat = 0.2944 + 0.0041*GDP_growth
(0.0278) (0.0088)

```
. regress inequality GDP_growth;

      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  1,    31) =    0.21
       Model |  .000761377     1  .000761377           Prob > F      =  0.6467
    Residual |  .110202263    31  .003554912           R-squared     =  0.0069
-------------+------------------------------           Adj R-squared = -0.0252
       Total |   .11096364    32  .003467614           Root MSE      =  .05962

------------------------------------------------------------------------------
  inequality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  GDP_growth |    .004094   .0088464     0.46   0.647    -.0139483    .0221363
       _cons |   .2944066   .0278435    10.57   0.000     .2376193    .3511939
------------------------------------------------------------------------------
```

(9) PCT_ATT measures the percent of classes that students attended for a course. MARK measures the percent grade students earn. Consider the following STATA data summaries and regression results.

```
                          PCT_ATT
-------------------------------------------------------------
      Percentiles      Smallest
 1%           30             30
 5%           45             40
10%           50             45       Obs                  55
25%           60             50       Sum of Wgt.          55

50%           70                      Mean            70.18182
                        Largest       Std. Dev.       14.93702
75%           80             90
90%           85             90       Variance        223.1145
95%           90             95       Skewness       -.3929561
99%          100            100       Kurtosis        2.641783

                           MARK
-------------------------------------------------------------
      Percentiles      Smallest
 1%           42             42
 5%           48             44
10%           49             48       Obs                  55
25%           60             49       Sum of Wgt.          55

50%           68                      Mean                 69
                        Largest       Std. Dev.       13.19652
75%           81             89
90%           85             90       Variance        174.1481
95%           90             93       Skewness       -.069775
99%           95             95       Kurtosis        2.15904


      Source |       SS       df       MS              Number of obs =      55
-------------+------------------------------           F(  1,    53) =   82.21
       Model |   5717.8752     1   5717.8752           Prob > F      =  0.0000
    Residual |   3686.1248    53  69.5495246           R-squared     =  0.6080
-------------+------------------------------           Adj R-squared =  0.6006
       Total |        9404    54  174.148148           Root MSE      =  8.3396

------------------------------------------------------------------------------
        MARK |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     PCT_ATT |   .6889006   .0759777     9.07   0.000     .5365086    .8412927
       _cons |    20.6517    5.44954     3.79   0.000     9.721309    31.58209
------------------------------------------------------------------------------
```

(a) What is the interpretation of the slope estimate: 0.6889?

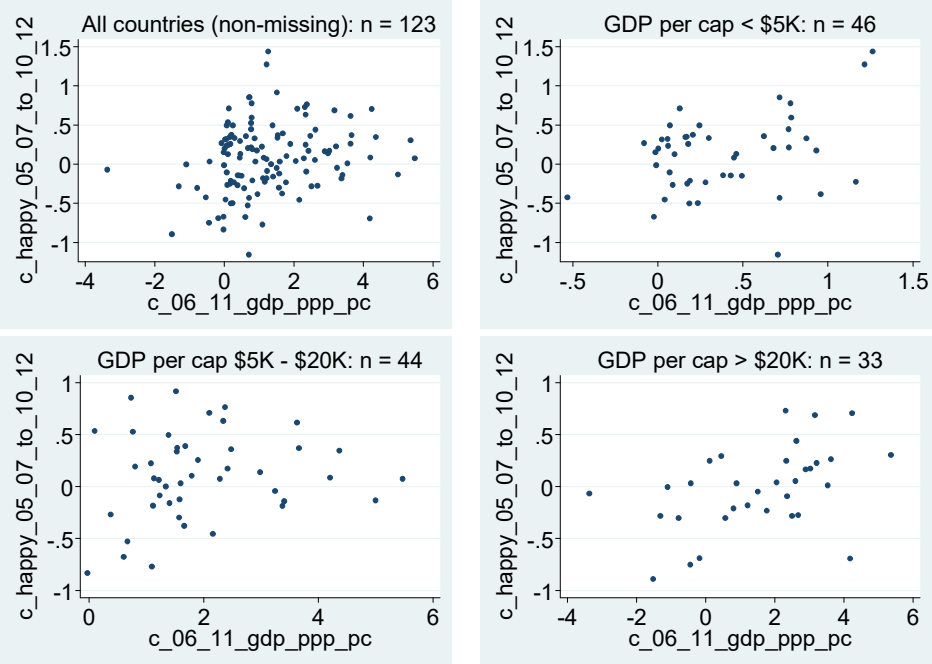
(b) What is the interpretation of the intercept estimate: 20.6517?

