

Biost 517: Applied Biostatistics I

Emerson, Fall 2007

Homework #4 Key

October 29, 2007

Written problems: To be handed in at the beginning of class on Monday, October 29, 2007.

*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.)*

Problems 1-4 makes use of the dataset related to estimating “normal” ranges for somatosensory evoked potentials (SEP) in healthy adults. (SEP.txt, with documentation in SEP.doc).

You will need to generate a new variable *p60* to represent the average of the measurements made using the left and right ankle for each individual. The following Stata code can be used to create this variable:

```
g p60= (p60R + p60L) / 2
```

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

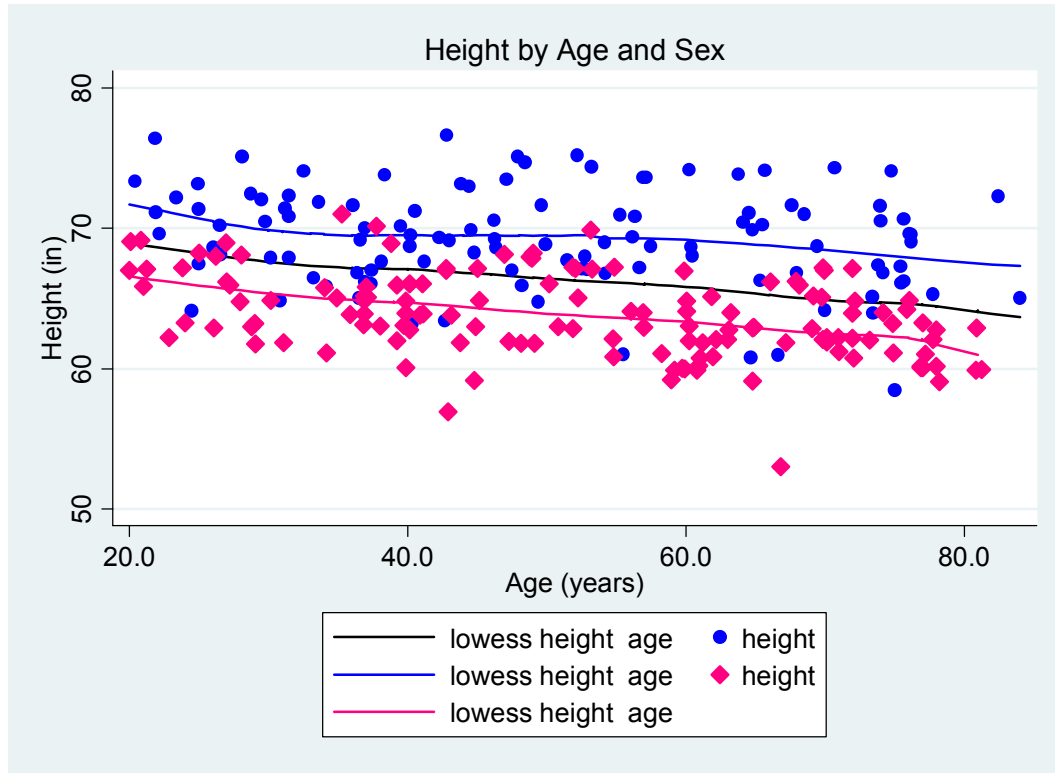
- scatterplots (which can be displayed in different colors and/or with different symbols)
- best fitting straight lines (which can be displayed in different colors and/or with different line types, e.g., solid, dashed, dotted)
- smoothed curves—we will focus most on “lowess” curves (which can be displayed 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 *p60* (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 p60 age, col(black) xtitle("Age (years)"))
      ytitle("p60 (msec)") t1("Time to p60 SEP by Age and Sex"))
      (lfit p60 age, col(black) lp("-"))
      (scatter p60 age if sex==1, jitter(2) col(blue))
      (lowess p60 age if sex==1, col(blue))
      (lfit p60 age if sex==1, col(blue) lp("-"))
      (scatter p60 age if sex==0, jitter(1) col(pink) msymb(D))
      (lowess p60 age if sex==0, col(pink))
      (lfit p60 age if sex==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”.

1. Produce a scatterplot of height (y axis) versus age (x axis), using a different symbol and color for each sex. Also display lowess curves for the entire sample as well as for each group separately.



- a. 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.

Ans: There does not appear to be any really extreme outlier, though there is one subject who is only 53 inches tall, which is noticeably smaller than others in her age cohort. The general tendency is for women to be shorter than men (the lowess smooth for women is below that for men), and for older subjects to be shorter than younger subjects (the slope across age groups is negative). A similar slope is observed for both men and women, and the curve would be well approximated by a straight line. (There is a very slight hint toward greater separation of the curves for the sexes at older ages, though this is not all that striking.) The variability of height measurements within age groups is fairly constant.

