Exercise 3.1

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.). However, we are concerned that the weight (pounds) of the car might be an important factor to also consider. They produced the following Stata output:

```
. regress mpg weight foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     69.75
       Model |   1619.2877         2  809.643849   Prob > F        =    0.0000
    Residual |  824.171761        71   11.608053   R-squared       =    0.6627
-------------+----------------------------------   Adj R-squared   =    0.6532
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.4071

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0065879   .0006371   -10.34   0.000    -.0078583   -.0053175
     foreign |  -1.650029   1.075994    -1.53   0.130      -3.7955    .4954422
       _cons |    41.6797   2.165547    19.25   0.000     37.36172    45.99768
------------------------------------------------------------------------------
```

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

$E(MPG) = \beta_0 + \beta_1 WEIGHT + \beta_2 FOREIGN$

(b) Interpret the value $-.0065879$ from the Stata output.

The estimated mean change in MPG for a 1 pound increase in weight is -0.0066 miles per gallon, controlling for foreign status. (increase weight, decrease mileage)

(c) Interpret the value -1.650029 from the Stata output.

The estimated mean mileage for foreign cars is 1.65 miles per gallon lower than the mean for U.S.-made cars, controlling for weight.

(d) Is there a significant difference in mileage between foreign and U.S.-made cars, controlling for the weight of the car? Cite specific evidence from the Stata output in your answer.

No, test of $H_0 : \beta_2 = 0$ has p-value=0.130

(e) Is there a significant effect of weight on mileage, controlling for foreign status? Cite specific evidence from the Stata output in your answer.

Yes, test of $H_0 : \beta_1 = 0$ has p-value $<$0.0005 (shown on Stata output as 0.000)

(f) How much of the variability in mileage is explained by foreign status and weight of the car together?

$R^2 = 0.6627 \rightarrow$ *66% of the variability in mileage is explained by foreign status and car weight.*

Exercise 3.2

A survey of a random sample of students at the University of New Hampshire was conducted. We are interested in predictors of grade point average (GPA), which is measured on a 4-point scale. In particular, we would like to know whether sex is an important predictor (`gender`; 1=male, 0=female), and also whether how far away a student lives from campus (`miles`; how many miles away from school the student lives) is an important predictor. We also believe that age (`age`; years) and how many hours per week a student studies on average (`study`) might be important predictors.

A regression model was fit to predict GPA. Use these results to answer the questions on the next few pages.

```
. regress gpa age gender study miles

      Source |       SS           df       MS      Number of obs   =       204
-------------+----------------------------------   F(4, 199)       =      9.30
       Model |    6.670995         4  1.66774875   Prob > F        =    0.0000
    Residual |  35.6766466       199  .179279631   R-squared       =
-------------+----------------------------------   Adj R-squared   =
       Total |  42.3476416       203  .208609072   Root MSE        =    .42341

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0384382   .0102536     3.75   0.000     .0182185    .0586578
      gender |  -.1236973   .0601494    -2.06   0.041    -.2423094   -.0050852
       study |   .0127933   .0034168     3.74   0.000     .0060555    .0195311
       miles |   .0051275   .0054843     0.93   0.351    -.0056873    .0159424
       _cons |   1.870604   .2215288     8.44   0.000     1.433759    2.307449
------------------------------------------------------------------------------
```

(a) Interpret the result of the overall F-test for this model (include reference to the p-value).

At least one of the predictors – age, gender, hours studied, and miles away from school – is significantly associated with GPA ($p<0.00005$).

(b) The R^2 value is missing from the output. Calculate it, and write a one sentence interpretation.

$R^2 = \frac{MSS}{TSS} = \frac{6.670995}{42.3476416} = 0.1575$

15.8% of the variability in GPA is collectively explained by age, gender, hours studied, and how far away from campus a student lives.

(c) The adjusted R^2 is also missing from the output. Calculate it.

Adjusted $R^2 = 1 - \frac{\textit{Mean Square Residual}}{\textit{Mean Square Total}} = 1 - \frac{0.179279631}{0.208609072} = 1 - 0.8594 = 0.1406$

(d) Interpret the coefficient for gender in this model.

$\hat{\beta}_{gender} = -0.124$
The estimated mean GPA for males is 0.124 points lower than the estimated mean for females, controlling for age, hours studied, and how far away from school a student lives.

(e) For the appropriate test to determine if there is a significant effect of gender on GPA, write down the null and alternative hypotheses, the test statistic value, and its distribution under the null hypothesis.

