Exercise 2.1

Information on 74 automobiles was collected (in 1978) to study the relationship between gas mileage (mpg) and various features of the cars. We would like to investigate the relationship between mileage and whether the car is made in the U.S. (`foreign`: 0=made in U.S.; 1=made outside U.S.).

A simple linear regression model is fit, with foreign as the explanatory variable:

$$E[MPG] = \beta_0 + \beta_1 FOREIGN$$

```
. regress mpg foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     13.18
       Model |  378.153515         1  378.153515   Prob > F        =    0.0005
    Residual |  2065.30594        72  28.6848048   R-squared       =    0.1548
-------------+----------------------------------   Adj R-squared   =    0.1430
       Total |  2443.45946        73  33.4720474   Root MSE        =    5.3558


------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   4.945804   1.362162     3.63   0.001     2.230384    7.661225
       _cons |   19.82692   .7427186    26.70   0.000     18.34634    21.30751
------------------------------------------------------------------------------
```

(a) Identify the value of $\hat{\beta}_1$ and interpret this value.

$\hat{\beta}_1 = 4.95.$

The estimated difference in mean mileage, foreign cars minus U.S.-made cars, is 4.95 MPG. (Mileage is higher for foreign cars.)

(b) Identify the value of $\hat{\beta}_0$ and interpret this value.

$\hat{\beta}_0 = 19.8$
The estimated mean mileage for U.S.-made cars is 19.8 miles per gallon.

(c) Is there a significant difference in mileage between foreign and domestic (not foreign) cars? Cite specific evidence from the Stata output in your answer.

Yes there is a significant difference in mean mileage, p-value $= 0.001$

(d) Suppose I am concerned about the normality assumption. I create a histogram of the mileage values and check to see if it is skewed. Is this an appropriate way to check the normality assumption? If not, describe a plot that could be used to check this assumption.

It would **not** *be appropriate to just look at a histogram of the mileage values themselves – the normality assumption says the* **errors** *are normally distributed. Thus we would need to create either a histogram or a Q-Q plot of the residuals in order to check this assumption.*

Exercise 2.2

A study compared the growth rate of 16 male and 16 female chicks. Growth, as measured by increase in weight in grams, was measured at day 7. Then a linear regression model was performed, using sex to predict weight gain. In the data set, the variable `male` takes the values 1 for male chicks and 0 for female chicks. Stata output is below.

```
. regress wtgain male

      Source |       SS           df       MS      Number of obs   =        32
-------------+----------------------------------   F(1, 30)        =      0.09
       Model |  3.78125262         1  3.78125262   Prob > F        =    0.7717
    Residual |  1323.27866        30  44.1092886   R-squared       =    0.0028
-------------+----------------------------------   Adj R-squared   =   -0.0304
       Total |  1327.05991        31  42.8083842   Root MSE        =    6.6415

------------------------------------------------------------------------------
      wtgain |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .6875002   2.348119     0.29   0.772
       _cons |   27.15625   1.660371    16.36   0.000     23.76532    30.54718
------------------------------------------------------------------------------
```

(a) Write the population regression line being estimated here (i.e., use βs not $\hat{\beta}$s).

$E(WTGAIN) = \beta_0 + \beta_1 MALE$

(b) In terms of the βs, what is the expected mean weight of female chicks?

β_0

(c) In terms of the βs, what is the expected mean weight of male chicks?

$\beta_0 + \beta_1$

(d) What are the null and alternative hypotheses to test whether there is a significant difference in weight gain between male and female chicks?

$H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$

(e) For the test of the hypotheses you stated in (d), report the test statistic value, the distribution of the test statistic under the null hypothesis, and the p-value.

$t = 0.29$*; Under* H_0, $t \sim t_{30}$*; p-value* $= 0.772$

(f) Write a one-sentence conclusion (in the context of the problem).

There is not a significant difference in mean weight gain between male and female chicks (p=0.77).

Exercise 2.3

A small study collected systolic blood pressure (SBP) from 32 men, along with several predictors of SBP. One predictor was smoking status, recorded as 1=smoker, 0=non-smoker. A two-sample t-test was performed to compare the mean SBP between smokers and non-smokers. Stata output from the t-test is 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
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.3981
Ho: diff = 0                                     degrees of freedom =       30

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0862         Pr(|T| > |t|) = 0.1723          Pr(T > t) = 0.9138
```

(a) Is there evidence of a significant difference in mean SBP for the two groups?

No – p-value = 0.1723, there is no evidence of a significant difference in means.

(b) A linear regression was also performed using the smoking status variable to predict SBP. The output is below – but part of the output is missing (missing values labeled with letters A-E). Fill in the missing values, using the t-test output if necessary.

```
. regress sbp smk

      Source |       SS           df       MS      Number of obs   =        32
-------------+----------------------------------   F(1, 30)        =      1.95
       Model |  393.098162         1  393.098162   Prob > F        =    __(D)_
    Residual |  6032.87059        30  201.095686   R-squared       =    __(E)_
-------------+----------------------------------   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 |   ___(A)__   5.023498   __(B)_  __(C)_    -3.235823    17.28288
       _cons |      140.8   3.661472    38.45   0.000     133.3223    148.2777
------------------------------------------------------------------------------
```

(A) 7.02 *— this is the estimated difference in mean SBP, smokers minus non-smokers. From the t-test output, the mean difference for non-smokers minus smokers is* -7.02*, so we just reverse the sign*
(B) 1.398 *— test statistic for the "slope" is the same as for the t-test, but with the sign reversed*
(C) 0.1723 *— p-value for the test of the "slope" is the same as the two-sided t-test p-value*
(D) 0.1723
The p-value for the overall F-test is the same as the p-value for the test of the "slope" when there is only one predictor variable.
(E) $R^2 = \frac{MSS}{TSS} = \frac{393.098162}{6425.96875} = 0.0612$