Biost 517: Applied Biostatistics I
Emerson, Fall 2005

Homework #4
November 2, 2005

A file containing the annotated Stata commands I used to solve this homework is available on the class web pages.

Written problems: To be handed in at the beginning of class on Monday, October 31, 2005.

> *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 - 3 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. Because this is a very large file, you will also have to tell Stata to increase the amount of memory it is using for data.

Salary raises and inflation are most often expressed on a multiplicative scale. That is, we talk about a percentage raise or percentage inflation. For this reason, it is often both scientifically and statistically preferable to analyze salary data after a logarithmic transformation. This is often equivalent to summarizing the salary distribution by geometric means and comparing distributions across groups by the ratio of geometric means. In Problems 1-3, I ask for statistical analysis on the scale of the log monthly salary. This can be effected in Stata by generating a new variable:

- `generate logsalary = log(salary)`

(in Stata, the log( ) function computes the natural log, which you may have previously encountered as ln ( )).

In the first three problems, you are asked to produce scatter plots with superimposed lowess smooths and/or least squares lines. The following command (which should all be typed into the Commands window prior to hitting ENTER) would produce a scatter plot of 1995 salary by year first hired at the university. On this graph, males and females would be displayed in different colors, and the lowess and least squares estimated lines for each sex would be displayed as solid and dashed lines, respectively, in the same color. I also include the lowess and least squares lines for the entire sample in black:

```
twoway (scatter salary startyr if year==95 & female==0, jitter(1) col("blue"))
       (lowess salary startyr if year==95 & female==0, col("blue"))
       (lfit salary startyr if year==95 & female==0, col("blue") lp("-"))
       (scatter salary startyr if year==95 & female==1, jitter(1) col("red"))
       (lowess salary startyr if year==95 & female==1, col("red"))
       (lfit salary startyr if year==95 & female==1, col("red") lp("-"))
       (lowess salary startyr if year==95, col("black"))
       (lfit salary startyr if year==95, col("black") lp("-"))
```

(The above graph is perhaps a bit busy, but I just gave all the commands so you could see what the commands do.)

You are also asked to find correlations, both in the entire sample and within strata. This can be effected through the use of the command `correlate` with and without the `bysort` prefix. For instance, the correlation between the logarithm of the 1995 monthly salary and the year first hired at the university could be obtained for the entire sample and within sex strata by:

```
cor logsalary startyr if year==95
bysort female: cor logsalary startyr if year==95
```

In solving Problems 1 – 3, 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 1995 monthly salary and year hired, I might choose to regard salary as the "response" and year hired as the "predictor" to examine:

- The correlation between salary and year hired using commands as given above.
- The variance of year hired using `summ startyr if year==95` to obtain the 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 salary startyr if year==95`, which would generate output looking like

```
. 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
------------------------------------------------------------------------------
```

From this voluminous output, we would (at this time) be interested in only two numbers. The least squares estimate of the slope is the number in the row labeled "startyr" (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 monthly salary averages $70.02 less for every year difference in starting year (with more recent hires earning less money). The estimated standard deviation in each group hired during the same year is labeled "Root MSE", and in the above table is estimated as $1,913.4. (I note that this estimates the standard deviation averaged across all starting years.) 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:

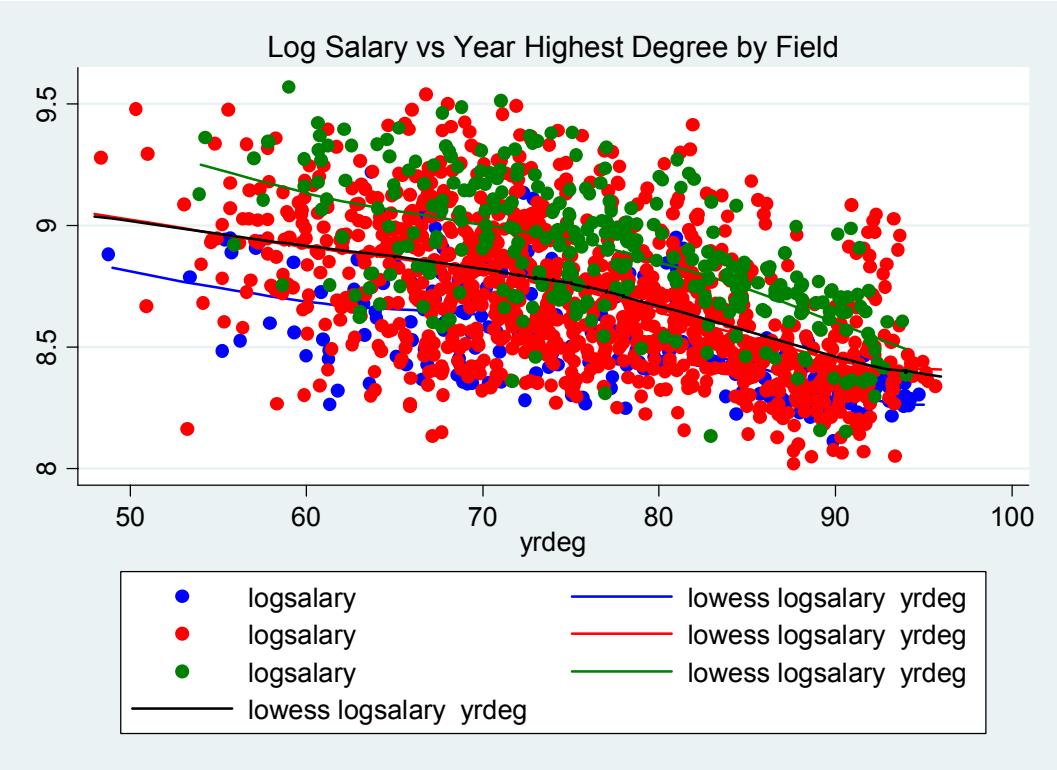
```
bysort female: regress salary startyr if year==95
```

