

## STATISTICS 210 – Final Exam Data Analysis Comments – Fall 2009

### SMOKING AND LUNG FUNCTION

- No detailed solution here; just some comments based on your analyses.
- A goal here was to have you write a coherent report. A listing of output with a sentence here or there is not a coherent report. You should start with an introduction, address key issues that come up in the modeling (nonconstant variance, interactions), interpret results, and discuss the results and the limitations of the analysis.
- The introductions were good.
- Some preliminary analysis is important. Examine the variables and their correlations. Make scatterplots. These will help you determine if a linear regression analysis is a sensible approach.
- Most everyone's initial regression identified a problem with non-constant variance and perhaps some evidence of a non-linear relationship. Smoking is not significant in this initial regression but should not be deleted from further consideration because it is the principal variable of interest – the goal here is not to predict lung function from the variables but to understand the impact of smoking on lung function.
- The first thing you might try to fix is the non-linearity (if you saw that in the residual plot); this would lead one to consider quadratic terms for age and/or height. This does help with the non-linearity but still leaves non-constant variance.
- A transformation (square root or logarithm) seems to fix the non-constant variance and create a more linear relationship. When it is a close call I prefer the logarithmic transformation because it makes interpretation easier.
- The question was explicit about wanting you to consider interactions (it asked whether the effect of smoking was additive or depends on other factors). Thus once you got your transformation you should try interactions. For the log transformation only the smoking-age interaction is a close call. Interestingly smoking by itself is significant, the smoking-age interaction by itself is significant, but when both are in the regression neither one looks significant. Any of the models would be a fine choice as long as you discuss what you found. (Interestingly with the square root transformation you end up with more interactions.)
- Diagnostics – everyone did a good job checking for outliers and unusual cases. It is important to do this but not necessary to report detailed results. Only report interesting findings about individual cases. In this data set there weren't any!
- Summary – You should summarize by repeating your final model and describing what it says about the effects of smoking. It was a bit disappointing that many of you never interpreted the final model in terms of what it said about the effect of smoking on lung function.
- Discussion – It is important to discuss limitations of your analysis. Here a couple of obvious points to make are (1) this is an observational study which limits the causal inferences that can be drawn; (2) the binary data about smoking is extremely limiting as we would prefer to know how long and how much people smoke. Another issue here is the question of whether it makes sense to include young children when they don't smoke. They provide useful information about the lung function vs age/height relationship but can't tell us anything about smoking.