

Biost 517
Applied Biostatistics I

Final Examination Key
December 16, 2009

Name: _____

Instructions: Please provide concise answers to all questions. 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. Should you not have a calculator available, write down the equation that you would plug into a calculator.

NOTE: When you need to make calculations, always use at least four significant digits in your intermediate calculations, and report at least three significant digits. (Example: 1.045 and 0.0001234 and 1234000 each have four significant digits.)

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.

PLEDGE:

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

Signed: _____

Problems 1-5 use the following data set.

Some diseases of the nervous system are characterized by decreased velocity of impulse conduction along the nerve fibers. One way in which this conduction velocity can be measured is to stimulate a peripheral nerve fiber, and then to measure the time to the production of a characteristic somatosensory evoked potential (SEP) in the spinal cord or brain as measured by electrodes placed over the spine or head. There are no references currently, however, which identify the normal range of conduction velocities for humans. The following data arise from 250 normal individuals who consented to participate in a study to determine normal ranges.

height: height in inches of the participant

age: age in years of the participant

sex: sex of the participant (0= female, 1= male)

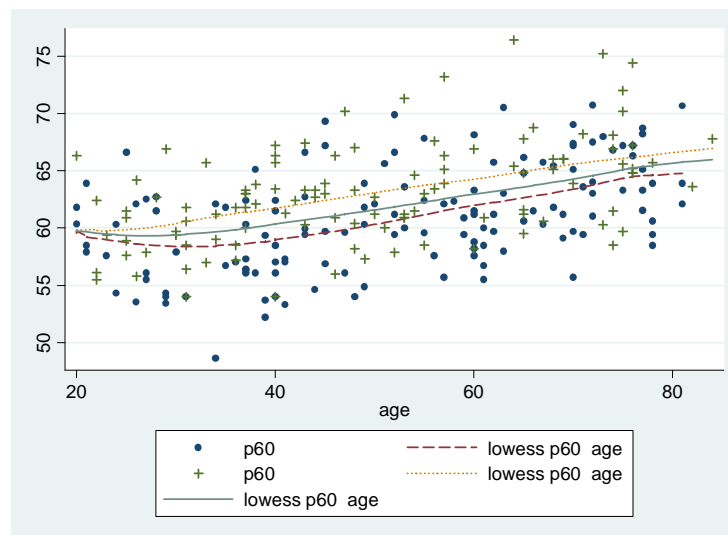
p60: delay time (in milliseconds) to detection of the second positive SEP following stimulation of the right posterior tibial nerve (average of right and left leg)

The following table contains descriptive statistics for these variables.

sex	N	Mean	SD	Min	p25	Mdn	p75	Max
<i>Females</i>								
Age (y)	137	52.3	17.3	20	39	53	68	81
Height (in)	137	63.8	2.9	53	62	63	66	71
p60 Delay time (msec)	137	61	4.5	48.6	57.6	61.1	63.9	70.7
<i>Males</i>								
Age (y)	113	49.8	17.2	20	36	48	65	84
Height (in)	113	69.3	3.5	59	67	69	72	77
p60 Delay time (msec)	113	63	4.4	54	60	62.7	65.7	76.4

1. (25 points) The following analysis was performed to examine the relationship between age and time to observation of the p60 SEP peak. Males and females are plotted using different symbols.

Scatterplot of time to p60 peak by age (symbols are + for males, circle for females; superposed lowess curves are dotted for males, dashed for females, solid for both sexes combined):



- a. What observations would you make about the above graph?

Ans: There are no obvious outliers. There does seem to be a trend toward higher *p60* in the older ages. Furthermore, that trend looks reasonably linear (*to me*). The variability of the data within age groups looks fairly constant as age varies. There does seem to be a tendency for males to have longer times to SEP than females at every age.

- b. What would you estimate the sample correlation to be in the combined sample including both sexes?

