MATH 4330
Applied Categorical Data Analysis

Midterm test
October 20, 2017

Duration: 50 minutes

Instructions: No aids are allowed except a non-progammable calculator. There are 10 questions each worth 10 marks.

1. [20] You are studying observational data on the relationship between a measure of Health and coffee consumption (measured in grams of caffeine consumed per day). Suppose you want to control for a possible confounding factor 'Stress'. Describe the consequences of measurement error in coffee consumption? Describe the consequences of measurement error in Stress? Compare the relative impact of each sourse of measurement error if a) your goal is a predictive inference and b) if your goal is causal inference on the health effects of coffee consumption.

2. [20] Discuss situations when a) it would be important in a regression to include a variable that is not significant and b) it would be important to exclude a variable that is highly significant?

3. Consider the following output

```
> library(car)
> library(spida2)
> head(Prestige)
                    education income women prestige type
gov.administrators      13.11  12351 11.16     68.8 prof
general.managers        12.26  25879  4.02     69.1 prof
accountants             12.77   9271 15.70     63.4 prof
purchasing.officers     11.42   8865  9.11     56.8 prof
chemists                14.62   8403 11.68     73.5 prof
physicists              15.64  11030  5.13     77.6 prof
> tab(Prestige, ~ type)
type
   bc  prof    wc Total
   44    31    23    98
> # women is the percentage of women in an occupation
> # type has three levels: prof, wc and bc for
> #       professional, white collar and blue collar respectively
> # education is the mean years of education for an occupation
> # income is the mean income for an coccupation
```

```
> fit <- lm(income ~ (women + education + type)^2, Prestige)
> summary(fit)

Call:
lm(formula = income ~ (women + education + type)^2, data = Prestige)
. . .
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          301.455   3607.274   0.084   0.9336
women                 23.746     83.827   0.283   0.7776
education            700.898    415.143   1.688   0.0949 .
typeprof           -2347.177   6296.157  -0.373   0.7102
typewc             -4494.487   8394.279  -0.535   0.5937
women:education       -8.276     10.302  -0.803   0.4240
women:typeprof        -5.102     63.458  -0.080   0.9361
women:typewc          12.666     40.847   0.310   0.7572
education:typeprof   369.593    532.130   0.695   0.4892
education:typewc     406.283    800.590   0.507   0.6131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2797 on 88 degrees of freedom
Multiple R-squared:  0.603,   Adjusted R-squared:  0.5623
F-statistic: 14.85 on 9 and 88 DF,  p-value: 2.314e-14
```

a. [10] Estimate the gender gap (the estimated difference between a job that is 0% female and a job that is 100% female), for a white collar job with 9 years of education.

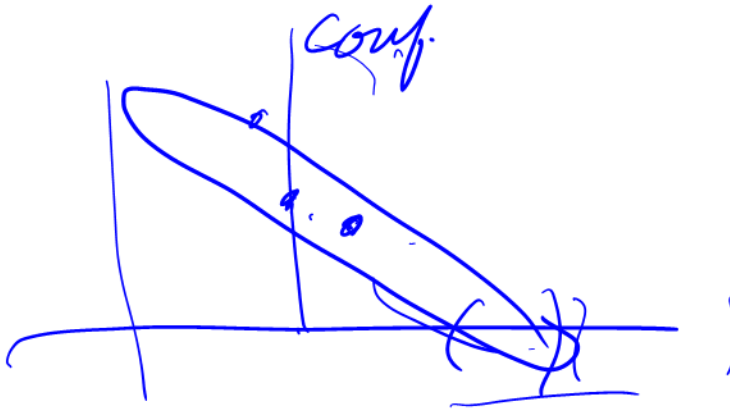
b. [10] Estimate the gap between professional and white collar jobs among jobs that are 50% female with 12 years of education.

4. [20] Describe the three basic formats for a data set consisting only of categorical data. Select two of the formats and describe how to transform data from one format to the other. Preferably, write a function in R. Alternatively, describe how such a function would work.

5. [20] Discuss how it could be possible for a regression with two linear predictors to produce two different final models when using forward stepwise versus backward stepwise variable selection algorithms. Explain how a confidence region for the coefficients of the two linear predictors is related to forward and backward stepwise selection.

Include not significant:

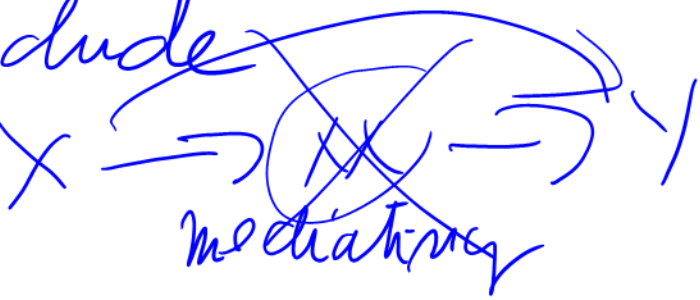
1) If a variable is a confounding factor!



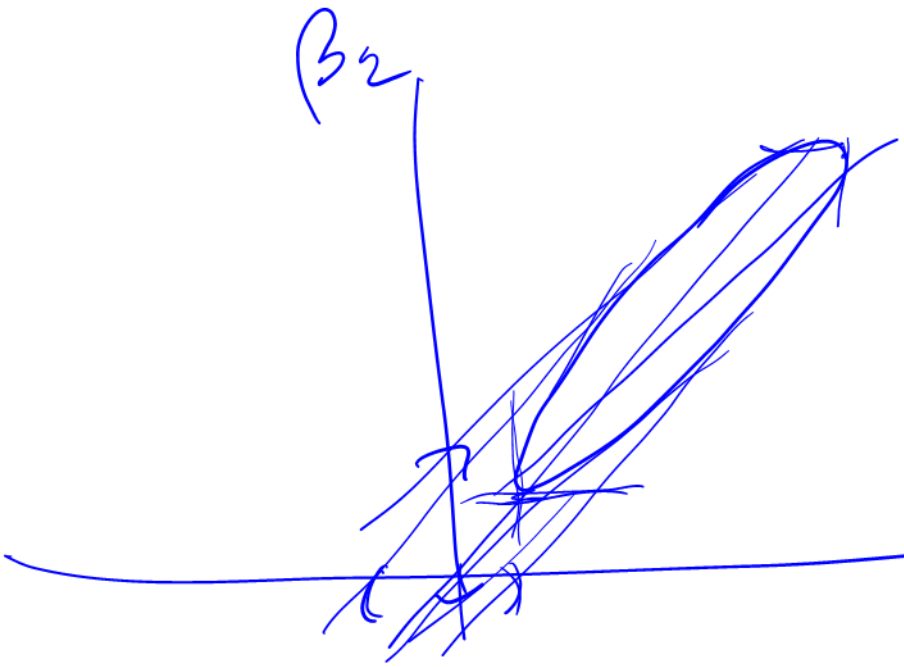
2) If dropping variable violates Prof M

Exclude highly sig. variable

Causal) Exclude



Predictive) Exclude variables not available.



Forward Stepwise: neither β_1 nor β_2 are sig

Backward " : Both β_1 & β_2 are sig.