Biost 517: Applied Biostatistics I
Emerson, Fall 2012

Homework #4 Key
October 19, 2012

Written problems: To be handed in at the beginning of class on Friday, October 26, 2012.

Problems make use of the university salary data (salary.txt). The class web pages contain an annotated Stata log file (initsalary.doc) illustrating the way in which this data can be input into Stata. In particular, I illustrate how string variables can be encoded and how labels can be associated with particular values of variables. However, for this homework, we have no real need for the encoded variables, so it will suffice to just use the following code (that identifies variables `sex`, `deg`, `field`, and `rank` as string variables):

```
infile case id str9 sex str9 deg yrdeg str9 field startyr year str9 rank admin
                          salary using salary.txt
```

You should drop the first line of the file, as that only contained the column labels

```
drop in 1
```

To read the data into R, you could use

```
SALARY <- read.table("http://www.emersonstatistics.com/datasets/salary.txt",
       header=TRUE, stringsAsFactors=FALSE)
```

A couple important points about this data set:

- We sample all non medical school faculty who are working for the university in 1995.
- We include a record for each of those faculty members for each year they worked for the university between 1976 and 1995.
- Obviously, every faculty member in the data set will have one record for 1995. Hence, by selecting `if year==95`, we will select only one row for each faculty member.

For this homework, we are in part interested in whether the university discriminates against women in promotion. Because there is a time clock for gaining tenure (assistant professors who are not granted tenure in their 6th year are typically not rehired), we choose to examine promotion from associate professor to full professor.

First we have to determine the earliest year in this data set that each faculty member was an associate professor. (The following code will have missing values for faculty who were never an associate professor in our data.)

In Stata:

- We first follow a two stage procedure using egen to identify the earliest time each person was an associate professor
    - `egen grbg= min(year) if rank=="Assoc", by(id)`
    - `egen fstAssoc= mean(grbg), by(id)`
- We then want to ignore all faculty who were hired as an associate professor or who were an associate professor in 1976, because we do not know how long they were associate professor prior to being hired

```
o  replace fstAssoc=. if fstAssoc==startyr | fstAssoc==76
```

- We then follow a two stage procedure using egen to identify the earliest time each person was a full professor

```
o  drop grbg
o  egen grbg= min(year) if rank==”Full”, by(id)
o  egen fstFull= mean(grbg), by(id)
```

- Now we create a variable that indicates whether we observed a faculty member to be promoted to associate and later promoted to full

```
o  g promoted= 0
o  replace promoted= 1 if fstAssoc!=. & fstFull!=.
```

- Now we want to create a variable that measures the (possibly censored) time to promotion from associate to full.
- First we compute the difference

```
o  g ttofull= fstFull - fstAssoc
```

- That variable will be missing for any subject who was missing either `fstFull` or `fstAssoc.` We need to register that anyone who was still employed as an associate professor in 1995 was censored for their time to promotion (note that `promote` will already have them as a 0)

```
o  replace ttofull= 95 – fstAssoc if fstAssoc!=. & fstFull==.
```

- Because Stata will just drop cases with missing data from all analyses, we can make sure that our survival analyses use only one case for each faculty member by setting missing data for everything but 1995

```
o  replace ttofull= . if year!= 95
```

- And now we “set” the survival variable to be used in analyzing the time to promotion

```
o  stset ttofull promoted
```

In R, the equivalent variables can be created using my functions with the following code

```
o  fstAssoc <- clusterStats(SALARY$year, SALARY$id, ”min”,
      subset=SALARY$rank==”Assoc”)
o  fstAssoc[fstAssoc==SALARY$startyr | fstAssoc==76] <- NA
o  fstFull <- clusterStats(SALARY$year, SALARY$id, ”min”,
      subset=SALARY$rank==”Full”)
o  promoted <- !is.na(fstAssoc) & !is.na(fstFull)
o  ttofull <- ifelse(is.na(fstFull),95,fstFull) – fstAssoc
o  ttofull <- ifelse(year==95,ttofull,NA)
o  ttofull <- Surv(ttofull, promoted)
```

Questions for Biost 514 <u>and</u> Biost 517:

1. For each faculty member in the data set, generate appropriate descriptive statistics by sex on the distribution of academic field, year in which he/she attained his/her highest degree, year he/she was hired at the university, academic rank in 1995, and monthly salary in 1995. Comment on any differences you observe between men and women faculty in these measurements.

<u>Answer</u>: Table 1a provides descriptive statistics for the year of degree, year hired at the university, and monthly salary in 1995 for each sex separately and for the sampled faculty combined. Women faculty tended to receive their degree 6.7 years later than men on average and to be hired by the university 5.9 years later than men on average. The variability of year of degree and year of hire were higher for men than for women. Such observations were consistent with a historical pattern in which women were more readily

hired than they are in current times, or a potentially persistent pattern in which men and women are hired in equal numbers, but women are fired more often.