$H_0 : \beta_{gender} = 0$ *vs.* $H_a : \beta_{gender} \neq 0$
test statistic: $t = -2.06$
Under H_0, $t \sim t_{199}$

(f) Is there a significant difference between men and women in mean GPA, controlling for the other factors in the model? (Cite a p-value in your answer.)

Yes; p-value = 0.041

(g) Is the distance a student lives away from school significantly associated with his/her GPA, controlling for age, gender, and hours studied? (Cite a p-value in your answer.)

No; p-value = 0.351

(h) Interpret the effect of hours studied on GPA, including reference to a p-value.

The estimated mean GPA increases by 0.038 points for each one year increase in age, adjusting for age, gender, and how far away from school a student lives; this is a significant association ($p<0.0005$).

(i) What is the estimated GPA for a male student who is 20 years old, lives 5 miles from campus, and studies 10 hours per week?

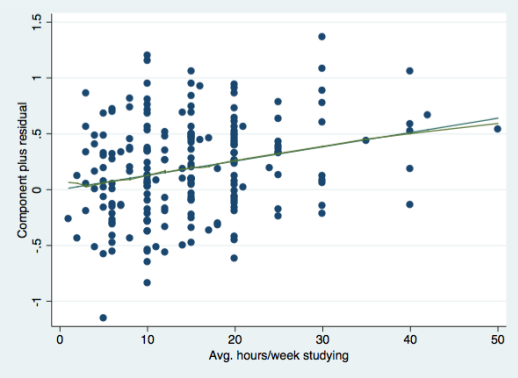
$$\begin{aligned}\widehat{GPA} &= 1.87+0.038\times AGE-0.124\times GENDER+0.0128\times STUDY+0.00513\times MILES \\ &= 1.87 + 0.038 \times 20 - 0.124 \times 1 + 0.0128 \times 10 + 0.00513 \times 5 \\ &= 2.67\end{aligned}$$

(j) What is the estimated difference in mean GPA between two female students who are the same age, and live the same distance from campus, but one studies 10 hours/week and the other studies 25 hours/week? (Also indicate which has the higher estimated GPA.)

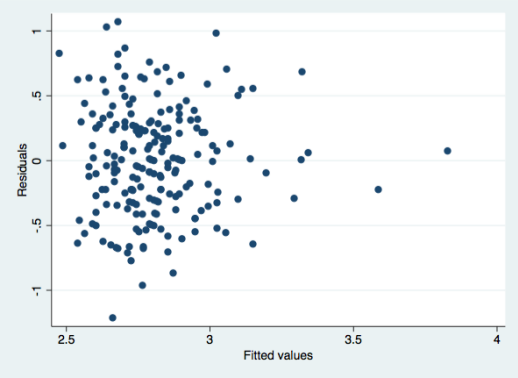
difference $= (25 - 10) \times \hat{\beta}_{study} = 15 \times 0.0127933 = 0.19$
Higher GPA for student who studies more (25 hours)

(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. (Carefully read the axes to make sure you understand what plot is being made.)

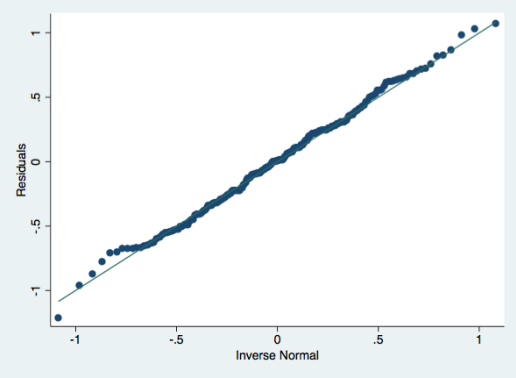
(A)



(B)



(C)



(A) Component-plus-residual (CPR) plot for the predictor STUDY – check for linearity for the predictor STUDY – no obvious violation since the two lines (green and blue) lie basically on top of each other

(B) Residual-vs-fitted (RVF) plot – check for equal variance (NOT linearity, this is a MLR model) – looks like equal variance probably okay since up-and-down spread is pretty constant along X-axis. There are not many data points at the far right, which makes it look like the variability is lower, but it's probably okay.

