STA 101 Homework 5 Solutions

Dr. Erin K. Melcon

1. (a) Using forward selection, the X's that were included were: X_1 and X_2.

   (b) Using backward selection, the X's that were included were: X_1 and X_2.

   (c) The resulting logistic model is: $logit(\hat{\pi}) = -1.4578 + (0.0779)X_1 + (-0.0955)X_2$
   With log-likelihood -52.8976936

   (d) The log-likelihood is: -52.8751729.

   (e) H_0 : The model without interactions fits better $\beta_3 = 0$. vs H_A : The model with interactions fits better $\beta_3 \neq 0$
   The test-statistic is: $LR = -2(LL_0 - LL_A) = -2(-52.8976936 - (-52.8751729)) = 0.045$, with corresponding d.f = 1, and p-value 0.832004 .

   (f) For any reasonable α, we fail to reject the null and conclude that the model without the interaction between age and health awareness score fits better.

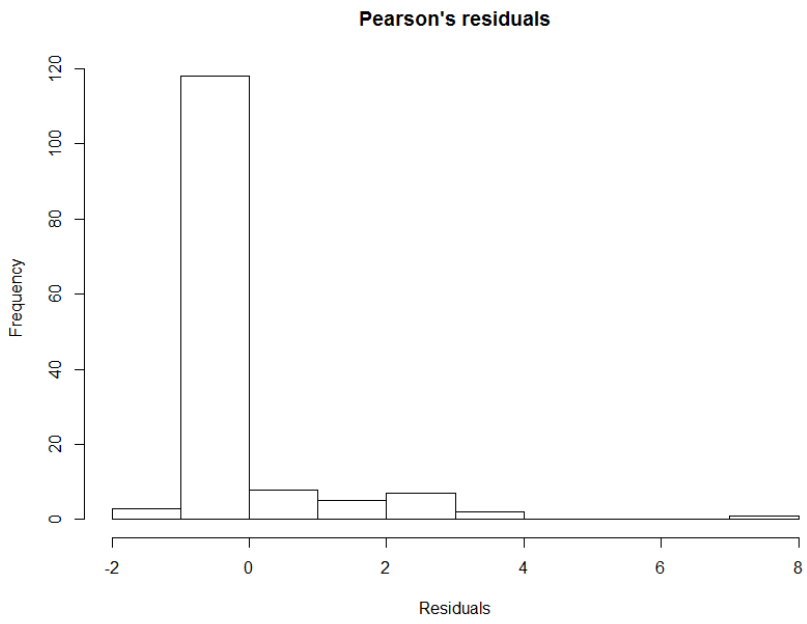
---

2. (a) When age increases by one year, the odds of getting a flu shot are multiplied by 1.081, holding all other variables contstant.

   (b) When health awareness score increases by one year, the odds of getting a flu shot are multiplied by 0.9089, holding all other variables contstant.

   (c) The prediction is: 0.1427452, in other words we would predict that they would not get a flu shot.

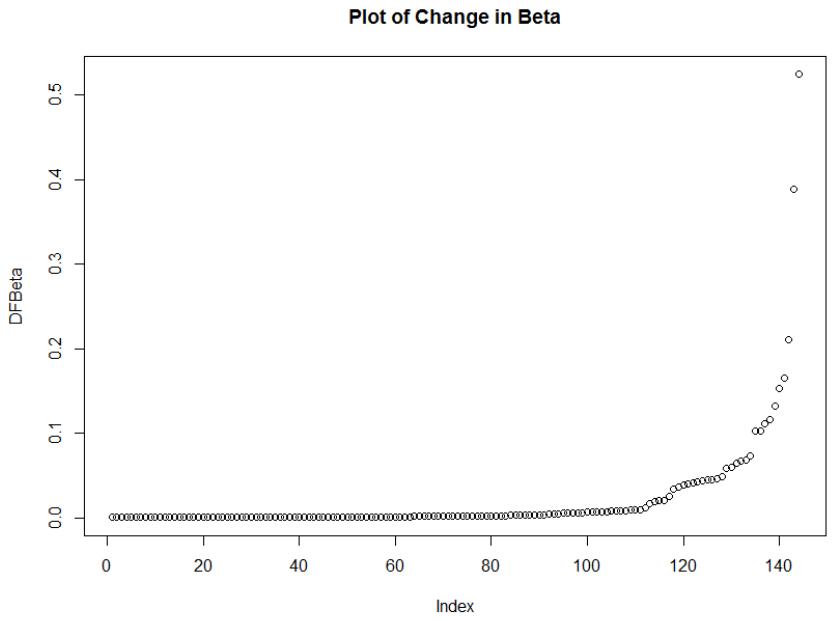
   (d) The histogram is:



   There is a value above 4, and the row is:

	y	aware	age	Pr
1	1.00	75.00	59.00	7.48

   (e) The index plot is:



There is a value above 0.30, and they are:

	y	aware	age	dBhat
1	1.00	42.00	51.00	0.39
2	1.00	75.00	59.00	0.52

---

3. (a) The best model by forward selection includes X_1 and X_2.

   (b) The best model by forward selection includes X_1 and X_2.

   (c) The two models are:
   $ln(\pi_{None}/\pi_{Long}) = -1.0317 + (0.0514)X_{age} + (-0.6055)X_G + (-0.1115)X_L + (-0.3609)X_M$
   $ln(\pi_{Short}/\pi_{Long}) = -3.6834 + (0.1053)X_{age} + (0.4272)X_G + (-10.6328)X_L + (-2.2721)X_M$

   (d) When age increases by one unit, the relative probability of a women using no contraceptive vs. long term contraceptive is multiplied by $exp(0.0514) = (1.0527)$, holding all other variables constant.

   (e) The relative probability of of a women using short term vs. long term contraceptive for the graduate group is $exp(-0.3609) = (0.697)$ times that of the advanced degree group, holding all other variables constant.

---

4. (a) The relative probability of of a women using no contraceptive vs. long term contraceptive for the graduate group is $exp(-0.6055 - -0.3609) = (0.783)$ times that of the high school degree group, holding all other variables constant.

   (b) The relative probability of of a women using short term vs. long term contraceptive for the below high school group is $exp(-10.6328 - -2.2721) = (2 \times 10^{-4})$ times that of the high school degree group, holding all other variables constant.

   (c) The null is: $H_0 : \beta_{None,2} = \beta_{None,3} = \beta_{None,4} = 0$ vs. H_A : At least one $\beta_{None,i} \neq 0$ for $i = 2, 3, 4$.
   .

   (d) The null is: The test-statistic is: $LR = -2(LL_0 - LL_A) = -2(-306.934011 - (-278.7485174)) = 56.3709871$, with corresponding d.f $= 6$, and p-value ≈ 0 .

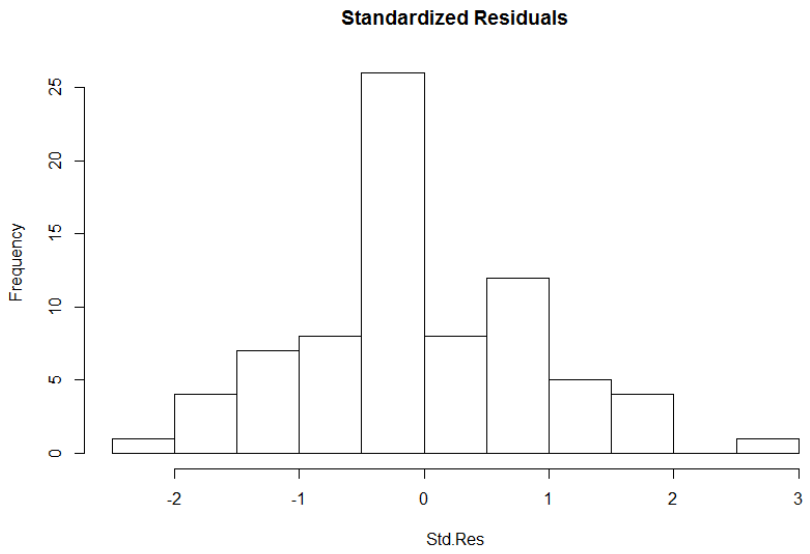
   (e) We reject the null, and conclude that the education level of the mother cannot be dropped from the model.

