Homework Solutions
Applied Regression Analysis

WEEK 5

Exercise Two

Complete the following six questions:

1. Generate the separate straight-line regressions of Y on X_1 (model 1) and Y on X_2 (model 2). Which of the two independent variables would you say is the more important predictor of Y? *Discuss your response in the homework forum.*

We will consider the two simple linear regression models separately. Type 'regress choles weight' in the command window.

Model 1

```
. regress choles weight

  Source |       SS       df       MS              Number of obs =      25
---------+------------------------------           F(  1,    23) =    1.74
   Model |  10231.7262     1  10231.7262           Prob > F      =  0.2000
Residual |  135145.314    23  5875.88321           R-squared     =  0.0704
---------+------------------------------           Adj R-squared =  0.0300
```

```
   Total |   145377.04    24  6057.37667           Root MSE      =  76.654

------------------------------------------------------------------------------
  choles |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  weight |   1.622343   1.229433      1.320   0.200      -.9209323    4.165618
   _cons |   199.2975   85.81792      2.322   0.029       21.76962    376.8254
------------------------------------------------------------------------------
```

Type 'regress choles age' in the command window.

Model 2

```
. regress choles age

  Source |       SS       df       MS              Number of obs =      25
---------+------------------------------           F(  1,    23) =   53.96
   Model |  101932.666     1  101932.666           Prob > F      =  0.0000
Residual |  43444.3743    23  1888.88584           R-squared     =  0.7012
---------+------------------------------           Adj R-squared =  0.6882
   Total |   145377.04    24  6057.37667           Root MSE      =  43.461

------------------------------------------------------------------------------
  choles |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   5.320676   .7242909      7.346   0.000       3.822367    6.818986
   _cons |   102.5751   29.63757      3.461   0.002       41.26516    163.8851
------------------------------------------------------------------------------
```

Age is a more important predictor. See the R^2 and the F test.

While making your decision regarding the independent variable, you need to take into consideration the p-value and the R^2 value. *Please discuss your response in the homework forum.*

2. Generate the regression model of *Y* on both X_1 and X_2.

We will now consider a multiple linear regression model. Type ‘regress choles weight age’ in the command window. From the output, you can obtain the coefficient for β_1 and β_2 as well as the intercept (β_0) in the bottom right corner of the output in the “Coef.” column.

```
. regress choles weight age

  Source |       SS       df       MS              Number of obs =      25
---------+------------------------------           F(  2,     22) =   26.36
   Model |  102570.815     2  51285.4073           Prob > F      =  0.0000
Residual |  42806.2253    22  1945.73752           R-squared     =  0.7056
---------+------------------------------           Adj R-squared =  0.6788
   Total |   145377.04    24  6057.37667           Root MSE      =  44.111

------------------------------------------------------------------------------
  choles |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  weight |   .4173621   .7287761     0.573   0.573      -1.094027    1.928751
     age |   5.216591   .7572445     6.889   0.000       3.646162     6.78702
   _cons |   77.98254   52.42964     1.487   0.151      -30.74988     186.715
------------------------------------------------------------------------------
```

3. For each of the models in questions 1 and 2, determine the predicted cholesterol level *(Y)* for patient 4 (with *Y* =263, X_1= 70, and X_2 = 30) and compare these predicted cholesterol levels with the observed value. *Comment on your findings in the homework forum.*

There are three models in total. Two simple linear regression models from question 1 and the multiple linear regression model that we just fit. We have the intercept and slope coefficients from the outputs. Recall that in order to obtain the predicted value we just substitute the value of the predictor variables in the regression equation.

Please discuss your response in the homework forum.

$Y=263$, $X_1 = 70$ and $X_2 = 30$

Model 1: y=199.298+1.62234WEIGT

=199.298+1.62234(70)

=312.8618

Model 2: y=102.575+5.32068AGE

=102.575+5.32068(30)

=262.1954

Model 3: y=77.9825+5.21659AGE+0.41736WEIGHT

=77.9825+5.21659(30)+0.41736(70)

=263.695

Models 2 and 3 yield predictions very close to the observed cholesterol value of 263 while Model 1 provides a very poor prediction. Model 3 is the closest to the observed value.

4. Carry out the overall *F* test for the two-variable model and the partial *F* test for the addition of X_1 to the model, given that X_2 is already in the model.

For this question we will consider the multiple linear regression model. To carry out the overall F test, we will test if the null hypothesis that β_1 and β_2 are simultaneously equal to zero. The F statistic and the p-value for the same can be obtained from the RHS of the output from question 2.

Overall F-Test

H_0: $\beta_{x1} = \beta_{x2} = 0$

H_A: At least one of the β's≠0

F=26.36, p-value<0.001 (from computer output)

Reject the null hypothesis. There is significant overall regression.

Partial F-Test for the addition of X_1 given that X_2 is already in the model

H_0: The addition of X_1 (Weight) to the model does not significantly improve the prediction of Cholesterol over and above that achieved by the model containing X_2 (Age).

H_A: The addition of X_1 adds to the prediction of Cholesterol

$$F(x_1 \mid x_2) = \frac{SS_{reg}(x_1, x_2) - SS_{reg}(x_2)}{MS_{residual}(x_1, x_2)} = \frac{102571 - 101933}{1945.74} = 0.3279, \text{ with 1,22 d.f.}$$

∴ Not Significant, Fail to reject the null hypothesis. The addition of X_1 to the model already containing X_2 does not add to the prediction of cholesterol.

5. Compute and compare the R^2-values for each of the three models considered in questions 1 and 2.

This question is similar to question B of Problem 1. We will make use of the three outputs of the regression models that we had obtained earlier. Make sure that you match the correct output to the models. The value of R^2 for each of the model can be obtained from the fourth line of the right hand side (RHS) of the outputs.

Model	R^2
1: WEIGHT	0.0704
2: AGE	0.7012
3: WEIGHT and AGE	0.7056

6. While concluding that a particular model is the best amongst the ones that you have fit, make sure you take into consideration the R^2 value and the corresponding p-value of the model. *Please discuss your response in the homework forum.*