STAT W4201 001, Homework 9

Brian Weinstein (bmw2148)

Apr 13, 2016

Code is attached here and also posted at https://github.com/BrianWeinstein/advanced-data-analysis. Where relevant, code snippets and output are are included in-line.

Problem 1: Ramsey 20.12

Duchenne Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her children. Doctors must rely on some kind of test to detect the presence of the disease. The data in Display 20.15 are levels of two enzymes in the blood, creatine kinase (CK) and hemopexin (H), for 38 known DMD carriers and 82 women who are not carriers. It is desired to use these data to obtain an equation for indicating whether a woman is a likely carrier.

(a) *Make a scatterplot of H versus* $\log(CK)$*; use one plotting symbol to represent the controls on the plot and another to represent the carriers. Does it appear from the plot that these enzymes might be useful predictors of whether a woman is a carrier?*

A coded scatterplot of H vs $\log(CK)$ is shown in Figure 1. Based on the scatterplot, it does appear that these enzymes might be useful predictors of whether a woman is a carrier — visually, at least, it looks like the carriers have higher levels of CK, and slightly higher levels of H.

```
> mdData <- Sleuth3::ex2012
> mdData$Group <- relevel(mdData$Group, ref = "Control")
```

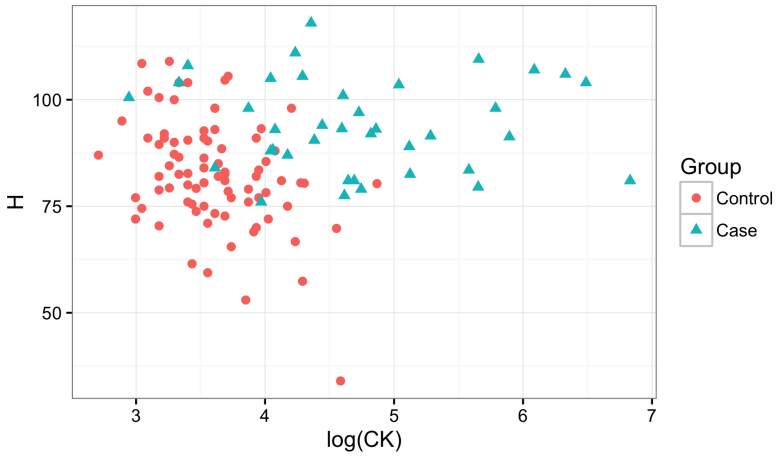

Figure 1: A coded scatterplot of H vs $\log(CK)$.

(b) *Fit the logistic regression of carrier on CK and CK-squared. Does the CK-squared term significantly differ from 0? Next fit the logistic regression of carrier on* $\log(CK)$ *and* $[\log(CK)]^2$*. Does the squared term significantly differ from 0? Which scale (untransformed or log-transformed) seems more appropriate for CK?*

The logistic regression of carrier on CK and CK-squared is shown below. The CK-squared term does not significantly differ from 0 (two-sided p-value 0.1219).

```
> glm_1b1 <- glm(formula = Group ~ CK + I(CK^2),
+                data = mdData, family = binomial)
> summary(glm_1b1)$coefficients
                  Estimate    Std. Error   z value          Pr(>|z|)
(Intercept) -4.17746146864 0.72637612139 -5.751100 0.000000008866482
CK           0.05797905485 0.01299217478  4.462614 0.000008096600851
I(CK^2)     -0.00005054336 0.00003267841 -1.546690 0.121938045015287
```

The logistic regression of carrier on $\log(CK)$ and $[\log(CK)]^2$ is shown below. The $[\log(CK)]^2$ term does not significantly differ from 0 (two-sided p-value 0.1737).

```
> glm_1b2 <- glm(formula = Group ~ log(CK) + I(log(CK)^2),
+                data = mdData, family = binomial)
> summary(glm_1b2)$coefficients
              Estimate Std. Error    z value  Pr(>|z|)
(Intercept)   9.735313  16.297521  0.5973493 0.5502742
log(CK)      -8.516251   8.358066 -1.0189261 0.3082381
I(log(CK)^2)  1.445731   1.062746  1.3603736 0.1737117
```

The log-transformed scale is more appropriate for CK, since it ranges over many orders of magnitude on the untransformed scale (from 15 to 925). A coded scatterplot of H vs CK, shown in Figure 2, further illustrates the need for the transformation.

