

Biost 514 / Biost 517: Applied Biostatistics I

Emerson, Fall 2011

Homework #4 Key

October 20, 2011

Note: The web pages contain a Stata file that could be used to answer the questions on this homework. However, for ease of discussing the restricted means, I chose to use R to actually produce the answers for problems 1 and 2. This is only relevant, because there are many competing approaches to estimate quantiles, thus there may be slight differences between the estimates obtained from Stata and the estimates that I gave in this homework. I think you will find that those differences are inconsequential as a rule. In any case, it should be remembered that quantiles are not necessarily unique, so different quantile estimates will even be produced in Stata when different commands are used.

Written problems: To be handed in at the beginning of class on Friday, October 28, 2011.

*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 and 2 make use of a dataset from a phase 2b clinical trial of an experimental tyrosine kinase inhibitor in non small cell lung cancer (NSCLC). The documentation file nsclc.doc and the data file nsclc.txt can be found on the class web pages.

Recall that when analyzing censored data in problems 1 and 2, descriptive statistics are obtained in Stata using its facility for Kaplan-Meier estimation:

- For problem 1, you are asked to describe the censoring distribution. You will need to create a variable that indicates subjects who were censored. This is easily effected by `g censored= 1 - death`, because all subjects who were not observed to die were censored. You will then need to declare the variables representing the censoring distribution: `stset obstime censored`.
 - For problem 2, you are asked to describe the distribution of times to death. You will need to declare the variables representing the possibly censored times to death: `stset obstime 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(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, the 6 month and 12 month survival probabilities would be obtained by `sts list, at(182 365)`. (Recall the observation time is measured in days. Prior to declaring the survival variables with `stset`, you might want to change the unit to months by `replace obstime=obstime / 30.4`.)
1. In studies with censored observations of time to some event, our ability to answer specific scientific questions will often depend upon the distribution of censoring times.

That is, we need to understand the times that we followed each patient. However, we only have partial information on this distribution. For instance, if we are ultimately investigating patient survival, we may want to understand how long we followed the patients: Was it 3 years, 30 years, 300 years? Was it 3 years for some patients and 300 years for others? When patients' survival times are censored, we know exactly the limits of our follow-up. But for patients who died, we do not know when we might have lost those patients to further follow-up. Luckily the Kaplan-Meier estimator comes to our rescue in this situation. By creating an indicator of censoring (0= not censored, 1= censored), we can use the KM estimates to describe the pattern of censoring.

- a. Provide suitable statistics for the distribution of times to censoring for observations of death. In particular, consider whether you can estimate the minimum time of follow-up for these patients.

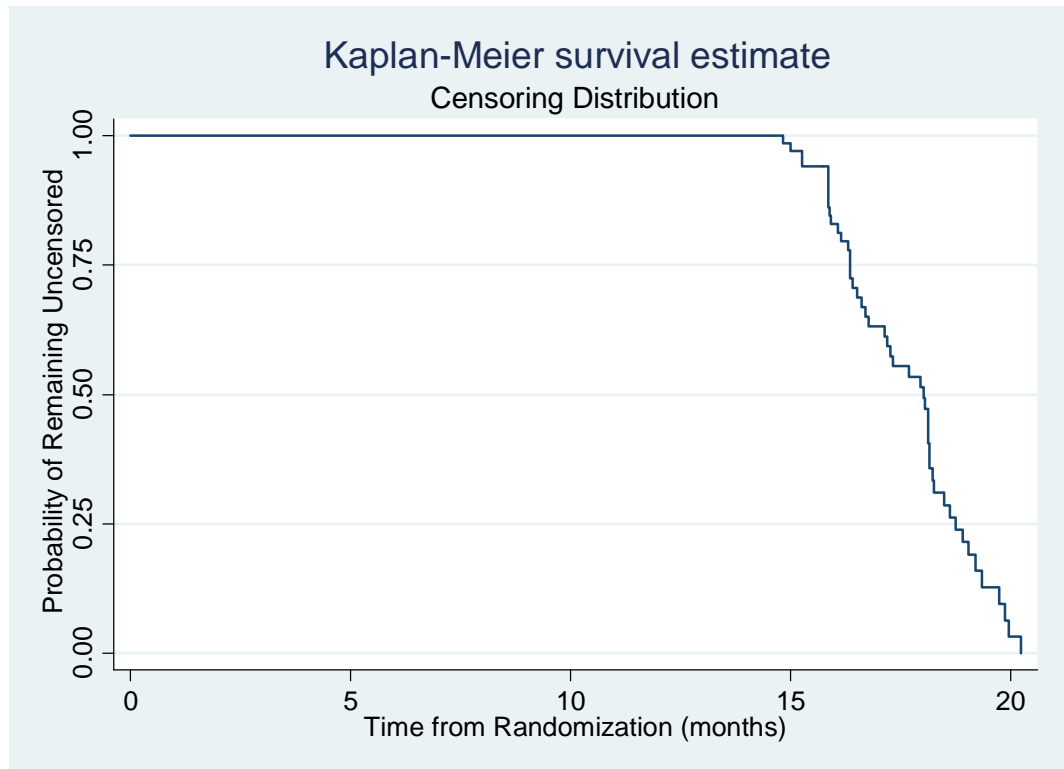


Table 1: Descriptive statistics for the distribution of time to censoring in months. Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being uncensored at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Uncens)
Censored	188(48)	17.7 (NA)	18.0 (16.3, 18.8)	(1.84*, 20.2)	(1.00, 1.00, 0.514)

Ans: The above figure and Table 1 display estimates of the censoring distribution based on Kaplan-Meier estimates. The earliest observed censoring is at 14.8 months, but we cannot know from the data when the subject who died at 1.84 months might have been censored. Hence, we cannot tell from the data when the minimum censoring time would have been. (From the experimental design of the study, censoring only occurred by “administrative censoring”, which arises due to subjects still being alive at the time of data analysis. Hence, the

minimum censoring distribution could be computed from the date that the last subject was randomized and the date of data analysis.)

- b. Suppose we want to divide individual patients into groups who die within 2 years and those who do not. On the basis of your answer to part a, will we be able to do so?

Ans: From Table 1, it can be seen that the Kaplan Meier estimate for the probability of remaining uncensored by 18 months is 0.514. Because we cannot know whether subjects censored prior to 2 years will survive 2 years or not, we cannot divide our sample into groups according to 2 year survival.