- b. These data represent cross-sectional sampling over from a population of healthy adults. Describe three distinct scientific mechanisms that might explain any linear trends in the data. (You need not restrict yourself to mechanisms that are known to be valid.)

Ans: Possibilities include

- that people shrink as they get older,
- that people born earlier in the century did not ever grow to as great a height as people born later in the century (under this hypothesis, we would expect that 20 year olds in 1940 were shorter than 20 year olds in 1990), or
- that taller people die earlier than shorter people, and thus the tendency for older people to be shorter reflects “survivorship” of the people who never grew as tall.

- c. Of the mechanisms that you listed in part b, which do you believe to be the most likely? What evidence is present in your data to support your belief?

Ans: In a cross-sectional study, we do not have any strong evidence to distinguish among the above hypotheses, though I will note that the fact that we do not appear to have too many short 20 year olds suggests that the third hypothesis is not as likely. From longitudinal studies (so repeat measurements on the same people), we do know that people shrink with age. However, we have also observed that the average height of 20 year olds has increased over the years.

In problems 2-4, you are also asked to find correlations, both in the entire sample and within strata. 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 `p60` and `age` could be obtained for the entire sample and within sex strata by:

```
cor p60 age
bysort sex: cor p60 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 `p60` and `age`, I might choose to regard `p60` as the “response” and `age` as the “predictor” to examine:

- The correlation between `p60` and `age` using commands as given above.
- The variance of `age` using `tabstat p60, 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 p60 age`, which would generate output looking like

```
. regress p60 age
```

Source	SS	df	MS			
Model	1081.52731	1	1081.52731	Number of obs =	250	
Residual	4112.46342	248	16.5825138	F(1, 248) =	65.22	
Total	5193.99073	249	20.8594005	Prob > F =	0.0000	
				R-squared =	0.2082	
				Adj R-squared =	0.2050	
				Root MSE =	4.0722	

p60	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1207971	.0149576	8.08	0.000	.0913369	.1502573
_cons	55.70264	.8078069	68.96	0.000	54.1116	57.29368

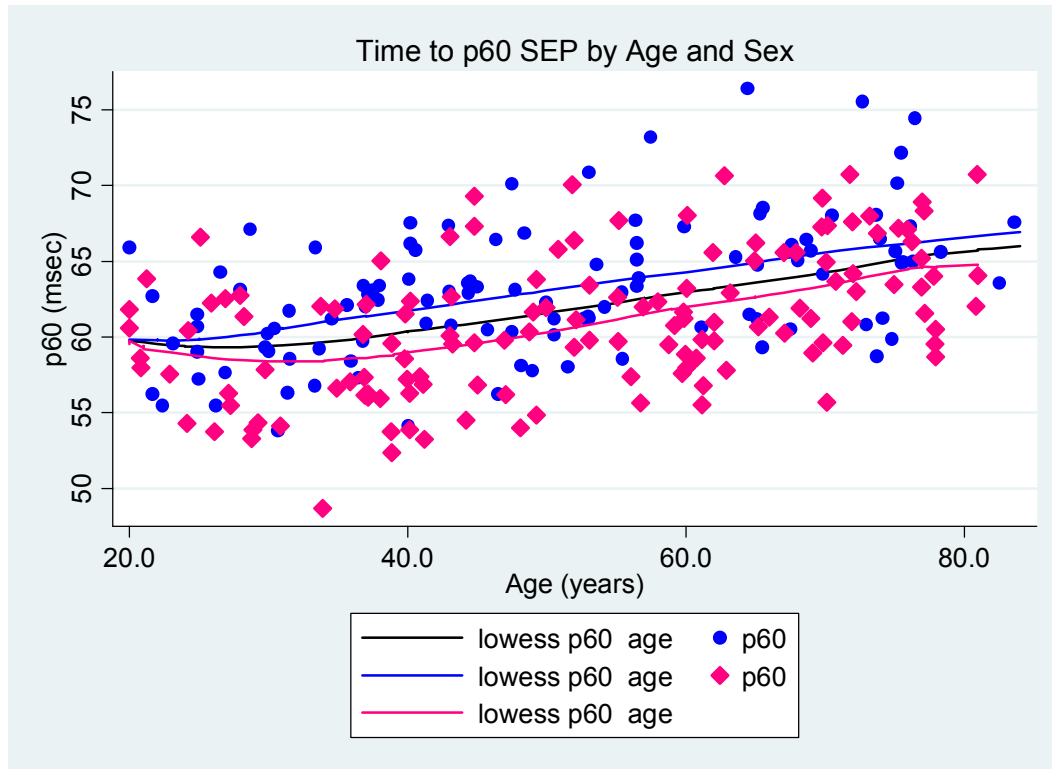
From this voluminous output, we would (at this time) be interested in only two numbers, which I have displayed in bold type. The least squares estimate of the slope is the number in the row labeled “age” (since that was the name of the variable we used as “predictor” or X variable) and column labeled “Coef.” in the bottom table. The slope estimate is that `p60` averages 0.1208 msec more for every year difference in age (with older participants tending toward higher `p60`). 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 4.0722 sec.

(I note that this estimates the standard deviation averaged across all ages.) We could then find $\text{Var}(Y | X)$ 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 sex: regress p60 age
```