Also shown in Table 1a is a tendency for women to have a salary that is $1,335 per month lower than men on average. Men's salaries were also more variable than the women's measurements, with quite similar minima monthly salaries between the sexes (though note the smaller sample size for women than for men), but seemingly increasing differences in salary percentiles in the highest paid faculty: The 25th, 50th, and 75th percentile of monthly salaries for women is $796, $1,297, and $1,800 lower than the respective percentiles for men.

		N	Mean (SD)	Mdn (IQR)	(Min, Max)
Year of degree	**Females**	409	81.1 (8.7)	82.0 (74.0, 89.0)	(54.0, 95.0)
	Males	1188	74.4 (9.6)	73.0 (67.0, 82.0)	(48.0, 96.0)
	Combined	1597	76.1 (9.9)	76.0 (69.0, 84.0)	(48.0, 96.0)
Start Year	**Females**	409	85.5 (8.0)	88.0 (80.0, 92.0)	(57.0, 95.0)
	Males	1188	79.6 (10.2)	80.0 (71.0, 89.0)	(48.0, 95.0)
	Combined	1597	81.1 (10.0)	83.0 (73.0, 90.0)	(48.0, 95.0)
Salary in 1995	**Females**	409	5,397 (1,481)	5,016 (4,292; 6,135)	(3,042; 11,036)
	Males	1188	6,732 (2,090)	6,313 (5,088; 7,935)	(3,131; 14,464)
	Combined	1597	6,390 (2,037)	5,962 (4,743; 7,602)	(3,042; 14,464)

Table 1b displays the sex distribution across academic fields and faculty ranks. Women tend to be more heavily represented in the fine arts and less represented in the professional fields. Women tend to be approximately evenly distributed across ranks, while 60% of male faculty are full professors.

		Females (Row %; Col %)	Males (Row %; Col %)	Total (Row %; Col %)
Field	**Arts**	80 (36.4%; 19.6%)	140 (63.6%; 11.8%)	220 (100%; 13.8%)
	Other	287 (26.9%; 70.2%)	780 (73.1%; 65.7%)	1,067 (100%; 66.8%)
	Professional	42 (13.6%; 10.3%)	268 (86.5%; 22.6%)	310 (100%; 19.4%)
Rank	**Assistant**	145 (46.0%; 35.5%)	170 (54.0%; 14.3%)	315 (100%; 19.7%)
	Associate	138 (31.6%; 33.7%)	299 (68.4%; 25.2%)	437 (100%; 27.4%)
	Full	126 (14.9%; 30.8%)	719 (85.1%; 60.5%)	845 (100%; 52.9%)
All		409 (25.6%; 100.0%)	1,188 (74.4%; 100.0%)	1,597 (100%; 100%)

2. We are interested in estimating the probability distribution of time to faculty promotion to full professor from the time of promotion to associate professor.

    a. If you followed my suggestions on how to process the variables for the analyses that might be used to analyze this question, some subjects will be missing data. How would you characterize the reason for the missing data (what term might you use)? How does omission of these subjects from the analysis affect the scientific interpretation of the analyses?

<u>Answer</u>: Some faculty had already been promoted to full professor at the earliest time data was available. These subjects were thus "left censored". Some subjects had already been promoted to associate professor prior to the earliest time data was available. If we had

known when they were promoted to associate professor, we could have considered "left entry" of these subjects by considering them as part of the risk set starting at the time we had data for them. It would be incorrect to just use the data from the time of a known promotion to associate professor for these faculty, because those who were promoted to full professor quickly would not be represented. That is, suppose a faculty member had been promoted to associate professor in 1970 and was still associate in 1976. We could use this subject in risk sets starting 6 years after promotion to associate, but we could not allow this subject in the risk sets for 1 – 5 years after becoming associate. This is because other faculty members who should have been in the risk set 1 – 5 years after becoming associate might have been promoted to full professor, and we would not have their data.

A difficulty with all of the data for this problem is that we only have data on subjects who were still employed in 1995. We do not know about faculty members who were fired or who were lured away to other universities.

b. Provide suitable descriptive statistics for the distribution of times to promotion for faculty in the dataset.

Answer: Table 2c provides standard descriptive statistics for the times to promotion as computed using Kaplan-Meier estimates owing to the observations subject to censoring. Over the first 18 years following promotion to associate, faculty averaged 9.31 years prior to being promoted. Median time to promotion was 7 years, with 16.2% of faculty promoted to full professor within 4 years, and 57.5% taking longer than 6 years to be promoted

c. 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 remaining unpromoted for 4, 5, and 6 years for each stratum. Are the estimates suggestive that sex is associated with time to promotion? Give descriptive statistics supporting your answer.

Answer: Figure 2c presents sex-stratified Kaplan-Meier estimates of the probability of of associate professors remaining unpromoted at the university. As a rule, women had higher probabilities of remaining unpromoted over the time period with available data. Table 2c provides standard descriptive statistics for the times to promotion as computed using Kaplan-Meier estimates owing to the observations subject to censoring. Over the first 18 years following promotion to associate, male faculty averaged 9.08 years prior to being promoted, while women averaged 9.45 years as associate during the first 16 years after becoming associate. The difference in probability of promotion by 4, 5, or 6 years was 13.9%, 15.8%, and 13.0%, with women having a lower probability of promotion at each of those time points. These observations are suggestive that faculty member sex is associated with the time to promotion to full professor.

