STA 312f22 Assignment Eight[1]

Please bring your R printout for Question 2 to the quiz. The non-computer questions are practice for the quiz on Friday Nov. 18th, and are not to be handed in.

1. Here is a logistic regression on the Titanic data, just for passengers. As you can see, the explanatory variable is `Class`, and the response variable is `Survived`. `Class2` is a dummy variable for second class, and `Class3` is a dummy variable for third class. We are not controlling for age and gender; we are ignoring them.

```
> modelT = glm(Survived ~ Class, family = binomial, data=tdat)
> summary(modelT)
Call:
glm(formula = Survived ~ Class, family = binomial, data = tdat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3999  -0.7623  -0.7623   0.9702   1.6600

Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)   0.5092     0.1146   4.445          0.000008793 ***
Class2       -0.8565     0.1661  -5.157          0.000000251 ***
Class3       -1.5965     0.1436 -11.114 < 0.0000000000000002 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1746.8  on 1315  degrees of freedom
Residual deviance: 1614.1  on 1313  degrees of freedom
AIC: 1620.1

Number of Fisher Scoring iterations: 4
> round(vcov(modelT),4)
            (Intercept)  Class2  Class3
(Intercept)      0.0131 -0.0131 -0.0131
Class2          -0.0131  0.0276  0.0131
Class3          -0.0131  0.0131  0.0206
```

   (a) Is there a difference in survival rate between first and second class at the 0.05 significance level? Give the test statistic (a number) and state your conclusion in plain, non-statistical language.

   (b) The estimated odds of survival for a passenger in second class are _____ times as great as the odds for a passenger in first class. The answer is a number.

   (c) Give a 95% confidence interval for that last number.

[1] Copyright information is at the end of the last page.

(d) Is there a difference in survival rate between first and third class at the 0.05 significance level? Give the test statistic (a number) and state your conclusion in plain, non-statistical language.

(e) Give the likelihood ratio test statistic for $H_0 : \beta_1 = \beta_2 = 0$. The answer is a number. What is the critical value at $\alpha = 0.05$? Do you reject the null hypothesis? Answer Yes or No. In plan, non-statistical language, what do you conclude?

(f) Estimate the probability of survival for a passenger in second class. The answer is a number.

(g) Give a 95% confidence interval for the probability of survival for a passenger in **first** class.

(h) You have enough information to test for a difference between second and third class. Report the test statistic, compare it to the critical value, and state your conclusion in plain, non-statistical language.

2. The *birth weight data* is a dataset containing information about a sample of new mothers. The response variable is a variable called `low`, an indicator of a baby with dangerously low weight at birth. To get the data and more information, type `library(MASS)`; `help(birthwt)` at the R prompt.

   Fit a logistic regression model with just the following explanatory variables: Age, mother's weight, race, smoking status during pregnancy, and an indicator for any first-trimester visits (1 = one or more, 0 = none). Make race a factor, and keep the default dummy variable setup. Look at `help(factor)` if you need to. I used the optional labels to make my output more readable.

   Not all the explanatory variables are significantly related to low birth weight when you control for the others. The non-significant variables could be removed from the model, but this is a matter of taste. We'll leave them in this time. Just to verify that we have the same model, my standard error for $\widehat{\beta}_0$ is 1.110641. If we don't agree, you might be using a different reference category for race, or you might have forgotten `family=binomial`.

   (a) Reproduce the standard errors from `summary` using the `vcov` function.

   (b) Controlling for all the other variables, is mother's weight at last period related to low birth weight at the 0.05 significance level? In plain, non-statistical language, what do you conclude?

   (c) Allowing for all the other variables, is smoking during pregnancy related to low birth weight at the 0.05 significance level? In plain, non-statistical language, what do you conclude?

   (d) Correcting for all the other variables, the odds of a low birth weight baby are an estimated _____ times as great for a mother who smokes during pregnancy.

   (e) We need to test for race, controlling for all other variables in the model.

i. Do a likelihood ratio test. Give the value of G^2, the degrees of freedom, and the p-value. In plain, non-statistical language, what do you conclude?

ii. Do a Wald test. Give the value of W_n, the degrees of freedom, and the p-value. In plain, non-statistical language, what do you conclude?

