Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Homework #1 Key

1. Forced Expiratory Volume (FEV) is a measure of lung capacity, with higher values indicating higher capacity. In this problem, you will perform an analysis to estimate the difference in mean FEV between smokers and non-smokers based on the data set `fev.csv`. As with ***all*** data sets used in this class, **please read/refer to the associated documentation**.

   (a) Create a table of descriptive statistics to characterize the distribution of FEV overall and stratified by smoking status. Does this finding surprise you? Further, do the descriptive statistics suggest unequal variances between groups?

   Ans: See table below.

	N (msg)	Mean (SD)	Median (IQR)	(Min, Max)
Overall	654 (0)	2.64 (0.867)	2.55 (1.98, 3.12)	(0.791, 5.79)
Smokers	65 (0)	3.28 (0.750)	3.17 (2.80, 3.75)	(1.69, 4.87)
Nonsmokers	589 (0)	2.57 (0.851)	2.47 (1.92, 3.05)	(0.791, 5.79)

   I probably would not expect smokers to have a higher FEV (a measure of lung capacity) *a priori*, so yes: this does surprise me. The sample standard deviation is higher among non-smokers, but it's not clear that the variances are markedly different between groups.

   (b) Perform an analysis to assess whether the difference in mean FEV between smokers and non-smokers is different from zero, assuming (in this problem) equal variances between groups.

   Ans: These data provide sufficient evidence of a difference in mean FEV between smokers and nonsmokers in children aged 3-19 years ($p < 0.001$). We estimate the difference in mean FEV to be 0.711 L, with the smokers having the higher estimated mean. Based on a 95% confidence interval, this estimate would not be judged unusual if the true mean difference were between 0.495 L and 0.927 L.

   (c) Repeat problem (b), this time allowing unequal variances between groups, and compare your results to those of problem (b).

   Ans: These data provide sufficient evidence of a difference in mean FEV between smokers and nonsmokers in children aged 3-19 years ($p < 0.001$). Based on this analysis, we estimate the difference in mean FEV to be 0.711 L, with the smokers having the higher estimated mean. Based on a 95% confidence interval, this estimate would not be judged unusual if the true mean difference were between 0.513 L and 0.908 L.

   (d) Which analysis—either that of (b) or (c)—would you prefer? Explain your answer in a sentence or two.

   Ans: As a general principle, I would choose the analysis that requires fewer assumptions to hold in order to be valid. In this case, that would mean the analysis allowing unequal variances, (c).

(e) In either analysis, do you reach the conclusion that smoking increases lung capacity? If not, restate the conclusion more accurately. Then, explore some of the other variables in the data and provide a logical explanation for this somewhat unexpected phenomenon.

Ans: Certainly not. We only are able to provide evidence that the mean difference in FEV between smokers and nonsmokers is nonzero, but not that smoking itself is causing an increase in FEV. Take a look at the age distribution in each smoking group; the smokers are much older. Hence, they tend to be larger and tend to have higher lung capacity.

2. Load the data set `mri.csv`. We seek to determine the association between serum low density lipoprotein (LDL) levels and all-cause death in healthy elderly subjects.

(a) Time to death is subject to censoring in these data. For now, though, dichotomize time to death according to whether death occurred within five years of study enrollment. Providing descriptive statistics that support your answer, explain why this is valid (*Hint*: Look at the earliest time of censoring).

Ans: The earliest time of censoring is 1827 days, which is the maximum number of days that five years can span (assuming leap years in the first and fifth year). The point is that at the five year mark, everyone's status (dead or alive) is known. If we were to try to dichotomize, say, at 2000 days, there would be subjects whose status would be unknown because censoring occurred prior to that time (and the resulting analysis would *not* be valid).

(b) Perform an analysis comparing mean LDL between groups defined by five-year vital status.

Ans: These data provide evidence that the difference in mean LDL differs between groups defined by five-year vital status ($p = 0.0186$). We estimate the difference in mean LDL to be 8.50 mg/dl, with the individuals dying within five years having the lower estimated mean LDL. These data would not be deemed atypical if the true mean difference were anywhere between 1.44 mg/dl and 15.6 mg/dl.

(c) Perform an analysis comparing geometric mean LDL between groups defined by five-year vital status.

Ans: These data provide evidence that the geometric mean ratio (GMR) comparing vital status groups is different than one ($p = 0.0128$). We estimate the geometric mean ratio to be 1.0965, with the individuals surviving five or more years having the higher estimated geometric mean LDL. These data would not be deemed atypical if the true GMR were anywhere between 1.0201 and 1.179.

OR

Ans: These data provide evidence that the GMR comparing vital status groups is different than one ($p = 0.0128$). We estimate the geometric mean ratio to be 0.9120, with the individuals dying within five years having the lower estimated geometric mean LDL. These data would not be deemed atypical if the true GMR were anywhere between 0.848 and 0.9803.

OR

Ans: These data provide evidence that the LDL geometric means differ between groups defined by five-year vital status (p = 0.0128). We estimate the geometric mean to be 9.65%, higher in individuals surviving five or more years. These data would not be deemed atypical if the true geometric mean were anywhere between 2.01% and 17.9% higher in the individuals surviving five or more years.

OR

Ans: These data provide evidence that the LDL geometric means differ between groups defined by five-year vital status (p = 0.0128). We estimate the geometric mean to be 8.80%, lower in individuals dying within five or more years having the higher estimated geometric mean LDL. These data would not be deemed atypical if the true geometric mean were anywhere between 1.97% and 15.2% higher in individuals dying within five or more years.

(d) Perform an analysis comparing the probability of death within five years between groups defined by whether the subjects had an LDL value exceeding 160 mg/dL.

Ans: These data do not provide sufficient evidence that the proportion of individuals dying within five years differs between groups defined by whether or not they had a LDL exceeding 160 mg/dl (p = 0.262). We estimate the difference in proportions to be 4.42%, with the individuals having the higher LDL having the lower estimated probability of death (95% CI: [-0.0264, 0.115]). Note that this confidence interval cannot be interpreted in the usual, convenient way as it does not invert the test. Instead, we say that if we were to repeatedly conduct this study under the same circumstances with the same sample size, constructing confidence intervals in this way each time, approximately 95% of those confidence intervals would contain the true difference in proportions; this is one of those confidence intervals, but that's about all we can say.

(e) Which of the analyses—(b), (c), or (d)—-would you have preferred *a priori* to answer the question about an association between mortality and serum LDL? Briefly explain.

Ans: A justification could be made for either (b) or (c) in my judgment. The analysis performed in (d) unnecessarily dichotomizes LDL, resulting in a loss of power. In large part, which analysis we prefer is tied to whether or not we are interested in the arithmetic or geometric mean. Part of this can be informed by the fact that LDL is a concentration; as such, it is not unreasonable to treat it on a multiplicative scale. In this particular case, contrasts on both the additive and multiplicative scale are trivially satisfied because the exposure only has two levels (either dead within five years or not). The major thing that is *not* acceptable in this problem is to state the preference of your analysis on the basis of the results obtained. Analytic decisions should generally be made *a priori*, and this is especially true for primary analyses.