1. Produce a scatterplot of the logarithm of monthly salary in 1995 (on the Y axis) versus the year that the highest degree was obtained (on the X axis). Use a different symbol or color for each field, and display stratified lowess smooths on the plot. (You could also display least squares fits to be able to assess the slope of the best fitting linear trend.

    a. What is the correlation between the log salary and year of highest degree?

    b. What is the correlation between log salary and year of highest degree for each field separately?

    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 lower in the combined sample than it was in each stratum defined by field? 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.



<u>Ans:</u> Above I present a scatterplot of log salary versus year of highest degree for the year 1995. Each field is displayed in a different color: Professional=green, Arts=blue, Other=red. I superimpose a lowess curve for all the data in black, as well as lowess curves for each field separately (in their respective colors). I note the following:

- **The black lowess curve is estimating the association between year of highest degree and log salary unadjusted for field (or any other variable). We see that there is an overall trend toward lower (log) salaries for faculty who received their degree more recently. Furthemore, the trend between logsalary and year of highest degree looks reasonably linear, suggesting that the geometric mean of**

salaries is higher by a relatively constant percentage with each year difference in the year since obtaining the highest degree. There does seem to be decreased spread of log salaries about the lowess curve for faculty receiving their degrees more recently than for faculty who received their degree a long time ago. (Neither the trend toward lower mean log salaries with more recent degrees nor the trend toward decreased variability of the log salaries within groups having the same year of degree are particularly surprising: Greater experience generally is rewarded with higher salaries, and there is quite reasonably less variation in starting salaries than for salaries after a faculty member has been at the university a while.

- **The stratified lowess curves estimate the association between year of highest degree and log salary within each field. From these lowess curves we can deduce**
  - **Each field shows a similar trend toward lower (log) salaries for more recent graduates.**
  - **The fact that these lines appear roughly parallel suggests that the association between log salary and year of degree is not modified by field.**
  - **The vertical separation of the three lowess curves suggests that there is an association between (log) salary and field after adjusting for year of highest degree. The degree of vertical separation would quantify that adjusted association.**
  - **The fact that there is vertical separation of the three curves also would suggest that some part of the variability of log salary within groups defined by year of degree can be "explained" by there being different fields. That is, the separation of the curves will dictate that the within yrdeg group variance of log salary for each field will be less than the within yrdeg group variance of log salary for the sample as a whole.**
  - **An average of the slopes of the three lowess curves would be a measure of the association between log salary and year of degree after adjusting for field.**
  - **The spread of the points around their respective lowess curves gives a clue as to variability of log salary for each field within groups having the same year of degree. In order to use these graphs to judge whether the variance is equal, we would of course have to make sure that we were judging the range of equal numbers of observations.**
  - **The range of year of degree observed for each stratum gives an idea about the variability of year of degree for each field.**

In the following table, I present the correlation, the slope of the least squares fitted line, the standard deviation of year of degree, and the estimated average standard deviation of log salary within groups having the same year of degree for the entire sample, as well as for each field separately.

	Correlation (r)	LS Slope (β)	SD(yrdeg)	SD(logsal | yrdeg)
Overall Sample	**-0.510**	**-0.0157**	**9.86**	**0.261**
Within Field:				
Professional	**-0.604**	**-0.0178**	**9.40**	**0.221**
Arts	**-0.583**	**-0.0128**	**10.3**	**0.183**
Other	**-0.529**	**-0.0160**	**9.90**	**0.255**

It can be seen that the correlation between log salary and year of degree is -0.510 for the combined sample, with more extreme negative correlations observed in each field considered separately (Professional -0.604; Arts -0.583; Other -0.529). This behavior can be explained as follows. Correlation is influenced by:

- **The slope β of the least squares line. The correlation will be the same sign as the slope, and a slope that is higher in absolute value will tend to increase the absolute value of the correlation.**
- **The variance Var(X) of the "predictor". Higher variability of the predictor variable will tend to cause the correlation to be higher in absolute value (closer to 1 or -1, according to the sign of the slope).**
- **The variability of the response Y in groups that have similar values of the predictor X: Var(Y | X). Correlation tends to decrease in absolute value (so get closer to 0) as the within group variance increases.**

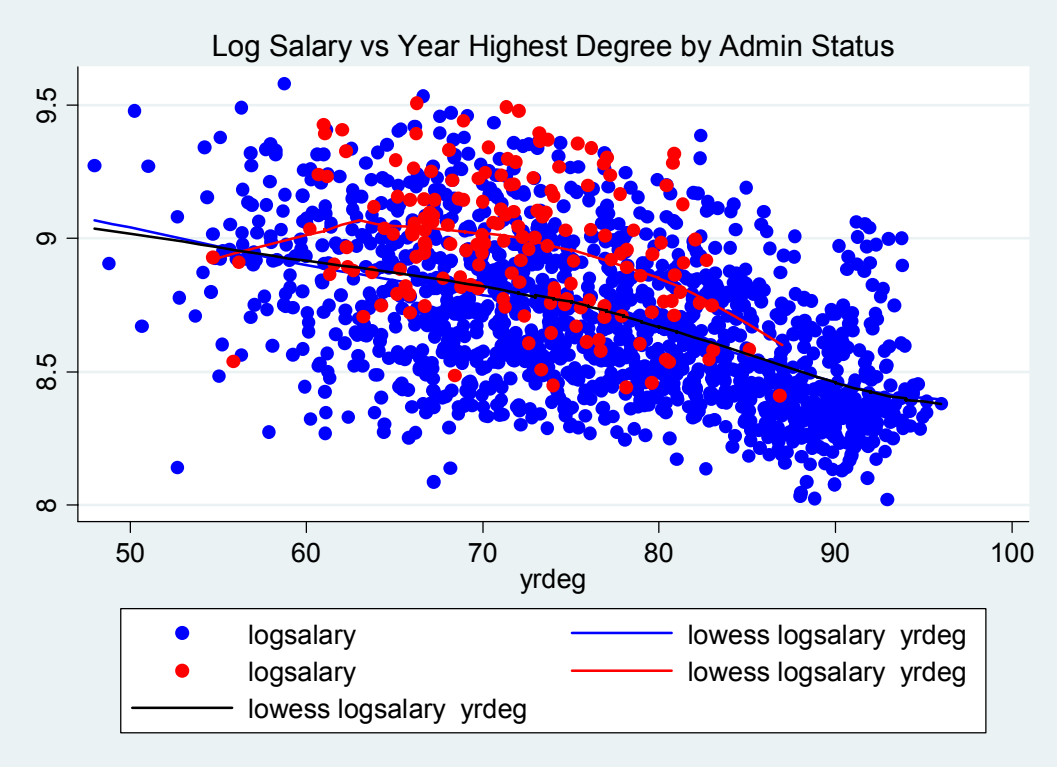
For these analyses, the slopes and variance of the predictors are nearly the same in the combined data and in each field separately. The within yrdeg group variance, however, tends to be smaller in the stratified analyses than in the combined sample, thus leading to the more extreme negative correlations within strata.

2. Produce a scatterplot of the logarithm of monthly salary in 1995 (on the Y axis) versus the year that the highest degree was obtained (on the X axis). Use a different symbol or color according to administrative duties, and display stratified lowess smooths on the plot. (You could also display least squares fits to be able to assess the slope of the best fitting linear trend.

    a. What is the correlation between the log salary and year of highest degree?

    b. What is the correlation between log salary and year of highest degree for administrators and non-administrators separately?

    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 higher or lower in the combined sample than it was in each stratum defined by administrative duties? 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: **Above I present a scatterplot of log salary versus year of highest degree for the year 1995. Each administrative group is displayed in a different color: nonAdmin=blue, Admin=red. I superimpose a lowess curve for all the data in black, as well as lowess curves for each field separately (in their respective colors). I note the following:**

- **(See the answer to problem 1 for comments about the results in the overall sample.)**
- **The stratified lowess curves estimate the association between year of highest degree and log salary within each admin group. From these lowess curves we can deduce**
  - **Each admin group shows a similar trend toward lower (log) salaries for more recent graduates.**
  - **The fact that these lines appear roughly parallel suggests that the association between log salary and year of degree is not modified by admin group. (We need to be aware that lowess curves are not too reliable in the end of the range, so I don't make too much of the curvature near the ends—but I could be wrong.)**
  - **The vertical separation of the two lowess curves suggests that there is an association between (log) salary and admin duties after adjusting for year of highest degree. The degree of vertical separation would quantify that adjusted association.**
  - **The fact that there is vertical separation of the two curves also would suggest that some part of the variability of log salary within groups defined by year of degree can be "explained" by there being different admin groups. That is, the separation of the curves will dictate that the within yrdeg group variance of log salary for each admin group will be less than the within yrdeg group variance of log salary for the sample as a whole.**