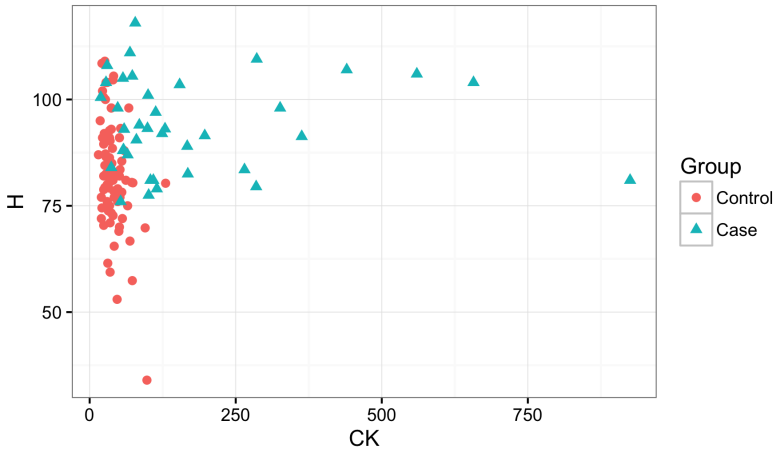

Figure 2: A coded scatterplot of H vs CK.

(c) *Fit the logistic regression of carrier on* $\log(CK)$ *and H. Report the coefficients and standard errors.*

The coefficients and standard errors in the logistic regression of carrier on $\log(CK)$ and H is shown below.

```
> glm_1c <- glm(formula = Group ~ log(CK) + H,
+               data = mdData, family = binomial)
> summary(glm_1c)$coefficients
              Estimate Std. Error   z value       Pr(>|z|)
(Intercept) -28.9134030 5.80016937 -4.984924 0.000000619862
log(CK)       4.0204252 0.82909534  4.849171 0.000001239784
H             0.1365189 0.03654202  3.735943 0.000187013358
```

(d) *Carry out a drop-in-deviance test for the hypothesis that neither* $\log(CK)$ *nor* H *are useful predictors of whether a woman is a carrier.*

We first fit a reduced model that includes only an intercept term.

```
> glm_1d <- glm(formula = Group ~ 1,
+               data = mdData, family = binomial)
> summary(glm_1d)$coefficients
              Estimate Std. Error   z value      Pr(>|z|)
(Intercept) -0.7691331  0.1962419 -3.919311 0.00008880244
```

We then compare the model from part (c) to this reduced model using a drop-in-deviance test (likelihood ratio test), testing the hypothesis that neither $\log(CK)$ nor H are useful predictors of whether a woman is a carrier.

```
> anova(glm_1c, glm_1d, test="LRT")
Analysis of Deviance Table

Model 1: Group ~ log(CK) + H
Model 2: Group ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       117     61.992
2       119    149.840 -2  -87.847 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

There is overwhelming evidence that either (1) one of the variables or (2) both of the variables are useful predictors of whether a woman is a carrier of DMD (two-sided p-value 2.2×10^{-16} from a drop-in-deviance test).

(e) *Typical values of* CK *and* H *are 80 and 85. Suppose that a suspected carrier has values of 300 and 100. What are the odds that she is a carrier relative to the odds that a woman with typical values (80 and 85) is a carrier?*

The odds of a woman being a carrier with values of 300 and 100 is 1575 times higher than the odds of a woman with values of 80 and 85.