Table 2c: Descriptive statistics for the distribution of time (in years) to promotion 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), and the probability of not being promoted (the survivor function probability) by 4, 5, and 6 years.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(4,5,6 yr Surv Prob)
Females	170 (70)	9.45 (16)	8 (6, 13)	(≤2, >16)	(0.939, 0.817, 0.669)
Males	398 (221)	9.08 (18)	7 (5, 13)	(≤2, >18)	(0.800, 0.659, 0.539)
TOTAL	568 (291)	9.31 (18)	7 (5, 13)	(≤2, >18)	(0.838, 0.702, 0.575)

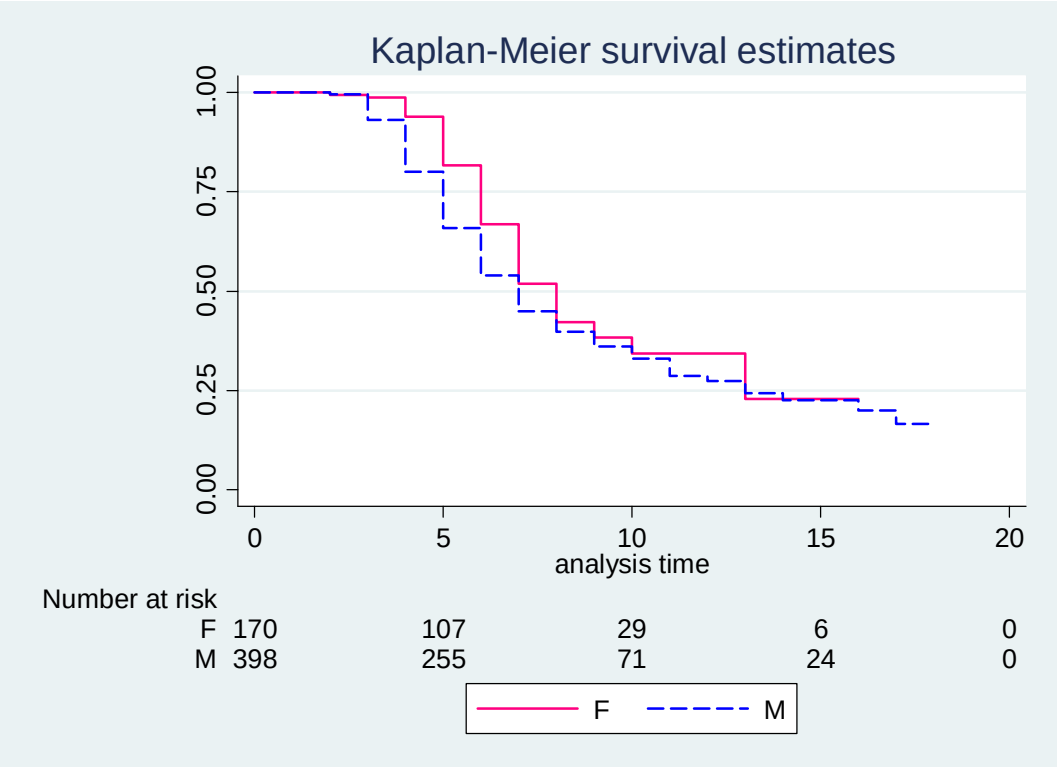

Figure 2c: Kaplan-Meier estimates of the probability of associate professors remaining unpromoted, stratified by sex.

d. Produce a plot of survival curves by the groups defined by academic field. 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 remaining unpromoted for 4, 5, and 6 years for each stratum. Are the estimates suggestive that academic field is associated with time to promotion? Give descriptive statistics supporting your answer.

Answer: Figure 2d presents field-stratified Kaplan-Meier estimates of the probability of of associate professors remaining unpromoted at the university. There is a clear tendency for faculty members in the professional fields to be promoted more rapidly than those in the fine arts, with faculty members in the other disciplines being intermediate to those extremes. Table 2d provides standard descriptive statistics for the times to promotion as computed using Kaplan-Meier estimates owing to the observations subject to censoring.

Over the first 18 years following promotion to associate, fine arts faculty averaged 10.6 years prior to being promoted and "other fields" faculty averaged 9.33 years prior to being promoted. Professional field faculty averaged 7.77 years as associate professor in the first 16 years after becoming an associate professor. ***(Note that the different time frame used for the restricted means complicates the comparison here: 7.77 years is an underestimate of the restricted mean based on 18 years. A good statistical program would make it easy for you to specify a time frame for the restricted mean, but Stata does not make it easy. You would have to censor all subjects at 16 years to make Stata provide the statistics you want.)*****, The probability of being promoted within 5 years is estimated to be 15.6%, 31.3%, and 39.4% for the fine arts, "other", and professional fields, respectively. These observations are suggestive that faculty member academic field is associated with the time to promotion to full professor.**

