Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

Midterm Examination Key
February 13, 2015

Name: ____

Instructions: **This exam is closed book, closed notes. You have 50 minutes. You may not use any device that is capable of accessing the internet.**

Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

NOTE: When you need to make calculations, always use at least four significant digits in your intermediate calculations, and report at least three significant digits. (Example: 1.045 and 0.0001234 and 1234000 each have four significant digits.)

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor.

PLEDGE:
On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:____

Problems 2 - 7 deal with data from an observational study of lung function and smoking in children. The appendices contain results from selected analyses:

Appendix A : Plots of simulated data (7 scenarios) **(problem 1)**
Appendix B : Description of the variables and descriptive statistics **(all problems 2- 7)**
Appendix C : Regression analyses of height by age **(problems 3 and 4)**
Appendix D : Regression analyses of smoking and sex **(problem 5)**
Appendix E : Regression analyses of smoking, age, and sex **(problem 6)**
Appendix F : Regression analysis of log transformed FEV by smoking, age, and height **(problem 7)**

1. (10 points) **Appendix A** displays 7 scatterplots labeled A - G. In the blanks below, list the plots in order according to lowest (most negative) to highest (most positive) correlation. (In all cases, the scale for the x and y axes are the same.)

Most……………………………………………………………………………Most
Neg Pos

	E		**B**		**D**		**F (G)**		**G (F)**		**A**		**C**
Slope:	***Mod neg***	=	***Mod neg***	=	***Mod neg***	<	***Flat***	=	***U shape***	<	***Mod pos***	<	***Steep pos***
Var(Y|X)	***Small***	<	***Moderate***	=	***Moderate***		*(Moderate)*		*(Small)*		***Moderate***	=	***Moderate***
Var(X)	***High***	=	***High***	>	***Low***		*(High)*		*(High)*		***High***	=	***High***
Corr	***Strong neg***		***Mod neg***		***Weak neg***		***Near 0***		***Near 0***		***Mod pos***		***Strong pos***

Correlation depends on 1) the slope of the best fitting straight line, 2) the variance of Y within a group having equal values of X (related to the vertical spread of the data for a single X), and 3) the variance of X (related to the horizontal spread of X).

- ***Correlation is of the same sign as the slope of the best fitting straight line. If the best fitting straight line is flat, then the correlation is zero, and the Var(Y|X) and Var(X) do not matter.***

- ***Correlation is lower in absolute value (so closer to 0) when Var(Y|X) is higher.***

- ***Correlation is lower in absolute value (so closer to 0) when Var(X) is lower.***

2. **Appendix B** presents descriptive statistics, including a scatterplot of child height versus age (the first of the three scatterplots presented).
   a. (10 points) What observations do you make from the scatterplot regarding the association between height and age?

   <u>Ans:</u> Major points:

   - **No obvious outliers**
   - **Upward trend**
   - **Curvilinear: up then leveling off**
   - **Heteroscedastic with higher variance among older**
   - *(Effect modification by sex: No separation in young, then separation in old)*

   b. (5 points) Is there evidence from this plot that sex modifies the height - sex association? **Briefly** explain your reasoning.

Ans: **Yes, lines not parallel.** *(Note that this effect modification will exist no matter the scale used to contrast height by age: There is essentially zero vertical separation of the lines at younger ages, with marked separation in the teenagers.)*

c. (5 points) Does sex confound the description of the height – age association in the population? **Briefly** explain your reasoning.

Ans: **No, the age distribution is similar by ses.** *(As seen in the descriptive statistics tables: the mean ages are quite similar. Sex is of course associated with height, but once we have ruled out an association with the predictor of interest in the sample, we have our answer.)*

3. Suppose we are interested in any association between height and age in pre-pubescent children. **Appendix C** contains the results of linear regression analyses exploring this question among the subjects who are 10 years old or younger..

a. (5 points) Based on the regression model, what is the best estimate for the mean height in subjects who are 5 years old? (**In this and all problems, whenever the precision of the analysis reports will allow, use at least four significant digits in your calculations, and provide at least three significant digits in your answer, or you will receive no credit.)**

Ans: **5 x 2.282997 + 39.82337 = 51.24**

b. (5 points) Based on the regression model, what is the best estimate for the mean height in subjects who are 6 years old?

Ans: **6 x 2.282997 + 39.82337 = 53.52**

c. (5 points) Based on the regression model, what is the best estimate for the mean height in subjects who are 10 years old?

Ans: **10 x 2.282997 + 39.82337 = 62.65**

d. (5 points) Based on the regression model, what is the best estimate for the difference in mean height between 6 year old subjects and 5 year old subjects?

