Homework Solutions
Applied Regression Analysis

WEEK 2

Exercise Four

1. **Determine the least-squares estimates of slope and intercept for the straight-line regression of SBP *(Y)* on QUET *(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 quet'.

   This will produce the output below.

```
. regress sbp quet

  Source |       SS       df       MS              Number of obs =      32
---------+------------------------------           F(  1,    30) =   36.75
   Model |  3537.94585     1  3537.94585           Prob > F      =  0.0000
Residual |   2888.0229    30  96.2674299           R-squared     =  0.5506
---------+------------------------------           Adj R-squared =  0.5356
   Total |  6425.96875    31  207.289315           Root MSE      =  9.8116

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    quet |   21.49167   3.545147      6.062   0.000       14.25151    28.73182
   _cons |   70.57641   12.32187      5.728   0.000        45.4118    95.74102
------------------------------------------------------------------------------
```

$$\hat{\beta}_0 = 70.576$$
$$\hat{\beta}_1 = 21.492$$

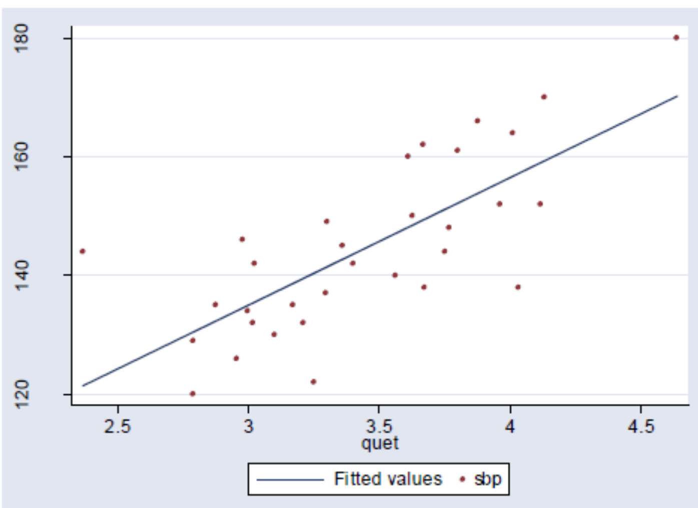
2. **Sketch the estimated regression line on the scatter diagram involving SBP and QUET.**

   In order to fit a regression line in STATA, you must first create a new variable in your dataset for of the predicted y given x under the regression model.

   You can do this simply by entering 'predict yhat' into the command window. Next, create a scatterplot with a line by entering 'scatter yhat sbp quet, c(1 .) s(i o)' into the command window.

   The commands 'c(1 .) s(i o)' specify that the yhat should be labeled with a line and data points with dots, respectively.

```
. predict yhat
(option xb assumed; fitted values)

. scatter yhat sbp quet, c(l .) s(i o)
```



3. **Test the hypothesis of zero slope.**

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

Reject the null hypothesis, $p < 0.001$. There is 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 less than 5%. Therefore there is sufficient evidence to reject the null.

4. Find a 95% confidence interval for $\mu_{y|\bar{x}}$.

To calculate confidence intervals, you need to know the descriptive statistics for the variables, including their mean values and standard deviations.

To get these values, use the 'sum' command by entering '.sum sbp quet age smk' into the command window.

Next we can calculate $\mu_{y|\bar{x}}$ by entering the mean value for quet within the regression equation using our previously estimated parameters. The confidence limits about $\mu_{y|\bar{x}}$ can then be estimated using the mean value and standard deviation of x.

```
. sum sbp quet age smk

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |        32    144.5313    14.39755        120        180
        quet |        32    3.441094    .4970781      2.368      4.637
         age |        32       53.25    6.956083         41         65
         smk |        32      .53125    .5070073          0          1
```

$$\hat{y}_{\bar{x}} = 70.57641 + 21.49167 * 3.44 = 144.508$$

$$s_{\hat{y}_{x_0}}^2 = s_{y|x}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)$$

$$s_{\hat{y}_{\bar{x}}}^2 = s_{y|\bar{x}}^2\left(\frac{1}{n}\right) = \frac{96.2674299}{32} = 3.008357$$

$$s_{\hat{y}_{\bar{x}}} = \sqrt{3.008357} = 1.7344616$$

95% CI: $\hat{y}_{\bar{x}} \pm t_{.975}(30)s_{\hat{y}_{\bar{x}}} = 144.508 \pm 2.042 \times 1.7344616 = (140.97, 148.05)$

Interpretation: We are 95% confident that the true value for the mean value of y is between 140.97 and 148.05 mm Hg.