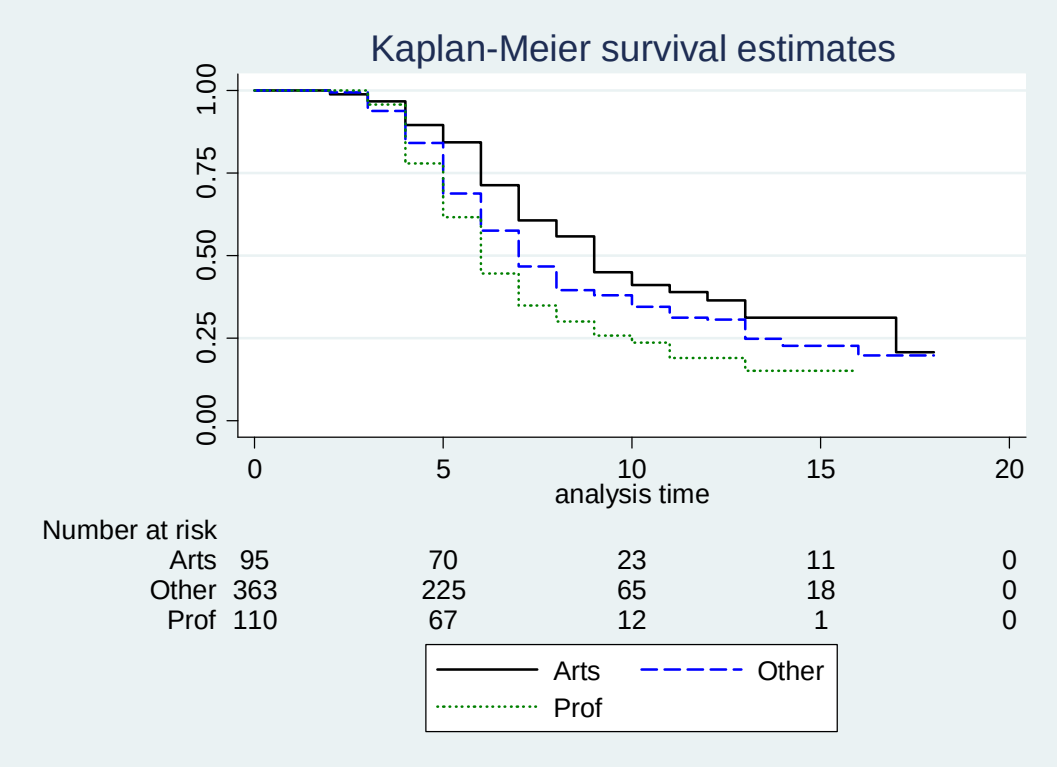

Figure 2d: Kaplan-Meier estimates of the probability of associate professors remaining unpromoted, stratified by sex.

Table 2d: Descriptive statistics for the distribution of time to promotion by academic field. 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), and the probability of not being promoted by 4, 5, and 6 years.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(4,5,6 yr Surv Prob)
Arts	95 (45)	10.6 (18)	9 (6, 17)	(≤2, >17)	(0.895, 0.844, 0.713)
Other	363 (183)	9.33 (18)	7 (5, 13)	(≤2, >16)	(0.840, 0.687, 0.575)
Prof	110 (63)	7.77 (16)	6 (5, 10)	(≤3, >16)	(0.779, 0.616, 0.446)
TOTAL	568 (291)	9.31 (18)	7 (5, 13)	(≤2, >18)	(0.838, 0.702, 0.575)

e. Produce a plot of survival curves by the groups defined both by sex and the academic field. 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 remaining unpromoted for 4, 5, and 6 years for each stratum. Are the estimates suggestive that field confounds the association between sex and time to promotion? Give descriptive statistics supporting your answer.

Answer: Figure 2e presents sex and field-stratified Kaplan-Meier estimates of the probability of of associate professors remaining unpromoted at the universityand table 2e provides standard descriptive statistics for the times to promotion as computed using Kaplan-Meier estimates owing to the observations subject to censoring. In each sex group, field is associated with monthly salary in 1995 as evidenced by the different median or restricted means across groups. ***(Note that different trends are observed for males and females, but I will discuss that in 2f.)*** **As noted in problem 1, there is also a different distribution of fields across the sexes (females are relatively more likely to be in fine arts than are males, and males are relatively more likely to be in the professional fields than females.). Thus, field could be judged a confounder when trying to judge discrimination against women in promotion, if we were not interested in the possibility that it was discrimination against women that led to the longer promotion times by field (A discriminatory view might be that there is no reason to rapidly promote in a field dominated by women) or the possibility that the preponderance of women in the arts was in fact the result of a discriminatory process.** ***(When I teach Biost 518/515, I use this data set as an example of a setting in which the identification of confounders is quite difficult. I typically come down on the side of treating field as a confounder in my primary analysis, because I think that the field-sex association might be partly due to historical (discminatory?) trends in the general population and partly due to personal preference of faculty members. But then I do some secondary analyses that regard field as more of a mediator of discrimination. The truth probably lies somewhere in between.)***