Ans: **2.283** *(This is just the slope, but you could of course have subtracted your answer in part a from that in part b.)*

e. (5 points) Based on the regression model, what is the best estimate for the difference in mean height between 10 year old subjects and 5 year old subjects?

Ans: **5 x 2.282997 = 11.41** *(This is just 5 times the slope, but you could of course have subtracted your answer in part a from that in part c.)*

f. (5 points) Which regression analysis presented in Appendix B would you have chosen *a priori* to make statistical inference about any associations between mean height and age? Why? (A very brief answer should suffice here.)

Ans: Model C2 because it allows for valid inference about the mean even if data is heteroscedastic *(Note my wording is "allows for heteroscedasticity" not "assumes heteroscedasticity". This model works even if data is homoscedastic.)*

g. (5 points) Using the regression analysis you identified in part (f), provide a 95% confidence interval for the difference in mean height between two populations who differ in age by 5 years. (Just the numbers, no interpretation necessary here.)

Ans: 5 x (2.120363, 2.445631) = (10.6, 12.2) *(We can just multiply the CI for the slope by 5.)*

h. (5 points) Provide an interpretation for the intercept in the regression model. What scientific use would you make of this estimate?

Ans: Estimated average height of newborns. Outside the range of our data, so not trustworthy scientifically.

i. (5 points) Provide an interpretation for the slope in the regression model. What scientific use would you make of this estimate?

Ans: Estimated difference in mean height per 1 year difference in age *(higher age minus lower age)***.**

j. (5 points) Is there evidence that there is an association between mean height and age? State your evidence.

Ans: P < 0.00005 from P value for slope. *(Independent variables would show a flat line.)*

k. (5 points) Is there evidence that there is a statistically significant correlation between height and age? State your evidence.

Ans: Yes, same as test for slope.

l. (5 points) Can you provide an estimate of the correlation between height and age in this sample? If so, do so. If not, explain why not.

Ans: r = sqrt (.6265) = 0.7915 *(In simple linear regression, the correlation is the square root of R^2 that has the same sign as the slope.)*

m. (5 points) Based on the regression model, what is the best estimate for the average standard deviation of height within a group that is homogeneous with respect to age?

Ans: Root MSE = 2.9221 *(This is SD (Y | X))*

4. Again using the scatterplot of height versus age in **Appendix A**, answer the following questions.

   a. (5 points) From that plot, comment on the reliability of your estimates of age group specific means in parts (a) through (c) of problem 1.

   <u>Ans:</u> We need the line to look pretty straight. It does up to age 10, so OK.

   b. (5 points) From this plot, comment on the reliability of your answers to the statistical inference you provided in parts (g), (i), and (j) of problem 1.

   <u>Ans:</u> We need to correctly handle any heteroscedasticity. We used the robust SE from Huber-White sandwich estimator, so OK.

   c. (5 points) From this plot, comment on the reliability of your answers to part (m) of problem 1.

   <u>Ans:</u> We would need homoscedasticity and a pretty straight line. Not OK because the data are heteroscedastic. *(So the RMSE is based on some sort of average Var(Y|X) across values of X)*

   d. (5 points) From this plot, comment on why the difference between the precision of the confidence intervals for the two analyses might have been anticipated.

   <u>Ans:</u> Groups with higher variance have larger sample size, so conservative inference when incorrectly presuming equal variance. *(There is lower variance among the 3-4 year olds than among the 9-10 year olds. We have way more 9-10 year olds (who have larger variance of height), so we expect conservative inference (higher P values, wider CI) when incorrectly presuming equal variances than when correctly handling heteroscedasticity.)*

5. (15 points) Now suppose we are interested in investigating any association between self reported smoking and sex. **Appendix D** contains three regression analyses addressing this question**.**

   a. Using **Model D1**, provide estimates of the probability and odds that boys would report as smokers, and the probability and odds that girls would report as smokers. (I want four numbers: the probability and odds estimates for each sex.)

   <u>Ans:</u> Saturated model, so can use sample proportions and odds (*from first page of Appendix B)***:**

   Males: Probability 7.74%, Odds= 26 / 310 = .0774 / (1 - .0774) = 0.08387

   Female: Probability 12.26%, Odds= 39 / 279 = .1226 / (1 - .1226) = 0.1398

   (You could of course have laboriously used the regression parameter estimates to calculate the proportions:

Males: intercept = p = .0774 ➔ odds = p / (1-p) = .08387

Females : p = .077381 + .0452606 = 0.1226 ➔ odds = .1226 / (1-.1226) = 0.1398)

b. Using **Model D2**, provide estimates of the probability and odds that boys would report as smokers, and the probability and odds that girls would report as smokers. (I want four numbers: the probability and odds estimates for each sex.)