(c) Verify that the Least Squares Line passes through the mean. (This means that the regression line actually goes though the point given by $(\bar{X}, \bar{Y})$.)

(d) Suppose that one student did not show up to the exam without a proper excuse and got a 0 mark in the course. Would this observation be an outlier? Should it be included in the regression analysis? Under what conditions would this outlier be an influential data point in terms of the slope estimate? Under what conditions would this outlier not be an influential data point in terms of the slope estimate?

(10) Recall the data from the 2012 and 2013 World Happiness Reports discussed Lecture 19 (slides 17 – 19). The table below summarizes some of the variables and includes two new variables not discussed in Lecture 19. Scatter diagrams explore some relationships for all data and then break things down by very poor countries, developing countries, and rich countries. Finally, it shows the STATA regression results for rich countries.

Variable name, variable description
mean_happy_10_12, Mean reply in a country to "Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" asked in 2010, 2011, and 2012 Note: "We average the three most recent years (2010–12). In looking for possible trends, we compare these most recent three years with average values in the earliest years (2005–07) of data available for each country" p. 9 of *The World Happiness Report 2013*.
real_gdp_ppp_2011_pc, Real GDP per capita at current PPP (purchasing power parity) in 2011 $1,000s US
c_happy_05_07_to_10_12, Difference (change) in mean happiness from 2005 – 2007 to 2010 – 2012 (= mean_happy_10_12 – mean_happy_05_07)
c_06_11_gdp_ppp_pc, Difference (change) in GDP per capita from 2006 to 2011 (= real_gdp_ppp_2011_pc – real_gdp_ppp_2006_pc)




```
. regress c_happy_05_07_to_10_12 c_06_11_gdp_ppp_pc if real_gdp_ppp_2011_pc>20

      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  1,    31) =    8.14
       Model |  1.04904956     1  1.04904956            Prob > F      =  0.0076
    Residual |  3.99406013    31  .128840649           R-squared     =  0.2080
-------------+------------------------------           Adj R-squared =  0.1825
       Total |   5.0431097    32  .157597178           Root MSE      =  .35894

------------------------------------------------------------------------------
c_happy_0~12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
c_06_11_gd~c |    .092246   .0323278     2.85   0.008     .0263131     .158179
       _cons |  -.1626193   .0800016    -2.03   0.051    -.3257837    .0005451
------------------------------------------------------------------------------
```

(a) What does 0.157597178 mean? Units? How should you interpret it?

(b) What does 0.35894 mean? Units? How should you interpret it?

(c) What does 0.092246 mean? How should you interpret it?

(d) What do 0.0323278 mean? 2.85? 0.008? How should you interpret them?

(e) What does -0.1626193 mean? How should you interpret it?

(f) What does 0.2080 mean? How should you interpret it?

(g) How does looking at the change in happiness versus the change in GDP *partially* address the violation of Assumption #6 (i.e. no lurking variables that affect *both* the y-variable and the x-variable)? What are some realistic and reasonable examples of lurking variables that would likely be taken care of by looking at changes as opposed to levels (in Lecture 18 we looked at the level of happiness versus the level of GDP)? What are some realistic and reasonable examples of variables that would likely still be lurking variables even if we look at changes instead of levels?

(11) Review the excerpt below from the "What Can Go Wrong?" section at the end of Chapter 18 (page 623). Make sure to notice the part that says "the evidence is in favor of the one-tailed test" before trying the multiple choice question.

> - **WATCH OUT FOR ONE-TAILED TESTS**. Because tests of hypotheses about regression coefficients are usually two-tailed, software packages report two-tailed P-values. If you're using that type of software and the evidence is in favour of the one-tailed test, you'll need to divide the reported P-value by two.

If a one-tailed t test gives a P-value of 0.90, what is the P-value for the two-tailed test? (Hint: Sketch the Student t distribution.)

(A) 0.05
(B) 0.10
(C) 0.20
(D) 0.45
(E) 1.80