Biost 517
Applied Biostatistics I

Midterm Examination Key
November 4, 2011

Name: _____ Disc Sect: M Tu W F

Instructions: Please provide concise answers to all questions. The exam is worth a total of 147 points.

Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

The examination is closed book and closed notes. You may use calculators, but you may not use any special programs written for programmable calculators.

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor on Monday.

PLEDGE:

On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:_____

1. (30 points total) Suppose we are interested in studying whether levels of the chemical lactate in the cerebrospinal fluid (CSF - the fluid surrounding the brain and spinal cord) can be used to diagnose bacterial meningitis (an infection of the membranes covering the brain and spinal cord).
    - Suppose we sample 120 subjects with culture proven bacterial meningitis and we measure the CSF lactate. Among these subjects we find that 96 subjects have a CSF lactate greater than 3 mmol/L.
    - Suppose we also sample 250 otherwise healthy subjects who were undergoing radiologic examinations due to low back pain. CSF samples from these subjects were measured for lactate levels, and 10 were found to have CSF lactate levels greater than 3 mmol/L.

    a. (5 points) Can the above data be used to estimate the probability of bacterial meningitis at the hospital where the study was performed? If so, provide the estimate. If not, briefly explain why not.

Ans: No. We controlled the sample size of those with or without meningitis by study design.

    b. (5 points) Can the above data be used to estimate the sensitivity of a high CSF lactate (greater than 3 mmol/L) in diagnosing bacterial meningitis? If so, provide the estimate. If not, briefly explain why not.

Ans: Yes: 96 / 120 = 80%.

c. (5 points) Can the above data be used to estimate the specificity of a high CSF lactate (greater than 3 mmol/L) in diagnosing bacterial meningitis? If so, provide the estimate. If not, briefly explain why not.

Ans: Yes: 240 / 250 = 96%

d. (5 points) Can the above data be used to estimate the predictive value of the positive of a high CSF lactate (greater than 3 mmol/L) in diagnosing bacterial meningitis? If so, provide the estimate. If not, briefly explain why not.

Ans: No. We controlled the sample size of those with or without meningitis by study design, and that might affect the relative proportions of meningtitis within groups defined by CSF lactate levels (we are hoping that it would).

e. (5 points) Can the above data be used to estimate the predictive value of the negative of a high CSF lactate (greater than 3 mmol/L) in diagnosing bacterial meningitis? If so, provide the estimate. If not, briefly explain why not.

Ans: No. We controlled the sample size of those with or without meningitis by study design, and that might affect the relative proportions of meningtitis within groups defined by CSF lactate levels (we are hoping that it would).

f. (5 points) Suppose at another hospital the probability of bacterial meningitis is 10% among all patients for whom a CSF sample is obtained. Can the above data be used to estimate the positive predictive value of high CSF lactate (greater than 3 mmol/L) at that hospital? If so, provide the estimate and explain how you derived your answer.

Ans: Yes. We use Bayes Rule with the sensitivity and specificity calculated above:

$$PPV = \Pr(D \mid +) = \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+ \mid D)\Pr(D) + \Pr(+ \mid H)\Pr(H)} = \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev)}$$

