Final Exam Stat 505 Fall 2012

December 10, 2012
100 pts total

Name:_____

1. In building a model to predict **log(earnings)**, I used indicator variables for female, black, and hispanic and a numeric variable for education. The `ed12` variable is 12 minus the highest grade completed (-4 for only eighth grade, 0 = graduated from high school, 4 for college degree, etc.). The following output was obtained.

	Estimate	Std. Error	t value
(Intercept)	9.89	0.04	224.20
ed12	0.11	0.01	11.19
female	-0.58	0.05	-10.83
black	-0.26	0.13	-1.98
hisp	-0.42	0.16	-2.58
female:black	0.34	0.17	2.02
female:hisp	0.37	0.22	1.70

Table 1: n = 1161 rank = 7 resid sd = 0.836 R-Squared = 0.185

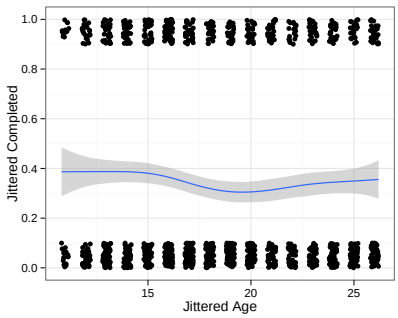
(a) Write out a complete interpretation (including any conditionality) for the first two lines of output: Intercept and ed12 in dollars. (10 pts)

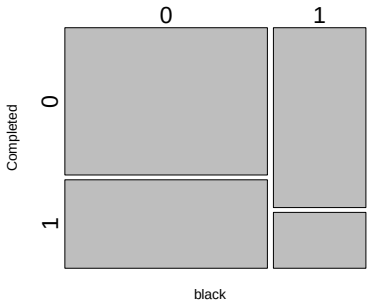
(b) The `display` function in the `arm` package leaves out p-values. If we want to compute them, say for the last line in the above table, what distribution would we use? (5 pts)

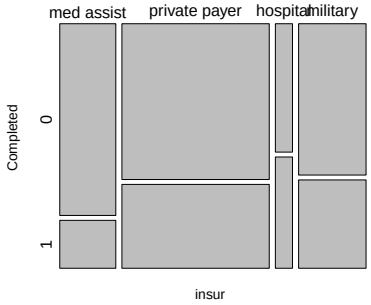
(c) Compare white female earnings to white male earnings in dollars. (10 pts)

(d) Estimate the difference between black female earnings and white male earnings. Explicitly describe any further information needed for a complete comparison. (10 pts)

2. Doctors at Johns Hopkins Medical School wish to encourage young women to get a series of three vaccinations against human pampilla virus, HPV. (With less than 3 shots, a woman is not completly protected.) They enrolled 1413 subjects from age 11 to 26 (mean age = 18.5), of whom 443 were black (most of the rest were white with a few "other race"). Four different types of insurance paid for the vaccinations (medical assistance, private payer, hospital, or military). A logistic regression was run to fit binary variable `Completed` (1 is getting all three shots, 0 for fewer than 3 shots) to Age, black and insurance type.







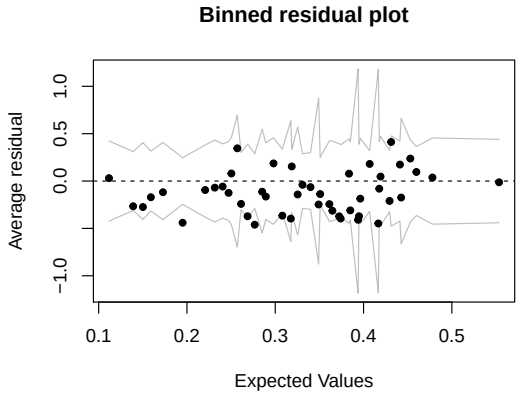
	Estimate	Std. Error	z value
(Intercept)	-1.23	0.17	-7.34
cAge	-0.05	0.01	-3.26
black	-0.58	0.13	-4.28
insurprivate payer	0.82	0.18	4.54
insurhospital	1.30	0.27	4.74
insurmilitary	0.74	0.20	3.75

Table 2: n = 1413 rank = 6

(a) Interpret the results shown for cAge (centered age), and black. (10 pts)

(b) Interpret the binned residual plot, explaining what we are looking for and what you conclude in this case. (10 pts)

```
binnedplot(fitted(gard1), resid(gard1), nclass = 50)
```



3. Why is it so much easier to make **causal** inference in a study where treatments are randomly allocated to the units than in a study where the treatment variable is simply observed? (10 pts)

4. The coin came up Heads, so here is your Gauss-Markov question.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0}$$

   (a) True or False: $\boldsymbol{\epsilon}$ must have a Gaussian distribution. (5 pts)

   (b) One of two assumptions about variance of $\boldsymbol{\epsilon}$ is needed to apply the G-M Theorem. What are the two assumptions and what does the G-M estimator look like in each case? (10 pts)

   i.

   ii.

   (c) For 10 points extra credit, show that the two assumptions are equivalent.