<u>Ans:</u> Same as part a.

(You could of course have laboriously used the regression parameter estimates to calculate the log proportions, accounting for the log links:

Males: log (p) = -2.559015 ➔ p = exp(-2.559015) = .0774 ➔ odds = p / (1-p) = .08387

Females : log(p) =-2.559015 + 0.4605249 = -2.0984901 ➔ p = exp(-2.0984901) =0.1226 ➔ odds = .1226 / (1-.1226) = 0.1398)

c. Using **Model D3**, provide estimates of the probability and odds that boys would report as smokers, and the probability and odds that girls would report as smokers. (I want four numbers: the probability and odds estimates for each sex.)

<u>Ans:</u> Same as part a.

(You could of course have laboriously used the regression parameter estimates to calculate the log proportions, accounting for the log links:

Males: log (odds) = -2.478476 ➔ odds = .08371 = ➔ p = odds / (1+odds) = .0774

Females : log(odds) =-2.478476 + 0.5108256 = -1.96765 ➔ odds = exp(-1.96765) = =0.1398 ➔ p = .1398 / (1+.1398) = 0.1226)

6. **Appendix E** presents regression analyses investigating any association between self reported smoking and age and sex.. Use the results of **Appendices B** and **E** to answer the following questions.

a. (5 points) Do the results of the available analyses suggest that sex modifies any association between smoking and age? Clearly explain your reasoning.

<u>Ans:</u> No. Estimated OR for males of 1.630 (Model E2) is similar to estimated OR for females of 1.669 (Model E3).

b. (5 points) Do the results of the available analyses suggest that sex confounds any association between smoking and age? Clearly explain your reasoning.

<u>Ans:</u> No, not much difference in age distribution by sex. *(Note that the unadjusted OR is closer to the null than is either stratum OR (but not by much). This is consistent with a precision variable, but can also happen with a confounder. So this criterion is not of use* *here. If the unadjusted OR were further from the null or on the opposite side of the null than the stratum ORs that would have to be confounding.)*

c. (Bonus: 15 points) **As briefly as you can do so while demonstrating your understanding of the scientific interpretation of the models,** describe the correspondences between the estimates and inference obtained in Models E2 and E3 as compared to the estimates and inference in Model E5, and explain how they differ from Models E1 and E4.

Ans: Considering the interpretation of the parameters we can confirm that Model E5 has sufficient parameters to model the age and slope for each sex separately:

- **Model E5 intercept = Model E2 intercept**
- **Model E5 age slope = Model E2 age slope**
- **Model E5 female slope = (Model E3 intercept) – (Model E2 intercept)**
- **Model E5 female-age slope = (Model E3 age slope) – (Model E2 age slope)**

Models E1 and E4 each borrow information across the sexes to estimate the age slope.

7. **Appendix F** presents results of regression analyses exploring an association between FEV and self-reported smoking, potentially adjusted for age and/or height.

a. (15 points) Provide full inference (methods and results) for the analysis represented by **Model F1**. (You may be brief, but make sure you include the relevant parts.)

Ans: ***Methods:*** **We investigate association between FEV and smoking by ratio of geometric mean (GM) FEV using Wald statistics from linear regression on log transformed FEV on binary smoking variable, and we use Huber-White sandwich estimator for standard errors.**

Results: **Geometric mean FEV in smokers is estimated 10.8% higher in smokers than nonsmokers. 95% CI suggests data typical of true GM ratio 1.004 to 1.18. This is highly statistically significant: two-sided P= .001.**

b. (15 points) Provide brief interpretations of each of the four regression parameters in the regression in **Model F4**. (You do not need to report or interpret the CI, but your answer should include the number and what it means.)

Ans:

- **intercept: Geometric mean FEV in newborn, 0 inch tall nonsmokers is estimated to be 0.131**

- **height slope: Geometric mean FEV is estimated to be 1.046 times as high per 1 inch difference in height when comparing groups of same age, same smoking behavior** *(taller tend to have higher FEV)*

- **age slope: Geometric mean FEV is estimated to be 1.022 times as high per 1 year difference in age when comparing groups of same height, same smoking behavior** *(older tend to have higher FEV)*

- **age slope: Geometric mean FEV for smokers is estimated to be only 0.9492 times as high as nonsmokers of same height, same age**

c. (15 points) Although we would truly choose one of the models before looking at the data, it is still instructive to consider the differences between them as they might relate to confounding and precision. Discuss how any confounding and precision might affect the differences between the conclusions you might reach when using **Models F1, F2, F3, and F4.** Be sure to justify your reasoning (in a word or two)

