Biost 517: Applied Biostatistics I
Emerson, Fall 2006

Homework #4 Key
October 28, 2006

Written problems: To be handed in at the beginning of class on Friday, October 27, 2006.

> *On this (as all homeworks) unedited Stata output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

Problem 1 makes use of the FEV and smoking data in children, problems 2 - 4 make use of the adult FEV data (adultfev.txt).

In the first four problems, you are asked to produce scatter plots with superimposed lowess smooths and/or least squares lines. The Stata function `twoway` allows you to "build" plots by overlaying

- scatterplots (which can be displayed in different colors and/or with different symbols)
- best fitting straight lines (which can be displayed in different colors and/or with different line types, e.g., solid, dashed, dotted)
- smoothed curves—we will focus most on "lowess" curves (which can be displayed in different colors and/or different line types)

As an example, the following command (which should all be typed into the Commands window prior to hitting ENTER) would produce a scatter plot of FEV (y axis) by age (x axis). On this graph, males and females would be displayed in different colors (blue is for males, pink is for females), and the lowess and least squares estimated lines for each sex would be displayed as solid and dashed lines, respectively, in the color chosen for each sex. I also include the lowess and least squares lines for the entire sample in black:

```
twoway (lowess fev age, col(black) xtitle("Age (years)")
              ytitle("FEV (l/sec)") t1("FEV by Age and Sex"))
       (lfit fev age, col(black) lp("-"))
       (scatter fev age if male==1, jitter(2) col(blue))
       (lowess fev age if male==1, col(blue))
       (lfit fev age if male==1, col(blue) lp("-"))
       (scatter fev age if male==0, jitter(1) col(pink) msymb(D))
       (lowess fev age if male==0, col(pink))
       (lfit fev age if male==0, col(pink) lp("-"))
```

The above graph is perhaps a bit busy, but I just gave all the commands so you could see what the commands do. I note that if you try to "cut and paste" the above command into a Stata window you may run into problems due to the font change of the quotation marks and the fact that the commands above have embedded "carriage returns".

In order, the subcommands to `twoway` (which are enclosed in parentheses) do the following:

1. Produce a lowess smooth of FEV on age using all the data. The lowess line will be black, and, because I did not specify a line pattern, it will be solid. I provided a label for the x-axis ("xtitle"), a label for the y-axis("ytitle"), and a title for the graph ("t1"). Note that no points are plotted by this command.

2. Produce the "best" fitting straight line for FEV on age using all the data. The "least squares fit" will be black and dashed. No points are plotted by this command.
3. Produce a scatterplot of FEV on age for males. The points will be jittered slightly. They will be plotted in blue, and, because I did not specify a symbol, they will be a solid circle.
4. Produce a lowess smooth of FEV on age for males. The lowess line will be blue, and, because I did not specify a line pattern, it will be solid.
5. Produce the "best" fitting straight line for FEV on age for males. The "least squares fit" will be blue and dashed.
6. Produce a scatterplot of FEV on age for females. The points will be jittered slightly. They will be plotted in pink, and I asked for them to be solid diamonds.
7. Produce a lowess smooth of FEV on age for females. The lowess line will be pink, and, because I did not specify a line pattern, it will be solid.
8. Produce the "best" fitting straight line for FEV on age for females. The "least squares fit" will be pink and dashed.

FEV is a 3 dimensional volume, while height is a linear dimension. Hence, we might expect FEV to be proportional to the cube of height. Because straight line relationships are easier to deal with, we often transform variables to achieve this. In this case, we can consider any of three options:

- FEV vs height^3
- FEV$^{1/3}$ vs height (i.e., the cube root of FEV vs height)
- log(FEV) vs log(height)

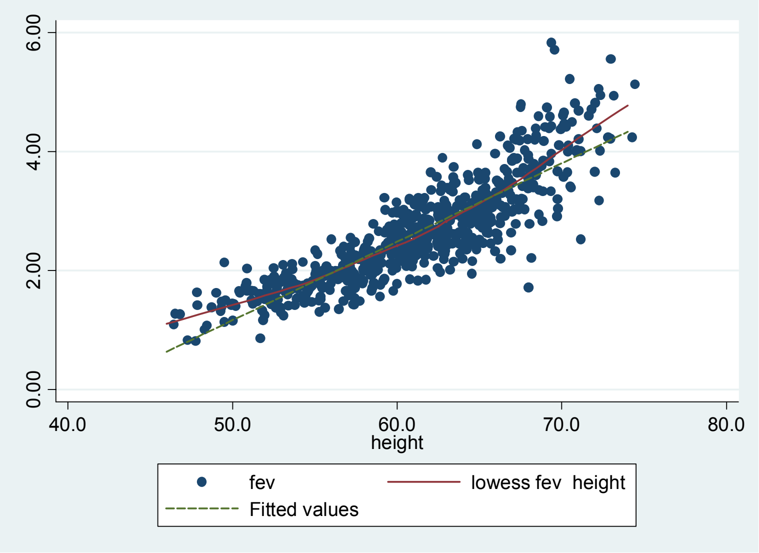
We can generate variable to be able to consider such transformations using the following Stata commands:

- `g htcub= height ^ 3`
- `g cubrtfev = fev ^ (1/3)`
- `g logfev = log(fev)`
- `g loght = log(ht)`

In Stata, the log( ) function computes the natural log, which you may have previously encountered as ln ( ). The class web pages contain a handout that reviews some of the basic properties of logarithms.