2. Produce a scatterplot of p60 (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.)



Ans: Both men and women appear to have been sampled over similar ranges of age. There does not appear to be any really extreme outlier, though there is one approximately 35 year old subject who has a noticeably smaller SEP time than others in her age cohort. The general tendency is for women to have shorter SEP delays than men (the lowess smooth for women is below that for men), and for older subjects to have longer delays than younger subjects (the slope across age groups is positive). A similar slope is observed for both men and women, and the curve would be well approximated by a straight line, except possibly at the youngest age groups. The variability of p60 measurements within age groups shows a slight hint toward greater variability with increasing age, though this is admittedly in the eye of the beholder.

- a. What is the correlation between p60 and age in the sample? Is this what you would expect? Why?

Ans: There is a positive correlation of 0.456 suggesting a trend toward longer SEP in older subjects. Such might be consistent with an aging process that slows nerve conduction.

- b. What is the correlation between p60 and age for each sex separately?

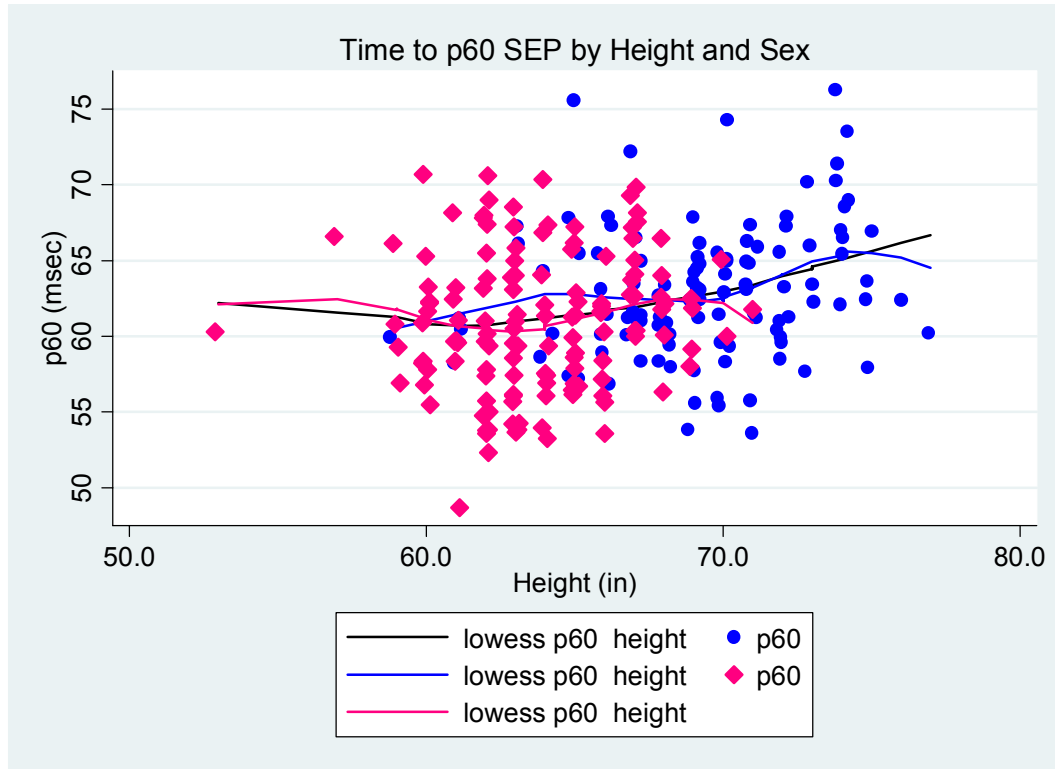
Ans: Relative to the combined sample, there is a slightly higher positive correlation of 0.491 in males and 0.482 in females.

- 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.