- c. Suppose we want to divide individual patients into groups who die within 1 year and those who do not. On the basis of your answer to part a, will we be able to do so?

Ans: From Table 1, it can be seen that the Kaplan Meier estimate for the probability of remaining uncensored by 12 months is 1.00. Because no subjects were censored prior to 1, we can easily divide our sample into groups according to 1 year survival.

2. We are interested in estimating the probability distribution of patient survival following accrual to the study.
- a. Provide suitable descriptive statistics for the distribution of times to death for patients in the dataset.

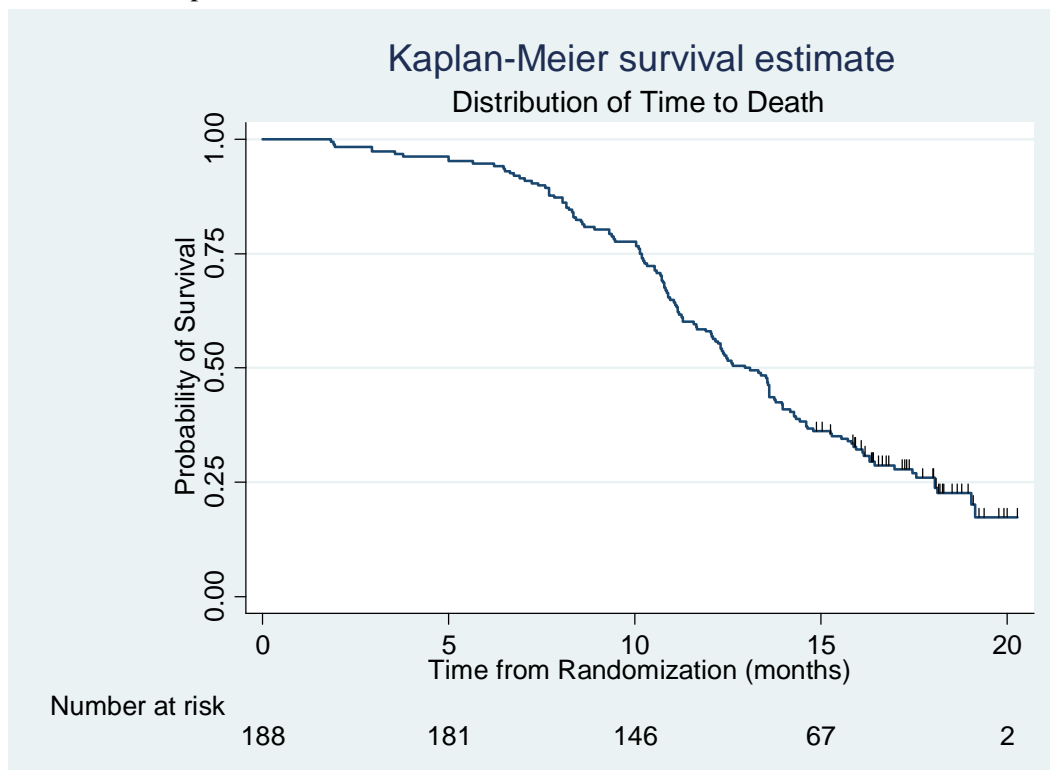


Table 2a: Descriptive statistics for the distribution of time to death in months. Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified

time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being alive at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Surv)
TOTAL	188 (140)	13.4 (20.2)	13.0 (10.2, 18.1)	(1.84, 20.2*)	(0.947, 0.580, 0.259)

Ans: From the above figure and Table 2a, it can be seen that the Kaplan Meier estimate for the median survival time is 13.0 months with an interquartile range from 10.2 to 18.1 months. The probability of 18 month survival is 25.9%. We can also compute the 20.2 month restricted mean as 13.4 months. *(In the absence of censoring, the sample mean can be computed as the area under the survival curve. Because we have censoring, we would estimate the mean by computing the area under the Kaplan-Meier curve. However, the censoring distribution does not allow estimation of the entire survival curve: The person observed for the longest period of time was censored. Hence the KM estimate does not drop all the way to zero. We thus compute the area under the curve up until some threshold. By default, most programs use the largest observation time. Hence, in this case we estimate the mean restricted to 20.2 months. This can be interpreted as the average number of months the patients survived during the first 20.2 months following randomization.)*

- b. Produce a plot of survival curves by the groups defined by sex. Produce a table of estimates of the 75th, 50th, and 25th percentiles of the survival distribution within each stratum. Also include in that table the estimated probabilities of surviving for 6, 12, and 18 months for each stratum. Are the estimates suggestive that sex is associated with survival? Give descriptive statistics supporting your answer.