1. For this problem, use the data set used for investigating associations between smoking and lung function in children (see fev.doc and fev.txt on the class web pages). For each of the following pairs of "response" (y axis) variables and "predictor" (x axis) variables, produce a scatterplot and comment on the presence of unusual (outlying) values, whether there appears to be a linear trend in the central tendency for response across groups having different values of the predictor, whether there is any curvilinear aspect (e.g., curved, U-shaped upward or downward, S-shaped) to the trends in the data across predictor groups, and whether there appear to be trends in the variability of response across predictor groups.
    a. FEV vs height

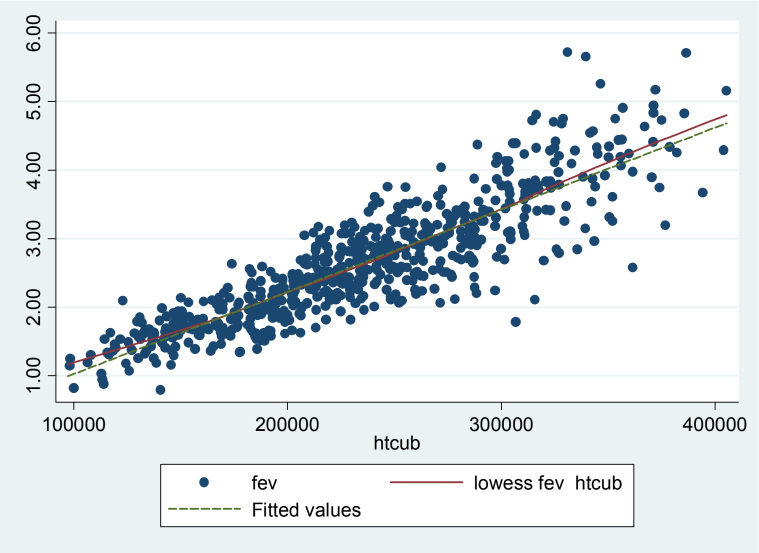
Answer:



The above scatterplot of FEV versus height (with superimposed lowess curve and least squares fit) shows no particular outlying values. There is evident an overall trend toward higher FEV in taller children. There does appear to be a curvilinear aspect to the relationship in which the taller children have even higher FEV than would have been predicted by a straight line relationship (the curve is "concave up"). There also appears to be a trend in which the variability of FEV is higher among the taller children (as assessed by looking at the range of values for two different height ranges having the same number of observations).

b. FEV vs height cubed

Answer:

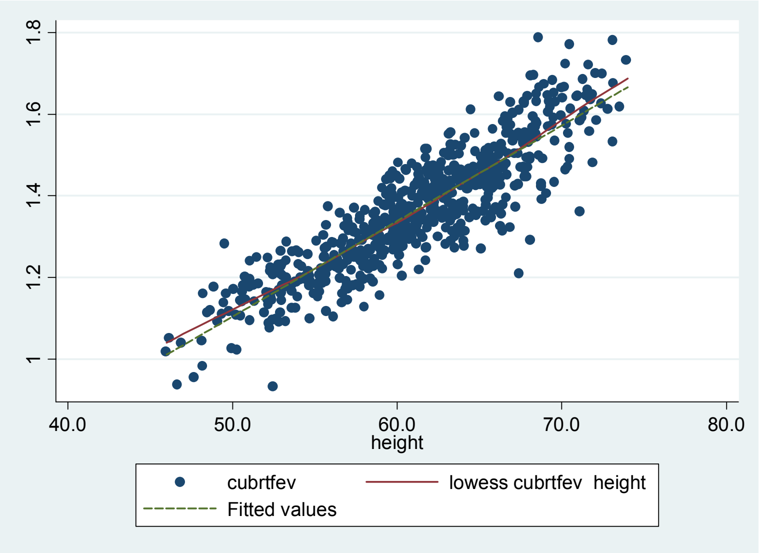


The above scatterplot of FEV versus height cubed (with superimposed lowess curve and least squares fit) shows no particular outlying values. There is evident an overall trend toward higher FEV in taller children. There does not appear to be a curvilinear aspect to

the relationship between FEV and the cube of height: The central tendency in the data seems to be very well approximated by a straight line. There also appears to be a trend in which the variability of FEV is higher among the taller children (as assessed by looking at the range of values for two different height ranges having the same number of observations).