Table 2e: Descriptive statistics for the distribution of time (in years) to promotion by sex within academic field. 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), and the probability of not being promoted (the survivor function probability) by 4, 5, and 6 years.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(4,5,6 yr Surv Prob)
Arts: Females	39 (19)	8.81 (16)	7 (6, 13)	($\leq$2, >16)	(0.910, 0.840, 0.630)
Arts: Males	56 (26)	11.4 (18)	10 (7, >18)	($\leq$2, >18)	(0.886, 0.846, 0.757)
Other: Females	117 (47)	9.23 (16)	8 (6, 13)	($\leq$3, >16)	(0.942, 0.794, 0.662)
Other: Males	246 (135)	9.18 (18)	7 (5, 14)	($\leq$2, >18)	(0.798, 0.644, 0.539)
Prof: Females	14 (3)	11.6 (14)	>14 (7, >14)	($\leq$2, >14)	(0.939, 0.817, 0.669)
Prof: Males	96 (60)	7.05 (16)	6 (4, 8)	($\leq$3, >16)	(0.748, 0.574, 0.393)
TOTAL	568 (291)	9.31 (18)	7 (5, 13)	($\leq$2, >18)	(0.838, 0.702, 0.575)

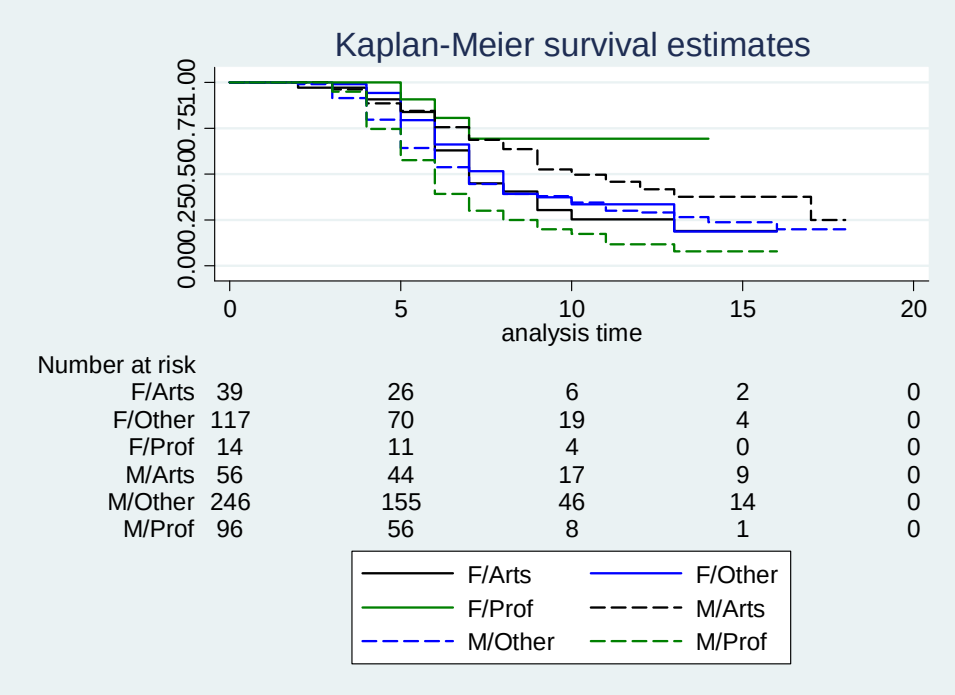

Figure 2e: Kaplan-Meier estimates of the probability of associate professors remaining unpromoted, stratified by sex and field.

f. Suppose we are interested in whether sex might be associated with time to promotion differently in the different academic fields (i.e., we are interested in whether academic field modifies the association between sex and time to promotion). Provide descriptive statistics addressing this question.

Answer: Looking again at Figure 2e and (especially) Table 2e, I note that when judging either the restricted means or medians, females tend toward a 3 year shorter promotion time than males in the fine arts (median 7 years for females, 10 years for males), females tend toward more than an 8 year longer promotion time than males in the professional fields (median greater than 14 years for females, 6 years for males), and there is a 1 year difference in median promotion times in the "other" academic fields (median 8 years for females, 7 years for males . Hence, there does seem to be some evidence that field modifies the association between sex and time to promotion.

3. The comparisons of time to promotion across sex groups might be potentially confounded by other variables.
   a. Would you *a priori* (before looking at the data) suspect that comparisons across sex groups of time to promotion might be confounded by year of degree? Explain. Provide descriptive statistics that explore the possibility of such confounding.

Answer: *A priori* I would have expected that men tended to receive their degree earlier than women, and I would guess that economic conditions and periods of university expansion might lead to differing patterns in timing of promotion over time. As I would be more interested in current (rather than historical) discrimination at the university, I would tend

to regard this as a confounder. Figure 3a and Table 3a present descriptive statistics exploring this possibility. In the subjects for whom we can assess time to promotion, we see that 70 of 170 (41%) female faculty members received their degree after 1980, while 130 of 398 (43%) male faculty members received their degree in that time frame. As I do not judge those proportions to be very different, I do not think the time of receiving degree confounds the analysis of associations between sex and time to promotion.

