FINAL PROJECT

COMPUTING FOR RESEARCH

Spring 2013

DUE MAY 3, 5PM**. Submit via email to garrettm@musc.edu**

Use Sweave and R to complete the questions below. Turn in both (1) the final report in either postscript (PS) or PDF format and (2) your Sweave file. You may contact the instructor with any questions. Note that there is not a ‘correct’ model. This project is designed to show your facility with using R, latex and Sweave and reasonable statistical logic.

The final project is based on the dataset used in the paper at the following link:

<http://www.ncbi.nlm.nih.gov/pubmed/18493865>

Data was collected from lung, breast and prostate cancer patients regarding their quality of life. In addition to a composite quality of life variable, there are 8 symptoms (fatigue, pain, nausea and vomiting, dyspnea, insomnia, appetite loss, constipation, diarrhea). In this analysis, you will only focus on 3 symptoms: fatigue, nausea and vomiting, and pain. See codebook on the next page for a description of the variables. **The goal of the analysis is to see which symptoms are associated with overall quality of life.** You should use the qol composite variable as your outcome. You will use linear regression to explore these associations. You will also need to consider adjusting for potential confounders such as gender, cancer type, age, ECOG status, and extent of disease.

1. Perform any data exploration that you feel is necessary, including figures and summary statistics.
2. Are any transformations necessary? Include a discussion/description of why or why not. If you decide to transform any variables, please use them in the questions below (e.g. if you decide to use log(qol) as your outcome, in the following questions instead of performing regressions of qol on covariates, you should perform regressions of log(qol) on covariates).
3. Perform simple linear regressions of qol on the potential confounders. Include a discussion of the relationship between gender and cancer type. Which of these variables appears to be related to qol and should be considered for inclusion in multiple regression models of qol?
4. Perform simple linear regressions of qol on the 3 symptoms. Which symptoms are associated with qol?
5. Build a multiple regression model. Describe some justifications of the process you use from deriving a “final” multiple regression model that includes symptoms and/or the other variables (age, gender, etc.)
6. Create a table that summarizes the results from your regression modeling. Interpret the findings in words.
7. Include one or more diagnostic plots from your regression model that highlights the distribution of residuals or other assumptions of linear regression modeling.
8. Provide a discussion that includes your findings, and any particular challenges you may have encountered in modeling qol and the symptoms.

**Codebook:**

ptid: numeric patient identifier

age: patient age in years

sex: gender. 1= male; 2 = female

ecog: ECOG performance status. See <http://ecog.dfci.harvard.edu/general/perf_stat.html>

type: type of cancer. 1= lung, 2 = breast, 3 = prostate

extent: extent of disease 1 = early stage; 2 = locoregional; 3 = metastatic

fa: fatigue score on a scale of 0 to 100

nv: nausea and vomiting score on a scale of 0 to 100

pa: pain score on a scale of 0 to 100

qol: qol score on a scale of 0 to 100