c. cube root of FEV vs height

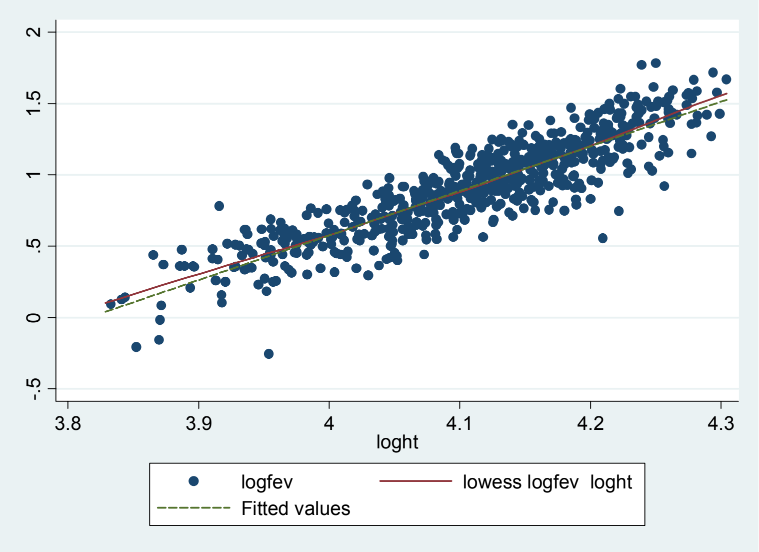
Answer:



The above scatterplot of the cube root of FEV versus height (with superimposed lowess curve and least squares fit) shows no extreme outlying values, though there is one child who is approximately 52 inches tall whose cube root of FEV might be a little lower than is the tendency in the other children. Similarly a 68 inch tall child has a slightly lower FEV than would be predicted by the bulk of the data. There is evident an overall trend toward higher cube root FEV in taller children. There does not appear to be a curvilinear aspect to the relationship between the cube root of FEV and height: The central tendency in the data seems to be very well approximated by a straight line. There does not appear to be a trend in which the variability of cube root FEV is higher among the taller children (as assessed by looking at the range of values for two different height ranges having the same number of observations).

d. log FEV vs log height

Answer:



The above scatterplot of the log of FEV versus log height (with superimposed lowess curve and least squares fit) shows no extreme outlying values, though there is one child who is approximately 52 inches tall whose log of FEV might be a little lower than is the tendency in the other children. Similarly a 68 inch tall child has a slightly lower FEV than would be predicted by the bulk of the data. There is evident an overall trend toward higher log FEV in taller children. There does not appear to be a curvilinear aspect to the relationship between the log FEV and log height: The central tendency in the data seems to be very well approximated by a straight line. There does not appear to be a trend in which the variability of log FEV is higher among the taller children (as assessed by looking at the range of values for two different height ranges having the same number of observations).

Debriefing Comments: Scientifically, FEV is a volume. Height is a linear dimension that is probably proportional to the height of the lung. Furthermore, as the width and depth of a lung is likely proportional to the height of the lung, it is quite likely that FEV is proportional to the cube of height. Algebra would say that the cube root of FEV is therefore proportional to height. We would also have by algebra that the log of FEV would be linearly related to the log of height (with slope of 3, if we could trust the algebra absolutely). Hence, the plots in parts b, c, and d should all "fix" the curvilinearity seen in part a. Of course, using height as a surrogate for lung height has some error associated with it. What ever variability there is in this error is then magnified as we cube height. We will find that transforming the response variable often works to reduce that magnification of the error. Hence, the plots in parts c and d have also removed the "heteroscedasticity" (which means "unequal variance across groups"). We tend to have greater precision when making estimates with "homoscedastic" data, so for technical reasons we often prefer to analyze data on a scale that does not have unequal variability. And for all of you who wish logarithms had never been invented, you may take my word for it that dealing with logarithms is way easier than dealing with cube roots. I will eventually come down on the side of analyzing the log FEV. This is of course tantamount to considering the geometric mean FEV rather than the mean FEV.

In problems 2-4, you are also asked to find correlations, both in the entire sample and within strata for the data related to FEV and smoking in elderly adults. Computation of correlations can be effected through the use of the Stata command `correlate` with and without the `bysort` prefix. For instance, the correlation between the FEV and age could be obtained for the entire sample and within sex strata by:

```
cor fev age
bysort male: cor fev age
```

In solving Problems 2 – 4, you should be considering the ways that correlation is influenced by the slope of a linear trend between two variables, the variance of the "predictor", and the within group variance of the "response" (where we are speaking of the variance of the "response" within groups which have identical values of the "predictor"). While it is sufficient for my purposes that you might consider these issues descriptively from the scatterplots, I note that we can also use Stata to give us numeric estimates of these quantities. For instance, if we were interested in the correlation between FEV and age, I might choose to regard FEV as the "response" and age as the "predictor" to examine:

- The correlation between FEV and age using commands as given above.
- The variance of age using `tabstat age, stat(n mean sd)` to obtain the mean and standard deviation (which is just the square root of the variance).
- The slope and within group variance of response using the linear regression command: `regress fev age`, which would generate output looking like

```
. regress fev age

      Source |       SS       df       MS              Number of obs =     725
-------------+------------------------------           F(  1,   723) =   36.82
       Model |   16.581065     1   16.581065           Prob > F      =  0.0000
    Residual |  325.577855   723  .450315152           R-squared     =  0.0485
-------------+------------------------------           Adj R-squared =  0.0471
       Total |   342.15892   724  .472595194           Root MSE      =  .67106

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0276619   .0045586    -6.07   0.000    -.0366117   -.0187122
       _cons |   4.269633   .3408604    12.53   0.000     3.600439    4.938827
------------------------------------------------------------------------------
```

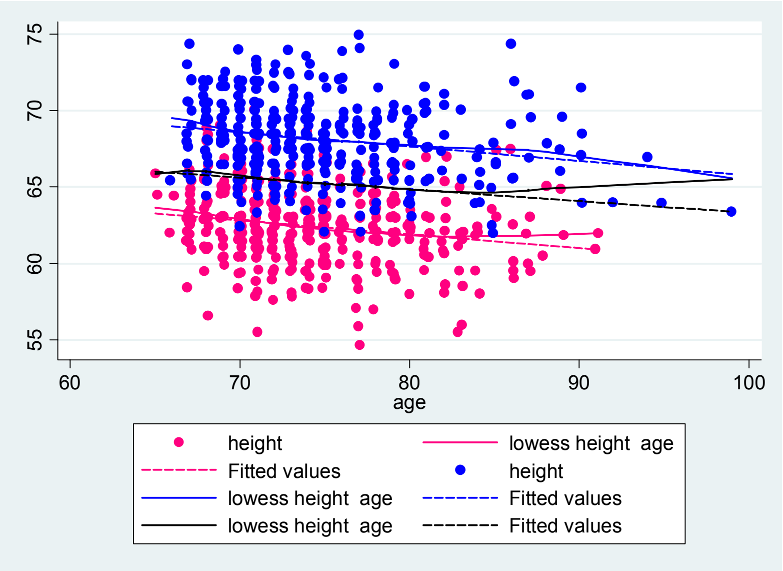
From this voluminous output, we would (at this time) be interested in only two numbers, which I have displayed in bold type. The least squares estimate of the slope is the number in the row labeled "age" (since that was the name of the variable we used as "predictor" or X variable) and column labeled "Coef." in the bottom table. The slope estimate is that FEV averages 0.0277 liters per second less for every year difference in age (with older participants tending toward lower FEV). The estimated standard deviation in each age group (people of the same age) is labeled "Root MSE", and in the above table is estimated as 0.671 l/sec. (I note that this estimates the standard deviation averaged across all ages.) We could then find Var (Y | X) as the square of the "Root MSE".

In order to get estimated slopes and within group SD for a stratified analysis, you can again use the `bysort` prefix. For instance, estimates within sex strata could be obtained by:

```
bysort male: regress fev age
```

2. For this problem, use the data set concerned with associations between FEV and smoking in elderly adults (see adultfev.doc and adultfev.txt on the class web pages). Produce a scatterplot of height (on the Y axis) versus age (on the X axis). Use a different symbol or

color for each sex, and display stratified lowess smooths on the plot. (You could also display least squares fits to be able to assess the slope of the best fitting linear trend.)



a. What is the correlation between height and age in the sample? Is this what you would expect?

Answer:

r = -0.11. A negative correlation is consistent with elderly adults “shrinking” with age due to compression of vertebral disks and/or osteoporosis.

b. What is the correlation between height and age for each sex separately?

Answer:

r = -0.21 in males and r = -0.19 in females.

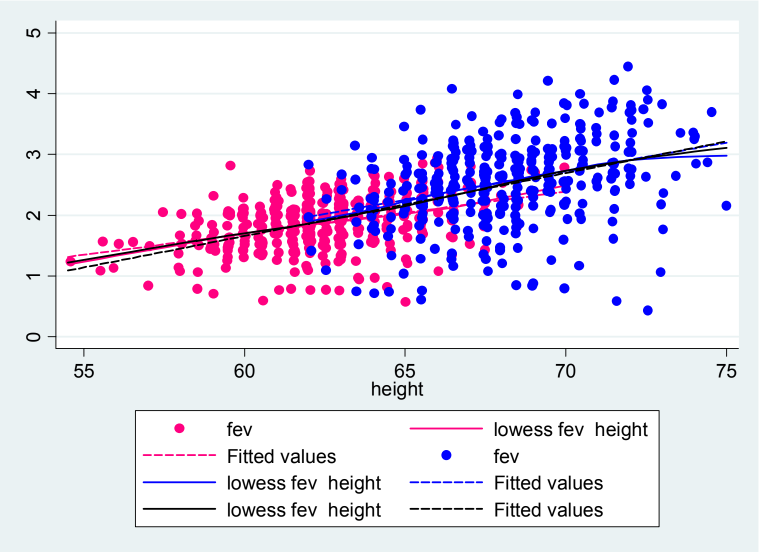
c. How do you explain any difference you observe in the answers to parts a and b? In particular, why might you expect the correlation to be less extreme in the combined sample than it was in each stratum defined by sex? Consider the statistical behavior of correlation as it relates to the slope of linear trend, the variance of the “predictor”, and the within group variance of response in groups homogeneous with respect to the “predictor”. Also consider the scientific issues that might lead to that statistical behavior.

