Exercise 6.1

A survey of a random sample of students at the University of New Hampshire was conducted. We are interested in whether how far away from campus a student lives (`miles`; number of miles away, 0 miles means student lives on campus) is associated with grade point average (GPA), which is measured on a 4-point scale.

A series of models were performed and Stata output is provided at the end of the problem. Use this to answer the questions below (note: output spans 2 pages).

(a) What is the unadjusted effect of miles from school on GPA? Is there evidence of a significant effect? ("unadjusted" = not adjusted for any other predictors)

Estimated mean GPA increases by 0.0124 points for each 1 mile further away from school. This is a significant association (p-value = 0.025).

(b) Is there evidence that age confounds the relationship between GPA and miles away from school?

A: predictor of interest (miles) is not a cause of potential confounder (miles) – definitely true
B: potential confounder (age) is associated with outcome (GPA) – yes, p-value $<$ 0.0005 from model of GPA predicted by age
C: potential confounder (age) is associated with predictor of interest (miles) – yes, p-value $<$ 0.0005 from model of miles predicted by age
D: coefficient for predictor of interest (miles) changes when potential confounder (age) is added – goes from 0.0124 to 0.00513 (much larger than a 10% change!)

Thus yes, age is a confounder of the relationship between GPA and miles from school.

```
. regress gpa age

      Source |       SS           df       MS      Number of obs   =       218
-------------+----------------------------------   F(1, 216)       =     17.64
       Model |  3.45507544         1  3.45507544   Prob > F        =    0.0000
    Residual |  42.2966637       216  .195817887   R-squared       =    0.0755
-------------+----------------------------------   Adj R-squared   =    0.0712
       Total |  45.7517391       217  .210837507   Root MSE        =    .44251

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0415214   .0098848     4.20   0.000     .0220383    .0610045
       _cons |   1.952825   .2058752     9.49   0.000     1.547044    2.358607
------------------------------------------------------------------------------


. regress gpa miles

      Source |       SS           df       MS      Number of obs   =       206
-------------+----------------------------------   F(1, 204)       =      5.08
       Model |  1.04763577         1  1.04763577   Prob > F        =    0.0253
    Residual |  42.1070429       204  .206407073   R-squared       =    0.0243
-------------+----------------------------------   Adj R-squared   =    0.0195
       Total |  43.1546787       205  .210510628   Root MSE        =    .45432

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       miles |   .0123695   .0054905     2.25   0.025     .0015442    .0231949
       _cons |   2.775878   .0338218    82.07   0.000     2.709193    2.842563
------------------------------------------------------------------------------
```

```
. regress miles age

      Source |       SS           df       MS      Number of obs   =       226
-------------+----------------------------------   F(1, 224)       =     31.98
       Model |  866.680685         1  866.680685   Prob > F        =    0.0000
    Residual |  6071.31932       224  27.1041041   R-squared       =    0.1249
-------------+----------------------------------   Adj R-squared   =    0.1210
       Total |        6938       225  30.8355556   Root MSE        =    5.2062

------------------------------------------------------------------------------
       miles |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .6477434   .1145489     5.65   0.000     .4220121    .8734748
       _cons |  -11.23861   2.366632    -4.75   0.000    -15.90232   -6.574904
------------------------------------------------------------------------------


. regress gpa age miles

      Source |       SS           df       MS      Number of obs   =       206
-------------+----------------------------------   F(2, 203)       =      9.66
       Model |  3.75001816         2  1.87500908   Prob > F        =    0.0001
    Residual |  39.4046605       203  .194111628   R-squared       =    0.0869
-------------+----------------------------------   Adj R-squared   =    0.0779
       Total |  43.1546787       205  .210510628   Root MSE        =    .44058

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0395682   .0106047     3.73   0.000     .0186587    .0604777
       miles |   .0051321   .0056668     0.91   0.366    -.0060412    .0163053
       _cons |   1.974672   .2172224     9.09   0.000      1.54637    2.402973
------------------------------------------------------------------------------
```

Exercise 6.2

We are interested in quantifying the relationship between percent body fat `pctfat` and the following predictors: triceps skin-fold thickness (`tricep`), thigh circumference (`thigh`), mid-arm circumference (`midarm`). A regression model was run using these three variables to predict percent body fat. Use the Stata output provided at the end of the problem to answer the questions below.

(a) Report the p-value from the overall F-test for this model, and write a one-sentence interpretation. (Assume $\alpha = 0.05$.)

p-value $<$ 0.00005 (Stata shows as 0.0000)
At least one of tricep circumference, thigh circumference, and mid-arm circumference is significantly associated with percent body fat.

(b) What are the results of the individual t-tests for the regression coefficients (ignoring the intercept)? Does this seem "right" given the result of the overall F-test?