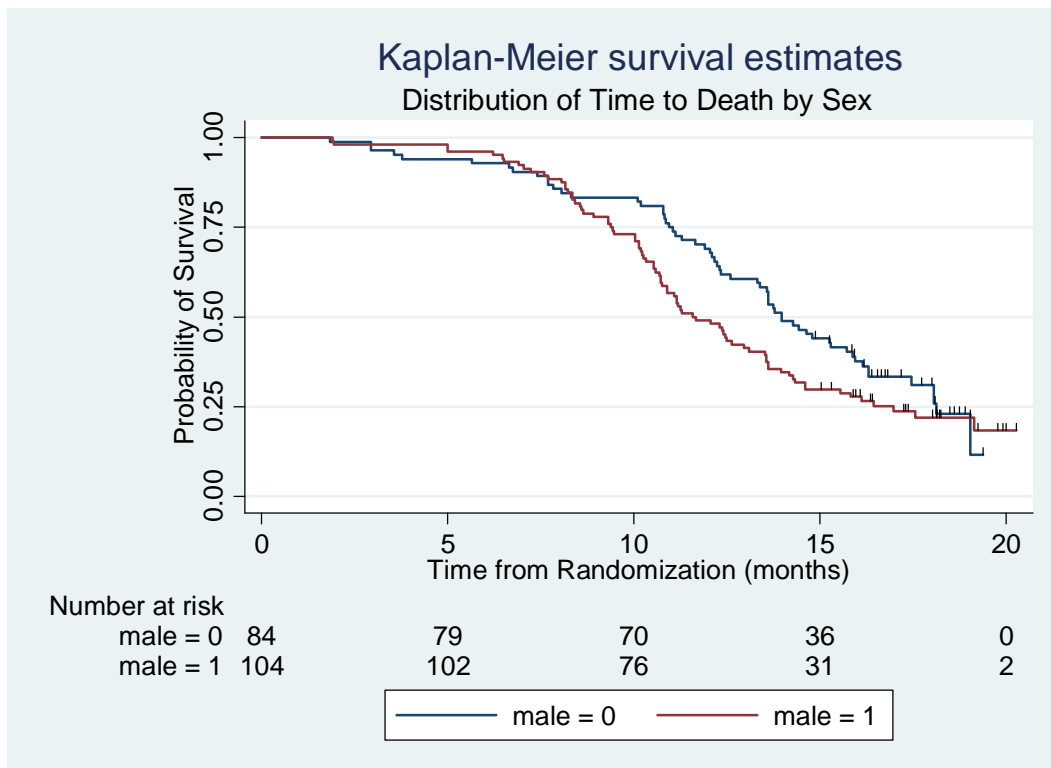


Table 2b: Descriptive statistics for the distribution of time to death by sex. Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified

time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being alive at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Surv)
Females	84 (60)	13.8 (19.3)	14.0 (11.0, 18.1)	(1.84, 19.3*)	(0.929, 0.691, 0.311)
Males	104 (80)	12.9 (20.2)	11.6 (9.4, 17.0)	(1.91, 20.2*)	(0.962, 0.490, 0.220)
TOTAL	188 (140)	13.4 (20.2)	13.0 (10.2, 18.1)	(1.84, 20.2*)	(0.947, 0.580, 0.259)

Ans: From the above figure and Table 2b, we estimate longer median survival for females (14.0 months) than for males (11.6 months). Females have higher estimated probabilities of surviving for 12 and 18 months, but lower estimates for 6 months. (Note that comparison of the restricted means is complicated by the differing time used for the restriction: The longest observation time for a female is 19.3 months, while it is 20.2 months for a male. A better approach would be to use the same time period to restrict the means for both sexes. Stata does not make this easy: We would have to create a new variable that censored all observations at, say, 18 months, and then compare the restricted means over that first 18 months post randomization. All of that having been said, the fact that the females have a higher 19.3 month restricted mean than the males' 20.2 month restricted mean does suggest better survival for females.)

- c. Produce a plot of survival curves by the groups defined by the indicator of advanced disease. Produce a table of estimates of the 75th, 50th, and 25th percentiles of the survival distribution within each stratum. Also include in that table the estimated probabilities of surviving for 6, 12, and 18 months for each stratum. Are the estimates suggestive that stage of disease is associated with survival? Give descriptive statistics supporting your answer.

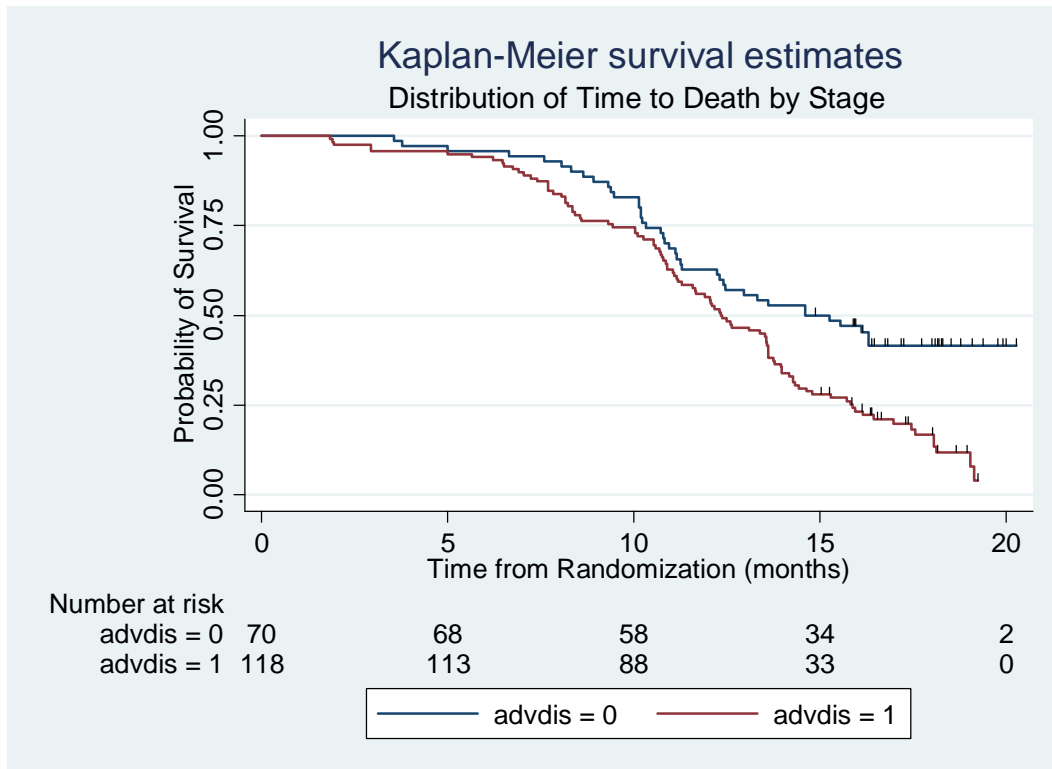


Table 2c: Descriptive statistics for the distribution of time to death by stage of disease (Early vs Advanced). Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being alive at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Surv)
Early	70 (40)	14.8 (20.2)	14.9 (10.3, NA)	(3.55, 20.2*)	(0.957, 0.629, 0.415)
Advanced	118 (100)	12.5 (19.2)	12.4 (9.4, 15.9)	(1.84, 19.2*)	(0.941, 0.551, 0.167)
TOTAL	188 (140)	13.4 (20.2)	13.0 (10.2, 18.1)	(1.84, 20.2*)	(0.947, 0.580, 0.259)

Ans: From the above figure and Table 2c, we estimate longer median survival for patients with early disease (14.9 months) than for those with advanced disease (12.4 months). Patients with early stage disease have higher estimated probabilities of surviving for 6, 12, and 18 months. *(Note again that the restricted means are not easily compared. Also note that the early stage patients were not observed long enough to be able to estimate when only 25% were still surviving. Hence the estimated 75th percentile of the time to death is censored for that group. It would be better to report the censored estimate, rather than just listing it as missing. Both Stata and my R functions, however, return missing for the estimated quantile. I will fix my routines to mirror the way I report the maximum.)*

- d. Produce a plot of survival curves by the groups defined both by sex and the indicator of advanced disease. Produce a table of estimates of the 75th, 50th, and 25th percentiles of the survival distribution within each stratum. Also include in that table the estimated probabilities of surviving for 6, 12, and 18 months for each stratum. Are the estimates suggestive that stage of disease confounds the association between sex and survival? Give descriptive statistics supporting your answer.