	All Subjects	**Males**	**Females**
Correlation (r)	-0.11	-0.21	-0.19
LS slope (β)	-0.077	-0.094	-0.091
SD (Height | Age)	3.80	2.50	2.45
SD (Age)	5.45	5.64	5.26

The correlation in the combined sample is smaller than that observed in each sex separately. This happens despite the similarity of the slope of the best fitting straight line across sexes and in the combined sample and despite the similarity in the distribution of ages sampled in each sex (see above table). We can attribute the higher correlation in the

groups restricted to a single sex to the decreased variability of height within age groups. In the combined sample, we estimate that a group of adults of a single age would have a standard deviation of heights of 3.80, while within each sex that age specific standard deviation of height would be 2.45 for females and 2.50 for males. This decreased variability of the "errors" (the difference between an individual's height and the mean height for the individual's age group) leads to a more extreme correlation when the slope of the linear trend and the variability of ages is held constant. The decreased "error" variability in sex specific groups makes sense, because sex is an important determinant of height.

3. For this problem, use the data set concerned with associations between FEV and smoking in elderly adults (see adultfev.doc and adultfev.txt on the class web pages). Produce a scatterplot of FEV (on the Y axis) versus height (on the X axis). Use a different symbol or color for each sex, and display stratified lowess smooths on the plot. (You could also display least squares fits to be able to assess the slope of the best fitting linear trend.)



a. What is the correlation between FEV and height in the sample? Is this what you would expect?

Answer:

r = 0.58. A positive correlation is consistent with the larger lung size expected with larger body size.

b. What is the correlation between FEV and height for each sex separately?

Answer:

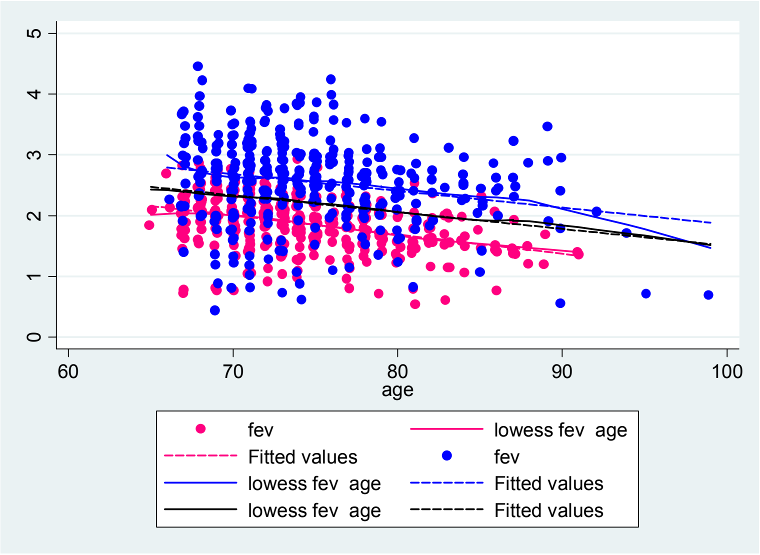
r = 0.34 in males and r = 0.39 in females.

c. How do you explain any difference you observe in the answers to parts a and b? In particular, why might you expect the correlation to be more extreme in the combined sample than it was in each stratum defined by sex? Consider the statistical behavior of correlation as it relates to the slope of linear trend, the variance of the "predictor", and the within group variance of response in groups homogeneous with respect to the "predictor". Also consider the scientific issues that might lead to that statistical behavior.

	All Subjects	Males	Females
Correlation (r)	0.58	0.34	0.39
LS slope (β)	0.104	0.094	0.069
SD (FEV | Height)	0.56	0.67	0.41
SD (Height)	3.82	2.56	2.41

The correlation in the combined sample is larger than that observed in each sex separately. This happens despite the similarity of the slope of the best fitting straight line across sexes and in the combined sample. We can attribute the higher correlation in the groups restricted to a single sex to the decreased variability of height in the sample restricted to a single sex compared to that in the combined sample. In the combined sample, the standard deviation of height is 3.82, while within each sex the standard deviation of height is 2.56 for males and 2.41 for females. There are slight differences in the least squares slopes, as well as larger differences in the standard deviation of the "error" distribution (the difference between an individual's FEV and the mean FEV for their height). These combine to make the correlation with females slightly higher than that for males. The decreased height variability in sex specific groups makes sense, because sex is an important determinant of height.