<u>Ans:</u> Because these are linear regression models, we can use the estimated slopes to assess confounding. Model F4 serves as a reference, because it includes smoking, age, and height. We find:

- **In Model F1, qualitative confounding by age leads to higher estimated geometric mean for smokers to nonsmokers: a statistically significant 10.77% higher.** *(This is a valid estimate: smoking children do tend to have higher FEV than nonsmoking children, but our best guess as to the reason is that the smokers are older and more likely to smoke.)*

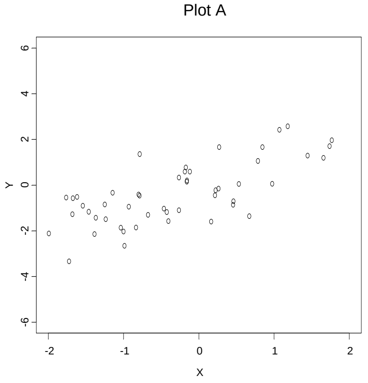
- **In Model F2, adjustment for age removes the confounding, but the correlation between smoking and age adds "variance inflation" that is not sufficiently countered by the added precision from adding age to the model: we know estimate that smokers have 5.00% lower geometric mean than nonsmokers of same age, but that estimate is not statistically significant.**

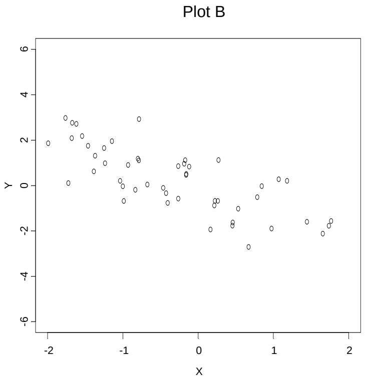
- **In Model F4, the adjustment for height adds precision to the analysis that was done in Model F2: We still estimate that smokers have 5% lower geometric mean FEV than nonsmokers of same age , but by holding height constant as well, the estimate is now statistically significant. Thus after adjusting for age there was no further confounding by height.**

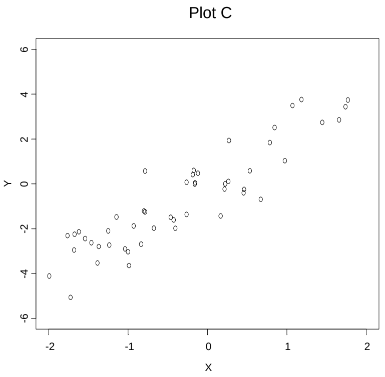
- **On the other hand, in Model F3, we adjusted for height, but there was still some confounding by age: In Model F3 we estimated that the geometric mean for smokers was only 1.34% lower than that in nonsmokers of same height. Compared to Model F4, this suggests age is still confounding the FEV-smoking association.**

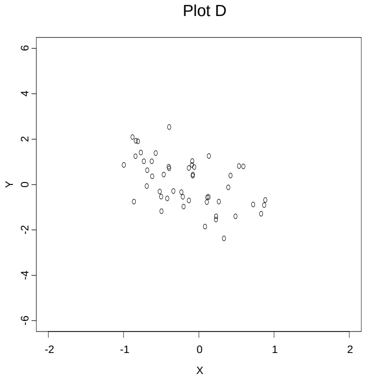
<u>APPENDIX A</u>: Simulated scatterplots.

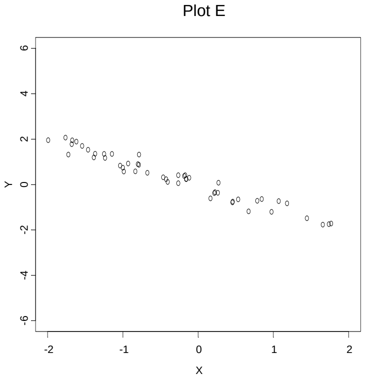
Below are 7 scatterplots labeled A - G. In all cases, the scale for the x and y axes are the same.