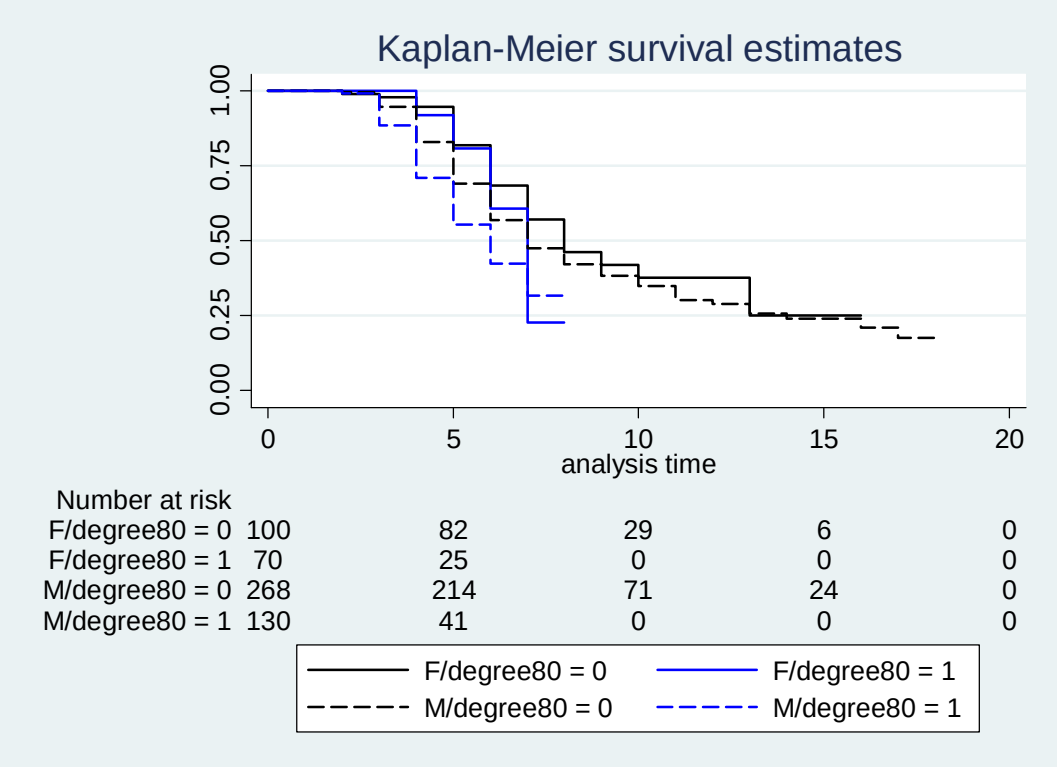

Figure 3a: Kaplan-Meier estimates of the probability of associate professors remaining unpromoted, stratified by whether the faculty member received his/her highest degree before or after 1980.

Table 3a: Descriptive statistics for the distribution of time (in years) to promotion by sex within groups defined by receiving degree after 1980. 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), and the probability of not being promoted (the survivor function probability) by 4, 5, and 6 years.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(4,5,6 yr Surv Prob)
Deg ≤80: Females	100 (55)	9.75 (16)	8 (6, >16)	(≤2, >16)	(0.946, 0.819, 0.685)
Deg ≤80: Males	268 (181)	9.37 (18)	7 (5, 14)	(≤2, >18)	(0.829, 0.690, 0.568)
Deg >80: Females	70 (15)	6.56 (8)	7 (6, 7)	(≤4, >8)	(0.919, 0.809, 0.607)
Deg >80: Males	130 (40)	5.88 (8)	6 (4, >8)	(≤2, >8)	(0.710, 0.554, 0.424)
TOTAL	568 (291)	9.31 (18)	7 (5, 13)	(≤2, >18)	(0.838, 0.702, 0.575)

b. Would you *a priori* (before looking at the data) suspect that comparisons across sex groups of time to promotion might be confounded by calendar year at the

time the faculty member first became an associate professor? Explain. Provide descriptive statistics that explore the possibility of such confounding.

Answer: *A priori* I would have expected that men tended to have been promoted to associate earlier than women, and I would guess that economic conditions and periods of university expansion might lead to differing patterns in timing of promotion over time. As I would be more interested in current (rather than historical) discrimination at the university, I would tend to regard this as a confounder. Figure 3b and Table 3b present descriptive statistics exploring this possibility. In the subjects for whom we can assess time to promotion, we see that 67 of 170 (40%) female faculty members became associate before 1985, while 198 of 398 (50%) male faculty members became associates in that time frame. As I do judge those proportions to be somewhat different, I then note that the medians and the probabilities of being promoted within 4, 5, or 6 years are relatively similar for groups defined by time of becoming associate, so there does not appear to be a strong association between time to promotion to full and date of becoming an associate. *(I did not want to use the restricted means in this instance, because they were estimated over such different time restrictions. With a little work I could have made Stata give me what I wanted, but the medians and survival curves gave me enough of a sense to go on..)* Hence, in the end, I do not think date of becoming associate is much of a confounder.