4. For this problem, use the data set concerned with associations between FEV and smoking in elderly adults (see adultfev.doc and adultfev.txt on the class web pages). Produce a scatterplot of FEV (on the Y axis) versus age (on the X axis). Use a different symbol or color for each sex, and display stratified lowess smooths on the plot. (You could also display least squares fits to be able to assess the slope of the best fitting linear trend.)



a. What is the correlation between FEV and age in the sample? Is this what you would expect?

Answer:

r = -0.22. A negative correlation is consistent with elderly adults "shrinking" with age due to compression of verterbral disks and/or osteoporosis (and thus having lower lung volume), as well as with loss of lung function with the aging process .

b. What is the correlation between FEV and age for each sex separately?

Answer:

r = -0.22 in males and r = -0.37 in females.

c. How do you explain any difference you observe in the answers to parts a and b? In particular, why might you expect the correlation to be less extreme in the combined sample than it was in each stratum defined by sex? What might explain the difference in correlations observed for men and women? Consider the statistical behavior of correlation as it relates to the slope of linear trend, the variance of the "predictor", and the within group variance of response in groups homogeneous with respect to the "predictor". Also consider the scientific issues that might lead to that statistical behavior.

	All Subjects	**Males**	**Females**
Correlation (r)	-0.22	-0.22	-0.37
LS slope (β)	-0.028	-0.028	-0.032
SD (Height | Age)	0.67	0.70	0.41
SD (Age)	5.45	5.64	5.26

The correlation in the combined sample is smaller than that observed in females, but not males separately. This happens despite the similarity of the slope of the best fitting straight line across sexes and in the combined sample and despite the similarity in the distribution of ages sampled in each sex (see above table). We can attribute the higher correlation in the groups restricted to a single sex to the decreased variability of FEV within age groups in women, though males have the same variability of FEV within age groups as does the combined samples. In the combined sample, we estimate that a group of adults of a single age would have a standard deviation of FEV of 0.67, while when restricted to a single sex that age specific standard deviation of FEV would be 0.41 for females and 0.70 for males. This decreased variability of the "errors" (the difference between an individual's height and the mean height for the individual's age group) in females leads to a more extreme correlation when the slope of the linear trend and the variability of ages is held constant. I do not have a definitive explanation for the greater variability of FEV within age groups for males than for females. Perhaps this reflects patterns of lung disease secondary to smoking.

The following problems make use of a dataset from a clinical trial exploring the therapeutic value of D-penicillamine in the treatment of primary biliary cirrhosis (see liver.doc and liver.txt on the class web pages.) I note that the data set contains data on 106 patients who were screened for the clinical trial, but who were not randomized for one reason or another. These patients have missing data for the variable indicating treatment arm.

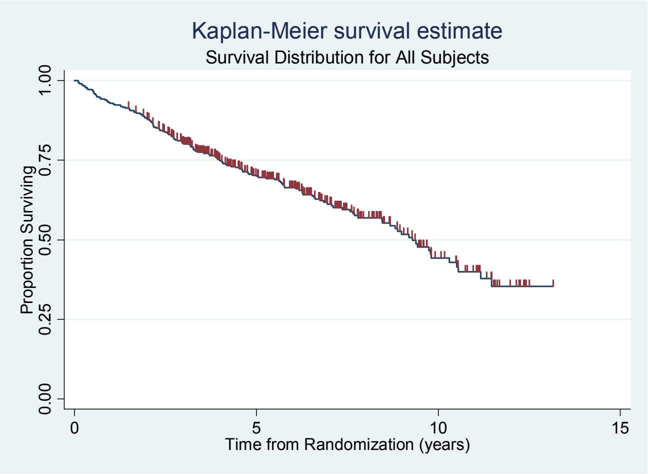
Recall that when analyzing censored data, descriptive statistics are obtained in Stata using its facility for Kaplan-Meier estimation:

- The variable `obstime` contains (right) censored observations of the time from accrual to the study to death or censoring according to the value of variable `status.`

- Note that variable `obstime` is measured in days in this dataset. You might find it more convenient to measure survival in weeks, months, or years. Using the "`replace`" command in Stata, you can easily obtain this. For instance, you can choose one of the following commands
  1. (weeks): `replace obstime= obstime / 7`
  2. (months): `replace obstime= obstime / 30.4`
  3. (years): `replace obstime= obstime / 365.25`
- You will need to declare the variables representing the possibly censored times to death: `stset obstime status`
- To obtain a graph of survival curves, you can then just use `sts graph`. (If you want stratified curves by, say, sex, you use the `by( )` option: `sts graph, by(sex)`.)
- To obtain numeric output of the estimated survivor function you use `sts list` with or without the `by( )` option. If you only want the survivor function at specific times, you can use the `at( )` option, as well. For instance, if your observation time were measured in months, the 6 month and 15 month survival probabilities would be obtained by `sts list, at(6 15)`.

5. We are interested in estimating the probability of a patient's survival following accrual to the study.

   a. Provide suitable descriptive statistics for the distribution of times to death among all patients with available data.