All three predictors are NOT significantly associated, p-values are $>$ 0.05 for all of them. This seems to be in conflict with the overall F-test result.

(c) Variance inflation factors for the model are also in the output. Do they indicate a problem? If so, what is the problem?

VIFs are all REALLY big ($>$ 100!!!). This indicates a problem with collinearity – the predictors are highly correlated with each other.

(d) What would your next step be to address the problem you identified in (c)?

Remove one of the predictors, probably tricep since it has the largest VIF – though might also consider which of the 3 is of most and least scientific interest.

```
. regress pctfat tricep thigh midarm

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(3, 16)        =     21.52
       Model |  396.984607         3  132.328202   Prob > F        =    0.0000
    Residual |  98.4049068        16  6.15030667   R-squared       =    0.8014
-------------+----------------------------------   Adj R-squared   =    0.7641
       Total |  495.389513        19  26.0731323   Root MSE        =      2.48

------------------------------------------------------------------------------
      pctfat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tricep |   4.334085   3.015511     1.44   0.170    -2.058512    10.72668
       thigh |  -2.856842   2.582015    -1.11   0.285    -8.330468    2.616785
      midarm |  -2.186056   1.595499    -1.37   0.190    -5.568362     1.19625
       _cons |   117.0844   99.78238     1.17   0.258    -94.44474    328.6136
------------------------------------------------------------------------------


. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      tricep |    708.84    0.001411
       thigh |    564.34    0.001772
      midarm |    104.61    0.009560
-------------+----------------------
    Mean VIF |    459.26
```

Exercise 6.3

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. A regression model was fit using the following predictors: age (age), year in school (year; 1=freshman, 2=sophomore, 3=junior, 4=senior), sex (gender; 1=male, 0=female), and how far away from school the student lives (miles). Note that students who live on campus would have a "0" for the miles variable.

The age variable was centered before including it in the model – the sample mean, 20, was subtracted from all values (resulting variable: age_20). Use the Stata output provided at the end of the problem to answer the questions below (note: output spans 2 pages).

(a) Interpret the estimated coefficient for age_20.

Estimated mean GPA increases by 0.0345 points for each 1 year increase in age, adjusting for year in school, sex, and how far away from school the student lives.

Note that the interpretation of a centered covariate is the same as a non-centered covariate when it's not involved in an interaction.

(b) Carefully interpret the intercept estimate. Is this meaningful?

The estimated mean GPA for students who are 20 years old, freshman, females, and live on campus (0 miles away) is 2.95. Yes, this is meaningful (because we centered age at 20!).

(c) Is there any problem with multicollinearity for this model?

No – all VIFs are below 5 (even the ones for the dummy variables for year in school)

(d) We would like to perform backwards selection starting from this model, with 0.05 as the removal criterion. What should be the first predictor removed?

`miles` *– with a p-value of 0.466. The p-value from the partial F-test for year in school is only 0.2917 so* `miles` *should be removed.*

```
. generate age_20 = age - 20
. generate year2 = (year==2) if !missing(year)
. generate year3 = (year==3) if !missing(year)
. generate year4 = (year==4) if !missing(year)

. regress gpa age_20 year2 year3 year4 gender miles

      Source |       SS           df       MS      Number of obs   =       206
-------------+----------------------------------   F(6, 199)       =      4.58
       Model |  5.23601098         6  .872668496   Prob > F        =    0.0002
    Residual |  37.9186677       199  .190546069   R-squared       =    0.1213
-------------+----------------------------------   Adj R-squared   =    0.0948
       Total |  43.1546787       205  .210510628   Root MSE        =    .43652

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_20 |   .0345186   .0111388     3.10   0.002     .0125534    .0564838
       year2 |  -.2075202    .126369    -1.64   0.102    -.4567144     .041674
       year3 |  -.1030474    .123832    -0.83   0.406    -.3472388     .141144
       year4 |   -.091817   .1276345    -0.72   0.473    -.3435067    .1598727
      gender |   -.121338   .0617188    -1.97   0.051    -.2430448    .0003687
       miles |   .0041283   .0056541     0.73   0.466    -.0070212    .0152779
       _cons |   2.950831   .1173551    25.14   0.000     2.719412     3.18225
------------------------------------------------------------------------------
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
       year3 |      3.79    0.263649
       year2 |      3.56    0.280596
       year4 |      3.56    0.280699
      age_20 |      1.27    0.785503
       miles |      1.15    0.870509
      gender |      1.02    0.975817
-------------+----------------------
    Mean VIF |      2.39

. test year2 year3 year4

 ( 1)  year2 = 0
 ( 2)  year3 = 0
 ( 3)  year4 = 0

       F(  3,   199) =    1.25
            Prob > F =    0.2917
```