Ans: The true correlation is **0.456**. (I gave credit so long as you quoted a positive number. I note that most people will not see a linear trend in a scatterplot if the correlation is between -0.25 and

0.25. A few students noted that the regression output included a value for R^2 , and as noted in class, when only one variable is included as a predictor, the sample correlation r is the square root of R^2 . I did find it interesting that quite a few of you felt that the observed linear trend was quite slight. In any case, you got 0 points if you intimated that the correlation was the same thing as the slope. They are related, but not at all the same thing.)

- c. How do you think that correlation in the combined sample would compare to correlations computed in each sex separately? Briefly explain why.

Ans: The slopes of the lowess curves are quite similar in both sexes and in the combined group. The distribution of ages is also the same in each sex and in the combined sample. However, as evidenced by the separation of the curves for males and females, there is a slightly decreased $Var(p60 | age)$ in the sex strata when compared to the combined sample. Hence we would expect a higher sample correlation in each sex group. (The sample correlation is 0.482 in females and 0.490 in males. Again, you lost all credit if you acted as though only the slope mattered.)

- d. How do you think that correlation in the entire combined sample (both sexes) compare to correlations computed in the subset of subjects who are between ages 40 and 60? Briefly explain why.

Ans: The slopes of the lowess curves appear to approximate a straight line quite well, so I am not too concerned about the slope of the line being different in the restricted age range. The data appears fairly homoscedastic, so I am also not too concerned with the within group variability effecting a change in the correlation. Hence, we are left with a great reduction in the variance of age, which will tend to drive the correlation closer to 0 than it was in the combined sample. (The sample correlation is 0.184 in subjects older than 40 and younger than 60. Some of you talked about the slope being flatter in the youngest and oldest patients. While I would urge you to be careful in overinterpreting the ends of a lowess curve, I did give a point for this answer, because if the difference in the slope had been drastic that could have overwhelmed the difference in the variance of age. But I think you can see that the effect of the restricted age range is quite dramatic on the correlation.)

- e. Does sex appear to confound the association between p60 and age? Explain your answer.

Ans: No. Given the generally parallel curves, there does not seem to be effect modification in the difference of means or medians. Given the similar distribution of ages for each sex, we do not think there would be confounding. So sex is a precision variable, if we believe that observed separation of the curves is real. (If you told me that you thought the two year difference in mean age between the sexes was important and you thought the separation between the mean p60 levels at each age was important, I also accepted an answer that you thought there was confounding. However, from the graph it is pretty apparent that a two year difference in mean age is not associated with a large difference in mean p60, so I would not consider this confounding. The age-adjusted sex effect on the mean p60 is highly significant, however, and is estimated to be of the same magnitude as an approximate 20 year difference in age. So we have much more evidence of a meaningful association between sex and the p60 measurement after adjusting for age, than we have for any meaningful difference in the age distribution between the sexes.)

2. (30 points) The following output provides a linear regression analysis of p60 regressed on age.

`. regress p60 age, robust`

Linear regression

Number of obs = 250
 F(1, 248) = 69.68
 Prob > F = 0.0000
 R-squared = 0.2082
 Root MSE = 4.0722

p60	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1207971	.014471	8.35	0.000	.0922953	.1492988
_cons	55.70264	.7798229	71.43	0.000	54.16672	57.23856

- a. Provide an interpretation of the intercept in the above model. What scientific use would you make of this estimate?

Ans: The intercept estimate would suggest that the average p60 in a population of newborns (age=0) would be 55.7 msec. This is extrapolating way outside our data, and I would not use this number to answer a scientific question. (To get full credit, you had to tell me that we were estimating the average p60 time delay. Something that I have tried to stress is that each regression model is directed toward different summary measures. We need to interpret our models according to the summary measure we were analyzing.)

- b. Provide an interpretation of the slope in the above model? What scientific use would you make of this estimate?

