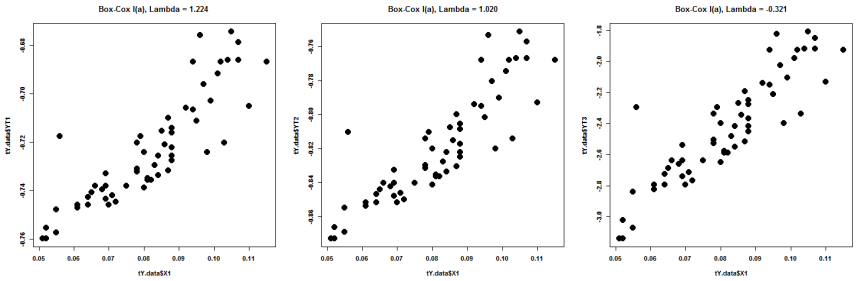
STA 101 Homework 4 Solutions

Dr. Erin K. Melcon

1. (a) The three possible values of λ are :

Method	QQ-plot	Shapiro-Wilks	Log-Liklihood
λ	1.224	1.021	-0.321

With the following corresponding plots:

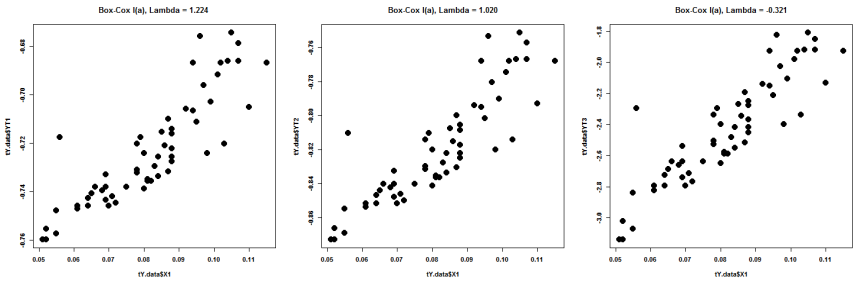


(b) The three possible p-values are: For any of the three, we would fail to reject the null hypothesis and conclude that

Method	QQ-plot	Shapiro-Wilks	Log-Liklihood
p-value	0.0556	0.0576	0.2610

the two groups of errors have equal variance.

(c) The three possible plots follow:



Notice they all look pretty similar, because the λ value for X was 1.0740246, which is fairly close to 1 (i.e, no transformation).

(d) The possible combinations are:

	QQ-Plot	Shapiro-Wilks	Log-Likelihood
FK Test	0.0515	0.0562	0.2454
SW test	0.0272	0.0280	0.0053

(e) If we use $\alpha = 0.01$, we can fail to reject both the null hypothesis for FK and SW for either the Shapiro-Wilks method or the QQ-plot method (but just barely).
When using the Log-Likelihood

2. (a) The estimated regression function is:

$$logit(\hat{\pi}) = -1.1772 + (0.0728)X_1 + (-0.099)X_2 + (0.434)X_{3,M}$$

(b) The estimated probability of a flu shot for a 55 year old is: 0.0162565

(c) Since the sign of $\hat{\beta}_1$ is positive, it suggests the probability increases as age does.

(d) The odds of a flu shot are multiplied by 1.0755154 for every one year increase in age, holding all other variables constant.

(e) A 90% confidence interval is found to be: (1.0149, 1.1446) (after we exponentiate). We are 90% confident that when age increases by 1 year, the odds of getting a flu shot are between 1.0149 and 1.1446 times what they were, holding all other variables constant.

(f) The odds of a flu shot if you are a male are 1.5434189 times that a of a female, holding all other variables constant.

(g) A 90% confidence interval is found to be: (0.5613, 4.4485) (after we exponentiate). We are 90% confident the odds of a flu shot for a male is between 0.5613 and 4.4485 times that of a female, holding all other variables constant.

---

3. (a) The plot follows:

Since there is not a steep slope, it does not appear that age has a significant effect on if you get a flu shot or not.

(b) Since the data range is from 48 to 84, and 0 is far beyond that age (and also infants do not get flu shots), there is no practical interpretation.

(c) Since the age 12 is well beyond the range of our data, we **would not** predict the probability of a flu shot at this age.

(d) The error matrix is: Then,

	$\hat{y}=0$	$\hat{y}=1$
$y=0$	132	3
$y=1$	22	2

i. the sensitivity is: $\frac{6}{24} = 0.25$

ii. the specificity is: $\frac{130}{135} = 0.962963$

iii. the error rate is: $1 - \frac{136}{159} = 0.1446541$

(e) : The AUC is: 0.8223765 with cooresponding 95% confidence interval: (0.7308155, 0.9139376). Since this confidence interval is above 0.60, it suggests our model fits well.

