Exercise 1.1

I have observations on systolic blood pressure (SBP) and age for a sample of 30 adults. A linear regression using age to predict SBP was conducted; Stata output is below.

```
. regress sbp age

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =   21.33
       Model |  6394.02269     1  6394.02269           Prob > F      =  0.0001
    Residual |  8393.44398    28  299.765856           R-squared     =  0.4324
-------------+------------------------------           Adj R-squared =  0.4121
       Total |  14787.4667    29  509.912644           Root MSE      =  17.314

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9708704   .2102157     4.62   0.000     .5402629    1.401478
       _cons |   98.71472   10.00047     9.87   0.000     78.22969    119.1997
------------------------------------------------------------------------------
```

(a) Interpret the intercept estimate. Is it meaningful?

The estimated mean SBP for a person who is 0 years old is 98.7 mmHg.
Not meaningful; these are adults and thus one can't be 0 years old.

(b) Interpret the slope estimate.

The estimated mean change in SBP for a 1 year increase in age is 0.97 mmHg.
Or, we estimate that mean SBP increases by 0.97 mmHg for each 1 year increase in age.

(c) What is the estimated mean SBP for a person who is 30 years old?

$98.71 + .971 \times 30 = 127.8$ *mmHg*

(d) What is the estimated difference in mean SBP for a person who is 45 years old compared to a person who is 30 years old? (And clearly state which mean is higher.)

$45 - 30 = 15$ *year difference*
$\rightarrow$ *difference in estimated mean SBP* $= 15 \times .971 = 14.6$ *mmHg, higher SBP for the 45 year old.*

Could also calculate the estimated mean for the 45 year old $= 98.71 + .971 \times 45 = 142.4$*mmHg*
And subtract: $142.4 - 127.8 = 14.6$

(e) Is there evidence of a significant association between age and SBP? Cite specific evidence from the output (i.e., a p-value).

Yes, there is evidence of a significant association, the p-value for the slope ("age" term) is < 0.0005 *(or, can look at the overall F-test since there is only one* X*: p-value* $= 0.0001$*)*

(f) What is the coefficient of determination for this model? Write a one sentence interpretation of this quantity.