   (f) The predicted probabilities are:

	Long	None	Short
$\hat{\pi}_i$	0.3729	0.3221	0.3049

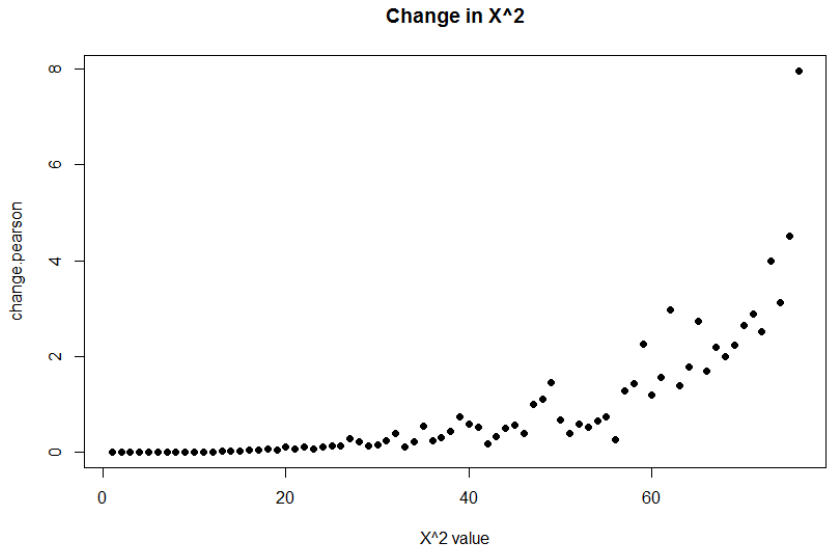
5. (a) The estimated **logistic** regression model is (where $y = 1$ means they were in the "short" category):
$logit(\pi) = -3.7968 + (0.1064)X_{age} + (0.4583)X_G + (-17.4359)X_L + (-1.94)X_M$

   (b) The histogram is:



   There are no values above 3 or below negative 3.

   (c) The index plot is:



   There are no values above 8, although one is very close!

   (d) There were no values that particuarlly stood out as outliers, so I would suggest thtat we do not remove any observations.

---

6. (a) FALSE. An influential point may also be a row of the dataset that showed up multiple times.

   (b) FALSE. The β's will in general be different for every sub-model.

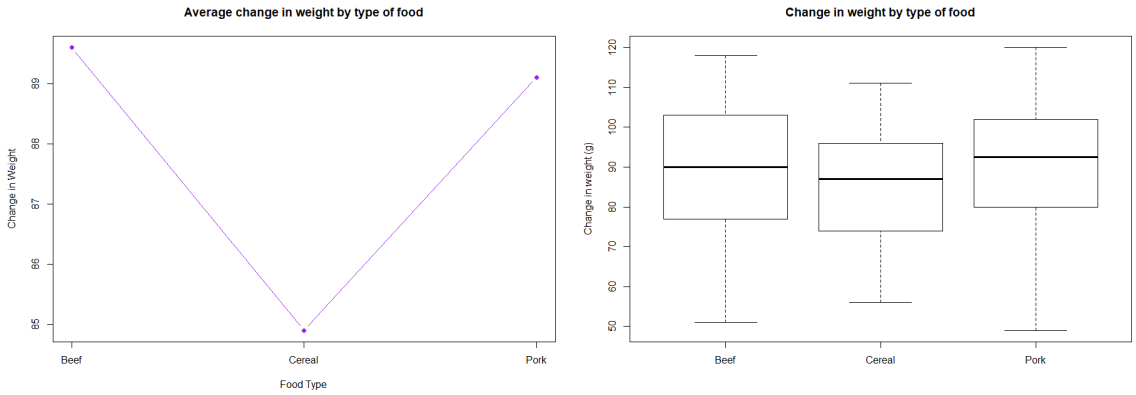
   (c) FALSE. The null hypothesis is the hypothesis that the smaller model is a better fit.

   (d) FALSE. These are distributed approximately standard normal, and 0.50 is a very common value for that distribution.

7. (a) The summary statistics are:

	Beef	Cereal	Pork
Means	89.6	84.9	89.1
Std. Dev	17.7123	14.9944	17.3202
Sample Size	20	20	20

(b) The plot follows:



Since the boxplots overlap quite a bit, there does not seem to be a significant differences in the means.