Ans: The table below presents relevant descriptive measures for the combined sample, as well as for each sex separately. Correlation would tend to be higher in absolute value for samples having a higher SD of age, a lower SD of p60 within age groups, and a more extreme slope. In this case, the variance of age is approximately the same in the combined sample and in each sex, so that does not contribute to the higher correlation in the sex strata. There is a slight tendency toward smaller variance of p60 within age groups when the analysis is restricted to a single sex. This is consistent with there being a tendency for males to have longer SEPs than women, because restricting an analysis to a single sex group would remove some variability due to the sex-p60 association. There is also a very slightly higher estimated slope in the groups restricted to a single sex

	All Subjects	Males	Females
Correlation (r)	0.456	0.491	0.482
LS slope (β)	0.121	0.125	0.126
SD (p60 Age)	4.07	3.84	3.98
SD (Age)	17.3	17.2	17.3

3. Produce a scatterplot of p60 (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.)



Ans: Men and women appear to have been sampled over dissimilar ranges of height. This is consistent with a known tendency for men to be taller than women. There does not appear to be any really extreme outlier, though there is one approximately 53 inch tall woman whose p60 measurement may be influential in determining the lowess curve for females. The general tendency is for men and women to have similar SEP delays for the same height (the lowess smooths are largely coincident to my eye), and for taller subjects to have longer delays than shorter subjects (the slope across height groups is positive). Apart from the shortest subject, a similar slope is observed for both men and women, and the curve would be well approximated by a straight line. The variability of p60 measurements within height groups shows a slight hint toward greater variability with shorter stature, though this is admittedly in the eye of the beholder.

- a. What is the correlation between p60 and height in the sample? Is this what you would expect?

Ans: There is a positive correlation of 0.260 suggesting a trend toward longer SEP in taller subjects. Such might be consistent with the fact that a longer nerve should be associated with a longer delay, if nerve conduction velocities are held constant.

- b. What is the correlation between p60 and height for each sex separately?

Ans: Relative to the combined sample, there is a slightly lower positive correlation of 0.207 in males and a markedly lower 0.110 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.

Ans: The table below presents relevant descriptive measures for the combined sample, as well as for each sex separately. Correlation would tend to be higher in absolute value for samples having a higher SD of height, a lower SD of p60 within height groups, and a more extreme slope. In this case, the variance of p60 within height groups is approximately the same no matter whether we consider the combined groups or the individual sex strata, so this aspect is not likely to be the cause of the different correlations. On the other hand, the variability of height is less in each of the sex strata (thereby leading to lower correlation) and the slope is lower for females than it is in males or the combined sample. The first of these explains the lower correlation in males relative to the combined sample, and both of these aspects will contribute to an even lower correlation among females than males.

	All Subjects	Males	Females
Correlation (r)	0.260	0.207	0.110
LS slope (β)	0.282	0.258	0.170
SD (p60 Height)	4.42	4.31	4.51
SD (Height)	4.2	3.5	2.9

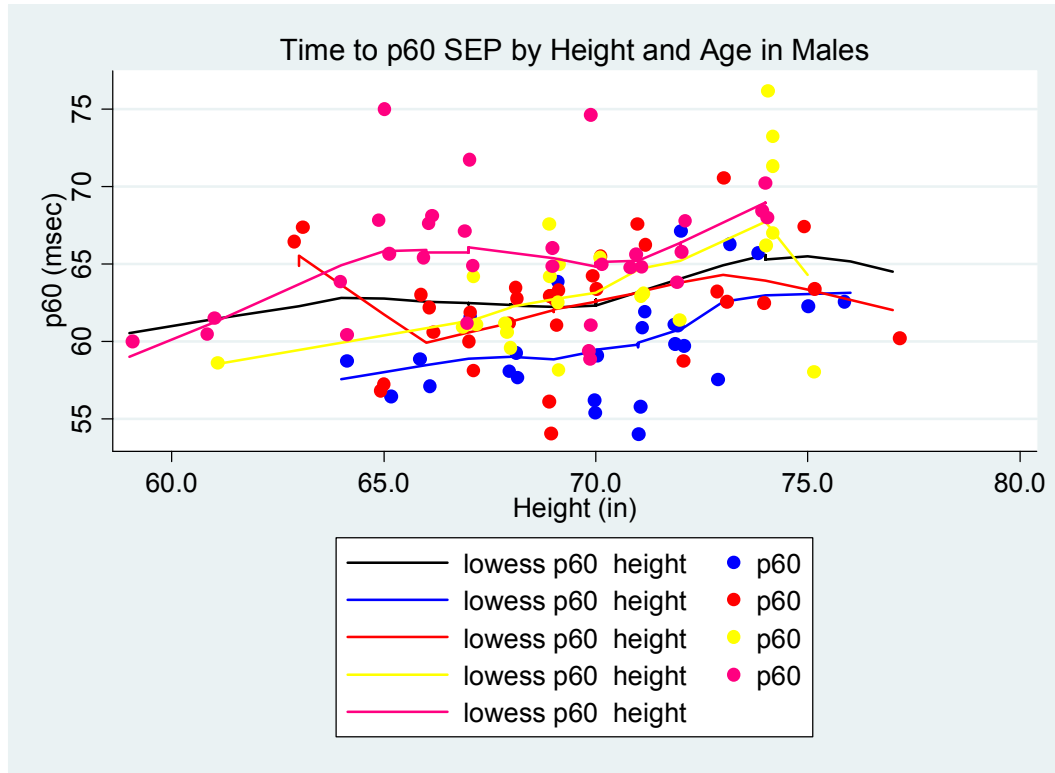
4. For this problem, you will need to create a variable indicating age within the categories 20 – 34, 35 - 49, 50 – 64, and 65 – 84. The following Stata commands can be used to create such a variable:

```
g agectg= age
recode agectg 20/34=27 35/49=42 50/64=57 65/84=74
```