- **An average of the slopes of the two lowess curves would be a measure of the association between log salary and year of degree after adjusting for admin.**
- **The spread of the points around their respective lowess curves gives a clue as to variability of log salary for each admin level within groups having the same year of degree. In order to use these graphs to judge whether the variance is equal, we would of course have to make sure that we were judging the range of equal numbers of observations.**
- **The range of year of degree observed for each stratum gives an idea about the variability of year of degree for each field. There is a clear trend toward decreased range of yrdeg for the admin group relative to the group having no administrative duties (though we do need to consider the sample sizes as we try to equate range with variance).**

In the following table, I present the correlation, the slope of the least squares fitted line, the standard deviation of year of degree, and the estimated average standard deviation of log salary within groups having the same year of degree for the entire sample, as well as for each field separately.

	Correlation (r)	LS Slope (β)	SD(yrdeg)	SD(logsal | yrdeg)
Overall Sample	**-0.510**	**-0.0157**	**9.86**	**0.261**
Within Field:				
nonAdmin	**-0.506**	**-0.0148**	**10.1**	**0.254**
Admin	**-0.287**	**-0.0111**	**6.30**	**0.235**

It can be seen that the correlation between log salary and year of degree is -0.510 for the combined sample, with less extreme negative correlations observed in the Admin group considered separately (nonAdmin -0.506; Admin -0.287). This behavior can be explained as follows. (See problem 1 for a discussion of the determinants of correlation.)

For these analyses, the slopes within admin groups are not very different, but, interestingly, the slope of the combined sample is more negative than the slope in either subgroup. This suggests a slight confounding. The within yrdeg group variance of log salary is not very different. I conclude that the major factor decreasing the absolute value of the correlation within the Admin group relative to the combined sample is the much lower standard deviation of yrdeg in the Admin group.

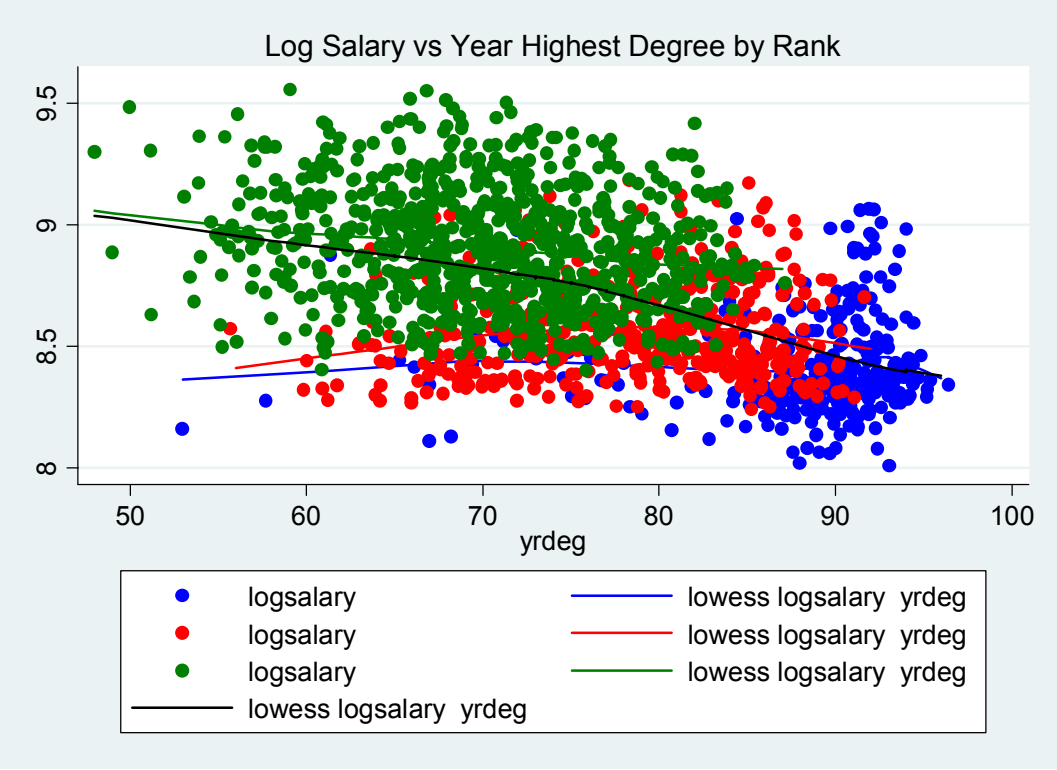
3. Produce a scatterplot of the logarithm of monthly salary in 1995 (on the Y axis) versus the year that the highest degree was obtained (on the X axis). Use a different symbol or color for each rank, and display stratified lowess smooths on the plot. (You could also display least squares fits to be able to assess the slope of the best fitting linear trend.

   a. What is the correlation between the log salary and year of highest degree?

   b. What is the correlation between log salary and year of highest degree for each rank separately?