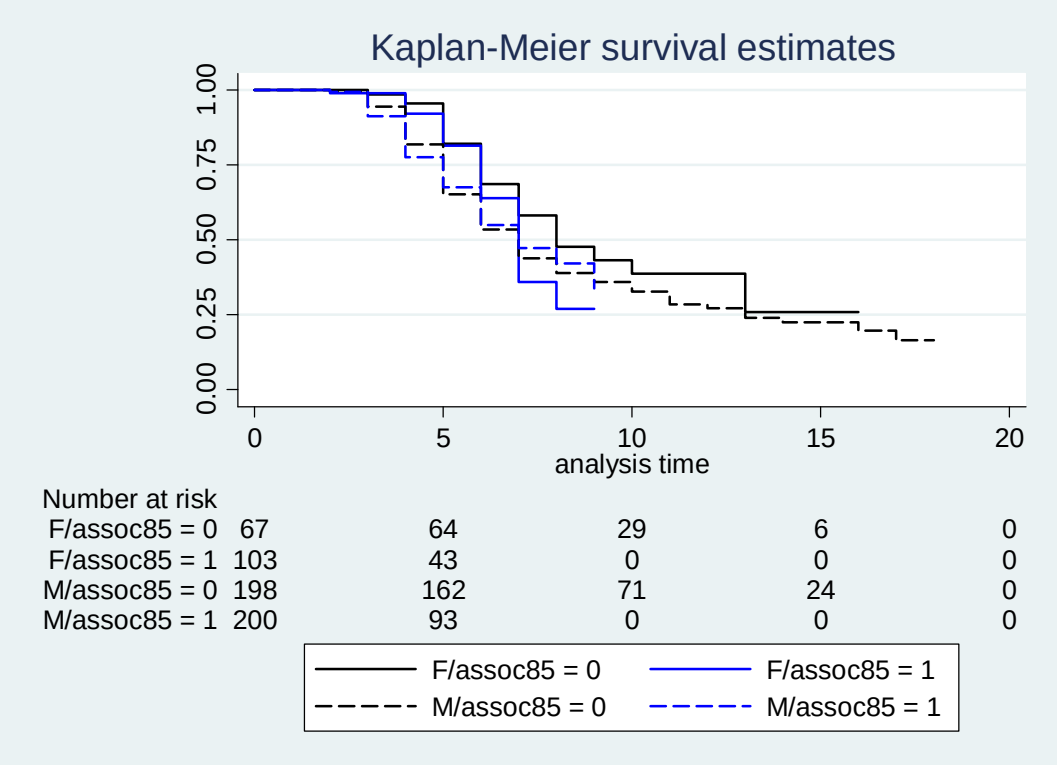

Figure 3b: Kaplan-Meier estimates of the probability of associate professors remaining unpromoted, stratified by whether the faculty member became an associate professor before or after 1985.

Table 3b: Descriptive statistics for the distribution of time (in years) to promotion by sex within groups defined by becoming associate after 1985. 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), and the probability of not being promoted (the survivor function probability) by 4, 5, and 6 years.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(4,5,6 yr Surv Prob)
Assoc ≤85: Females	67 (46)	9.88 (16)	8 (6, >16)	(≤3, >16)	(0.955, 0.821, 0.687)
Assoc ≤85: Males	198 (152)	9.07 (18)	7 (5, 13)	(≤2, >18)	(0.818, 0.652, 0.535)
Assoc >85: Females	103 (24)	6.99 (9)	7 (6, >9)	(≤2, >9)	(0.922, 0.815, 0.640)
Assoc >85: Males	200 (69)	6.80 (9)	7 (5, >9)	(≤2, >9)	(0.776, 0.676, 0.549)
TOTAL	568 (291)	9.31 (18)	7 (5, 13)	(≤2, >18)	(0.838, 0.702, 0.575)

4. Provide the sample correlation between year of hire at the university and monthly salary in 1995 for all faculty members in the dataset, as well as separately for men and women. How might you explain the differences between the overall correlation and the stratum specific correlation? Justify your answer with descriptive statistics.