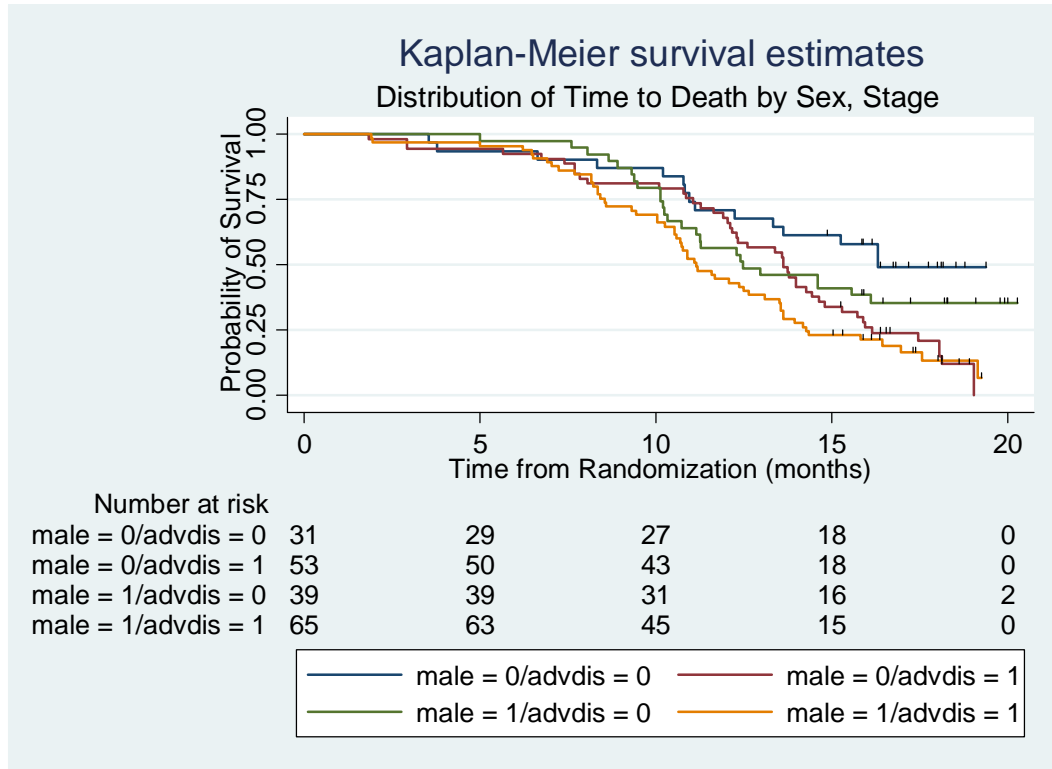


Table 2d: Descriptive statistics for the distribution of time to death by sex and stage of disease (Early vs Advanced). Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified time (Restr)), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being alive at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Surv)
F Early	31 (15)	15.2 (19.3)	16.3 (11.0, NA)	(3.55, 19.3*)	(0.936, 0.710, 0.490)
F Adva	53 (45)	13.1 (NA)	13.6 (11.1, 16.2)	(1.84, 19.1)	(0.925, 0.679, 0.210)
Females	84 (60)	13.8 (19.3)	14.0 (11.0, 18.1)	(1.84, 19.3*)	(0.929, 0.691, 0.311)
M Early	39 (25)	14.2 (20.2)	12.5 (10.1, NA)	(5.00, 20.2*)	(0.974, 0.564, 0.353)
M Adva	65 (55)	11.9 (19.2)	11.2 (8.55, 14.3)	(1.91, 19.2*)	(0.954, 0.446, 0.133)
Males	104 (80)	12.9 (20.2)	11.6 (9.4, 17.0)	(1.91, 20.2*)	(0.962, 0.490, 0.220)
Early	70 (40)	14.8 (20.2)	14.9 (10.3, NA)	(3.55, 20.2*)	(0.957, 0.629, 0.415)
Adva	118 (100)	12.5 (19.2)	12.4 (9.4, 15.9)	(1.84, 19.2*)	(0.941, 0.551, 0.167)
TOTAL	188 (140)	13.4 (20.2)	13.0 (10.2, 18.1)	(1.84, 20.2*)	(0.947, 0.580, 0.259)

Ans: From the above figure and Table 2d, we estimate longer median survival for patients with early disease than for advanced disease within each sex. Similarly patients with early stage disease have higher estimated probabilities of surviving for 6, 12, and 18 months within each sex. Hence stage of disease is associated with survival after adjusting for sex. While we have no proof from this analysis that the association is causal, it would seem that causation is highly likely and it is certainly believed to be causal by the greater scientific community.

When we consider associations between sex and stage of disease in the sample, we note that 31 of 84 females (36.9%) have early stage disease, while 39 of 104 males (37.5%) have early

stage disease. Given the similar distribution of stage of disease across the sexes, there does not seem to be compelling evidence for confounding, even though stage of disease does appear causally associated with survival.

(Note that we could come up with something of an estimate of the association between sex and median survival after adjustment for stage of disease. In the early stage disease, the median survival is 16.3 months for females and 12.5 months for males, for a difference in median survival of 3.8 months. In the advanced stage disease, the median survival is 13.6 months for females and 11.2 months for males, for a difference in median survival of 2.4 months. For an adjusted estimate, we might take an average of those two differences to arrive at an adjusted difference of median survival of 3.1 months. Alternatively, we could choose to weight the early and late stage groups according to their sample size to obtain $(70 \times 3.8 + 118 \times 2.4) / 188 = 2.92$ months difference in median survival. The unadjusted estimate for the difference in median survival was $14.0 - 11.6 = 2.4$ months.)