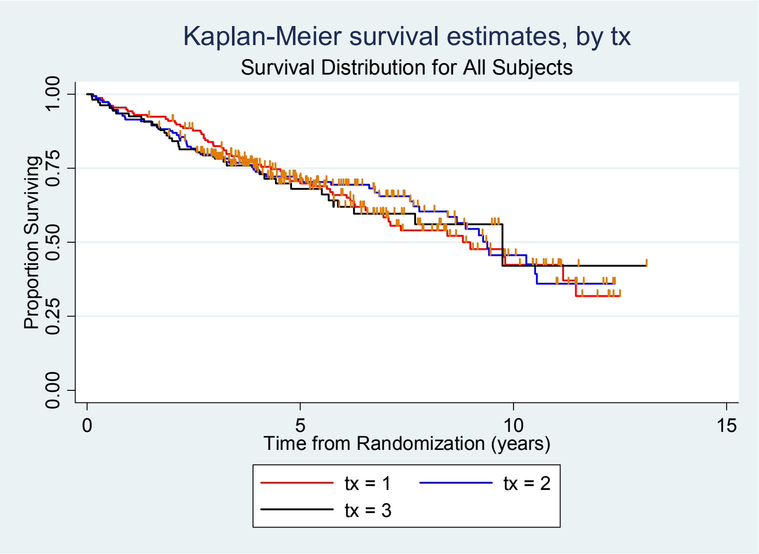


	Initially At Risk	Total Number Of Deaths	75th %ile Surv Distn	50th %ile Surv Distn	25th %ile Surv Distn	3 Year Surv Prob	6 Year Surv Prob	9 Year Surv Prob
Combined	418	161	4.005	9.295	NA	0.801	0.664	0.517
D-penicill	157	65	4.315	8.986	NA	0.825	0.661	0.477
Placebo	153	60	3.907	9.385	NA	0.790	0.694	0.545
Unrndmzd	108	36	4.003	9.749	NA	0.783	0.621	0.562

The above graph and table provide descriptive statistics of the survival distribution. In order to allow an estimate of the amount of data contributing to the estimates, the table includes the number of subjects initially at risk in each group, as well as the total number of

events observed during the study. Note that the length of follow-up does not allow estimation of the time at which only 25% of subjects are expected to still be alive.

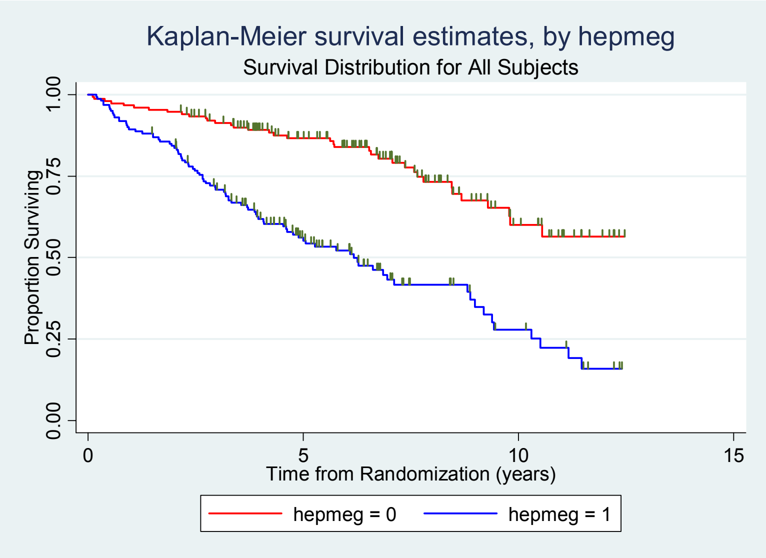
b. Produce a plot of survival curves by the groups defined by treatment group (D-penicillamine, placebo, or not randomized). (Note that you will have to assign a code to the group not randomized, as Stata will not otherwise plot a separate curve for the cases missing a value for variable `treatment`.) Produce a table of estimates of the times at which 75%, 50%, and 25% of the subjects are estimated to still be surviving. Also include a table of the estimated probabilities of surviving 3, 6, or 9 years for each stratum. Are the estimates suggestive that the patients randomized to the study are similar to those not randomized? Are the estimates suggestive that treatment with D-penicillamine affects survival? Give descriptive statistics supporting your answer.



From the above plot, we see that the distribution of survival times is quite similar across the treatment arms for the randomized clinical trial. We also see that the subjects not randomized have a distribution of survival times quite similar to that for the placebo arm (as well as the D-penicillamine arm). For instance, the median survival time is estimated to be 8.99 years with D-penicillamine, 9.39 years for placebo, and 9.75 years for patients not randomized to the clinical trial. The probability of surviving 3, 6, or 9 years is also quite similar across the treatment arms and unrandomized group as shown in the table given in the answer to part a. (Note that the problem asked for the estimated time at which only 25% of subjects are still expected to be surviving, but that is not estimable with the available data. This is not an unusual setting in the presence of censored data. We often cannot estimate the descriptive statistics we might usually want. We would only have been able to estimate the mean if we had estimated survival curves dropping to 0.)