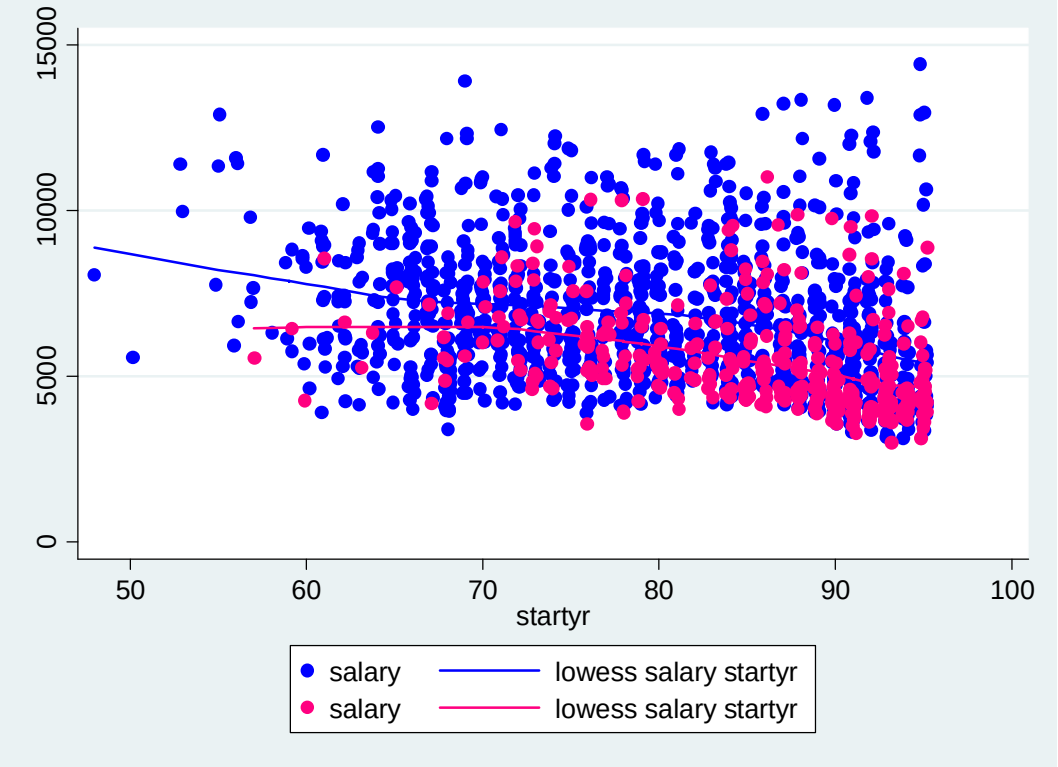

Figure 4: Scatterplot (and superimposed lowess smooths) of monthly salary in 1995 by year hired at the university. Males are denoted in blue, and females in pink.

Table 4: Correlations between monthly salary in 1995 and the year hired at the university for all faculty as well as separately for males and females. Also included are the descriptive statistics that contribute to the variance.

	All Subjects	Males	Females
Correlation (r)	- 0.344	- 0.271	- 0.403
LS slope (β)	-70.0	-55.6	-74.5
SD (Salary | Start Year)	1913	2013	1357
SD (Start Year)	9.99	10.2	8.02

Answer: Correlations and relevant descriptive statistics are displayed in Table 4, with a scatterplot displayed in Figure 4. While we could make guesses about the slopes and variances from the graph, it is useful to look at the numbers. In Stata we could get the slopes and "error" standard deviation using the "regress" command. For instance, for the combined group I used

```
. regress salary startyr if year==95

      Source |       SS       df       MS              Number of obs =    1597
-------------+------------------------------           F(  1,  1595) =  213.43
       Model |   781407281     1   781407281           Prob > F      =  0.0000
    Residual |  5.8395e+09  1595  3661133.64           R-squared     =  0.1180
-------------+------------------------------           Adj R-squared =  0.1175
       Total |  6.6209e+09  1596  4148443.26           Root MSE      =  1913.4

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     startyr |   -70.01917   4.792764   -14.61   0.000    -79.41995   -60.61839
       _cons |   12069.47   391.7064    30.81   0.000     11301.16    12837.79
------------------------------------------------------------------------------
```

I have put in bold the two numbers I would want from all this output

- **The least squares slope of -70.0 is in the column labeled "Coef" and the row named "startyr". It is interpretable as the estimated difference in mean salary per 1 year difference in starting year.**
- **The estimated standard deviation of 1913 in a group that has the same starting year is labeled "Root MSE" (square root of the mean squared error), and it is computed by pretending that that standard deviation would be the same in every starting year group.**

Results for the two sexes separately were obtained using the "bysort sex:" prefix with the "regress" command.

We find that the correlation in the combined group is -0.344, with a less extreme correlation of -0.271 seen in males and a more extreme correlation of -0.403 in females. Factors that contribute to the more extreme correlation in females compared to males is the more negative slope (about 50% more negative) and the lesser "error" variance (standard deviation of salary in groups that are homogeneous with respect to start year in females is about 65% of that in males). While the 20% lower variation of starting year among females would tend to lead to a less extreme correlation than for males, that is not enough to counter the more marked differences on the slope and the error variance. Note that in the combined group, the slope is in close agreement to that for females, but the variability of starting year and error variance is about the same as that for males. Hence the combined group has a correlation intermediate to those for the two sexes.

Questions for Biost 514 only:

5. Consider a continuous random variable *X* having density *f(x)* and cumulative distribution function *F(x)*. Define survivor function *S(x) = 1 – F(x)*. Further suppose that *E(X)* is finite. Show that

$$E[X] = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty} S(x)dx$$

Assignment for Discussion Sections
Mon Oct 22 – Fri Oct 26

We will be discussing descriptive statistics for the dataset on FEV and smoking in children. You should come to discussion section prepared to talk about your findings as you describe the sample univariately and bivariately. (You need not spend more than an hour looking at the descriptive statistics, but you should have looked at them.)