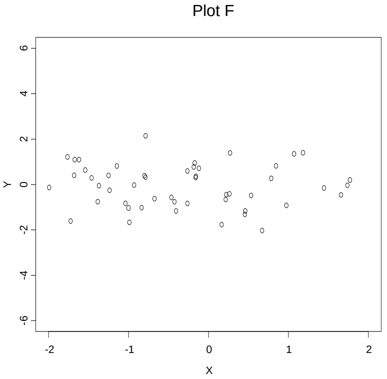


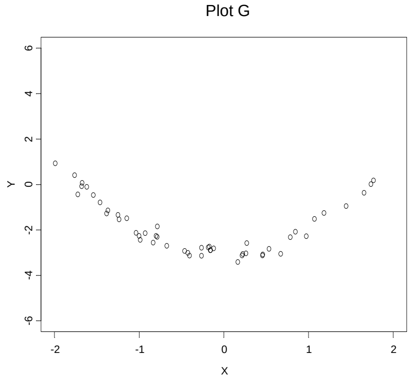












APPENDIX B: Description of variables and descriptive statistics

These data come from an observational study of lung function in a sample of **N= 654** healthy children. Of particular interest is how lung function might vary with respect to self-reported smoking behavior.

age: Age in years of the subject at the time of study enrolment
female: Indicator that the subject is male (**0**= male, **1**= female)
height*:* Height in age of the subject at the time of study enrolment.
smoker: Indicator that the subject self-reports as a smoker (**0**= nonsmoker, **1**= smoker)
fev*:* Forced expiratory volume (FEV) in liters/sec (FEV= the volume of air that can be expired in 1 second with maximal effort.

The following table presents cross tabulation of the children's self reported smoking behavior by sex. The table contains counts, as well as percentage calculated both by row and column.

```
. tabulate smoker female, row col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

           |        female
    smoker |         0          1 |     Total
-----------+----------------------+----------
         0 |       310        279 |       589
           |     52.63      47.37 |    100.00
           |     92.26      87.74 |     90.06
-----------+----------------------+----------
         1 |        26         39 |        65
           |     40.00      60.00 |    100.00
           |      7.74      12.26 |      9.94
-----------+----------------------+----------
     Total |       336        318 |       654
           |     51.38      48.62 |    100.00
           |    100.00     100.00 |    100.00
```

APPENDIX B (cont.): **Description of variables and descriptive statistics**

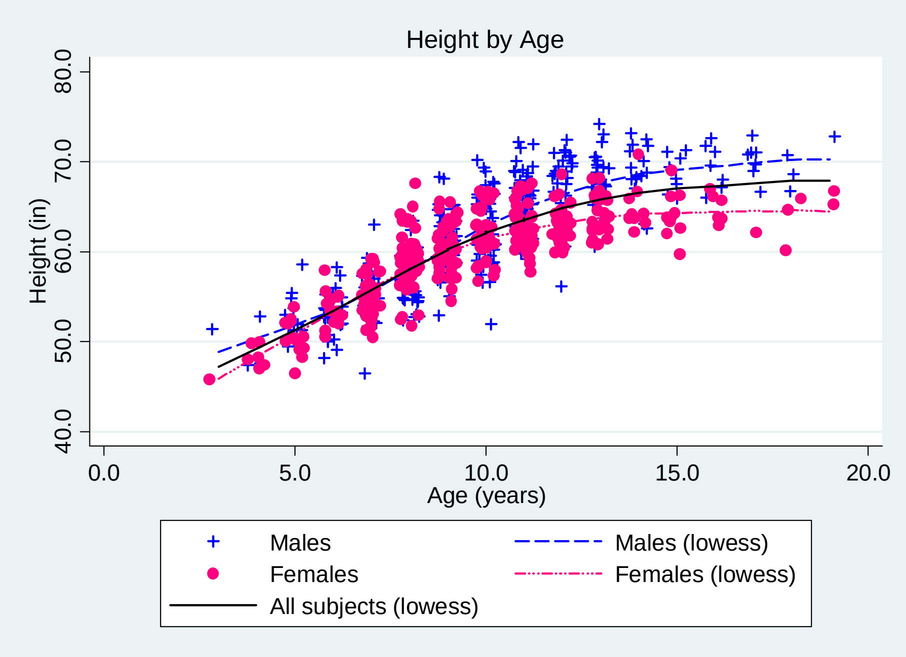
The following tables present descriptive statistics for the above variables within strata defined by subject sex, self-reported smoking behavior, and for the entire sample. There is no missing data for any variable. Descriptive statistics include the sample size (N), sample mean, standard deviation (sd), minimum (min), 25th percentile (p25), median (p50), 75th percentile (p75) and maximum (max).:

```
. tabstat age height fev, by(female) stat(n mean sd min q max) col(stat) long

female vrbl |        N   mean     sd     min       p25       p50       p75       max
------------+--------------------------------------------------------------------
0       age |      336  10.01   2.976      3         8        10        12        19
     height |      336  62.03   6.33      47        57        62      67.5        74
        fev |      336  2.812   1.004   .796     2.007     2.606     3.540     5.793
------------+--------------------------------------------------------------------
1       age |      318  9.843   2.933      3         8        10        12        19
     height |      318  60.21   4.792     46      57.5        61      63.5        71
        fev |      318  2.451   .6457   .791     1.947     2.486     2.993     3.835
------------+--------------------------------------------------------------------
Total   age |      654  9.931   2.954      3         8        10        12        19
     height |      654  61.14   5.704     46        57      61.5      65.5        74
        fev |      654  2.637   .8671   .791     1.979     2.548     3.12      5.793
---------------------------------------------------------------------------------
```