I note that some people try to assess confounding by seeing whether the adjusted and unadjusted estimates differ. That works if we are looking at means, but not if we are looking at other summary measures. The safest way to assess confounding is to decide whether there is a causal association between the third variable and the outcome within a group that is homogeneous with respect to the predictor of interest, and to see whether there is an association between the predictor of interest and the third variable in our sample. I further note that confounding is more or less a product of those two associations, so a slight imbalance in the distribution of the third variable across predictor of interest groups can still cause major confounding if that third variable is strongly associated with the outcome variable.)

- e. Produce a plot of survival curves by the groups defined by the indicator of treatment arm. Produce a table of estimates of the 75th, 50th, and 25th percentiles of the survival distribution by the treatment strata. Also include in that table the estimated probabilities of surviving for 6, 12, and 18 months for each stratum. Are the estimates suggestive that treatment is associated with survival? Give descriptive statistics supporting your answer.

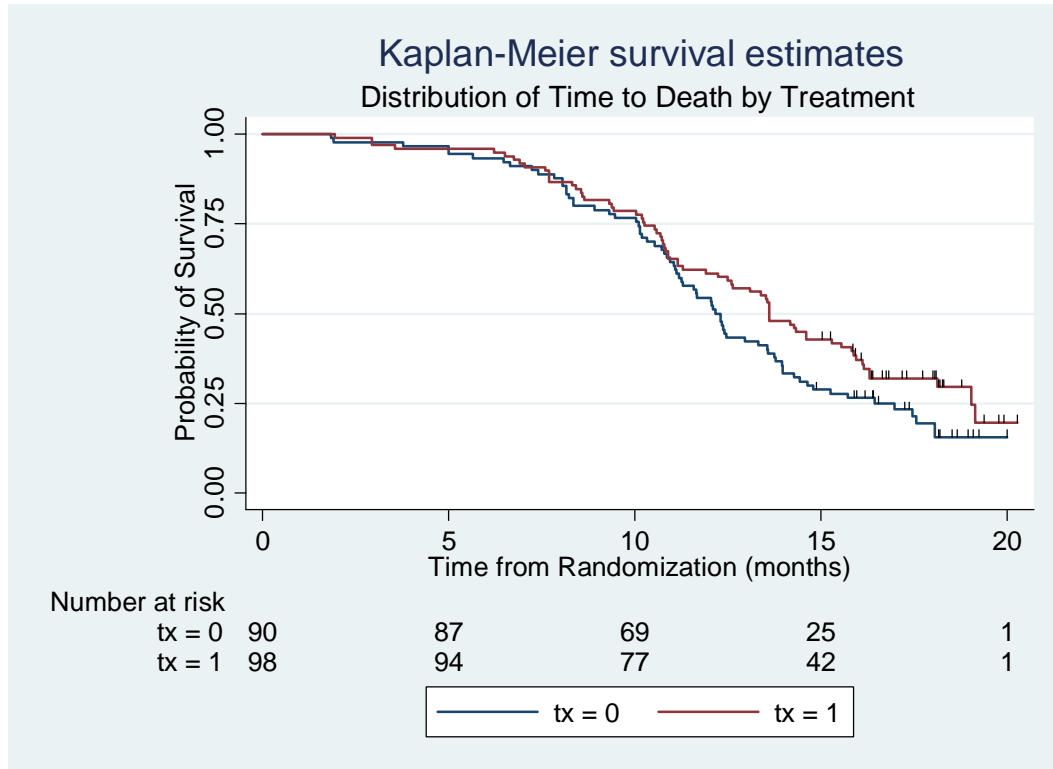


Table 2e: Descriptive statistics for the distribution of time to death by treatment arm (placebo vs treatment with TFD725). Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being alive at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Surv)
Placebo	90 (72)	12.8 (20.0)	12.2 (10.1, 17.0)	(1.84, 20.0*)	(0.933, 0.544, 0.195)
Treatment	98 (68)	13.9 (20.2)	13.6 (10.3, 19.1)	(1.94, 20.2*)	(0.959, 0.612, 0.320)
TOTAL	188 (140)	13.4 (20.2)	13.0 (10.2, 18.1)	(1.84, 20.2*)	(0.947, 0.580, 0.259)

Ans: From the above figure and Table 2e, we estimate slightly longer median survival for patients receiving the experimental treatment (13.6 months) than for those receiving placebo (12.2 months). Patients receiving treatment also have higher estimated probabilities of surviving for 6, 12, and 18 months. (Note here that the restricted means are a little more easily compared, because randomization tended to make the censoring distribution constant across treatment groups. Nevertheless, it would be better to have a common time of restriction for the restricted means.)

- f. Suppose we are interested in whether treatment might be associated with survival differently in the less advanced diseased patients than in the more advanced disease patients (i.e., we are interested in whether stage of NSCLC modifies the effect of treatment). Provide descriptive statistics addressing this question.