(C) 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),

Exercise 3.3

A study collected data from 89 high school seniors about substance abuse. One response variable was marijuana use, which was measured on a continuous scale where a higher number indicated more use of marijuana. A multiple regression analysis used grade point average (GPA), popularity, and depression score to predict marijuana use. Higher values on the popularity score indicate a student is more popular; higher scores on the depression scale mean a student is more depressed. The results were summarized as:

Covariate	$\hat{\beta}$	t	p-value
GPA	-0.597	4.55	$<$0.001
Popularity	0.340	2.69	$<$0.01
Depression	0.030	2.69	$<$0.01

(a) The overall F-statistic reported was 14.83. Write the regression model being estimated, then state the null and alternative hypotheses for this test statistic, give the test statistic's degrees of freedom under the null, and draw a conclusion (you will have to find a p-value or critical value to do this).

$E(M) = \beta_0 + \beta_1 GPA + \beta_2 POP + \beta_3 DEPRESS$
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ *versus* H_a *: at least one of these* $\beta_j \neq 0$
Under H_0, $F \sim F(p, n-p-1) = F(3, 89-3-1) = F(3, 85)$
p-value $= P(F_{(3,85)} > 14.83) =< 0.00001$
(Stata code to calculate: `.display Ftail(3, 85, 14.83)`
At least one of GPA, popularity, and depression is significantly associated with marijuana use.

(b) What do the signs of each of the regression coefficients say about the covariates' relationships with marijuana use?

GPA: coefficient is negative; as GPA goes up, marijuana use score tends to go down
Popularity: coefficient is positive; as popularity increases, marijuana use score tends to go up
Depression: coefficient is positive; as depression score increases, marijuana use score tends to go up

(c) Interpret the coefficient for GPA.

$\hat{\beta}_{GPA} = -0.597$*; A one point increase in GPA is associated with a 0.597 point decrease in marijuana use score, controlling for popularity score and depression score (p<0.001).*

(d) The variables in this study were measured by face-to-face interviews of the students, by trained adult interviewers. How might this affect the data/results? (This is not per se a question about statistics. . . .)

Might be underreporting of marijuana use; possibly underreporting of depression.

Exercise 3.4

We have measured the weight (lbs), height (inches), and age (years) of 12 children (ages 6-12). We wish to use height (H) and age (A) to predict weight (W). Stata output (with blanks) is below for the model:

$$W = \beta_0 + \beta_1 H + \beta_2 A + e \quad e \sim N(0, \sigma^2)$$

```
. regress weight height age

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F( ,  )         =     15.95
       Model |  692.822607         2  346.411303   Prob > F        =    0.0011
    Residual |  195.427393         9  21.7141548   R-squared       =    0.7800
-------------+----------------------------------   Adj R-squared   =    0.7311
       Total |      888.25        11       80.75   Root MSE        =    4.6598

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    .722038   .2608051     2.77   0.022
         age |   2.050126   .9372256     2.19   0.056
       _cons |   6.553048   10.94483     0.60   0.564
------------------------------------------------------------------------------
```

(a) Write the null and alternative hypotheses for the overall F-test, the test statistic value, its distribution under the null, the p-value, and write a one sentence conclusion.

$H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$
test statistic: $F = 15.95$; Under H_0, $F \sim F(p, n-p-1) = F(2, 12-2-1) = F(2, 9)$
p-value = 0.0011; At least one of height and age are significantly associated with child weight.

(b) Interpret the intercept estimate for this model. Is this a meaningful quantity?

The estimated mean weight of children who are 0 years old and 0 inches tall is 6.55 lbs.
Not meaningful; cannot have age=0 years or height=0 inches.

(c) Interpret the effect of height in this model. Is the effect significant? (Cite a p-value.)

The estimated mean weight increases by 0.72 lbs for each 1 inch increase in height, controlling for age. The effect is significant (p=0.022).

(d) What is the estimated change in weight for a 1 foot increase in height?

1 foot = 12 inches; $12 \times 0.722 = 8.7$ *lbs*

(e) Is the effect of age on weight significant after controlling for height? (Cite a p-value.)

No (but just barely non-significant); p-value = 0.056

(f) The 95% confidence intervals are missing from the output. For each parameter (the β_js), state whether the confidence interval will include 0 or not include 0.

β_0*: CI will include 0 (p-value* > 0.05*)*
β_1*: CI will not include 0 (p-value* < 0.05*)*
β_2*: CI will include 0 (p-value* > 0.05*)*

(g) If I remove age as a predictor and re-run the model, what will happen to the R^2 value? What will happen to the adjusted R^2 value?

R^2 *stays the same or go down (can't go up); Adjusted* R^2 *might go up or down, don't know for sure*