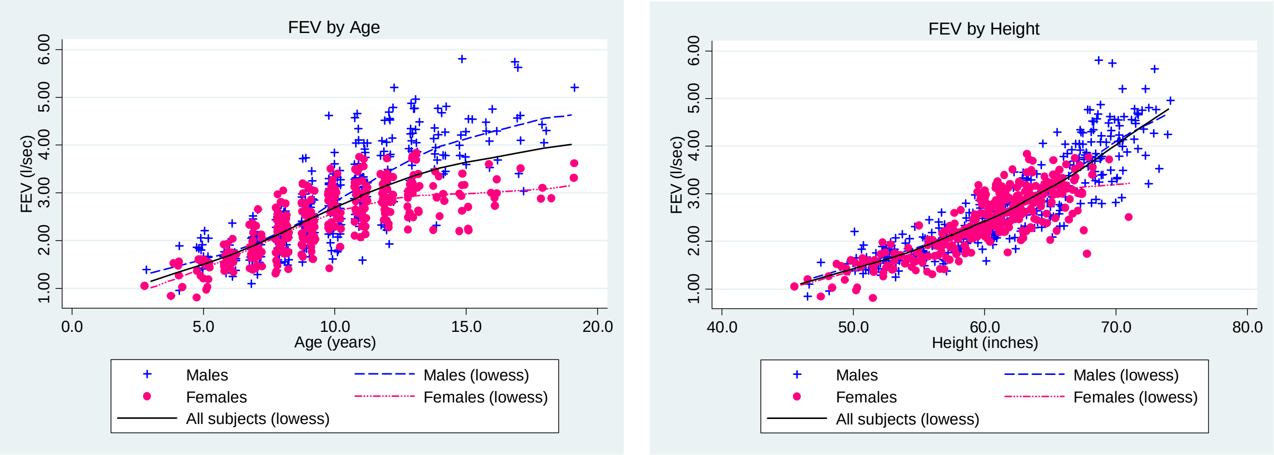
```
. tabstat age height fev, by(smoker) stat(n mean sd min q max) col(stat) long

smoker vrbl |        N   mean     sd     min       p25       p50       p75       max
------------+--------------------------------------------------------------------
0       age |      589  9.535  2.741       3         8         9        11        19
     height |      589  60.61  5.672      46        57        61      64.5        74
        fev |      589  2.566  .8505    .791      1.92     2.465     3.048     5.793
------------+--------------------------------------------------------------------
1       age |       65  13.52  2.339       9        12        13        15        19
     height |       65  65.95  3.193      58      63.5        66        68        72
        fev |       65  3.277  .7500   1.694     2.795     3.169     3.751     4.872
------------+--------------------------------------------------------------------
Total   age |      654  9.931   2.954      3         8        10        12        19
     height |      654  61.14   5.704     46        57      61.5      65.5        74
        fev |      654  2.637   .8671   .791     1.979     2.548     3.12      5.793
---------------------------------------------------------------------------------
```

<u>APPENDIX B (cont.)</u>:

Scatterplot of height versus age within sex strata (with superimposed lowess smooths by sex and overall).



Scatterplots of FEV versus age (left panel) and height (right panel) within sex strata (with superimposed lowess smooths by sex and overall).



APPENDIX C: Linear regression analyses of height by age in children who are 10 years old or younger.

```
######## MODEL C1

. regress height age if age <= 10

      Source |       SS       df       MS              Number of obs =     390
-------------+------------------------------           F(  1,   388) =  650.69
       Model |  5556.01834     1  5556.01834           Prob > F      =  0.0000
    Residual |  3313.00025   388  8.53866044           R-squared     =  0.6265
-------------+------------------------------           Adj R-squared =  0.6255
       Total |  8869.01859   389  22.7995336           Root MSE      =  2.9221

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2.282997    .089499    25.51   0.000     2.107033    2.458961
       _cons |   39.82337   .7306722    54.50   0.000      38.3868    41.25995
------------------------------------------------------------------------------
```

```
######## MODEL C2

. regress height age if age <= 10, robust

Linear regression                                      Number of obs =     390
                                                       F(  1,   388) =  761.73
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6265
                                                       Root MSE      =  2.9221

------------------------------------------------------------------------------
             |               Robust
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2.282997   .0827192    27.60   0.000     2.120363    2.445631
       _cons |   39.82337   .6410044    62.13   0.000      38.5631    41.08365
------------------------------------------------------------------------------
```