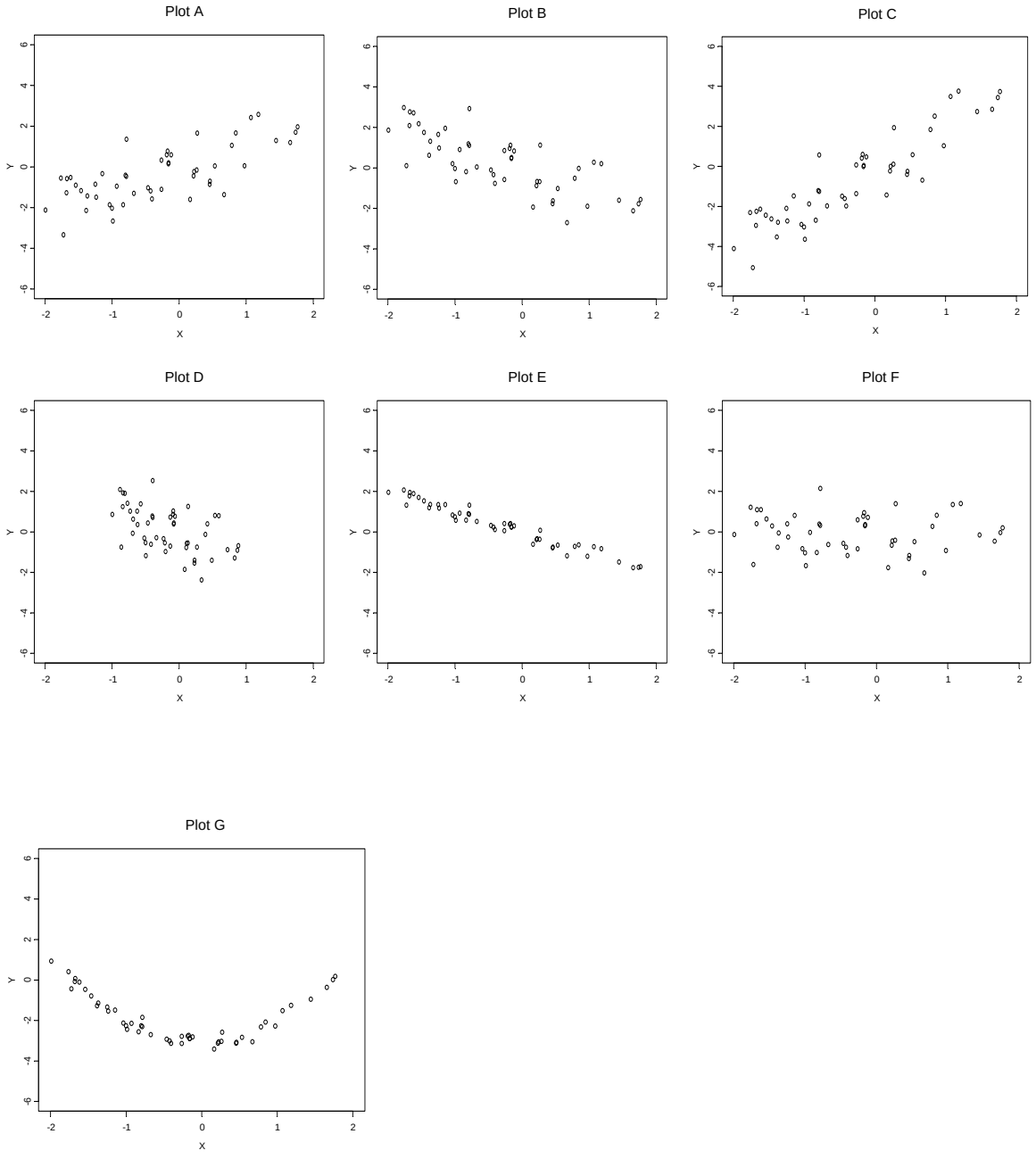
$$= \frac{0.80 \times 0.10}{0.80 \times 0.10 + (1 - 0.96) \times (1 - 0.10)} = 69.0\%$$

2. (15 points) Below are 7 scatterplots labeled A - G. List the plots in order according to lowest (most negative) to highest (most positive) correlation. (In all cases, the scale for the x and y axes are the same.)

Most……Most
Neg Pos

E < **B** < **D** < **G or F** = **F or G** < **A** < **C**

(E, B, D have the same negative slope; E has less vertical variation about the line than B, D; B is sampled over a wider variance of X. Both F and G have no linear trend, so these both have similar correlations. A,C have the same vertical variation about the line and are sampled over similar variance of X; C has a more positive slope.)



Problems 3 - 4 relate to a longitudinal observational study examining relationships between selected biomarkers of inflammation and cardiovarscular disease in 5,000 elderly, generally healthy subjects. The following variables are available:

Name	Description
Site	Indication of geographic site (1, 2, 3, or 4)
Age	Subject age at study accrual (years)
Smoker	Indicator of subject's current smoking status at study accual (0= no, 1= yes)
BMI	Subject's body mass index (weight / height^2) at time of study accrual (kg / m^2)
SBP	Subject's systolic blood pressure at time of study accrual (mm Hg)
Cholesterol	Subject's serum cholesterol at time of study accrual (mg/dl).
CRP	Subject's blood level of C reactive protein—a biomarker of inflammation (mg/l)
Ttodth	Observation time from randomization to death or data analysis, whichever came first (months)
Death	Indicator that subject was observed to die while on study

The following table contains descriptive statistics on the sample.