   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 higher or lower in the combined sample than it was in each stratum defined by field? 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.



<u>Ans:</u> Above I present a scatterplot of log salary versus year of highest degree for the year 1995. Each rank is displayed in a different color: Full=green, Associate=red, Assistant=blue,). I superimpose a lowess curve for all the data in black, as well as lowess curves for each field separately (in their respective colors). I note the following:

- **(See the answer to problem 1 for comments about the results in the overall sample.)**
- **The stratified lowess curves estimate the association between year of highest degree and log salary within each rank. From these lowess curves we can deduce**
  - **The ranks show slightly different patterns with respect to the trend in (log) salaries for more recent graduates. Assistant and associate professors have a slight hint of a U-shaped function in which the most recent graduates and the most distant graduates have lower salaries than faculty with intermediate time since highest degree. Full professors show a more linear trend toward lower salaries for the professors having obtained their degrees most recently. (I note that these results are not that unexpected: It is somewhat unusual for faculty to stay as assistant or associate professors for long periods of time, so I would surmise that the junior faculty who received their degrees a long time ago may not be as marketable.)**
  - **The fact that different patterns are observed across the ranks suggests some element of effect modification.**
  - **The vertical separation of the three lowess curves suggests that there is an association between (log) salary and rank after adjusting for year of**

highest degree. The degree of vertical separation would quantify that adjusted association. Because the amount of the separation differs by year of degree, the strength of the "rank effect" also differs by year of degree.

- **The fact that there is vertical separation of the three curves also would suggest that some part of the variability of log salary within groups defined by year of degree can be "explained" by there being different ranks. That is, the separation of the curves will dictate that the within yrdeg group variance of log salary for each rank will be less than the within yrdeg group variance of log salary for the sample as a whole.**
- **An average of the slopes of the three lowess curves would be a measure of the association between log salary and year of degree after adjusting for field. Of course, in the presence of effect modification, this averaging of the slopes may not be desirable.**
- **The spread of the points around their respective lowess curves gives a clue as to variability of log salary for each field within groups having the same year of degree. In order to use these graphs to judge whether the variance is equal, we would of course have to make sure that we were judging the range of equal numbers of observations.**
- **The range of year of degree observed for each stratum varies markedly across ranks. This means that the most recent graduates tend to be assistant professors, and the most distant graduates tend to be associate professors. Because rank is also strongly associated with (log) salary, we obtain a much stronger decreasing trend in log salary with year of degree in the combined group than is observed in any of the strata. To the extent that rank is not within the causal pathway of interest and to the extent that any effect modification is not of primary concern, we would consider rank a confounder of the association between log salary and year of degree.**

In the following table, I present the correlation, the slope of the least squares fitted line, the standard deviation of year of degree, and the estimated average standard deviation of log salary within groups having the same year of degree for the entire sample, as well as for each field separately.

	Correlation (r)	LS Slope (β)	SD(yrdeg)	SD(logsal | yrdeg)
Overall Sample	**-0.510**	**-0.0157**	**9.86**	**0.261**
Within Rank:				
Assistant	**0.0457**	**0.00147**	**6.37**	**0.205**
Associate	**0.0520**	**0.00141**	**7.20**	**0.196**
Full	**-0.1978**	**-000660**	**7.23**	**0.237**

It can be seen that the correlation between log salary and year of degree is -0.510 for the combined sample, with correlations much closer observed in each rank considered separately (Assistant 0.0457, Associate 0.0520, Full -0.1978). This behavior can be explained as follows. (See problem 1 for a discussion of the determinants of correlation.)

For these analyses, the variance of the predictors are nearly the same in each of the ranks, and that variance is much smaller than in the combined sample. This will tend to make the correlation in each rank closer to zero than it is for the combined sample. The

within yrdeg group variance, however, tends to be smaller in the stratified analyses than in the combined sample, thus leading to the more extreme negative correlations within strata. Finally, the slope estimate is near zero for the assistant and associate professors and only moderate for the full professors. None of the slopes are as negative as the slope in the combined group, because of the confounding: Assistant professors (who have lower salaries) tend to have received their degree more recently, and full professors (who have the highest salaries(tend to have received their degree long ago. The major influence here, then, is probably the confounding leading to the steeper slope in the combined data.

The following problems make use of a dataset exploring the prognostic value of prostate specific antigen (PSA) on hormonally treated prostate cancer. The documentation file psa.doc and the data file psa.txt can be found on the class web pages. (Note that the variable `inrem` is a string variable and there are several variables containing missing data.)