APPENDIX D: Regression analyses of self reported smoking behavior score by sex.

```
######## MODEL D1
. regress smoker female, robust

Linear regression                                      Number of obs =     654
                                                       F(  1,   652) =    3.71
                                                       Prob > F      =  0.0546
                                                       R-squared     =  0.0057
                                                       Root MSE      =  .29878
------------------------------------------------------------------------------
             |               Robust
      smoker |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
      female |   .0452606   .0235061     1.93   0.055    -.0008962    .0914173
       _cons |    .077381    .014599     5.30   0.000     .0487142    .1060477
------------------------------------------------------------------------------
```

```
######## MODEL D2
. poisson smoker female, robust

Poisson regression                                Number of obs   =        654
                                                  Wald chi2(1)    =       3.65
                                                  Prob > chi2     =     0.0560
Log pseudolikelihood = -213.37548                 Pseudo R2       =     0.0079
------------------------------------------------------------------------------
             |               Robust
      smoker |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
      female |   .4605249   .2409782     1.91   0.056    -.0117838    .9328336
       _cons |  -2.559015   .1885197   -13.57   0.000    -2.928507   -2.189523
------------------------------------------------------------------------------
```

```
######## MODEL D3
. logit smoker female

Logistic regression                               Number of obs   =        654
                                                  LR chi2(1)      =       3.75
                                                  Prob > chi2     =     0.0527
Log likelihood = -209.84678                       Pseudo R2       =     0.0089
------------------------------------------------------------------------------
      smoker |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
      female |   .5108256   .2662942     1.92   0.055    -.0111014    1.032753
       _cons |  -2.478476   .2041748   -12.14   0.000    -2.878651     -2.0783
------------------------------------------------------------------------------
```

```
. logistic smoker female

Logistic regression                               Number of obs   =        654
                                                  LR chi2(1)      =       3.75
                                                  Prob > chi2     =     0.0527
Log likelihood = -209.84678                       Pseudo R2       =     0.0089
------------------------------------------------------------------------------
      smoker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
      female |   1.666667   .4438236     1.92   0.055       .98896    2.808787
       _cons |    .083871   .0171243   -12.14   0.000     .0562105    .1251427
------------------------------------------------------------------------------
```

APPENDIX E: Regression analyses of self reported smoking behavior by age and sex.

```
######## MODEL E1 : FEMALES

. logistic smoker age

Logistic regression                               Number of obs   =        654
                                                  LR chi2(1)      =     104.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -159.28248                       Pseudo R2       =     0.2477
      smoker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
         age |   1.621967   .0894154     8.77   0.000     1.455852    1.807036
       _cons |   .0004334   .0003072   -10.92   0.000      .000108    .0017389
------------------------------------------------------------------------------
```

```
######## MODEL E2 : MALES

. logistic smoker age if female==0

Logistic regression                               Number of obs   =        336
                                                  LR chi2(1)      =      46.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -68.118663                       Pseudo R2       =     0.2555
      smoker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
         age |   1.629525   .1337891     5.95   0.000     1.387314    1.914024
       _cons |   .0002709   .0002982    -7.46   0.000     .0000313    .0023434
------------------------------------------------------------------------------
```

```
######## MODEL E3 : FEMALES
. logistic smoker age if female==1

Logistic regression                               Number of obs   =        318
                                                  LR chi2(1)      =      61.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -87.699186                       Pseudo R2       =     0.2590
      smoker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
         age |   1.669165    .132741     6.44   0.000     1.428259    1.950704
       _cons |   .0004415   .0004361    -7.82   0.000     .0000637    .0030605
------------------------------------------------------------------------------
```

```
######## MODEL E4 : Adjusted for sex
. logistic smoker age female

Logistic regression                               Number of obs   =        654
                                                  LR chi2(2)      =     111.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -155.83995                       Pseudo R2       =     0.2639
      smoker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
         age |   1.650156   .0941799     8.78   0.000     1.475516    1.845465
      female |   2.209876   .6810666     2.57   0.010     1.207909     4.04298
       _cons |   .0002296   .0001804   -10.66   0.000     .0000492    .0010711
------------------------------------------------------------------------------
```