---

4. (a) The estimated regression function is:

$$logit(\hat{\pi}) = -5.3095 + (0.1109)x$$

(b) The estimated probability of a CHD for a 69 year old is: 0.9124646

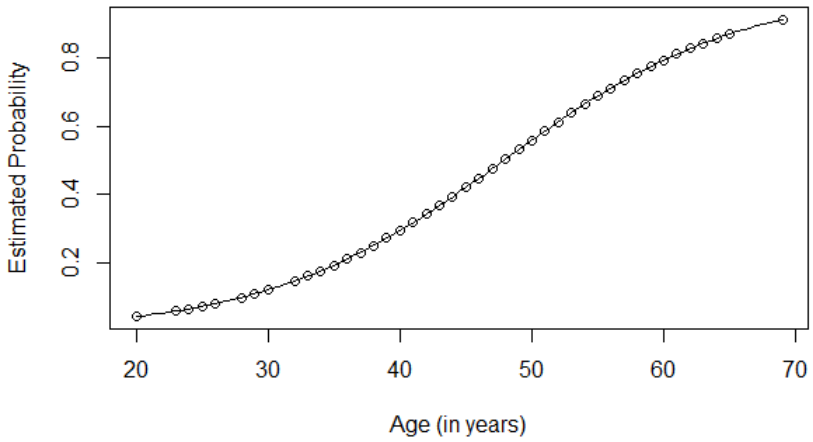
(c) Since the sign of $\hat{\beta}_1$ is positive, it suggests the probability increases as age does.

(d) The odds of a flu shot are multiplied by 1.1172832 for every one year increase in age.
Or, when age increases by 1 year, the odds of getting CHD shot are 1.1172832 times what they were.

(e) A 99% confidence interval is found to be: (1.0692, 1.1759) (after we exponentiate). We are 99% confident that when age increases by 1 year, the odds of getting CHD are between 1.0692 and 1.1759 times what they were.

---

5. (a) The plot follows:



Since there is not a steep slope, it does not appear that age has a significant effect on if you get CHD or not.

(b) Since the data range is from 48 to 84, and 0 is far beyond that age (and also aged 0 people do not get CHD), there is no practical interpretation.

(c) Since the age 44 is well within the range of our data, we **would** predict the probability of CHD at this age.

(d) The error matrix is: Then,

	$\hat{y}=0$	$\hat{y}=1$
$y=0$	45	12
$y=1$	14	29

i. the sensitivity is: $\frac{29}{43} = 0.6744186$

ii. the specificity is: $\frac{45}{57} = 0.7894737$

iii. the error rate is: $1 - \frac{74}{100} = 0.26$

(e) : The AUC is: 0.7998776 with cooresponding 95% confidence interval: (0.7113503, 0.8884049). Since this confidence interval is above 0.67, it suggests our model fits relatively well.

---

6. (a) TRUE. If $\beta_1 = 0$, there is no effect on the logit of the probability of success, and $\exp(\beta_1) = 1$.

(b) FALSE. $\exp \beta_1$ what the odds are multiplied by, not the odds themselves.

(c) FALSE. α only has meaning when $x = 0$ is a legitimate value of X.