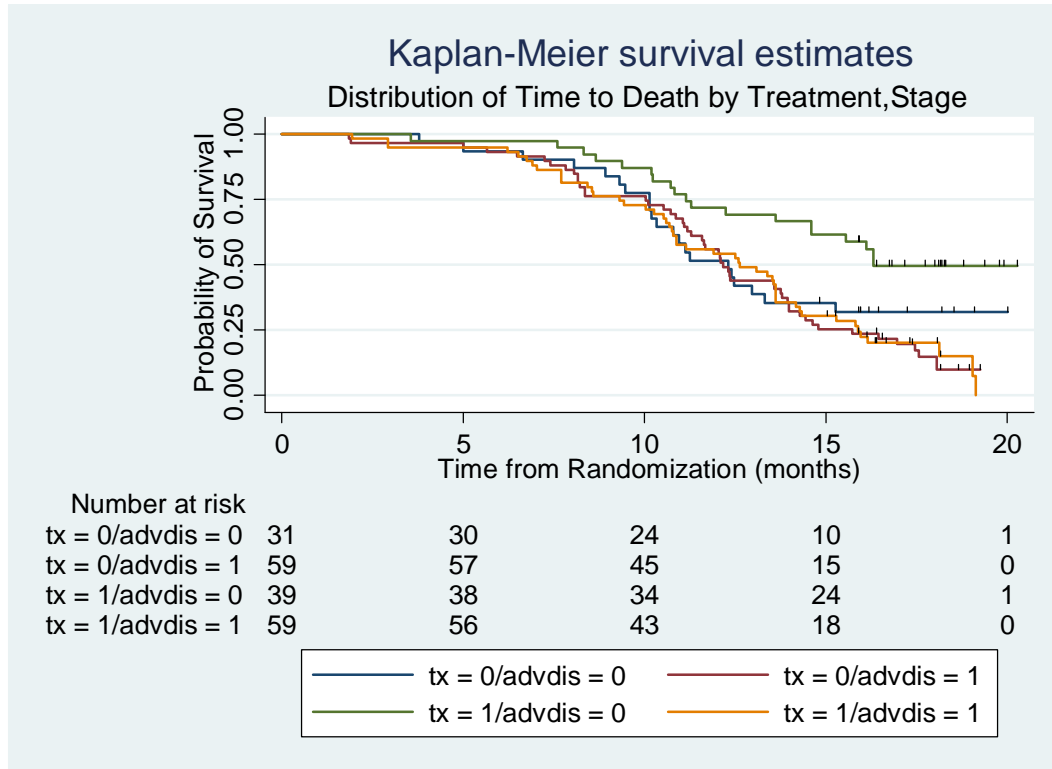


Table 2f: Descriptive statistics for the distribution of time to death by stage of disease (E=Early vs A=Advanced) and treatment arm (Plc=Placebo, Tx=Treatment with TFD725). Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min) and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being alive at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(6,12,18 mo Surv)
E Plc	31 (21)	13.4 (20.0)	12.3 (10.1, NA)	(3.78, 20.0*)	(0.936, 0.516, 0.319)
E Tx	39 (19)	16.0 (20.2)	16.3 (11.2, NA)	(3.55, 20.2*)	(0.974, 0.718, 0.497)
Early	70 (40)	14.8 (20.2)	14.9 (10.3, NA)	(3.55, 20.2*)	(0.957, 0.629, 0.415)
A Plc	59 (51)	12.5 (19.2)	12.2 (10.0, 15.7)	(1.84, 19.2*)	(0.932, 0.559, 0.147)
A Tx	59 (49)	12.5 (NA)	12.6 (9.31, 15.9)	(1.94, 19.1)	(0.949, 0.542, 0.201)
Advanced	118 (100)	12.5 (19.2)	12.4 (9.4, 15.9)	(1.84, 19.2*)	(0.941, 0.551, 0.167)
Placebo	90 (72)	12.8 (20.0)	12.2 (10.1, 17.0)	(1.84, 20.0*)	(0.933, 0.544, 0.195)
Treatment	98 (68)	13.9 (20.2)	13.6 (10.3, 19.1)	(1.94, 20.2*)	(0.959, 0.612, 0.320)
TOTAL	188 (140)	13.4 (20.2)	13.0 (10.2, 18.1)	(1.84, 20.2*)	(0.947, 0.580, 0.259)

Ans: From the above figure and Table 2f, we estimate longer median survival for patients receiving the experimental treatment among those patients with early stage disease (16.3 months versus 12.3 months on placebo), while among patients with more advanced disease there is very little difference (12.6 months on experimental treatment versus 12.2 months on placebo). Similar patterns are observed when comparing the KM survival curves: Among advanced disease patients, the two survival curves overlap substantially, while in the early stage disease the survival curve for the experimental group is higher than that for the placebo. Given the observed difference in treatment effect across the strata defined by stage of disease, we might suspect that there is some effect modification. (Eventually, we would

want to assess whether this evidence was merely a coincidental observation, or whether we could conclude a difference beyond that that could be explained by chance. We will cover such methods of testing for effect modification in Biost 518/515.)

Problems 3-6 make 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:

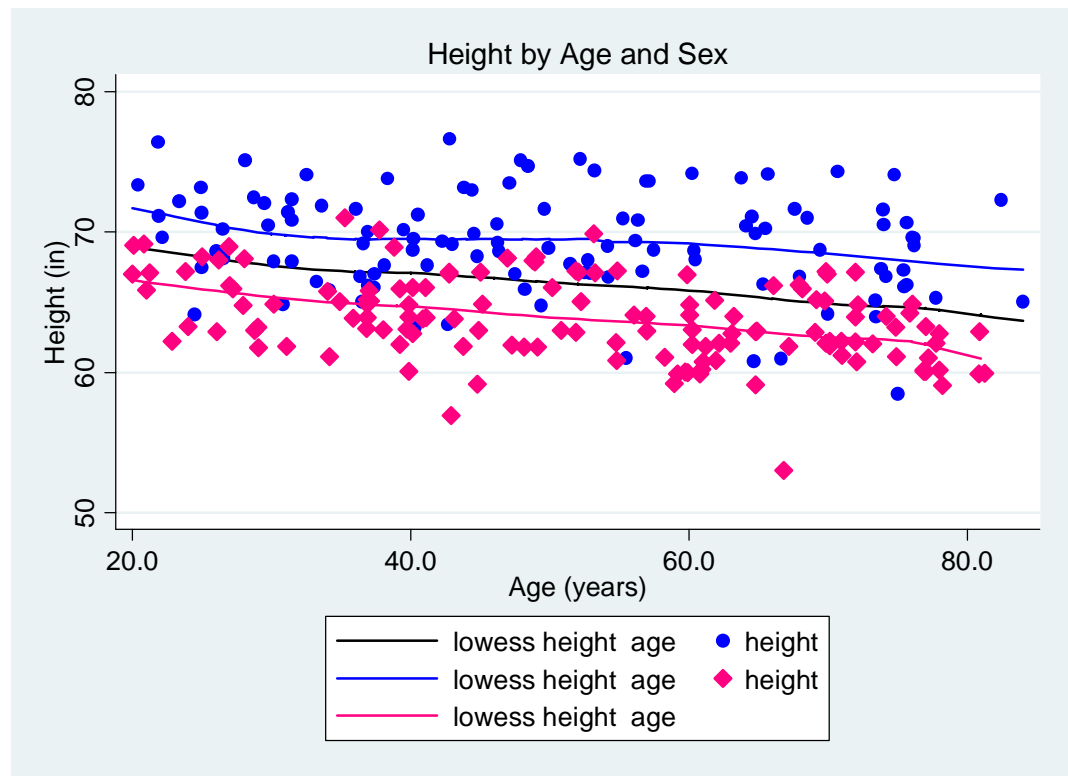
```
twoway (lowess p60 age, col(black) xtitle("Age (years)")
       ytitle("p60 (msec)") t1("Time to p60 SEP by Age and Sex"))
       (lfit p60 age, col(black) lp("-"))
       (scatter p60 age if sex==1, jitter(2) col(blue))
       (lowess p60 age if sex==1, col(blue))
       (lfit p60 age if sex==1, 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”.

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

- Produce a lowess smooth of $p60$ 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.
- Produce the “best” fitting straight line for $p60$ on age using all the data. The “least squares fit” will be black and dashed. No points are plotted by this command.
- Produce a scatterplot of $p60$ 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.

- Produce a lowess smooth of $p60$ on age for males. The lowess line will be blue, and, because I did not specify a line pattern, it will be solid.
 - Produce the “best” fitting straight line for $p60$ on age for males. The “least squares fit” will be blue and dashed.
 - Produce a scatterplot of $p60$ 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.
 - Produce a lowess smooth of $p60$ on age for females. The lowess line will be pink, and, because I did not specify a line pattern, it will be solid.
 - Produce the “best” fitting straight line for $p60$ on age for females. The “least squares fit” will be pink and dashed.
3. 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 4-6, 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	

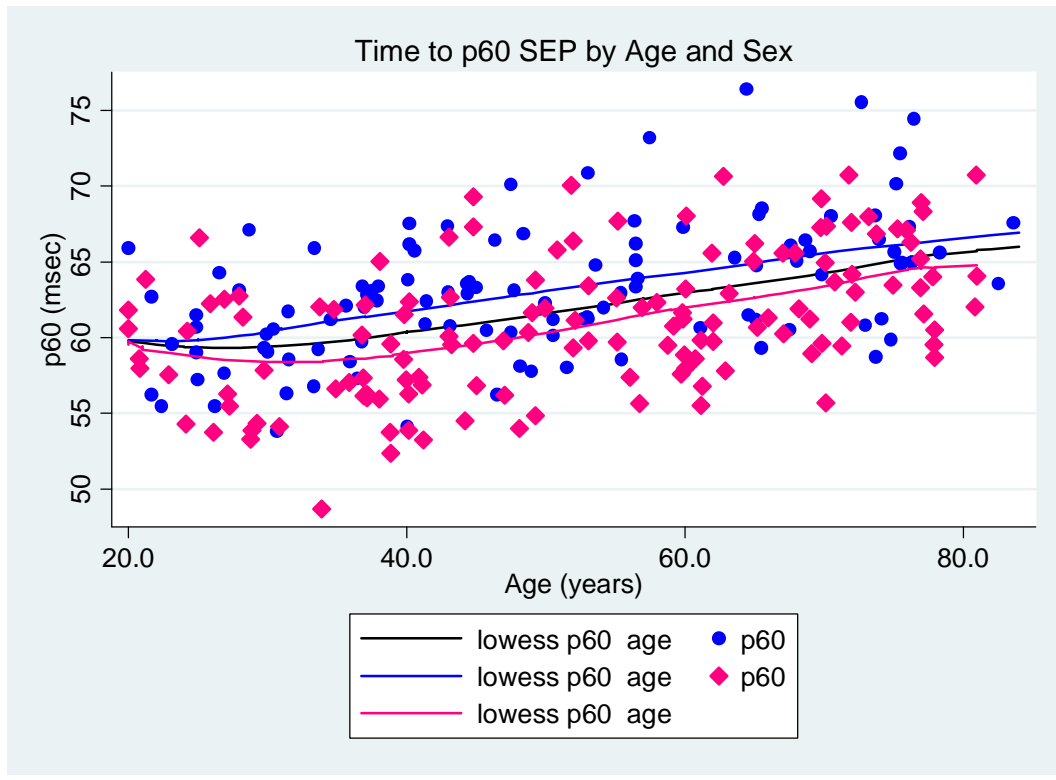
p60	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1207971	.0149576	8.08	0.000	.0913369	.1502573
_cons	55.70264	.8078069	68.96	0.000	54.1116	57.29368

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 $Var(Y | X)$ 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
```

4. 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 lowest 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.

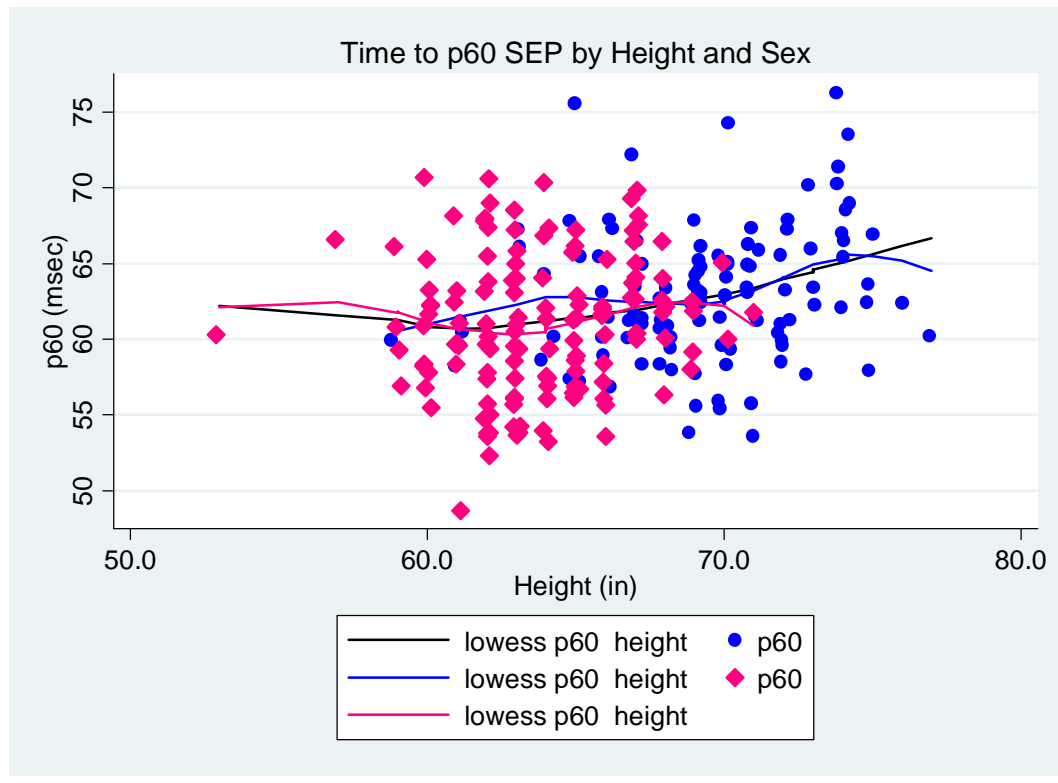
- c. How do you explain any difference you observe in the answers to parts a and b? In particular, why might you expect the correlation to be less extreme in the combined sample than it was in each stratum defined by sex? Consider the statistical behavior of correlation as it relates to the slope of 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.

d.