Note that I used a coding that indicates the midpoint of each of the ranges. You might examine the descriptive statistics (mean, median) for age within each of these groups in order to check that my coding is at all reasonable as a description of the central tendency for age in each group. You can do this using the Stata command:

```
bysort agectg: tabstat age, stat(n mean sd min med max)
```

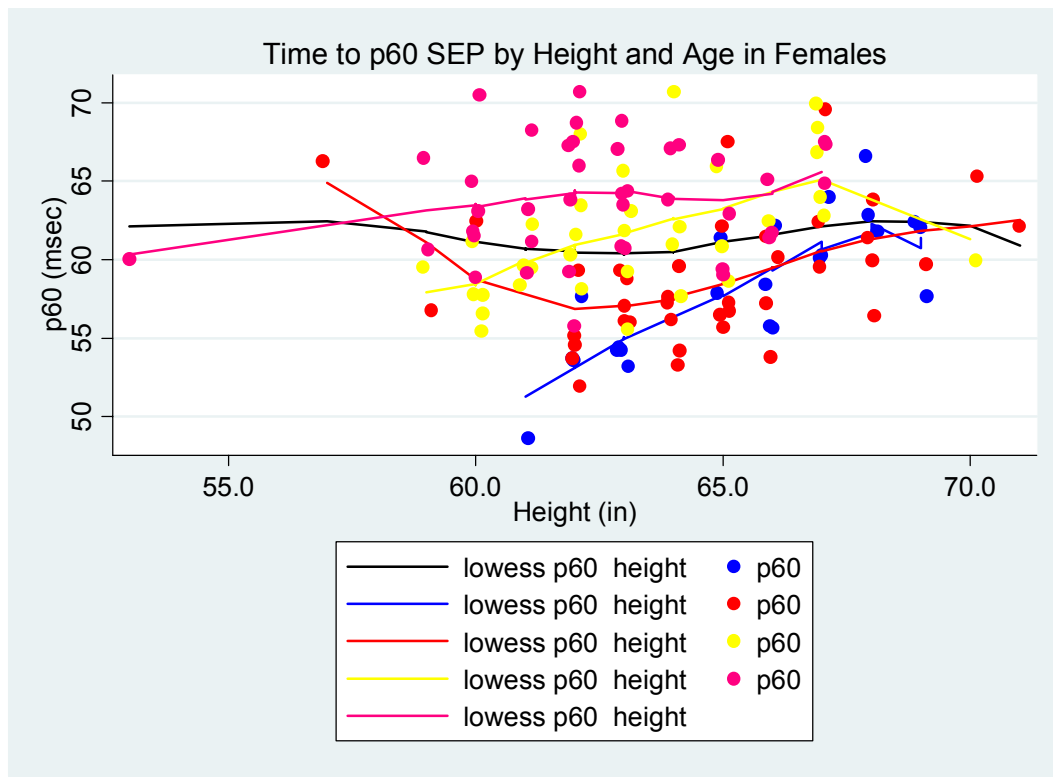
- a. For each sex separately, produce a scatterplot of p60 (on the Y axis) versus height (on the X axis). Use a different symbol or color for each age category, and display stratified lowess smooths on the plot. You could also display least squares fits to be able to assess the slope of the best fitting linear trend. (Note that when we are able to consider the description of the relationship between p60 and height, age, and sex by using the two graphs.) Does the relationship between p60 and height differ across the age and sex strata? Do any such differences seem biologically plausible?



Ans: The above plot for men shows lowess curves that are not as smooth as they might be with a larger sample size. None the less, there is a clear trend toward longer p60 delays in taller men for every age stratum. Furthermore, the vertical separation of the lowess curves for the strata suggests a tendency for older men to have longer p60 delays than younger men of the same height (the strata are ordered from youngest to oldest: blue red yellow gray). Judging whether the curves are parallel is hard when the lowess curves are this “wiggly”, but I note that the separation between the lowest and highest curves is not that different for the shorter heights and the taller heights. Thus, my overall impression is that there is no great tendency for age to modify the effect of height on p60 delays in males.

The plot for women (shown below) again shows fairly “wiggly” lowess curves. Furthermore, there does seem to be a single woman in the second lowest age stratum (represented in red here) that pulls the lowess curve up. Otherwise, the lowess curves would look reasonably straight with a positive slope in each stratum, though they do have different slopes (note the smaller separation between the youngest and oldest age strata among taller women relative to the corresponding separation among shorter women). This lack of similarity across the age strata suggest that age does modify the association between p60 and height in women.