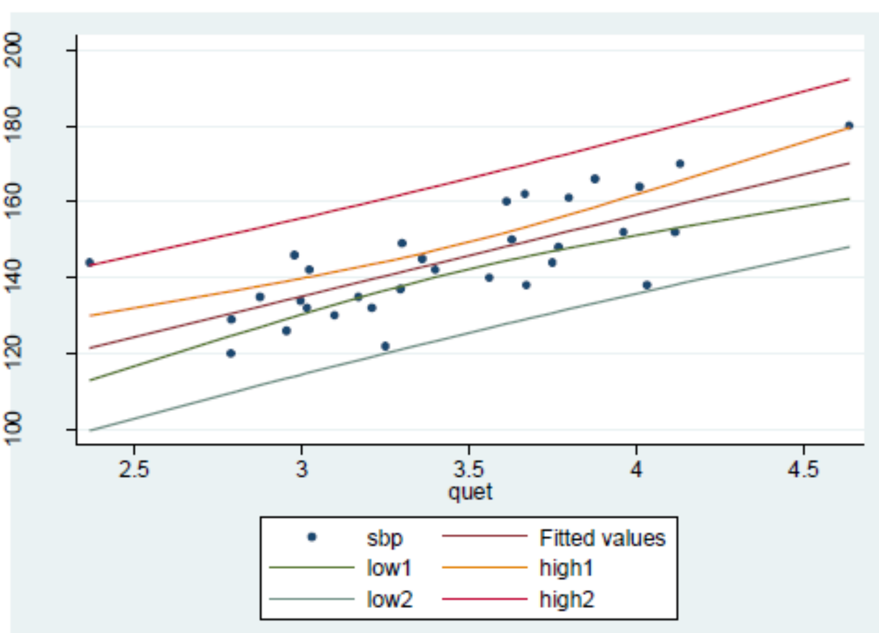
5. Calculate 95% prediction bands.

For this problem, we are asking for a plot of the prediction bands using STATA, not for hand-calculations. To do this, we must enter 'predict sepred, stdf' into the command window to generate a variable- 'sepred'- for the standard deviation used within the prediction interval.

Next, we can calculate the value for the lower limit of the prediction interval by entering 'generate low=yhat-invttail(30,0.025)*sepred' and the upper limit of the prediction interval by entering 'generate high=yhat+invttail(30,0.025)*sepred' (note: invttail(30,0.025)= $t_{0.975}(30)$).

From here you can create a plot of the prediction intervals with the regression line by entering 'scatter sbp yhat low high, sort connect (. 1 1 1) symbol (o i i i)'. The code and plot below includes both the confidence and prediction intervals, however you only need to graph the prediction intervals for this question.

```
. predict yhat
(option xb assumed; fitted values)

. predict seyhat, stdp

. display invttail(30,0.025)
2.0422724

. generate low1= yhat-2.0422724* seyhat

. generate high1= yhat+2.0422724* seyhat

. predict sepred, stdf

. generate low2= yhat-invttail(30,0.025)* sepred

. generate high2= yhat+invttail(30,0.025)* sepred

. scatter sbp yhat low1 high1 low2 high2 quet,sort connect(. 1 1 1 1 1)
symbol(o i i i i i)
```



6. **Based on the above, would you conclude that blood pressure increases as body size increases?**

   Yes, because the fitted regression line, as well at the confidence and prediction band, appear to have an upward slope.

7. **Are any of the assumptions for straight-line regression clearly not satisfied in this example?**

Simple Linear Regression Assumptions:

Linearity: SBP and SMK appear to be linearly related based on the above scatterplot
Independence: The study design does not suggest that the observations are not independent
Normality: The variables appear to be normally distributed (there are no significant outliers)
Equal Variance (homoscedasticity): The variances along the regression line appear to remain similar as you move across the line

There are no apparent violations of homoscedasticity, normality, or independence. Formal tests of these assumptions are possible but are not included here.