$R^2 = 0.4324$ *43% of the variability in SBP is explained by age.*

Exercise 1.2

Information on 74 automobiles was collected (in 1978) to study the relationship between gas mileage (mpg) and various features of the cars. In particular, we are interested in the relationship between mileage and weight of the car, measured in pounds. A linear regression produced the (incomplete) Stata output below.

```
. regress mpg pounds

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =    134.62
       Model |   1591.9902         1   1591.9902   Prob > F        =
    Residual |  851.469256        72  11.8259619   R-squared       =
-------------+----------------------------------   Adj R-squared   =    0.6467
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pounds |  -.0060087   .0005179
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

(a) Write the estimated regression equation.

$$\widehat{MPG} = 39.4 - 0.006 \times POUNDS$$

(b) Interpret the number 39.44 in the above output.

The estimated mean MPG for a car that weighs 0 pounds is 39.4 miles per gallon.

(c) Interpret the number -0.0060087 in the above output.

For every 1 pound increase in the weight of a car, the estimated mean MPG decreases by 0.006 miles per gallon.

(d) What is the estimated change in MPG for a 1 U.S. ton increase in car weight? (Note: 1 U.S. ton = 2000 pounds)

$2000 \times \hat{\beta}_1 = 2000 \times -0.00601 = -12.012$ *MPG, i.e., 1 ton increase in weight corresponds to a 12 MPG decrease.*

(e) Test whether there is a significant effect of weight on mileage.

Model: $E(MPG) = \beta_0 + \beta_1 \times MPG$
Hypotheses:
$H_0 : \beta_1 = 0$
$H_1 : \beta_1 = 1$
Test statistic $= \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} = \frac{-.0060087}{.0005179} = -11.6$
Under H_0, $t \sim t_{n-2} = t_{74-2} = t_{72}$
p-value $= 2 \times P(t_{72} > |-11.6|) = 2 \times < 0.0001 = \quad < 0.0002$
Stata code to find $P(t_{72} > |-11.6|)$*:* `display ttail(72, 11.6)`
Reject H_0
There is evidence of a significant effect of car weight on mileage $(p < 0.0002)$*.*

(f) Calculate a 95% confidence interval for β_1. Do you expect 0 to be in the interval?

Because we rejected $H_0 : \beta_1 = 0$, we do not expect 0 to be in the interval.
95% CI: $\hat{\beta}_1 \pm t^*_{n-2} \times \widehat{SE}(\hat{\beta}_1)$
$t^*_{n-2} =$ *critical value from t distribution with 72 DF that has 0.05/2 = 0.025 in the upper tail*
$t^*_{n-2} = t^*_{72} = 1.99$
Stata code to find critical value: `display invttail(72, 0.025)`
95% CI: $-.0060087 \pm 1.99 \times .0005179 = -.0060087 \pm .00103062 = (-0.0070, -0.0050)$

(g) The R^2 value is missing from the output. Calculate it.

$$R^2 = \frac{MSS}{TSS} = \frac{1591.9902}{2443.45946} = 0.65$$

(h) Calculate the correlation between MPG and pounds.

$$r = sign(\hat{\beta}_1) \times \sqrt{R^2} = -\sqrt{0.65} = -0.81$$

(i) If I convert the weight of the car from pounds to kilograms and rerun the regression model (i.e., use weight in kilograms to predict MPG), will the slope estimate change? Will the R^2 change?

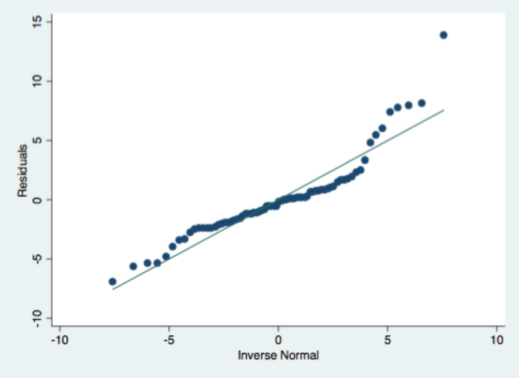
Slope will change, R^2 will not change

(j) The p-value for the overall F-test is missing from the output. What is it?

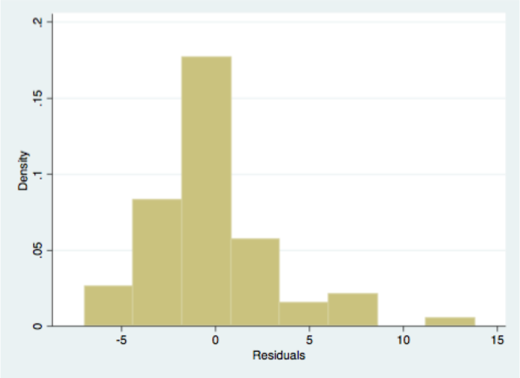
p-value from overall F-test = p-value from test of slope (when there is just one X)
→ *p-value for overall F-test* $= < 0.0002$

(k) Three plots generated after running the regression model are shown below (labeled A-C). For each plot, state the name of the plot (or what is being plotted), which assumption(s) of the model can be checked with the plot, and whether or not you see any problems with the assumptions.

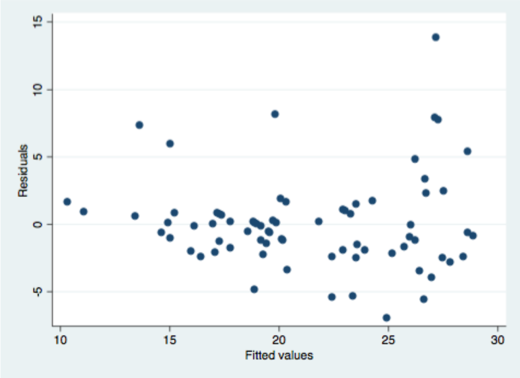
(A)



(B)



(C)



(A) Q-Q plot of residuals – check for normality – looks like there might be some skewness since the points don't lie along the line; also there looks like there might be an outlier in the upper tail (the point way off the line at the top)

(B) Histogram of residuals – check for normality – looks like there is a little right skew, with possibly an outlier in the upper (right) tail

(C) Residual-vs-fitted (RVF) plot – check for linearity and for equal variance – looks like linearity is okay (no real obvious pattern); maybe a slight problem with unequal variance since there is sort-of a fan shape (up-and-down spread gets larger as you move right along the X-axis).

Exercise 1.3

An economist is interested in the relationship between money flowing to stock mutual funds and money flowing into bond mutual funds. She collects data on the net new money flow into stocks and bonds for each year from 1985 to 2000 (in billions of dollars) and adjusts for inflation. She then uses Stata to run the regression model: $E(BONDS) = \beta_0 + \beta_1 \times STOCKS$

```
. regress bonds stocks

      Source |       SS           df       MS      Number of obs   =        16
-------------+----------------------------------   F(1, 14)        =      1.60
       Model |  5749.11511         1  5749.11511   Prob > F        =    ______
    Residual |  50200.7429        14  3585.76735   R-squared       =    ______
-------------+----------------------------------   Adj R-squared   =    0.0387
       Total |   55949.858        15  3729.99053   Root MSE        =    59.881

------------------------------------------------------------------------------
       bonds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      stocks |  -.1962223   .1549669    -1.27   _____     -.5285933    .1361486
       _cons |   53.40959    22.9925     2.32   0.036      4.09557    102.7236
------------------------------------------------------------------------------
```

(a) The p-value for the test of $H_0 : \beta_1 = 0$ is missing from the output. Based only on the confidence interval, what do you know about the missing p-value? (Hint: would you reject or fail to reject H_0?)

Since the 95% CI contains 0, we would fail to reject $H_0 : \beta_1 = 0$, thus the p-value must be > 0.05.

(b) The R^2 value is also missing. Calculate it.

$$R^2 = \frac{MSS}{TSS} = \frac{5749.11511}{55949.858} = 0.103$$