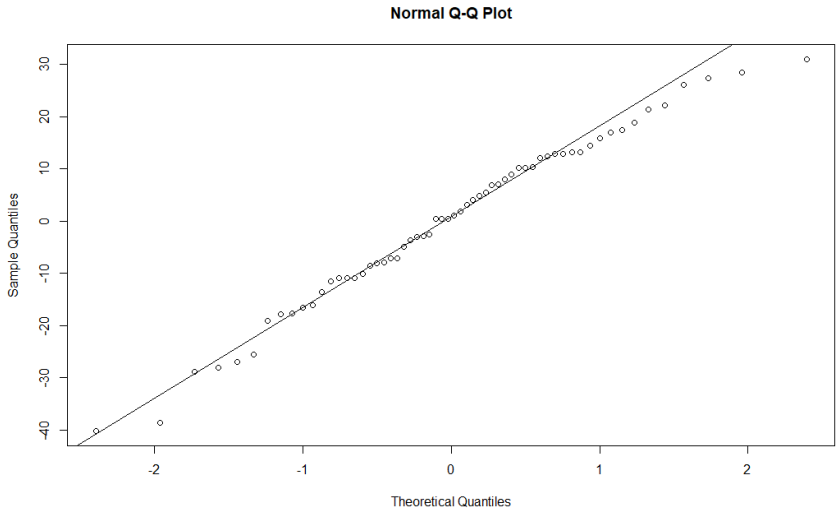
(c) The null is: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ vs. H_A : At least one $\alpha_i \neq 0$.

(d) The test-statistic is: 0.4767769, with corresponding p-value 0.6232343.

(e) Since the p-value is larger than α, we fail to reject the null, and conclude that there is no significant difference in the average weight change for the three groups (or there is no significant group effect).

---

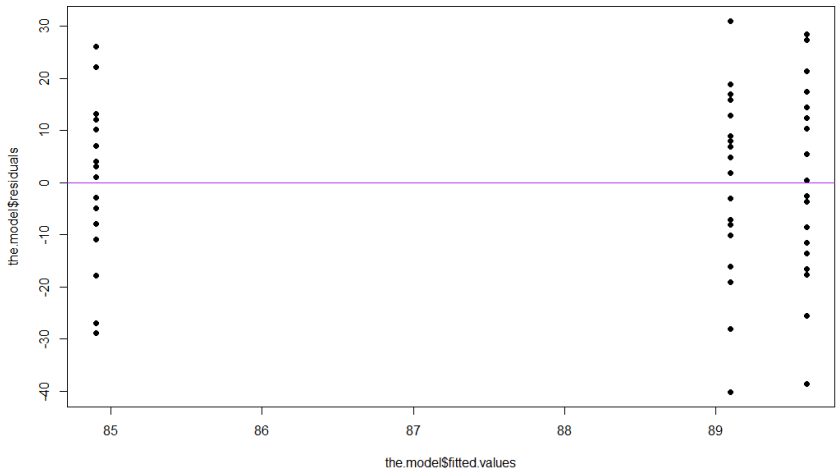
8. (a) The plot follows:



The points are close to the line, so that the data looks approximately normal.

(b) The associated p-value is : 0.5160281. Since this p-value is larger than α, we fail to reject the null, and conclude that the data is approximately normal.

(c) The plot follows:



The variance of the groups appear to be approximately equal, since at each group the points seem to have the same spread.

(d) The associated p-value is : 0.8357656, NA. Since this p-value is larger than α, we fail to reject the null, and conclude that the variances of each group are approximately equal.