Because there is no huge age-height interaction in males, but there is in women, this suggests that there is a three-way interaction between height, age, and sex in the p60 delay. This actually makes scientific sense: Longer nerves should be associated with longer time nerve conduction times, as might aging. But height is merely a surrogate for nerve length, and short, old women (who might suffer from skeletal compression due to osteoporosis) might have much longer nerves than their height would suggest. Men do not suffer as much from osteoporosis, so there is not as much of an age-height interaction for them.



b. What is the correlation between p60 and height in the sample? Is this what you would expect?

Ans: There is a positive correlation of 0.260 suggesting a trend toward longer SEP in taller subjects. Such might be consistent with the fact that a longer nerve should be associated with a longer delay, if nerve conduction velocities are held constant.

c. What is the correlation between p60 and height for each age category separately?

Ans: Relative to the combined sample, there is a higher positive correlation in each of the age strata, as shown in the table below.

d. How do you explain any difference you observe in the answers to parts a and b? Identify how differences in the distribution of heights across age groups, differences in the slope of p60 versus height across age groups, and differences in the within height variation of p60 across age groups might contribute to these differences in correlations.

Ans: As in problems 2 and 3, we can explain differences in the stratum specific correlations relative to the combined sample by examining for each stratum the LS slope, the variability of p60 within height groups, and the variability of height. In this problem, the major explanation for the discrepancies in correlation probably relates to the differences in the p60-height slopes across the age strata. And as noted above, this is probably driven by the prevalence of osteoporosis in the older women, though I do not really have that data available to be able to confirm the hypothesis.

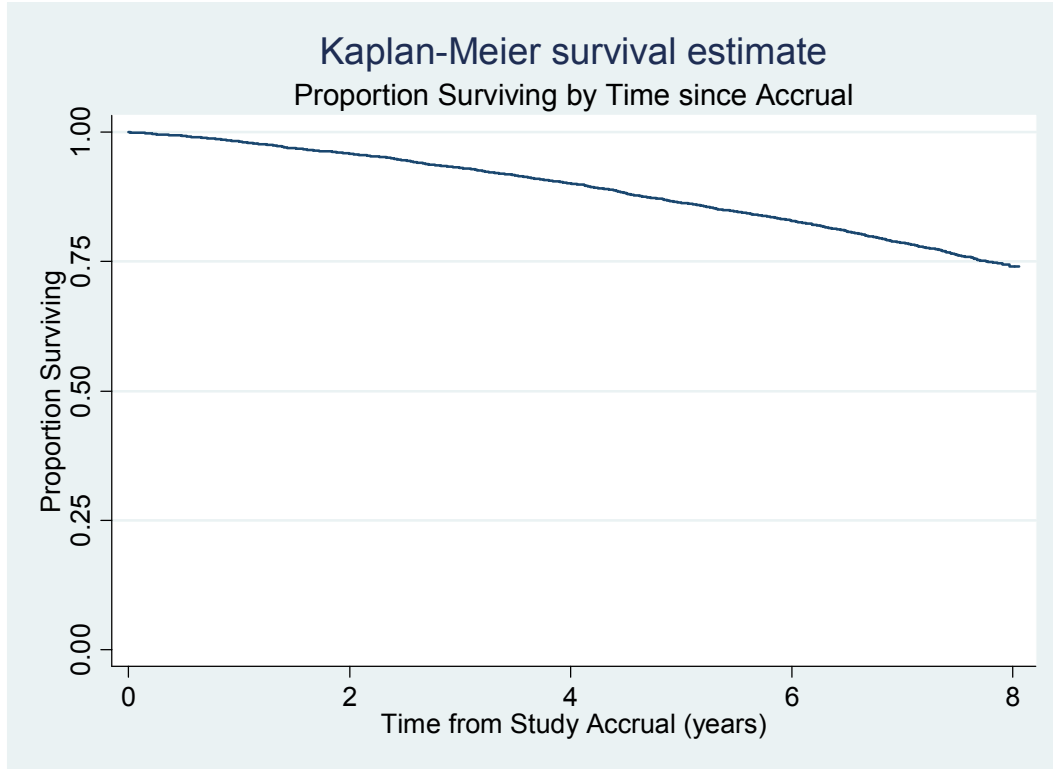
	All Subjects	20-34 yo	35-49 yo	50-64 yo	65-84yo
--	--------------	----------	----------	----------	---------

Correlation (r)	0.260	0.589	0.433	0.538	0.301
LS slope (β)	0.282	0.624	0.445	0.532	0.272
SD (p60 Height)	4.42	3.21	3.84	3.70	3.70
SD (Height)	4.2	3.7	3.9	4.4	4.3

The following problems make use of the dataset related to markers of inflammation (see `inflamm.doc` and `inflamm.txt` on the class web pages.)