```
> # calculate odds and probability of having DMD at CK=80, H=85
> odds1 <- exp(predict(glm_1c, data.frame(CK=80, H=85)))[[1]] ; odds1
[1] 1.361127
> 1 / (1 + exp(-odds1))
[1] 0.7959428
> # calculate odds and probability of having DMD at CK=300, H=100
> odds2 <- exp(predict(glm_1c, data.frame(CK=300, H=100)))[[1]] ; odds2
[1] 2143.332
> 1 / (1 + exp(-odds2))
[1] 1
> # calculate the odds ratio
> odds2/odds1
[1] 1574.675
```

Problem 2: Ramsey 21.16

Twenty tanks of rainbow trout embryos were exposed to one of five doses of Aflatoxicol for one hour. The data in Display 21.20 (from George Bailey and Jerry Hendricks) represent the numbers of fish in each tank and the numbers of these that had liver tumors after one year. Describe the relationship between dose of Aflatoxicol and odds of liver tumor. It is also of interest to determine the dose at which 50% of the fish will get liver tumors. (Note: Tank effects are to be expected, meaning that tanks given the same dose may have slightly different πs. Thus, one should suspect extra-binomial variation.)

A jittered scatterplot of logit(proportion of trout with liver tumors) vs log(Dose) is shown in Figure 3. The scatterplot reveals a possible quadratic parameter.

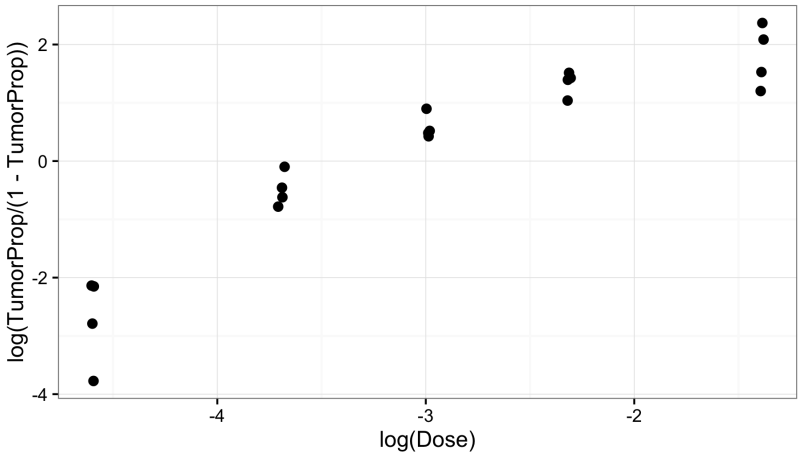

Figure 3: A jittered scatterplot of logit(proportion of trout with liver tumors) vs log(Dose).

Based on the study design, there's no guarantee that, at each Dose, the π's from each tank will be the same. As such, extra-binomial variation is suspected.

To test for extra-binomial variation, we first fit a rich model with both log(Dose) and $[\log(\text{Dose})]^2$ as explanatory variables, as shown below.

```
> glm2 <- glm(formula = TumorProp ~ log(Dose) + I(log(Dose)^2),
+ data = troutData, family = binomial, weights = Total)
> summary(glm2)

Call:
glm(formula = TumorProp ~ log(Dose) + I(log(Dose)^2), family = binomial,
    data = troutData, weights = Total)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1349  -0.6860  -0.1067   1.0382   1.8863

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.02921    0.49343   2.086  0.03699 *
log(Dose)      -1.03048    0.35743  -2.883  0.00394 **
I(log(Dose)^2) -0.39195    0.06136  -6.388 1.68e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 667.195  on 19  degrees of freedom
Residual deviance:  26.048  on 17  degrees of freedom
AIC: 119.45

Number of Fisher Scoring iterations: 4
```

The residual deviance of 26.048 on 17 degrees of freedom gives a p-value of 0.0736 from a goodness of fit test, as shown below. The p-value is low enough to indicate that the model is inadequate, and suggests the possible presence of extra-binomial variation.