(f) Allowing for all the other variables, is there a difference between Black and White mothers in their chancees of having a low birth weight baby? In plain, non-statistical language, what do you conclude? Use the the 0.05 significance level.

(g) Correcting for all the other variables, the odds of a low birth weight baby are an estimated _____ times as great for a Black mother, compared to a White mother.

(h) Give a 95% confidence interval for that last number.

(i) For a 26-year-old, 130 pound, Black non-smoking mother with no first trimester visits to the doctor,

i. Estimate the probability of a low birth weight baby.

ii. Give a 95% confidence interval for the probability. There are two good wys to do this. You only need to do one of them.

(j) This question is about comparing Black and Other mothers in the chances of a low birth weight baby, controlling for all other variables.

i. Carry out a Wald test. What do you conclude?

ii. Correcting for all the other variables, the odds of a low birth weight baby are an estimated _____ times as great for a Black mother, compared to an Other mother.

iii. Give a 95% confidence interval for the last odds ratio. This question might require a bit of thought. To help you along, what is the approximate large-sample distribution of the difference between the two $\widehat{\beta}$ values? The hypothesis matrix $\mathbf{L}$ from the Wald test should be useful.

3. The U.S. Census Bureau divides the United States into small pieces called census tracts; lots of information is collected about each census tract. The census tracts are grouped into four geographic regions: North Central, Northeast, South and West. In one study, the cases were census tracts, the explanatory variables were Region and average income, and the response variable was crime rate, defined as the number of reported serious crimes in a census tract, divided by the number of people in the census tract.

(a) Write $E(y|\mathbf{x})$ for a regression model with *no intercept* and parallel regression lines. You do not have to say how your dummy variables are defined. You will do that in the next part.

(b) Make a table showing how your dummy variables are set up. There should be one row for each region, and a column for each dummy variable. Add a wider column on the right, in which you show $E(y|\mathbf{x})$.

(c) For each of the following questions, give the null hypothesis in terms of the β_j parameters of your regression model. Also, give the $\mathbf{L}$ and $\mathbf{h}$ matrices for $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$. We are not doing one-tailed tests, regardless of how the question is phrased.

i. Controlling for income, does average crime rate differ by geographic region?

ii. Controlling for income, is average crime rate different in the North Central and Northeast regions?

iii. Controlling for income, is average crime rate different in the Northeast and Western regions?

iv. Controlling for income, is the crime rate in the South more than the average of the other three regions?

v. Controlling for income, is the average crime rate in the Northeast and North Central regions different from the average of the South and West?

vi. Controlling for geographic region, is crime rate connected to income?

(d) Write $E(y|\mathbf{x})$ for a regression model in which the regression lines might not be parallel. This time, use a model with an intercept. Make North Central the reference category; that's what R would do, since it's alphabetically first.

(e) Make a table showing how the dummy variables are set up. There should be one row for each region, and a column for each dummy variable. Add a wider column on the right, in which you show $E(y|\mathbf{x})$.

(f) For this new model with possibly unequal slopes, give the null hypothesis you would test in order to answer each question. First, write it in scalar form, in terms of the β_j parameters. Then, give the $\mathbf{L}$ and $\mathbf{h}$ matrices for $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

i. Are the four regression lines parallel in the population?

ii. Is there an interaction between average income and geographic region?

iii. Does the relationship of average income to crime rate depend on geographic region?

iv. Do regional differences in average crime rate depend on the average income in the census tract?

v. Is the slope of the line relating average income to expected crime rate different for the North Central and Northeast regions?

vi. Is the slope of the line relating average income to crime rate different for the North Central and South regions?

vii. Is the slope of the line relating average income to crime rate different for the North Central and West regions?

viii. Is the slope of the line relating average income to crime rate different for the Northeast and South regions?

ix. Is the slope of the line relating average income to crime rate different for the Northeast and West regions?

x. Is the slope of the line relating average income to crime rate different for the South and West regions?

xi. Is average income related to crime rate for the South region? This is equivalent to asking if the slope of the regression line for the South region is different from zero.

xii. Is average income related to crime rate for the Northeast or South region (or both)? This is one test.