	N	Mean	SD	Min	q25	Median	q75	Max
Females								
Site	2904	2.45	1.13	1	1	2	3	4
Age	2904	72.6	5.52	65	68	71	76	100
Smoker	2901	0.126	0.332	0	0	0	0	1
BMI	2895	26.9	5.31	14.7	23.2	26.1	29.6	58.8
SBP	2897	137	22.3	77	122	135	151	235
Cholesterol	2870	221	38.9	88	195	219	245	430
CRP	2861	3.63	5.49	0	1	2	4	86
Ttodth	2904	79.9	20.2	0.8	80.1	88.3	92.3	96.8
Death	2904	0.170	0.376	0	0	0	0	1
Males								
Site	2096	2.50	1.13	1	1	2	4	4
Age	2096	73.2	5.69	65	69	72	77	95
Smoker	2093	0.114	0.318	0	0	0	0	1
BMI	2092	26.4	3.78	15.6	23.9	26.1	28.5	46.2
SBP	2093	136	21.3	79	121	133	148	219
Cholesterol	2083	198	35.7	73	174	197	221	407
CRP	2072	3.59	6.97	0	1	2	3	108
Ttodth	2096	75.2	24.6	0.2	54.7	87.6	92.3	96.7
Death	2096	0.299	0.458	0	0	0	1	1

3. (3 points each part) For each of the following variables, circle the descriptive statistics that you **WOULD NOT CONSIDER USING** to provide a scientifically meaningful description of a possible association between sex and the relevant scientific variable. Very briefly explain your reasons (just a few words should suffice to justify your entire answer).
   a. Consider **site**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

Ans: This is an unordered nominal variable.

b. Consider **age**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean *Std Dev* **Minimum** *25th Pctile* *Median* *75th Pctile* **Maximum**

Ans: This is an continuous variable, so only the minimum and maximum would be of no interest owing to their heavy dependence on the sample size. The standard deviation would be measuring spread rather than central tendency.

c. Consider **smoker**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

Ans: This is a binary variable. The mean (proportion) captures everything.

d. Consider **BMI**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean *Std Dev* **Minimum** *25th Pctile* *Median* *75th Pctile* **Maximum**

Ans: This is continuous uncensored variable, so same answer as in b.

e. Consider **SBP**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean *Std Dev* **Minimum** *25th Pctile* *Median* *75th Pctile* **Maximum**

Ans: This is continuous uncensored variable, so same answer as in b.

f. Consider **cholesterol**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean *Std Dev* **Minimum** *25th Pctile* *Median* *75th Pctile* **Maximum**

Ans: This is continuous uncensored variable, so same answer as in b.

g. Consider **CRP**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean *Std Dev* **Minimum** *25th Pctile* *Median* *75th Pctile* **Maximum**

Ans: This is continuous uncensored variable, so same answer as in b.

h. Consider **ttodth**. Circle Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

Ans: This is a continuous variable that is subject to right censoring. Comparisons should not be made on the basis of these sample descriptive statistics. We would use estimates based on K-M.

i. Consider **death**. Circle the descriptive statistics that WOULD NOT be useful to describe associations between that variable and sex. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

Ans: This is a binary variable indicating whether or not ttodth is a censored observation. Comparisons should not be made on the basis of these sample descriptive statistics. We would use estimates based on K-M.

4. (10 points) Do the above descriptive statistics provide evidence for associations with sex for any of the variables? If so, list the variable and provide the descriptive estimate that you used to make your decision. (I want to know what you looked at.)

Ans: Of the variables where it was appropriate to compare the means (Age, Smoker, BMI, SBP, Cholesterol, and CRP), only cholesterol seemed to have a very different mean: The average cholesterol was 221 mg/dl for females and 198 mg/dl for males.

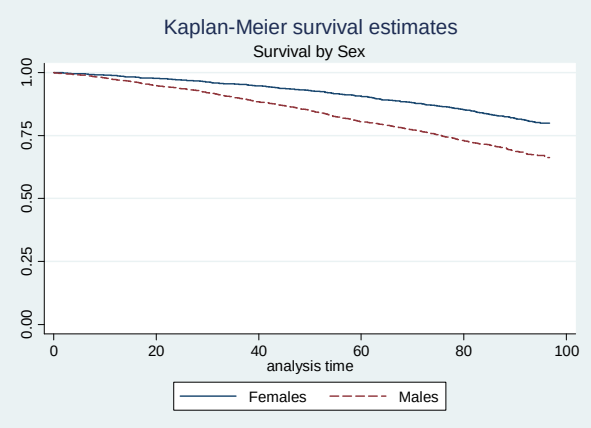
5. (10 points) For which of the variables is it of scientific interest to assess skewness caused by outlying values? Briefly list all the possible evidence for any such variable that you believe is affected by outliers.