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

- You will need to create a numeric variable indicating which observation times are not censored. For instance:
```
gen relapse= 0
replace relapse= 1 if inrem=="no"
```
- You will need to declare the variables representing the possibly censored times to relapse: `stset obstime relapse`
- To obtain a graph of survival curves, you can then just use `sts graph`. (If you want stratified curves by, say, tumor grade, you use the `by( )` option: `sts graph, by(grade).`)
- 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 15 month survival probabilities would be obtained by `sts list, at(6 15).`

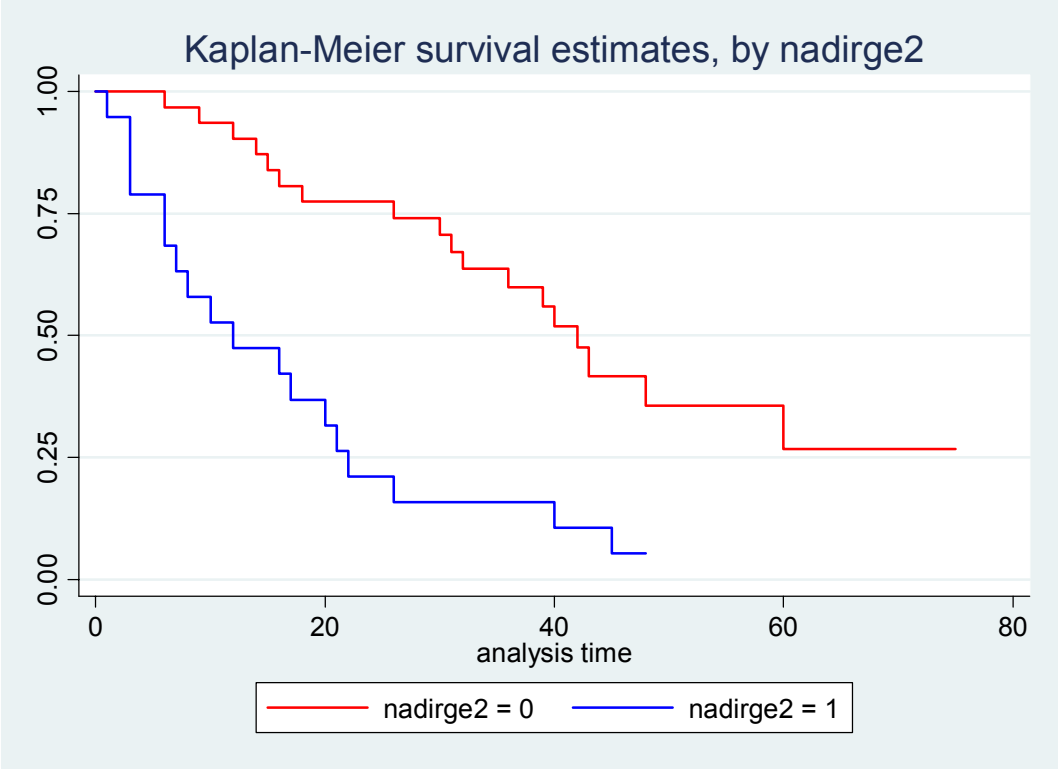
4. We are interested in estimating the probability of a patient remaining in relapse following hormonal treatment for prostate cancer.

    a. Provide suitable descriptive statistics for the distribution of times in remission for men receiving hormonal treatment for prostate cancer.

<u>Ans:</u> Fifty men were followed for signs of relapse for a minimum of 24 months following hormonal treatment of prostate cancer. The estimated time at which 75%, 50%, and 25% of men remain in remission is 12 months, 30 months, and 48 months, respectively. The estimated probability of remaining in remission for 1, 2, 3, or 4 years is 0.74, 0.56, 0.43, and 0.23, respectively.

    b. Produce a plot of relapse free survival curves by the groups defined by whether the nadir PSA value was less than 2 ng/ml or not. Produce a table of estimates of the 75th, 50th, and 25th percentiles of the survival distribution by nadir PSA strata. Also include in that table the estimated probabilities of surviving in remission for 12, 24, 36, and 48 months for each stratum. Are the estimates suggestive that nadir PSA level affects relapse free survival? Give descriptive

statistics supporting your answer.