Ans: When comparing groups of individuals of different ages, the mean time delay is estimated to be 0.121 msec longer for each year difference in ages, with the older group having the longer average time delay. (Again, to get full credit, you had to tell me that we were estimating the average p60 time delay.)

- c. Based on the above analysis, is there evidence of an association between age and the time to observation of the p60 peak? Explain your reasoning.

Ans: Yes, the p value for the slope ($P < .0005$) suggests a statistically significant trend toward different mean p60 time delays across age groups.

- d. Based on the above analysis, is there evidence of a statistically significant correlation between age and the time to observation of the p60 peak? Explain your reasoning.

Ans: Yes. A test for a nonzero slope in linear regression is exactly equivalent to a test for a nonzero correlation under comparable assumptions. (This holds exactly when presuming homoscedasticity and using classical linear regression and the most common test for nonzero correlation. But the extension of that result to the setting of possible heteroscedasticity is valid.)

- e. According to the above model, what would be your estimate of the average time to observation of a p60 peak in a 50 year old? Do you feel that this estimate is trustworthy? Explain your reasoning.

Ans: Using the estimated regression parameters: $55.70 + 0.1208 * 50 = 61.7$ msec. Given the reasonably linear appearance of the lowess curves, this estimate is probably not too bad.

- f. According to the above model, what would be your best estimate of the average difference in the time to observation of a p60 peak in a 50 year old compared to that in a 45 year old? Do you feel that this estimate is trustworthy? Explain your reasoning.

Ans: Using the estimated slope parameter: $5 * 0.1208 = 0.604$ msec average difference. Again, the straight line model does not appear to be a bad approximation to this data, so I think the estimated difference is probably reasonable. (You could, of course, have estimated the mean p60 time delay for 45 year olds and then subtracted it from you answer to part e. This way is faster.)

3. (25 points) The following output provides a t test analysis of p60 by sex.

```
. ttest p60, by(sex) unequal
Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	137	60.96642	.3866458	4.525573	60.20181	61.73104
1	113	63.00088	.4123848	4.383711	62.1838	63.81797
combined	250	61.886	.2888557	4.567209	61.31709	62.45491
diff		-2.034462	.565293		-3.147992	-.9209314

```
diff = mean(0) - mean(1) t = -3.5990
Ho: diff = 0 Satterthwaite's degrees of freedom = 241.666

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.0002 Pr(|T| > |t|) = 0.0004 Pr(T > t) = 0.9998
```

- a. Does the above analysis suggest evidence of an association between sex and time to observation of a p60 peak? Justify your answer.

Ans: Yes. The observed difference of 2.03 msec mean delay time (with males averaging higher values than females) is statistically significant ($P = 0.0004$).

- b. Provide an interpretation of the confidence interval.

Ans: The mean delay time for females is estimated to be 61.0, and that for males is estimated to be 63.0 msec. Based on the 95% confidence interval, we note that the observed difference in means of 2.03 msec is not unexpected if the true difference were such that the mean delay in men was between 0.921 msec higher and 3.15 msec higher than the mean delay for women.

- c. If I were to perform a linear regression of p60 (outcome variable) on sex, what would be the estimated intercept?

Ans: Because sex is coded 0 for females and 1 for males, the intercept would be the sample mean for women of 60.96642.

- d. If I were to perform a linear regression of p60 (outcome variable) on sex (predictor), what would be the estimated slope?

Ans: Because sex is coded 0 for females and 1 for males, the slope would be the difference in sample means (men minus women) of **2.034462**. (It would be positive, and you only got full credit if your answer was similarly positive.)

- e. How would the p value from a classical linear regression (no robust standard errors) compare to the p value obtained in the t test above? Briefly explain your reasoning.

Ans: In classical linear regression, the p value will be the exact same as the p value from the t test that presumes equal variances. The t test that presumes equal variances tends to be conservative (p values too large) when the larger sample has larger variance. In this data set, the females have a very slightly larger sample size (137 vs 113) and a very slightly larger standard deviation (4.53 vs 4.38). Thus I might think the p value from classical linear regression will be ever so slightly less than the p value given above. (In fact, the difference is not noticeable to the fourth decimal place in the p value, though the t statistic for the classical regression is -3.59 rather than -3.60 as given above.)

4. (25 points) The following output provides logistic regression analysis of sex regressed on p60. (Note the use of both the logit and logistic commands from Stata.)