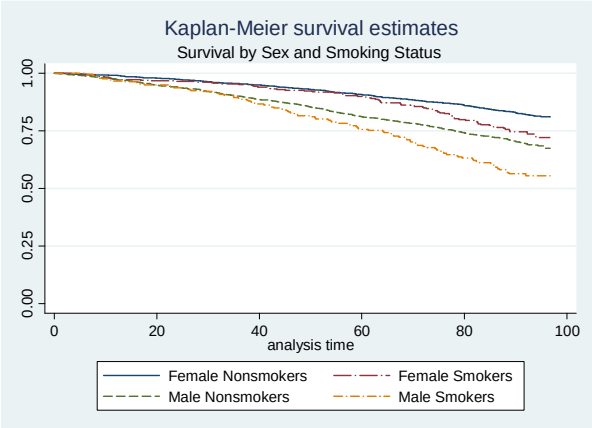
Ans: It was appropriate to consider skewness of the continuous, uncensored variables (Age, BMI, SBP, Cholesterol, and CRP). For CRP we note a high likelihood of a skewed distribution with many outliers, because

- **The SD is larger than the mean suggesting outliers for this positive variable.**
- **The mean is larger than the median suggesting skewness.**
- **The median is not midway between the minimum and maximum.**

For Age, BMI, SBP, and Cholesterol, the maximum is somewhat further from the mean/median than is the minimum, but the other criteria for marked skewness are not met. So there may be one value that is somewhat off by itself, but I would not find these so compelling.

6. (35 points) The following are results from Kaplan-Meier analyses of the time to death within strata defined by sex, smoking status, and all combinations of those two variables.





	Females			Males			Both Sexes		
		Survival Probability			Survival Probability			Survival Probability	
	n	60 mos	84 mos	n	60 mos	84 mos	n	60 mos	84 mos
Nonsmokers	2536	0.907	0.847	1854	0.811	0.728	4390	0.866	0.796
Smokers	365	0.899	0.778	239	0.757	0.612	604	0.843	0.713
All Patients	2901	0.906	0.838	2093	0.805	0.716	4994	0.864	0.787

a. (5 points) Under what conditions will the Kaplan-Meier estimates provide unbiased estimation of the distribution of times to death?

Ans: The censoring must be "noninformative" in that the censored subjects must be neither more nor less likely to die than the other subjects who were still at risk at the time of that subject's censoring.

b. (10 points) Based on the above statistics, would you conclude that there is overall an association between sex and the probability of survival? Provide statistics to quantify your answer.

Ans: Yes. The probability of surviving for 84 months is different between the sexes: 83.8% for females and 71.6% for males.

c. (10 points) Based on the above statistics, would you conclude that smoking behavior confounds any association between sex and survival? Provide statistics to quantify your answer.

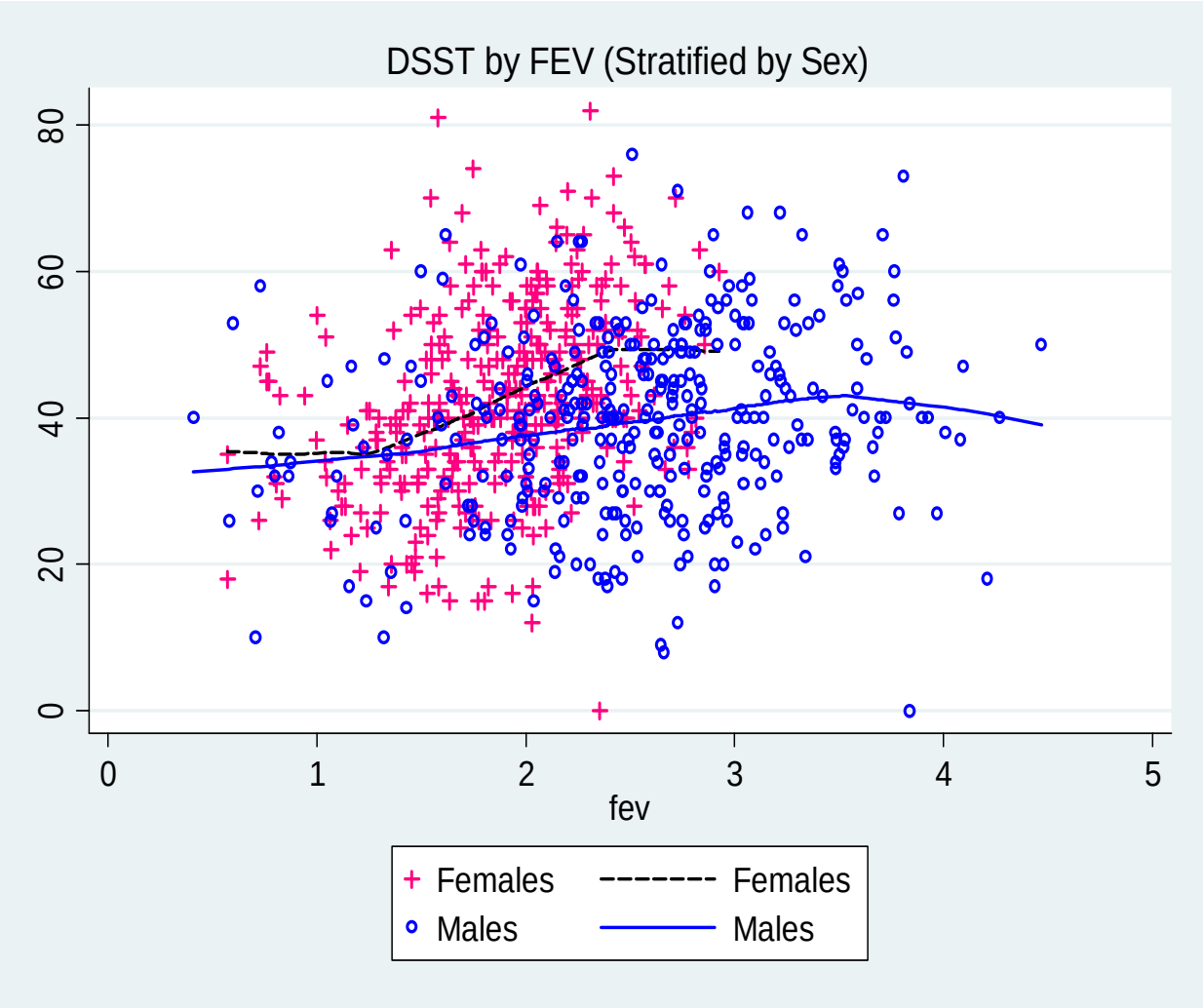
Ans: No. The proportion of smokers in the sample is 365 / 2901 = 12.6% among females and 239 / 2039 = 11.4% among males. These proportions are sufficiently similar for me to regard that there was no confounding.