(d) TRUE. This means that $logit(\pi(x))$ does not change with X_1, and neither does $\pi(x)$.

```r
library(car)
library(EnvStats)
alcohol <- read.csv("C:/Github/Teaching-Materials/STA-108-2017-Fall/Datasets/HW02/alcohol.csv")
names(alcohol) = c("X2","X1","Y")
the.model = lm(Y ~ X1, data = alcohol)

L1 =boxcox(the.model ,objective.name = "PPCC",optimize = TRUE)$lambda
L2 = boxcox(the.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda
L3 = boxcox(alcohol$Y,objective.name = "Log-Likelihood",optimize = TRUE)$lambda
LX1 = boxcox(alcohol$X1,objective.name = "Log-Likelihood",optimize = TRUE)$lambda

YT1 = (alcohol$Y^(L1)-1)/L1
YT2 = (alcohol$Y^(L2)-1)/L2
YT3 = (alcohol$Y^(L3)-1)/L3

X1T =  (alcohol$X1^(LX1)-1)/LX1
tY.data = data.frame(YT1,YT2,YT3, X1 = alcohol$X1, X2 = alcohol$X2)

# par(mfrow= c(1,3))
#plot(tY.data$X1,tY.data$YT1,pch = 19,cex = 2, main = "Box-Cox I(a), Lambda = 1.224 ",font = 2,font.lab = 2)
#plot(tY.data$X1,tY.data$YT2,pch = 19,cex = 2, main = "Box-Cox I(a), Lambda = 1.020",font = 2,font.lab = 2)
#plot(tY.data$X1,tY.data$YT3,pch = 19,cex = 2, main = "Box-Cox I(a), Lambda = -0.321",font = 2,font.lab = 2)


#(b)
equal.var = function(Y, ei){
  Group = rep("Lower",length(Y)) #Creates a vector that repeats "Lower" n times
  Group[Y > median(Y)] = "Upper" #Changing the appropriate values to "Upper"
  Group = as.factor(Group)
  the.FKtest= fligner.test(ei, Group)
  p.value = the.FKtest$p.value
  return(p.value)
}
ei1 = lm(YT1 ~ X1, data = tY.data)$residuals
ei2 = lm(YT2 ~ X1, data = tY.data)$residuals
ei3 = lm(YT3 ~ X1, data = tY.data)$residuals

pval1 = equal.var(YT1, ei1)
pval2 = equal.var(YT2, ei2)
pval3 = equal.var(YT3, ei3)

#(c)
tY.data$X1 = X1T

# par(mfrow= c(1,3))
#plot(tY.data$X1,tY.data$YT1,pch = 19,cex = 2, main = "Box-Cox I(a), Lambda = 1.224 ",font = 2,font.lab = 2)
#plot(tY.data$X1,tY.data$YT2,pch = 19,cex = 2, main = "Box-Cox I(a), Lambda = 1.020",font = 2,font.lab = 2)
#plot(tY.data$X1,tY.data$YT3,pch = 19,cex = 2, main = "Box-Cox I(a), Lambda = -0.321",font = 2,font.lab = 2)

#(d)


ei1 = lm(YT1 ~ X1, data = tY.data)$residuals
ei2 = lm(YT2 ~ X1, data = tY.data)$residuals
ei3 = lm(YT3 ~ X1, data = tY.data)$residuals

BF.pval1 = equal.var(YT1, ei1)
```

```
BF.pval2 = equal.var(YT2, ei2)
BF.pval3 = equal.var(YT3, ei3)

SW.pval1 = shapiro.test(ei1)$p.val
SW.pval2 = shapiro.test(ei2)$p.val
SW.pval3 = shapiro.test(ei3)$p.val

results = rbind(c(BF.pval1, BF.pval2, BF.pval3), c(SW.pval1,SW.pval2,SW.pval3))

library("pROC")
#Problem 1
flu <- read.csv("C:/Github/Teaching-Materials/STA-101-2018-Spring/Datasets/HW-4/flu.csv")
logit.model = glm(shot~age +aware+gender,data = flu,family=binomial)
CI = round(exp(as.numeric(confint(logit.model)[2,],level = .90)),4)
CI2= round(exp(as.numeric(confint(logit.model)[4,],level = .90)),4)

options(scipen = 8)
#################Question 2
#plot(flu$age,logit.model$fitted.values,xlab = "Age (in years)",ylab = "Estimated Probability")
#curve(predict(logit.model, data.frame(age=x), type="response"), add=TRUE)
pi0 =0.50
truth = logit.model$y
predicted = ifelse(fitted(logit.model)>pi0,1,0)
my.table = table(truth,predicted)
sens = sum(predicted == 1 & truth ==1 )/sum(truth == 1)
spec = sum(predicted == 0 & truth ==0 )/sum(truth == 0)
error = sum(predicted != truth)/length(predicted)
library(pROC)
the.roc = roc(logit.model$y, logit.model$fitted.values,auc = TRUE, ci = TRUE,plot=FALSE, legacy.axes = TRUE)
AUC= auc(the.roc)
CI.AUC = ci(the.roc,plot = FALSE)
#################Question 3
cheart <- read.csv("C:/Github/Teaching-Materials/STA-138/STA-138-2017-Winter/Datasets/HW-6/cheart.csv")
logit.model = glm(CHD~AGE,data = cheart,family=binomial)
CO = round(logit.model$coefficients,4)
CI = round(exp(as.numeric(confint(logit.model)[2,],level = .99)),4)
#################Question 4
#plot(heart$AGE,logit.model$fitted.values,xlab = "Age (in years)",ylab = "Estimated Probability")
#curve(predict(logit.model, data.frame(AGE=x), type="response"), add=TRUE)
pi0 =0.50
truth = logit.model$y
predicted = ifelse(fitted(logit.model)>pi0,1,0)
my.table = table(truth,predicted)
sens = sum(predicted == 1 & truth ==1 )/sum(truth == 1)
spec = sum(predicted == 0 & truth ==0 )/sum(truth == 0)
error = sum(predicted != truth)/length(predicted)
library(pROC)
the.roc = roc(logit.model$y, logit.model$fitted.values,auc = TRUE, ci = TRUE,plot=FALSE, legacy.axes = TRUE)
AUC= auc(the.roc)
CI.AUC = ci(the.roc)
```