```
. logistic sex p60, robust
Logistic regression                Number of obs   =          250
                                   Wald chi2(1)     =          12.95
                                   Prob > chi2      =          0.0003
Log pseudolikelihood = -165.82337   Pseudo R2      =          0.0367
```

sex	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf.Int]
p60	1.108	.0316	3.60	0.000	1.048 1.172

```
. logit sex p60, robust
Logistic regression                Number of obs   =          250
                                   Wald chi2(1)     =          12.95
                                   Prob > chi2      =          0.0003
Log pseudolikelihood = -165.82337   Pseudo R2      =          0.0367
```

sex	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
p60	.10266	.02853	3.60	0.000	.04675 .1586
_cons	-6.5541	1.7690	-3.71	0.000	-10.021 -3.087

- a. Provide an interpretation of the intercept in the above “logit” model. What scientific use would you make of this estimate?

Ans: The odds of being male is estimated to be $e^{-6.5541} = 0.00142$ for a subject who has a p60 of 0. (The probability of being male is thus estimated to be $0.00142 / (1+0.00142) = 0.00142$.) There is no such thing as a person having an SEP of 0—the relativistic bound on transfer of

information imposed by the speed of light argues that any distance between a person's ankle and his/her head means the SEP time delay must be greater than 0. Hence this estimate is of no scientific value.

- b. Provide an interpretation of the slope in the above model? What scientific use would you make of this estimate?

Ans: When comparing two groups differing in their p60 time delay, the odds of being male is 1.108-fold higher for every 1 msec difference in the p60 time delays between the groups, with the higher odds of being male in the group with the higher p60 time delay. (I just took the odds ratio from the "logistic" output, but you could have used $e^{0.10266} = 1.108$ to get the same thing.)

- c. Based on the above analysis, is there evidence of an association between sex and the time to observation of the p60 peak? Explain your reasoning.

Ans: Yes. The slope for the p60 variable has a p value $P < 0.0005$, indicating a statistically significant first order trend toward higher odds of being male with higher p60 time delay.

- d. How do the result of this analysis compare to that in problem 3?.

Ans: In problem 3, we compared the distribution of p60 across groups defined by sex, and in this problem we compared the distribution of sex across groups defined by p60 values. Remarkably, the statistical significance was virtually identical: A t statistic of 3.599 in problem 3 versus a Z statistic of 3.60 in this problem.

- e. According to the above model, what would be your estimate of the probability that a patient with a time to p60 peak of 65 is a male? What would you want to know before you trusted this estimate? Explain your reasoning.

Ans: The log odds of being male in such a group is estimated to be $-6.544 + 65 * 0.1027 = 0.1315$. Hence the odds of being male is estimated to be $e^{0.1315} = 1.1405$, and the probability of being male in such a group is estimated to be $1.1405 / (1 + 1.1405) = 0.533$. This estimate is only reliable to the extent that the log odds of being male is linear in the p60 measurement. (This is a difficult thing to assess graphically, although in multiple logistic regression, we could consider fitting a model that was a polynomial in p60, and see if there is evidence of a nonlinear trend. We will learn more about such an approach next quarter. I do note that if I test for nonlinearity using a quartic (fourth order) polynomial, the P value for the nonlinear trend is 0.165, so I do not have enough evidence to "prove" that the relationship is nonlinear in the log odds. Perhaps next quarter you will explore this data set in a bit more detail. Of particular interest is the interrelationship of sex, age, and height.)

5. (30 points) The following analyses were performed to examine the relationship between height and time to observation of the p60 SEP peak for males and females separately.