d. (10 points) Based on the above statistics, would you conclude smoking behavior modifies any association between sex and survival? Provide statistics to quantify your answer.

Ans: Yes. The difference in the sexes with respect to the probability of surviving for 84 months is 84.7% - 72.8% = 11.9% for nonsmokers and 77.8% - 61.2% = 16.6% for smokers. If we regard these different enough, this would be effect modification. *(For what it is worth, this difference is not statistically significant. I accepted either yes or no, so long as you looked at the "difference of the differences" and justified your statement.)*

Problem 7 pertains to a study of dementia in 735 elderly adults. Available data include

- **male** = Patient sex (0= female, 1= male)
- **fev** = Forced expiratory volume (l/sec).
- **dsst** = Cognitive function as measured by the Digit Symbol Substitution Test (DSST) on a scale from 0 (worst) to 100 (best)



7. (20 points total) The above graph displays a scatterplot of DSST versus patient FEV. Data is stratified by sex, with superimposed lowess curves.

   a. (5 points) Briefly summarize the observations you would make from this graph.

   Ans: Oveall, there are no outliers, there seems to be a very slight tendency toward higher mean DSST with higher FEV. There is no substantial curvature evident, nor does there appear to be substantial heteroscedasticity. Within sex groups, there seems to be more of a positive slope in females than in males (I still don't make too much of the curvilinearity in the lowess curves, but this is a matter of judgement) .

   b. (5 points) What would you estimate the correlation of these two variables to be in the combined sample?

   Ans: Given the very slight positive slope on the lowess, with little ability to detect it without the lowess, I would guess about 0.15. (*We can usually see linear trends in scatterplots only if the correlation is above .25, unless we look at a smooth. In this case, the correlation was 0.16.)*

   c. (5 points) How might you expect the correlation within each sex to differ from the correlation in the combined sample? Why?

   Ans: Because of the steeper slope in females, I might guess 0.30 within the females—not as much as a difference because the steeper slope was accompanied by less variation in FEV. When only

considering the males, there might be slightly less variation about the line, but overall the correlation would not change by much. Maybe 0.2 or so. *(In this case, the correlation was 0.18 for males and 0.36 for females.)*

d. (5 points) Is there evidence suggesting that sex modifies any association between FEV and DSST? Briefly explain your reasoning.

Ans: Yes. The difference in the slopes suggests a difference in the association between FEV and DSST. *(Now that we are coming to inference, we can consider whether any difference we see in the slopes of the smooths is statistically suggestive of a difference in the population. The answer is that the difference is highly statistically significant.)*

Grades:

Maximum Possible: **147**
Highest Achieved: **147**
Mean (SD) **120 (18.4)**

Percentile	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Grade	**96**	**107**	**115**	**119**	**126**	**128**	**132**	**135**	**138**