```
> pchisq(q = summary(glm2)$deviance, df = summary(glm2)$df.residual,
+        lower.tail = FALSE)
[1] 0.07358942
```

Examining a plot of the deviance residuals vs fitted values in Figure 4, only one residual — for the tank at Dose 0.010, where only 2 of the 89 fish had tumors — is extreme. There's no pattern in the residual plot, so this likely isn't the cause of the low p-value from the goodness of fit test and can likely be ignored. We thus assume that cause of the low p-value is due to the presence of extra-binomial variation.

Next, taking extra-binomial variation into account, we adjust the standard errors, t-statistics, and associated p-values for the model estimates.

The dispersion parameter is estimated to be $\hat{\psi} = \frac{26.048}{17} = 1.5322$, the quasi-likelihood standard error is sqrt($\hat{\psi}$) times the maximum likelihood standard error, the t-value is the the estimate divided by the quasi-likelihood standard error, and the associated two-sided p-value (now based on the t distribution, with 17 degrees of freedom) is defined in the normal way. The computation, and summary of the maximum-likelihood and quasi-likelihood statistics are shown below.

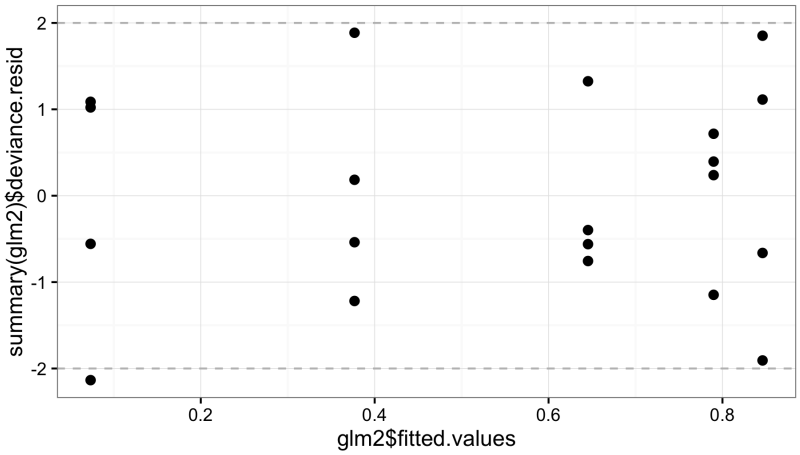

Figure 4: Scatterplot of deviance residuals vs fitted values, for `glm2`.

```
> glm2QuasiSummary <- data.frame(summary(glm2)$coefficients) %>%
+   mutate(Term=row.names(.)) %>%
+   select(Term, Estimate, ML_StdError=Std..Error, ML_ZValue=z.value, ML_PValue=Pr...z..) %>%
+   mutate(QL_StdError=ML_StdError * sqrt(dispersion_param),
+          QL_TValue=Estimate/QL_StdError)
> glm2QuasiSummary$QL_PValue <- 2 * pt(q = -1 * abs(glm2QuasiSummary$QL_TValue),
+                                      df = as.integer(summary(glm2)$df.residual))
> glm2QuasiSummary[, -1] <- round(glm2QuasiSummary[, -1], 5)
> glm2QuasiSummary
           Term Estimate ML_StdError ML_ZValue ML_PValue QL_StdError QL_TValue QL_PValue
1   (Intercept)  1.02921     0.49343   2.08585   0.03699     0.61078   1.68508   0.11024
2      log(Dose) -1.03048     0.35743  -2.88306   0.00394     0.44243  -2.32911   0.03244
3 I(log(Dose)^2) -0.39195     0.06136  -6.38761   0.00000     0.07595  -5.16031   0.00008
```

The final model, after taking the extra-binomial variation into account, is

```
           Term Estimate QL_StdError QL_TValue QL_PValue
1   (Intercept)  1.02921     0.61078   1.68508   0.11024
2      log(Dose) -1.03048     0.44243  -2.32911   0.03244
3 I(log(Dose)^2) -0.39195     0.07595  -5.16031   0.00008
```

and the odds ω of liver tumor as a function of Dose is given by

$$\omega = \frac{\pi}{1-\pi} = \mathrm{e}^{1.02921 - 1.03048 \cdot \log(\text{Dose}) - 0.39195 \cdot [\log(\text{Dose})]^2}.$$

The dose at which it's estimated that 50% of fish will get liver tumors is found by setting the probability of liver tumor π equal to 0.5, and then (numerically) solving for Dose:

$$\pi = 0.5 = \frac{1}{1 + \mathrm{e}^{-(1.02921 - 1.03048 \cdot \log(\text{Dose}) - 0.39195 \cdot [\log(\text{Dose})]^2)}}$$
$$\Rightarrow \text{Dose} \approx 0.0333345.$$

To verify this result, we plug the value into the model:

```
> # verify that ~50% of fish are expected to get tumors at Dose=0.0333345
> 1/(1+exp(-1 * predict(glm2, data.frame(Dose=0.0333345))))
       1
0.5000034
```

Problem 3:

Suppose that a population of individuals is partitioned into sub-poplations or groups, G_1 and G_2. It may be helpful to think of G_1 in an epidemiological context as the carriers of a particular virus, comprising $100\pi_1\%$ of the population, and G_2 as the non-carriers. Measurements Z made on individuals have the following distributions in the two groups:

$$\begin{aligned} G_1 &: \quad Z \sim N(\mu_1, \Sigma) \\ G_2 &: \quad Z \sim N(\mu_2, \Sigma) \end{aligned}$$

Let z be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_1 are $\pi_1/(1-\pi_1)$. Show that the posterior odds given z are

$$\frac{\pi_1}{1-\pi_1}\exp(\alpha + \beta^T z)$$

and give the form of α and β.

Let $\omega(A)$ denote the odds of A.

$$\begin{aligned} \omega(G_1|z) &= \frac{\mathrm{P}(G_1|z)}{1-\mathrm{P}(G_1|z)} \\ &= \frac{\mathrm{P}(G_1|z)}{\mathrm{P}(G_2|z)} \\ &= \frac{\mathrm{P}(z|G_1)}{\mathrm{P}(z|G_2)} \cdot \frac{\mathrm{P}(G_1)}{\mathrm{P}(G_2)} \\ &= \frac{\exp\left(\frac{-(z-\mu_1)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(z-\mu_2)^2}{2\sigma^2}\right)} \cdot \frac{\pi_1}{1-\pi_1} \\ &= \exp\left(\left(\frac{\mu_1-\mu_2}{\sigma^2}\right)^T z - \frac{\mu_1^2-\mu_2^2}{2\sigma^2}\right) \cdot \frac{\pi_1}{1-\pi_1} \\ &= \frac{\pi_1}{1-\pi_1} \cdot \exp\left((\mu_1-\mu_2)^T\Sigma^{-1}z - \frac{1}{2}\left(\mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1\right)\right) \\ &= \frac{\pi_1}{1-\pi_1} \cdot \exp\left(\beta^T z + \alpha\right) \end{aligned}$$

where $\alpha = \frac{1}{2}\left(\mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1\right)$ and $\beta = \Sigma^{-1}(\mu_2 - \mu_1)$.

```
################################################################################

# Brian Weinstein - bmw2148
# STAT W4201 001
# Homework 9
# 2016-04-13

# set working directory
setwd("~/Documents/advanced-data-analysis/homework_09")

# prevent R from printing large numbers in scientific notation
options(scipen=5)

# load packages
library(Sleuth3) # Data sets from Ramsey and Schafer's "Statistical Sleuth
(3rd ed)"
library(ggplot2); theme_set(theme_bw())
library(dplyr)



# Problem 1: Ramsey 20.12
#####################################################################

# load data
mdData <- Sleuth3::ex2012
mdData$Group <- relevel(mdData$Group, ref = "Control")

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###

# scatterplot
ggplot(mdData, aes(x=log(CK), y=H)) +
  geom_point(aes(color=Group, shape=Group), size=2)
ggsave(filename="writeup/1a.png", width=6.125, height=3.5, units="in")

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###

# fit a logistic regression model on CK and CK^2
glm_1b1 <- glm(formula = Group ~ CK + I(CK^2),
               data = mdData, family = binomial)
summary(glm_1b1)$coefficients