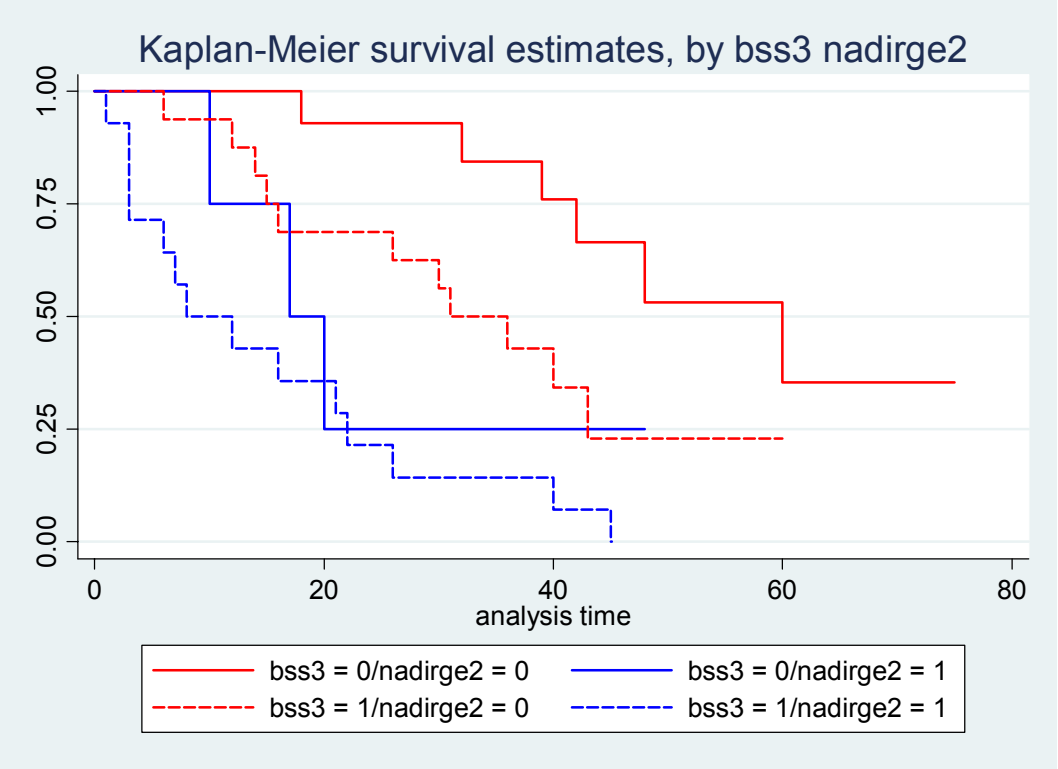


Ans: **The above graph displays the probability of remaining in remission by time since receiving hormonal treatment for prostate cancer. Estimates are stratified according to whether the nadir PSA was less than 2 ng/ml (red line) or greater than 2 ng/ml (blue line). The following table presents estimated quartiles of the times in remission (expressed for the times at which 75%, 50%, and 25% of the population would still be in remission), as well as the estimated probabilities of remaining in remission for 1, 2, 3, or 4 years. Estimates are provided for the strata defined by whether the nadir PSA exceeded 2 ng/ml. From this table, it is clear that there is a trend toward longer time in remission for subjects with a nadir PSA below 2 ng / ml than for subjects with a nadir PSA in excess of 2 ng / ml. For instance, the estimated probability of remaining in remission for 3 years is 0.599 if the nadir PSA is less than 2 ng / ml (95% CI: 0.401 to 0.750) but only 0.158 if the nadir PSA is more than 2 ng / ml (95% CI: 0.039 to 0.349). Furthermore, because the 95% CI do not overlap, we can with 95% confidence reject the null hypothesis that the probability of remaining in remission for three years is equal across these two strata.**

Nadir PSA (ng/ml)	Quartiles for Remission Time			Prob of Remaining in Remission			
	75%	50%	25%	1 Year	2 Years	3 Years	4 Years
< 2.0	26 mos	42 mos	> 75 mos	0.903	0.774	0.599	0.357
≥ 2.0	6 mos	12 mos	22 mos	0.474	0.211	0.158	0.053

c. Suppose we are interested in whether the nadir PSA provides information about time to relapse independent of bone scan score. How would you assess whether your answer to part b was merely reflective of confounding by bone scan score? Perform such an analysis and provide descriptive statistics addressing the

possibility of an association between time to relapse and nadir PSA that is independent of the bone scan measurements.



Ans: In order to assess whether nadir PSA is associated with time in remission independent of any association between nadir PSA and bone scan score, we merely need to do an analysis which compares the association between time in remission and nadir PSA within groups that have similar bone scan scores. Because of relatively few observed relapses among patients having the lowest bone scan score, I consider two strata according to whether bone scan score is less than 3 or equal to 3. The above figure and the following table present the estimated probability of remaining in remission for groups defined both by bone scan score strata and nadir PSA strata. From comparisons made within bone scan score strata, we see that a nadir PSA greater than 2 is associated with shorter times in remission in both strata: For bone scan score less than 3, the probability of remaining in remission for at least 24 months is 0.929 when the nadir PSA is less than 2 and 0.250 when the nadir PSA is greater than 2. For bone scan score equal to 3, the probability of remaining in remission for at least 24 months is 0.688 when the nadir PSA is less than 2 and 0.214 when the nadir PSA is greater than 2. (In later homeworks, we will perform statistical hypothesis tests by combining these results across strata.)