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

5. 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

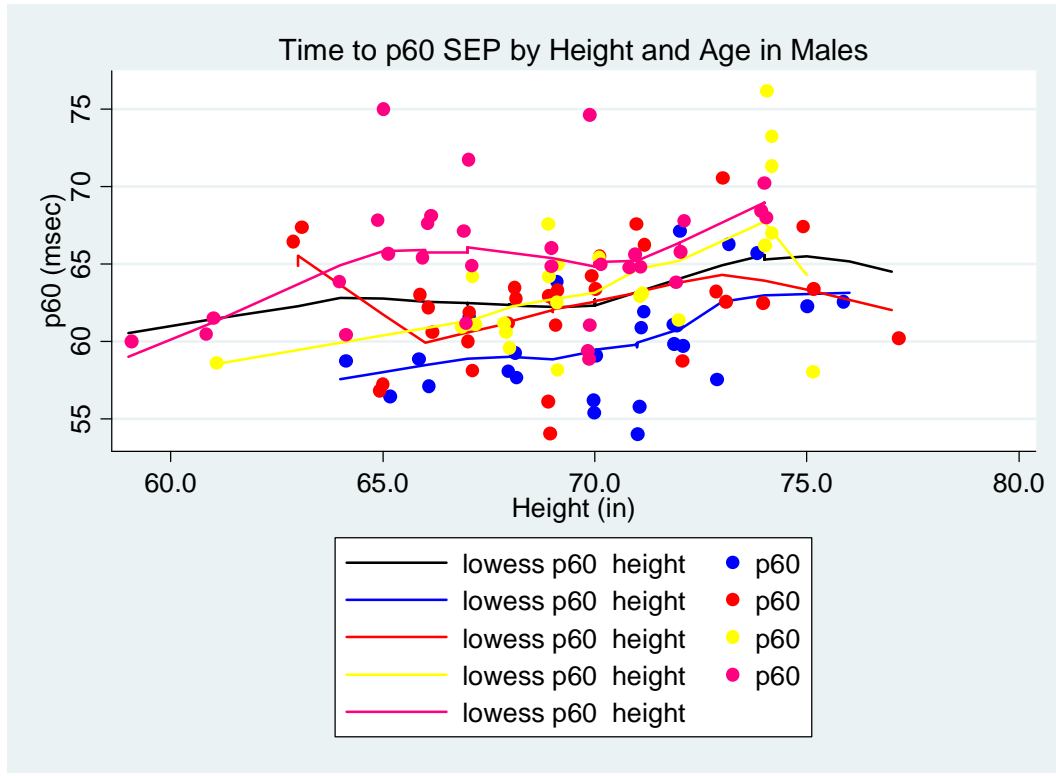
6. For this problem, you will need to create a variable indicating age within the categories 20 – 34, 35 – 49, 50 – 64, and 65 – 84. The following Stata commands can be used to create such a variable:

```
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?

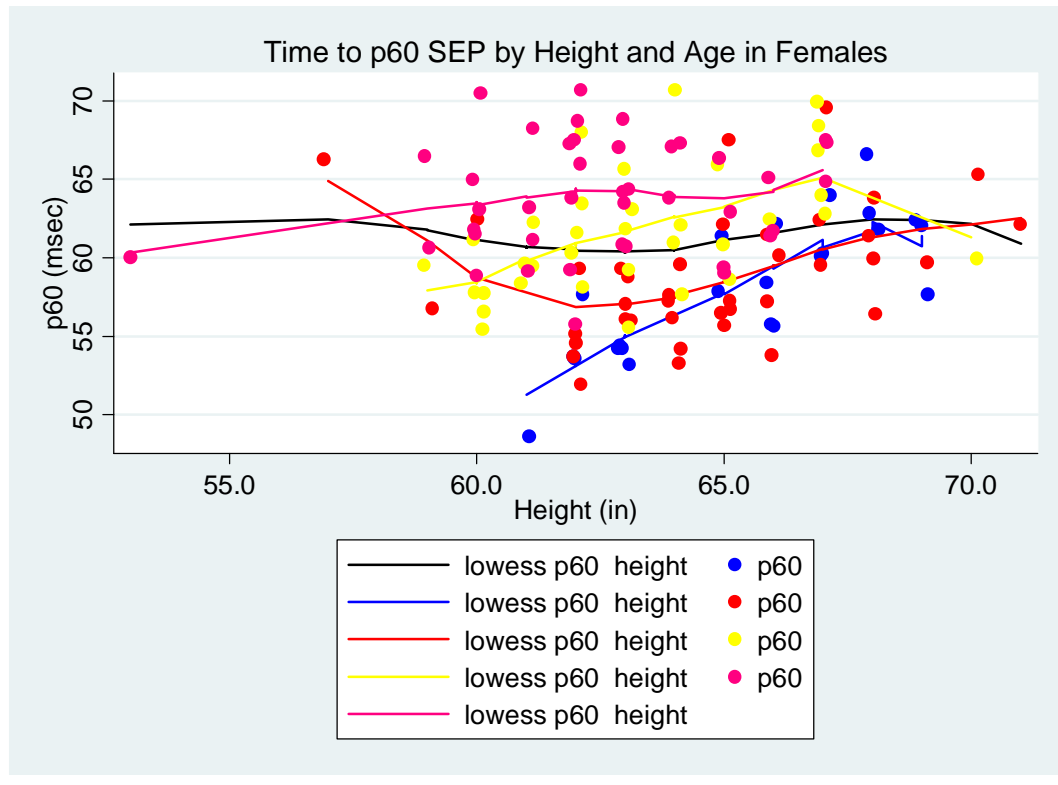


Ans: The above plot for men shows lowess curves that are not as smooth as they might be with a larger sample size. None the less, there is a clear trend toward longer p60 delays in taller men for every age stratum. Furthermore, the vertical separation of the lowess curves for the strata suggests a tendency 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 4 and 5, 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.