-> sex = 0

Linear regression

Number of obs	=	137
F(1, 135)	=	2.28
Prob > F	=	0.1334
R-squared	=	0.0121
Root MSE	=	4.5148

| Robust

p60	Coef.	Std. Err.	t	P> t	[95% Conf. Int]	
height	.1699	.1125	1.51	0.133	-.0526	.3924
_cons	50.1276	7.267	6.90	0.000	35.75	64.50

-> sex = 1

Linear regression

Number of obs = 113
 F(1, 111) = 5.00
 Prob > F = 0.0274
 R-squared = 0.0429
 Root MSE = 4.3078

p60	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Int]	
height	.2580	.1154	2.24	0.027	.0293	.4867
_cons	45.13	7.958	5.67	0.000	29.36	60.90

- a. Is there statistical evidence of sex modifying the association between height and time delay of the p60 SEP? Provide numerical evidence for or against.

Ans: We look at the difference in slopes. The point estimate is: $0.2580 - 0.1699 = 0.0881$. With independent strata, the standard error is found as the square root of the sum of the two squared standard errors: $(0.1154^2 + 0.1125^2)^{0.5} = 0.1612$.

We can form an approximate 95% confidence interval as $0.0881 \pm 1.96 * 0.1612$. This yields a 95% confidence interval for the difference of slopes of -0.228 msec to 0.404 msec. As this CI includes 0, we cannot with high confidence rule out the possibility that the linear association between height and p60 is the same for the two sexes.

(The following discussion is based on the estimates, even though there is no statistical significance on the difference. Next quarter, this dataset might be used to illustrate high order effect modications.

Interestingly, we see more of a trend in height for males than we do for females. Given that length of nerve is expected to be a strong determinant of the time delay until p60 SEP is observed, it at first seems surprising that the two sexes are not more similar in their estimates. Upon a little more reflection, we might consider that men and women might differ in the ability of height to predict nerve length: Osteoporosis, and the resulting decrease in height from verterbral collapse, is more common in women than in men. Hence, height might be a better surrogate for nerve length in men than it is in older women. To model this more appropriately, we might therefore need to consider the interaction of sex, age, and height.)

- b. How could you use the above results to estimate a sex adjusted estimate of the association between height and the p60 SEP time delay? Provide such an estimate and inference.

Ans: A sex adjusted effect might use some average of the stratum specific estimates. As the proportion of men and women in the population is 50% - 50%, it might seem reasonable to weight the two estimates equally, rather than in proportion to the sample sizes, which slightly favored females.

Hence our stratified estimate will be: $0.5 * 0.2580 + 0.5 * 0.1699 = 0.2140$.

With independent groups, the standard error for the weighted average will be (recall that we have to square the constants by which we multiply our estimates):

$$(0.5^2 * 0.1154^2 + 0.5^2 * 0.1125^2)^{0.5} = 0.08058$$

(this is of course just half the standard error we saw in part a, because we multiplied each estimate by 0.5).

A 95% confidence interval for the sex adjusted effect of height on the time delay to the p60 SEP is thus 0.0561 msec higher to 0.3719 msec higher for every 1 inch difference in heights between two groups. Because that CI does not include 0, we can with high confidence rule out the possibility that there is no linear trend in time delay of p60 SEP by height.

(Note that this analysis would adjust for any confounding that might be due to sex effects on the nerve conduction velocity—sex is certainly associated with height in our sample. This is in spirit what we get when we adjust for a variable in multiple regression. There will be some differences due to the possibility of effect modification and weighting due to sample sizes.)

Grades:

Maimum Possible: 135
Highest Achieved: 132
Mean (SD) 94 (17.7)

Percentile	10%	20%	30%	40%	50%	60%	70%	80%	90%
Grade	69	79	84	90	95	100	105	109	116