Homework Solutions
Applied Regression Analysis

Exercise Three
Comparing Blood Pressure with Smoking History

1. **Determine the least-squares estimates of slope (β_1) and intercept (β_0) for the straight-line regression of SBP *(Y)* on SMK *(X)*.**

   We can determine the least squares estimates for the parameters in simple linear regression by regressing Y on X. In the command window, enter '.regress sbp smk'. This will produce the output below.

```
. regress sbp smk

  Source |       SS       df       MS                  Number of obs =      32
---------+------------------------------                F(  1,    30) =    1.95
   Model |  393.098162     1  393.098162                Prob > F      =  0.1723
Residual |  6032.87059    30  201.095686                R-squared     =  0.0612
---------+------------------------------                Adj R-squared =  0.0299
   Total |  6425.96875    31  207.289315                Root MSE      =  14.181

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     smk |   7.023529   5.023498      1.398   0.172      -3.235823    17.28288
   _cons |      140.8   3.661472     38.454   0.000       133.3223    148.2777
```

$$y = \beta_0 + \beta_1 x$$

$$= 140.8 + 7.02(\text{smk})$$

Note: When entering data into STATA, always list the dependant variable first (sbp, in this case) and then the independent variable (smk, in this case).

The slope for smk is determined by the value listed in the "Coef." column of the table above, and the value for the intercept is determined in a similar fashion in the _cons row.

How can you interpret the slope?

"Under this model, current or previous smokers have an average systolic blood pressure 7.02 mm Hg higher than that of non-smokers"

2. **Compare the value of $\hat{\beta}_0$ with the mean SBP for nonsmokers. Compare the value of $\hat{\beta}_0 + \hat{\beta}_1$ with the mean SBP for smokers. Explain the results of these comparisons.**

   To compare the mean of one variable across different categories of another variable in STATA, you must first sort the data by the second categorizing variable (in this case, smk).

   In the command window, enter '.sort smk'.

   You much then use the 'sum' command to get descriptive statistics on your variable of interest (in this case, sbp), but first you must use the 'by' command to split the results by smoking status.

   In the command window, enter '.by smk:sum sbp'.

   This will produce the output below.

```
. sort smk
. by smk:sum sbp

-------------------------------------------------------------------------------
-> smk = 0

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |      15       140.8    12.90183       120        164

-------------------------------------------------------------------------------
-> smk = 1

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |      17    147.8235    15.21198       126        180
```

The mean value of SBP for nonsmokers (140.8) is equal to the value of $\hat{\beta}_0$. The mean value of SBP for smokers is 147.82 which is equal to $\hat{\beta}_0 + \hat{\beta}_1$.

In simple linear regression, the intercept can be interpreted as the value for Y when X=0. Given that smoking is a binary variable, and is coded (0,1) (0 for non-smokers, 1 for smokers), then the intercept is the mean value for non-smokers (Y when X=0), and the intercept plus the slope is the mean value for smokers (Y when X=1).

3. **Test the hypothesis that the true slope (β_1) is 0.**

$H_o : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

The null hypothesis cannot be rejected, $p = 0.172$. There is not sufficient evidence to conclude that the slope is significantly different from 0.

Note: You can test for the significance of the slope by looking at the p-value for the t-test in the table for the regression in problem 1. The p-value tells us that the probability of rejecting the null when the null is true is 17.2%, which exceeds 5%. Therefore there is insufficient evidence to reject the null.

4. **Is the test in part (3) equivalent to the usual two-sample *t* test for the equality of two population means assuming equal but unknown variances? Demonstrate your answer numerically.**

   To perform a t-test to compare the mean sbp across the populations in the different smoking categories, enter '.ttest sbp, by(smk)' into the command window.

   This will produce the output below.

```
. ttest sbp, by(smk)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      15       140.8    3.331237    12.90183    133.6552    147.9448
       1 |      17    147.8235    3.689448    15.21198    140.0022    155.6448
---------+--------------------------------------------------------------------
combined |      32    144.5313    2.545151    14.39755    139.3404    149.7221
---------+--------------------------------------------------------------------
    diff |           -7.023529    5.023498               -17.28288    3.235823
------------------------------------------------------------------------------
Degrees of freedom: 30

                    Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -1.3981                 t =  -1.3981                  t =  -1.3981
   P < t =   0.0862           P > |t| =   0.1723           P > t =   0.9138
```

The t-test gives the same t-value and p-value as the test for the hypothesis that the true slope, β_1, is 0.

The p-value for question 3 is the same at the two-sided p-value for the two-sample t-test. In both tests, you are testing whether smoking has a significant impact on systolic blood pressure by determining if the sbp for smokers is significantly different than that of non-smokers.