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

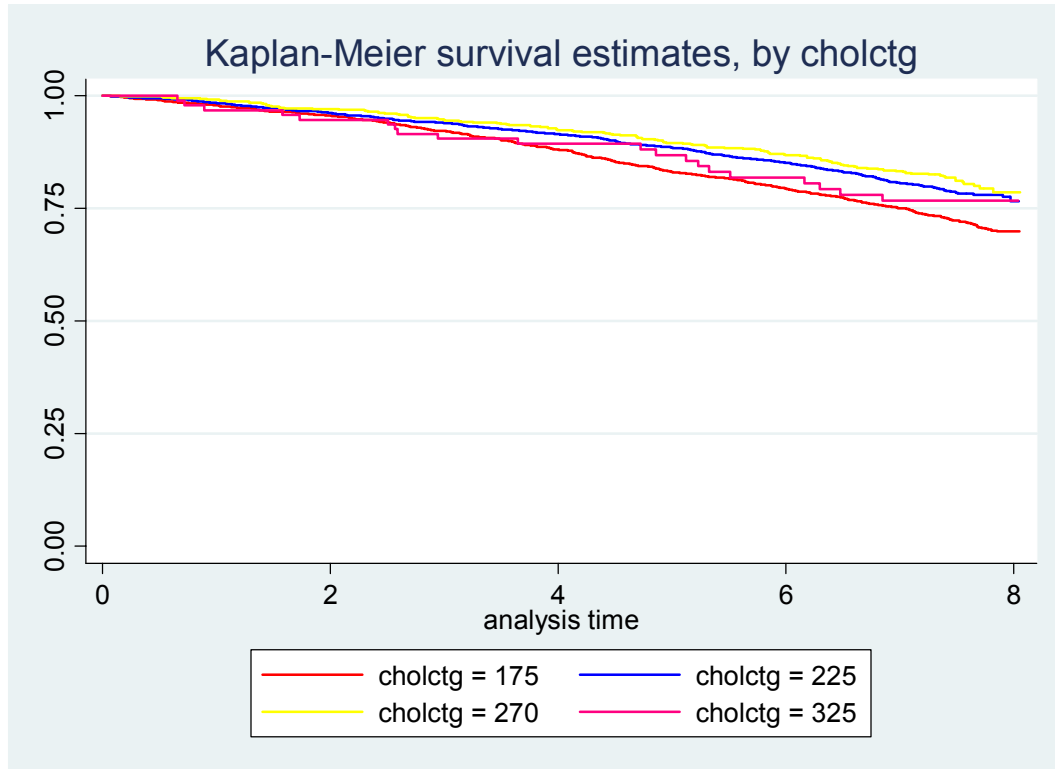
- The variable `ttodth` contains (right) censored observations of the time from accrual to the study to death or censoring according to the value of variable `death`.
 - Note that variable `ttodth` is measured in days in this dataset. You might find it more convenient to measure survival in weeks, months, 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 ttodth= ttodth / 7`
 2. (months): `replace ttodth= ttodth / 30.4`
 3. (years): `replace ttodth = ttodth / 365.25`
 - You will need to declare the variables representing the possibly censored times to death: `stset ttodth death`
 - 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(male)`.)
 - 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 12 month and 24 month survival probabilities would be obtained by `sts list, at(12 24)`.
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.



	n	Percentile of Survival Distribution (y)			Proportion Surviving		
		90 th %ile	80 th %ile	75 th %ile	2 Years	5 years	8 years
All	5000	4.02	6.68	7.78	0.959	0.863	0.741

Ans: The above Kaplan-Meier curve and tables of quantiles and survival probabilities provide some descriptive statistics. The interpretations of these descriptive statistics does not differ from what might be obtained with uncensored data, though the method of computing the descriptive statistics does differ.

- b. Produce a plot of survival curves by the groups defined by cholesterol level categorized into groups less than 200, 201- 250, 251 – 300, and greater than 300 mg/dl. Produce a table of estimates of the times at which 90%, 80%, and 75% of the subjects are estimated to still be surviving. Also include a table of the estimated probabilities of surviving 2, 5, or 8 years for each stratum. Are the estimates suggestive that the subjects’ survival is similar regardless of their cholesterol levels? Are the estimates suggestive that survival is associated with cholesterol level? Give descriptive statistics supporting your answer.



