**Chapter 4 practice problems**

The answers given here are sometimes only partial answers. Please see the answer keys for projects and tests for examples of full answers.

Note: Many of the practice problems are based on exercises given in Alan Agresti’s “Introduction to Categorical Data Analysis” book.

1. Derive the limits of the score interval for μ. Students are NOT responsible for this proof other than perhaps as an extra credit problem.

The limits of the score confidence interval come from “inverting” the score test for μ. This means finding the set of μ0 such that



is satisfied. Working with an equality produces,

 .

Then



Using the quadratic formula produces



Thus, the limits of score interval are .

1. Some of the same calculations performed with the stoplight data can be performed using the glm() function.
	1. Find the MLE of μ using glm().

To estimate a Poisson regression model without an explanatory variable (i.e., log(μ) = β0), use a “~ 1” in the formula argument:

> stoplight <- read.csv(file = "C:\\data\\stoplight.csv")

> head(stoplight)

 Observation vehicles

1 1 4

2 2 6

3 3 1

4 4 2

5 5 3

6 6 3

> mod.fit <- glm(formula = vehicles ~ 1, data = stoplight, family = poisson(link =

 log))

> summary(mod.fit)

Call:

glm(formula = vehicles ~ 1, family = poisson(link = log), data = stoplight)

Deviance Residuals:

 Min 1Q Median 3Q Max

-2.78388 -1.05091 0.06316 0.54674 1.82985

Coefficients:

 Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.35455 0.08032 16.86 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

 Null deviance: 49.191 on 39 degrees of freedom

Residual deviance: 49.191 on 39 degrees of freedom

AIC: 172.73

Number of Fisher Scoring iterations: 4

The estimated model is .

> exp(mod.fit$coefficients)

(Intercept)

 3.875

> mean(stoplight$vehicles)

[1] 3.875



* 1. Find a confidence interval for the mean.

> #Wald for exp(beta\_0) - see how this matches interval using log(mu) in the notes

> exp(confint.default(object = mod.fit, parm = "(Intercept)", level = 0.95))

 2.5 % 97.5 %

(Intercept) 3.310575 4.535655

> #profile LR for exp(beta\_0)

> # Don't actually need parm = "(Intercept)" since only one parameter is being

 estimated

> exp(confint(object = mod.fit, level = 0.95))

Waiting for profiling to be done...

 2.5 % 97.5 %

3.296533 4.517434

1. This is part of a Poisson regression problem from a past project.

This problem uses the data from Table 14.14 on p. 622 of Kutner, Nachtsheim, and Neter (2004).



The full data set can be downloaded from the course website. Use ONLY store distance (X5) to predict number of customers (Y). Below is an example of how I read the data into R.

> ch14ta08 <- read.csv(file = "CH14TA08.csv")

> head(ch14ta08)

 y x1 x2 x3 x4 x5

1 9 606 41393 3 3.04 6.32

2 6 641 23635 18 1.95 8.89

3 28 505 55475 27 6.54 2.05

4 11 866 64646 31 1.67 5.81

5 4 599 31972 7 0.72 8.11

6 4 520 41755 23 2.24 6.81

Answer the following questions.

* 1. (2 points) Why should Poisson regression be investigated here instead of using normal linear regression models that assume a normal distribution for Y?

Y is a count. The Poisson distribution is used to model counts because its random variable is a non-negative integer. The normal distribution is for continuous random variables.

* 1. (4 points) Estimate and state the estimated Poisson regression model.



> mod.fit< - glm(formula = y ~ x5, data = ch14ta08, family = poisson(link = log))

> summary(mod.fit)

Call:

glm(formula = y ~ x5, family = poisson(link = log), data = ch14ta08)

Deviance Residuals :

 Min 1Q Median 3Q Max

-3.6359 -0.8458 -0.1534 0.7114 3.5154

Coefficients :

 Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.54934 0.07158 49.59 <2e-16 \*\*\*

x5 -0.17915 0.01134 -15.79 <2e-16 \*\*\*

(Dispersion parameter for poisson family taken to be 1)

 Null deviance: 422.22 on 109 degrees of freedom

Residual deviance: 184.41 on 108 degrees of freedom

AIC: 632.45

Number of Fisher Scoring iterations: 4

* 1. (4 points) Construct a scatter plot of the data with an estimated Poisson regression model plotted.

> plot(x = ch14ta08$x5, y = ch14ta08$y, xlab = "Store Distance (miles)", ylab = "Number of Customer")

> curve(expr = exp(mod.fit$coefficients[1]+mod.fit$coefficients[2]\*x), lty = 1, col = "red", add = TRUE)



* 1. (4 points) Estimate μ for 5 miles. Perform the calculations using both “by-hand” calculations (i.e., use the estimated model equation yourself) and using the predict() function.



> lin.pred <- mod.fit$coefficients[1]+mod.fit$coefficients[2]\*5

> exp(lin.pred)

(Intercept)

 14.20466

> lin.pred.hat <- predict(object = mod.fit, newdata = predict.data, type = "link", se = TRUE)

> alpha <- 0.05

> lower <- exp(lin.pred.hat$fit-qnorm(1-alpha/2)\*lin.pred.hat$se)

> upper <- exp(lin.pred.hat$fit+qnorm(1-alpha/2)\*lin.pred.hat$se)

> data.frame(predict.data, mu.hat = round(exp(lin.pred.hat$fit),4), lower =

 round(lower,4), upper = round(upper,4))

 x5 mu.hat lower upper

1 5 14.2047 13.3972 15.0608

* 1. (2 points) Describe the relationship between store distance and the number of customers.

Because  is negative, there is a negative relationship between store distance and number of customers. Thus, as the store distance increases, the estimated mean number of customers decreases.

The estimated percentage reduction in the mean number of customers that results from a 1-mile increase in distance is 16.40. The corresponding 95% confidence interval is (14.52, 18.24)

> 100\*(exp(mod.fit$coefficients[2]) - 1)

 x5

-16.40235

> #Profile likelihood interval

> beta.ci <- confint(object = mod.fit, parm = "x5", level = 0.95)

Waiting for profiling to be done...

> beta.ci

 2.5 % 97.5 %

-0.2013509 -0.1568734

> 100\*(exp(beta.ci) - 1)

 2.5 % 97.5 %

-18.23746 -14.51878

* 1. (4 points) Perform a hypothesis test to determine if there is a relationship between the store distance and the number of customers. Use both a Wald and likelihood ratio test with α = 0.05.

H0: β1=0 vs. Ha: β1≠0

ZW = -15.79 from the R output in part b). Because -15.79 < -1.96, reject Ho. There is a relationship between store distance and the number of customers.

-2log(Λ) = 237.81 – Because the p-value is approximately 0, reject Ho. There is a relationship between store distance and the number of customers.

> library(car)

> Anova(mod.fit)

Analysis of Deviance Table (Type II tests)

Response: y

 LR Chisq Df Pr(>Chisq)

x5 237.81 1 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* 1. (3 points) Why isn’t the rate data format good to use for this data (assuming only Y and X5 are the variables of interest)?

There are few observations with the same x5. Converting it to the rate data format is not really going to change much.