# fit a logistic regression model on log(CK) and log(CK)^2
glm_1b2 <- glm(formula = Group ~ log(CK) + I(log(CK)^2),
               data = mdData, family = binomial)
summary(glm_1b2)$coefficients

# scatterplot
ggplot(mdData, aes(x=CK, y=H)) +
  geom_point(aes(color=Group, shape=Group), size=2)
ggsave(filename="writeup/1b.png", width=6.125, height=3.5, units="in")

# Part c ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
```

```
# fit a logistic regression model on log(CK) and H
glm_1c <- glm(formula = Group ~ log(CK) + H,
              data = mdData, family = binomial)
summary(glm_1c)$coefficients

# Part d ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###

# fit a reduced model
glm_1d <- glm(formula = Group ~ 1,
              data = mdData, family = binomial)
summary(glm_1d)$coefficients

# perform the likelihood ratio test (drop-in-deviance test)
anova(glm_1c, glm_1d, test="LRT")

# Part e ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###

# calculate odds and probability of having DMD at CK=80, H=85
odds1 <- exp(predict(glm_1c, data.frame(CK=80, H=85)))[[1]] ; odds1
1 / (1 + exp(-odds1))

# calculate odds and probability of having DMD at CK=300, H=100
odds2 <- exp(predict(glm_1c, data.frame(CK=300, H=100)))[[1]] ; odds2
1 / (1 + exp(-odds2))

# calculate the odds ratio
odds2/odds1

rm(list = ls()) # clear working environment



# Problem 2: Ramsey 21.16
##########################################################################

# load data and create a tumor proportion variable
troutData <- Sleuth3::ex2116 %>%
  mutate(TumorProp=Tumor/Total)

# scatterplot
set.seed(1)
ggplot(troutData, aes(x=log(Dose), y=log(TumorProp/(1-TumorProp)))) +
  geom_jitter(size=2, width=0.05)
ggsave(filename="writeup/2_scatter.png", width=6.125, height=3.5, units="in")

# fit a binomial counts logistic regression model on a rich model
glm2 <- glm(formula = TumorProp ~ log(Dose) + I(log(Dose)^2),
            data = troutData, family = binomial, weights = Total)
summary(glm2)

# compute the goodness of fit p value
pchisq(q = summary(glm2)$deviance, df = summary(glm2)$df.residual,
       lower.tail = FALSE)
```

```
# examine deviance residuals
qplot(x=glm2$fitted.values, y=summary(glm2)$deviance.resid) +
  geom_point(size=2) +
  geom_hline(yintercept = c(-2, 2), linetype="dashed", color="gray")
ggsave(filename="writeup/2_devresid.png", width=6.125, height=3.5, units="in")

# estimate the dispersion parameter
dispersion_param <- summary(glm2)$deviance / summary(glm2)$df.residual
dispersion_param
sqrt(dispersion_param)

# compute the quasi-likelihood standard errors, t-statistics, and pvalues
glm2QuasiSummary <- data.frame(summary(glm2)$coefficients) %>%
  mutate(Term=row.names(.)) %>%
  select(Term, Estimate, ML_StdError=Std..Error, ML_ZValue=z.value,
ML_PValue=Pr...z..) %>%
  mutate(QL_StdError=ML_StdError * sqrt(dispersion_param),
         QL_TValue=Estimate/QL_StdError)
glm2QuasiSummary$QL_PValue <- 2 * pt(q = -1 * abs(glm2QuasiSummary$QL_TValue),
                                     df =
as.integer(summary(glm2)$df.residual))
glm2QuasiSummary[, -1] <- round(glm2QuasiSummary[, -1], 5)
glm2QuasiSummary

# final model
glm2QuasiSummary %>%
  select(Term, Estimate, QL_StdError, QL_TValue, QL_PValue)

# solved numerically in mathematica that Dose=0.0333345:
# verify that ~50% of fish are expected to get tumors at Dose=0.0333345
1/(1+exp(-1 * predict(glm2, data.frame(Dose=0.0333345))))

rm(list = ls()) # clear working environment



# Problem 3
#########################################################################

# no code needed
```