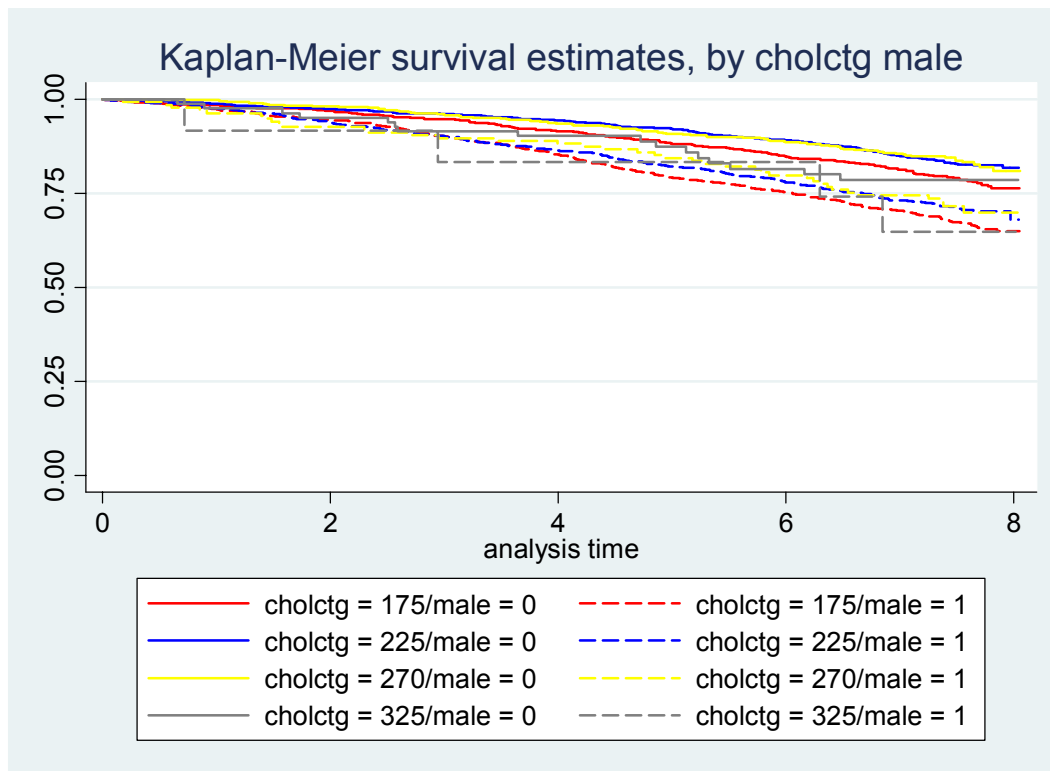
Ans: The above plot shows a tendency for the probability of survival to differ across the cholesterol strata. This suggests that there is an association between survival and cholesterol (although it is admittedly difficult to be sure that the differences in the curves is not due to random sampling error—I will note that we will find that we can with high confidence rule out that the differences are due to sampling error alone)..

Cholesterol	n	Percentile of Survival Distribution (y)			Proportion Surviving		
		90 th %ile	80 th %ile	75 th %ile	2 Years	5 years	8 years
<= 200	1997	3.56	5.87	7.02	0.955	0.831	0.698
201 - 250	2191	4.50	7.20	> 8.05	0.962	0.885	0.765
251-300	671	4.92	7.66	> 8.06	0.970	0.895	0.786
>= 301	94	3.64	6.30	> 8.04	0.947	0.868	0.768

- c. Based on your analysis in part b, how would you characterize the first order trend in survival across cholesterol strata? Is there any evidence of a curvilinear association between cholesterol and survival? Explain your answer.

Ans: The above plot shows a tendency for the lowest cholesterol stratum to have the worst survival (the lowest curve on the plot). Next worst is the highest stratum, with the second highest cholesterol stratum (i.e., cholesterol levels between 251 and 300 mg/dl) having the best estimated survival. The best fitting linear trend would likely suggest improved survival with higher cholesterol, though as noted from the curves and the descriptive statistics given in the table, the true relationship between survival and cholesterol appears to be a U-shaped trend with best survival in the 251 – 200 mg/dl range..

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



	N	Proportion Surviving					
		Females			Males		
Cholesterol		2 Years	5 years	8 years	2 Years	5 years	8 years
<= 200	1997	0.969	0.882	0.764	0.945	0.792	0.650
201 – 250	2191	0.974	0.921	0.817	0.940	0.822	0.679
251-300	671	0.981	0.908	0.810	0.927	0.844	0.699
>= 301	94	0.951	0.873	0.786	0.917	0.833	0.648

Ans: From the above stratified analyses, we see that males tend to survive less well than females having similar cholesterol values, thus suggesting that sex is associated with survival after adjustment for cholesterol. Furthermore, a comparison of cholesterol levels across sex groups finds that males average cholesterol levels of 198 mg/dl, while women average levels of 221 mg/dl. Thus we find that sex is associated with cholesterol level, and we would anticipate that sex would confound our ability to quantify an association between survival and cholesterol, unless we perform an adjusted analysis. Because women tend to have higher cholesterol levels in our sample, and because women tend to survive better than men, an unadjusted analysis might overstate any protective effect of high cholesterol on survival. I do note, however, that the U-shaped trend in survival is evident in each sex individually, and thus the observed protective effect of high cholesterol is not wholly explained by confounding by sex.