6. For this problem, we will ignore treatment. We are primarily interested in possible associations between survival and hepatomegaly.

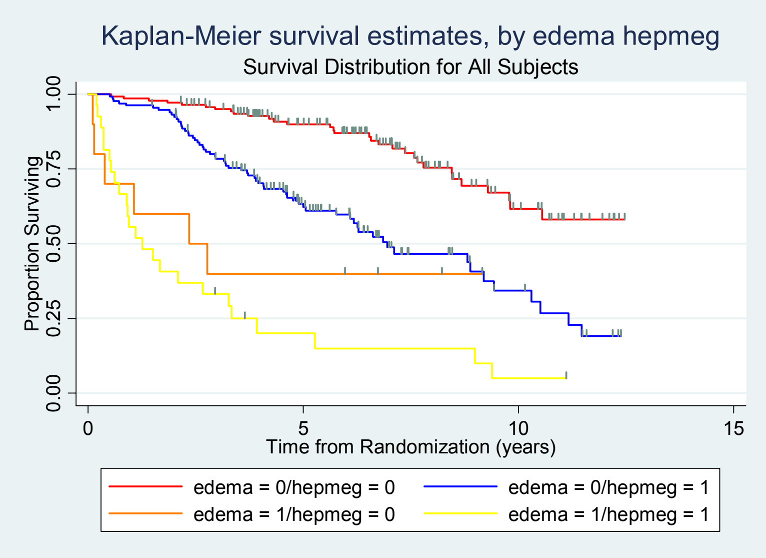
a. Produce a plot of survival curves by the groups defined by presence or absence of hepatomegaly (two stratified curves, one for each group). Produce a table of estimates of the times at which 75%, 50%, and 25% of the subjects are estimated to still be surviving. Also include a table of the estimated probabilities of surviving 3, 6, or 9 years for each stratum. Are the estimates suggestive of an association between prevalence of hepatomegaly and survival? Give descriptive statistics supporting your answer.



	Initially At Risk	Total Number Of Deaths	75th %ile Surv Distn	50th %ile Surv Distn	25th %ile Surv Distn	3 Year Surv Prob	6 Year Surv Prob	9 Year Surv Prob
No Hepmeg	152	37	7.655	NA	NA	0.913	0.839	0.676
Hepmegaly	160	88	2.658	6.177	10.51	0.709	0.522	0.348

The above graph and table provide descriptive statistics of the survival distribution within groups defined by absence or presence of hepatomegaly. Again, the number of subjects initially at risk and the total number of events are included in order to judge the amount of information available in the data. It is clear that subjects with hepatomegaly have worse survival: 83.9% of subjects without hepatomegaly are estimated to survive 6 years, while only 52.2% of subjects with hepatomegaly are estimated to survive 6 years. This difference in survival distributions is also apparent as we consider different timeframes.

b. Suppose we are interested in whether hepatomegaly is associated with survival beyond that which might be due to possible confounding by presence of edema. Perform an analysis to see whether edema might confound your analysis in part a. Is edema a confounder? Give descriptive statistics supporting your answer.



	Initially At Risk	Total Number Of Deaths	75th %ile Surv Distn	50th %ile Surv Distn	25th %ile Surv Distn	3 Year Surv Prob	6 Year Surv Prob	9 Year Surv Prob
No Edema No Hepmg	142	31	8.449	NA	NA	0.950	0.869	0.695
No Edema, Hepmeg	133	64	3.551	6.954	11.17	0.785	0.598	0.407
Edema No Hepmg	10	6	0.383	2.561	NA	0.400	0.400	0.400
Edema, Hepmeg	27	24	0.523	1.259	3.629	0.333	0.150	0.100

The above graph and table provide descriptive statistics of the survival distribution within groups defined by edema and hepatomegaly. By comparing these descriptive statistics with those in part a, we can see evidence that edema did in fact confound the association between survival and hepatomegaly. For instance, if we use the 6 year survival probability as our measure of association, we see that comparing those with edema to those without edema in subjects who have hepatomegaly, there is clearly an association between survival and edema: 59.8% 6 year survival with no edema, 15% survival if edema is present. We further can see that there is an association between edema and hepatomegaly in this sample: Of 275 subjects with no edema, 133 (or 48.6%) have hepatomegaly, while of the 37 subjects with edema, 27 (or 73.0%) have hepatomegaly. Now within those with no edema, presence of hepatomegaly is associated with 27.1% lower 6 year survival (59.8% with hepatomegaly and 86.9% with no hepatomegaly), and within those who do have edema, presence of hepatomegaly is associated with 25.0% lower 6 year survival (15.0% with hepatomegaly and 40.0% with no hepatomegaly). An "edema adjusted" estimate of the association between hepatomegaly and survival might then use some weighted average of the 25.0% and 27.1% observed differences. (Compare this with the "unadjusted" estimated difference of 31.7%.)