	Probability of Remaining in Remission		
Months Post Treatment	Estimate	95% Conf Interval (Low Bound)	95% Conf Interval (High Bound)
Bone Scan Score < 3; Nadir PSA < 2 ng/ml			
12	1.000	.	.
24	0.929	0.591	0.990
36	0.844	0.504	0.959
48	0.532	0.196	0.783
Bone Scan Score < 3; Nadir PSA > 2 ng/ml			

12	0.750	0.128	0.961
24	0.250	0.009	0.665
36	0.250	0.009	0.665
48	0.250	0.009	0.665
Bone Scan Score = 3; Nadir PSA < 2 ng/ml			
12	0.875	0.586	0.967
24	0.688	0.405	0.856
36	0.429	0.188	0.651
48	0.229	0.047	0.491
Bone Scan Score = 3; Nadir PSA > 2 ng/ml			
12	0.429	0.177	0.660
24	0.214	0.052	0.448
36	0.143	0.023	0.366
48	.	.	.

5. Suppose we are interested in using the nadir PSA to predict whether a patient will still be in remission two years after receiving hormonal treatment.

General comments: Because all men were followed for a minimum of 24 months, we are able to answer this question without worrying about censoring. The annotated Stata log file posted on the class web pages show the way that I computed variables to be able to answer these questions.

a. What is the prevalence of relapse within 24 months in our sample?

Ans: Twenty-two of the 50 men relapsed within 24 months, thus the estimated prevalence of relapse is 44%.

b. Suppose we consider a threshold of a nadir PSA greater than 2 ng/ml to be a "positive" test result. What are the sensitivity and specificity of such a diagnostic criterion? Briefly explain how these were calculated.

Ans: As detailed in the Stata log file (and in the class notes), we can estimate the sensitivity by considering the proportion of men who had a nadir PSA greater than 2 among all men who relapsed within 24 months. There were 22 men who relapsed within 24 months, and of those, 68.2% had a nadir PSA greater than 2, so the estimated sensitivity is 68.2%.

There were 28 men who did not relapse within 24 months, and of those, 14.3% had a nadir PSA greater than 2. Thus the estimated specificity is 85.7%.

c. If the sample accurately reflects the patient population of interest, what are the positive and negative predictive values of such a diagnostic criterion? Briefly explain how these were calculated.

Ans: If the sample accurately reflects the patient population of interest, then the positive and negative predictive values can be computed directly from the cross-sectional study.

There were 19 men who had a nadir PSA greater than 2, and of those, 78.9% relapsed within 24 months. The positive predictive value is 78.9%.

There were 31 men who had a nadir PSA less than 2, and of those 22.6% relapsed within 24 months. The negative predictive value is thus 77.4%.

d. Suppose instead that the sample that we obtained oversampled patients who would actually have relapsed. If the true prevalence of relapse in the target population were 40%, what would be the positive and negative predictive values of the diagnostic criterion based on a PSA greater than 2 ng/ml? Briefly explain how these were calculated.

Ans: If we want to use a different prevalence than that observed in the study, we need to use Bayes' rule to calculate the positive and negative predictive values:

$$PVP = \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev)}$$

$$= \frac{0.682 \times 0.400}{0.682 \times 0.400 + 0.143 \times 0.600} = 0.761$$

$$PVN = \frac{Spec \times (1 - Prev)}{Spec \times (1 - Prev) + (1 - Sens) \times Prev}$$

$$= \frac{0.857 \times 0.600}{0.857 \times 0.600 + 0.318 \times 0.400} = 0.802$$

e. Repeat parts b-d using a threshold of a PSA greater than 4 ng/ml.

Ans: As detailed in the Stata log file (and in the class notes), we can estimate the sensitivity by considering the proportion of men who had a nadir PSA greater than 4 among all men who relapsed within 24 months. There were 22 men who relapsed within 24 months, and of those, 68.2% had a nadir PSA greater than 4, so the estimated sensitivity is 68.2%.

There were 28 men who did not relapse within 24 months, and of those, 7.1% had a nadir PSA greater than 4. Thus the estimated specificity is 92.9%.

If the sample accurately reflects the patient population of interest, then the positive and negative predictive values can be computed directly from the cross-sectional study.

There were 17 men who had a nadir PSA greater than 4, and of those, 88.2% relapsed within 24 months. The positive predictive value is 88.2%.

There were 33 men who had a nadir PSA less than 4, and of those 21.2% relapsed within 24 months. The negative predictive value is thus 78.8%.

If we want to use a different prevalence than that observed in the study, we need to use Bayes' rule to calculate the positive and negative predictive values:

$$PVP = \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev)}$$
$$= \frac{0.682 \times 0.400}{0.682 \times 0.400 + 0.071 \times 0.600} = 0.865$$
$$PVN = \frac{Spec \times (1 - Prev)}{Spec \times (1 - Prev) + (1 - Sens) \times Prev}$$
$$= \frac{0.929 \times 0.600}{0.929 \times 0.600 + 0.318 \times 0.400} = 0.814$$