APPENDIX E: Regression analyses of self reported smoking behavior by age and sex. (cont'd)

```
######## MODEL E5 : Adjusted for sex and a multiplicative age-sex interaction

. g ageF= age*female
. logistic smoker age female ageF

Logistic regression                                 Number of obs   =        654
                                                    LR chi2(3)      =     111.81
                                                    Prob > chi2     =     0.0000
Log likelihood = -155.81785                         Pseudo R2       =     0.2641
      smoker | Odds Ratio   Std. Err.       z    P>|z|      [95% Conf. Interval]
         age |   1.629525   .1337891     5.95   0.000      1.387314    1.914024
      female |   1.629672   2.410434     0.33   0.741      .0897626    29.58727
        ageF |   1.024326   .1170837     0.21   0.833      .8187345    1.281543
       _cons |   .0002709   .0002982    -7.46   0.000      .0000313    .0023434
------------------------------------------------------------------------------


. logit smoker age female ageF


Logistic regression                                 Number of obs   =        654
                                                    LR chi2(3)      =     111.81
                                                    Prob > chi2     =     0.0000
Log likelihood = -155.81785                         Pseudo R2       =     0.2641
      smoker |      Coef.   Std. Err.       z    P>|z|      [95% Conf. Interval]
         age |   .4882888   .0821031     5.95   0.000      .3273696     .649208
      female |   .4883787   1.479091     0.33   0.741     -2.410587    3.387344
        ageF |   .0240346   .1143032     0.21   0.833     -.1999955    .2480647
       _cons |   -8.21374   1.100836    -7.46   0.000     -10.37134   -6.056142
----------------------------------------------------------------------------------
```

APPENDIX F: Regression analyses of FEV by self-reported smoking behavior, age, and height among children ages 9 and older. In all cases, the response variable is a logarithmic transformation of FEV, and regression coefficients are exponentiated:

```
. g logfev= log(fev)


######## MODEL F1 : Unadjusted

. regress logfev smoker if age >= 9, robust eform("exp(Beta)")

Linear regression                                      Number of obs =     439
                                                       F(  1,   437) =   10.45
                                                       Prob > F      =  0.0013
                                                       R-squared     =  0.0212
                                                       Root MSE      =  .24765

------------------------------------------------------------------------------
             |               Robust
      logfev |  exp(Beta)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   1.107722   .0350638     3.23   0.001     1.040907    1.178825
       _cons |   2.881095   .0372627    81.82   0.000     2.808782     2.95527
------------------------------------------------------------------------------

######## MODEL F2 : Adjusted for age

. regress logfev smoker age if age >= 9, robust eform("exp(Beta)")

Linear regression                                      Number of obs =     439
                                                       F(  2,   436) =   82.28
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3012
                                                       Root MSE      =  .20949

------------------------------------------------------------------------------
             |               Robust
      logfev |  exp(Beta)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   .9499466   .0326613    -1.49   0.136     .8878744    1.016358
         age |   1.065661   .0054776    12.37   0.000      1.05495    1.076482
       _cons |   1.421648   .0817463     6.12   0.000     1.269728    1.591744
------------------------------------------------------------------------------
```

APPENDIX F: (cont'd)

```
######## MODEL F3 : Adjusted for height

. regress logfev smoker height if age >= 9, robust eform("exp(Beta)")

Linear regression                                      Number of obs =     439
                                                       F(  2,   436) =  378.04
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6458
                                                       Root MSE      =  .14914

------------------------------------------------------------------------------
             |               Robust
      logfev |  exp(Beta)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   .9865713   .0228252    -0.58   0.559     .9427149    1.032468
      height |   1.051825    .001985    26.77   0.000     1.047931    1.055734
       _cons |   .1155044   .0138017   -18.06   0.000     .0913281    .1460807
------------------------------------------------------------------------------
```

```
######## MODEL F4 : Adjusted for age, height

. regress logfev smoker age height if age >= 9, robust eform("exp(Beta)")

Linear regression                                      Number of obs =     439
                                                       F(  3,   435) =  278.58
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6695
                                                       Root MSE      =  .14424

------------------------------------------------------------------------------
             |               Robust
      logfev |  exp(Beta)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   .9492467   .0229181    -2.16   0.032     .9052547    .9953766
         age |   1.021596   .0035316     6.18   0.000     1.014679    1.028561
      height |   1.045851   .0021099    22.22   0.000     1.041713    1.050006
       _cons |   .1309207   .0153144   -17.38   0.000     .1040306    .1647613
------------------------------------------------------------------------------
```