(e) Yes, since the variance of the errors appears to be constant, and the data appear to be normal.

```r
#Problem 1
flu = read.csv("C:/Github/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-5/flu.csv")
full.model = glm(shot ~ ., data = flu,family = binomial)
null.model = glm(shot ~ 1, data = flu,family = binomial)
forward.model = step(null.model, scope = list(lower = null.model, upper = full.model), direction = "forward",t
backward.model = step(full.model, scope = list(lower = null.model, upper = full.model), direction = "backward"
B = round(forward.model$coefficients,4)
inter.model = glm(shot ~ age + aware+ age*aware, data = flu,family = binomial)
best.model = forward.model
LLA = logLik(inter.model)
LL0 = logLik(best.model)
LR  = round(-2*(LL0-LLA),4); d.f = length(inter.model$coefficients) - length(best.model$coefficients)
p.val = pchisq(LR,d.f, lower.tail=F)
#Problem 2
EB = round(exp(B),4)
pi.hat = predict(best.model,data.frame(age = 57,aware=50),type = "response")
library(LogisticDx)
good.stuff = dx(best.model)
pear.r = good.stuff$Pr #Pearsons Residuals
#hist(pear.r,main = "Pearson's residuals", xlab = "Residuals")
outliers = good.stuff[pear.r >4,c("y","aware","age","Pr") ]
df.beta = good.stuff$dBhat #DF Beta for removing each observation
#plot(df.beta,ylab = "DFBeta",main = "Plot of Change in Beta")
outliers = good.stuff[df.beta >0.30,c("y","aware","age","dBhat") ]

#problem 3
library(nnet)
control = read.csv("C:/Github/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-5/control.csv")
full.model = multinom(con ~+age  +(edu) + (working),data = control,trace = FALSE)
null.model = multinom(con ~ 1,data = control,trace = FALSE)
forward.model = step(null.model, scope = list(lower = null.model, upper = full.model), direction = "forward",t
backward.model = step(full.model, scope = list(lower = null.model, upper = full.model), direction = "backward"
best.model = backward.model
B = round(coef(best.model),4)
#Problem 4
options(scipent = 8)
large.model = best.model
small.model = multinom(con ~ +age ,data = control,trace = FALSE)
LLA = as.numeric(logLik(large.model))
LL0 = as.numeric(logLik(small.model))
LR  = -2*(LL0-LLA); d.f = large.model$edf - small.model$edf
p.val = pchisq(LR,d.f, lower.tail=F)
x.star = data.frame(age = 29, edu = "G")
pi.hats =predict(large.model,x.star,type = "probs")
library(LogisticDx)
split.data = split(control, control$con)
ShortLong = rbind(split.data[[3]], split.data[[1]])
ShortLong$con = ifelse(ShortLong$con =="Short",1,0)
SL.model = glm(con ~+age  +(edu) ,data = ShortLong,family = binomial)
B = round(SL.model$coefficients,4)
good.stuff = dx(SL.model)
std.r = good.stuff$sPr #Standardized residuals (Pearson)
#hist(std.r,main = "Standardized Residuals",xlab = "Std.Res")
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation
#plot(change.pearson,main = "Change in X^2",xlab = "X^2 value")
outliers = which(change.pearson > 8)
outlier.values = good.stuff[outliers,c("y","age","eduG","eduL","eduM","sPr") ]
```

```
options(scipen = 8)
library(asbio)
rat = read.csv("C:/Github/Teaching-Materials/STA-101/STA-101-2017-Spring/Datasets/HW-6/rat.csv")
#Problem 1
#(a)
group.means = by(rat$Weight,rat$Type,mean)  # First argument is Y, second is grouping column/s
#par(mfrow = c(1,2))
#plot(group.means,xaxt = "n",pch = 19,col = "purple",xlab = "Food Type",ylab = "Change in Weight",main = "Aver
#axis(1,1:length(group.means),names(group.means))
#boxplot(Weight ~ Type,data = rat, main = "Change in weight by type of food",ylab = "Change in weight (g)")
#(b)
group.means = by(rat$Weight,rat$Type,mean)
group.sds = by(rat$Weight,rat$Type,sd)
group.nis = by(rat$Weight,rat$Type,length)
the.summary = rbind(group.means,group.sds,group.nis)
the.summary = round(the.summary,digits = 4)
colnames(the.summary) = names(group.means)
rownames(the.summary) = c("Means","Std. Dev","Sample Size")
#(c)
the.model = aov(Weight ~ Type, data = rat)
anova.table = anova(the.model)
Fs = anova.table[1,4]; p.val = anova.table[1,5]
#Problem 2
#(a)
#qqnorm(the.model$residuals)
#qqline(the.model$residuals)
#(b)
shap.test = shapiro.test(the.model$residuals)
p.val = shap.test$p.value
#(c)
#plot(the.model$fitted.values,the.model$residuals,pch = 19)
#abline(h= 0 , col = "purple")
#(d)
ML.test = modlevene.test(the.model$residuals,rat$Type)
p.val.ml = ML.test$'Pr(>F)'
```