# STAT W4201 001, Homework 19

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")
```

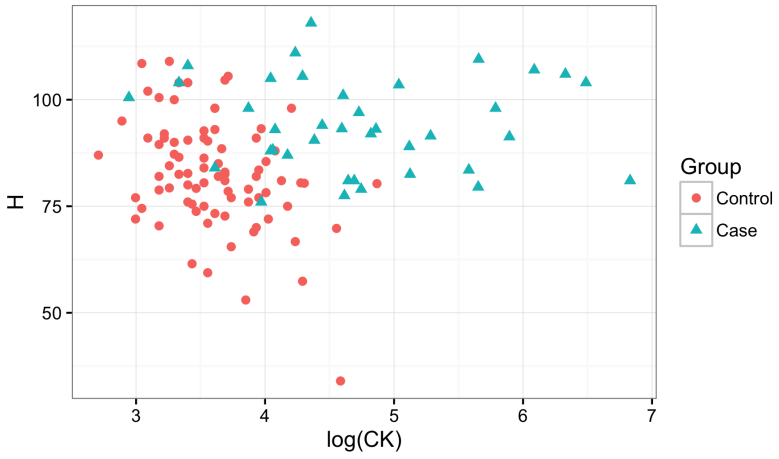


Figure 1: A coded scatterplot of $H$ vs $\log(CK)$.

(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.

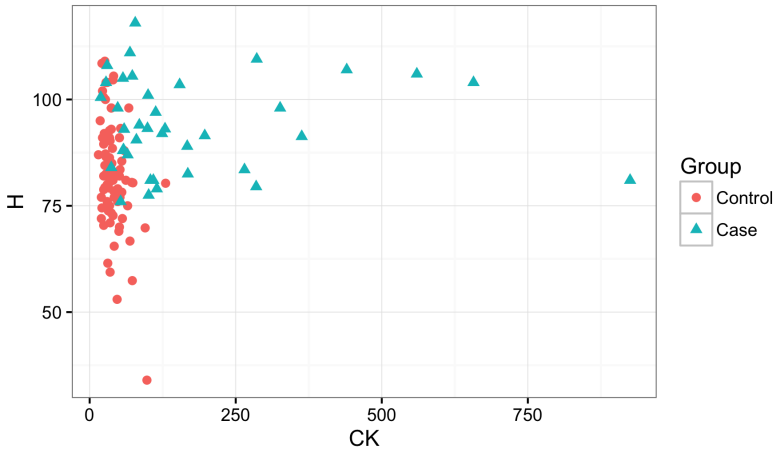


Figure 2: A coded scatterplot of $H$ vs $CK$.

(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 $2.2 \times 10^{-16}$ 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**

**Problem 3:**