Applied Logistic Regression: Homework - Week 3

September/2016

Contents

1 Exercise One

For this exercise, we need the Myopia Study dataset. The data is available by Professor Lemeshow's course on Canvas Network and it has been studied in Exercise One of Week One and Week Two.

Complete the following:

1. Using the results from Week Two, Exercise 1, compute 95 % confidence intervals for the slope coefficient SPHEQ. Write a sentence interpreting this confidence.
2. Use analysis software tool to obtain the estimated covariance matrix. Then compute the logit and estimated logistic probability for a subject with SPHEQ = 2. Then evaluate the endpoints of the 95 % confidence intervals for the logit and estimated logistic probability. Write a sentence interpreting the estimated probability and its confidence interval.

Solving:

For getting data in R[2]:

```
filename = "MYOPIA.csv"
if (!file.exists(filename)) {
fileURL="https://learn.canvas.net/courses/1179/files/461762/download?download_frd=1"
        download.file(fileURL, filename)}
data = read.csv("MYOPIA.csv", header = T, sep = ",")
```

1. Codes for slope:

```
mod = glm(MYOPIC ~ SPHEQ, data = data, family = binomial(link = "logit"))
mod1 = xtable(mod,
caption = "Results of Fitting the Logistic Regression Model to the MYOPIA Data, n = 618.",
label = "tab01")
print(mod1, caption.placement = 'top')
```

Table 1: Results of Fitting the Logistic Regression Model to the MYOPIA Data, n = 618.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0540	0.2067	0.26	0.7940
SPHEQ	-3.8331	0.4184	-9.16	0.0000

The estimated logit $\hat{g}(x)$ for the fitted model in Table 1 is shown in equation 1

$$\hat{g}(x) = 0.05 + (-3.83) \times SPHEQ. \tag{1}$$

The sentence for 95 percent confidence intervals for the slope $\hat{\beta_1}$ of coefficient SPHEQ is given in equation 2,

$$\hat{\beta_1} \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}(\hat{\beta_1}), \tag{2}$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $100(1-\alpha)$ % point from the standard normal distribution and $\widehat{SE}(\cdot)$ denotes a model-based estimator of the standard error of the $\hat{\beta_1}$.

```
test = summary(mod)
(test$coefficients[2]+c(-1,1)*abs(qnorm(.975,lower.tail=F))*test$coefficients[4])
```

[1] -4.653087 -3.013108

We have from equation 2 and Table 1,

$$\hat{\beta_1} \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}(\hat{\beta_1}) = -3.83 \pm 1.96 \times (0.42) = (-4.65, -3.01) \tag{3}$$

and we can be 95 % confident that true value of β_1 is between -4.65 and -3.01.

2. The covariance matrix for model in Table 1 is given in Table 2.

```
A = xtable(test$cov.scaled,
caption = "Estimated Covariance Matrix of the Estimated Coefficients in Table 1.",
label = "tab02")
print(A, caption.placement = 'top')
```

The estimated logit from equation 1 for a subject with SPHEQ = 2 is given for

Table 2: Estimated Covariance Matrix of the Estimated Coefficients in Table 1.

	(Intercept)	SPHEQ
(Intercept)	0.04	-0.06
SPHEQ	-0.06	0.18

```
(g2 = test$coefficients[1]+test$coefficients[2]*2)
```

[1] -7.612222

this is,

$$\hat{g}(2) = 0.05 + (-3.83) \times 2 = -7.61. \tag{4}$$

The estimated probability of having myopia at SPHEQ = 2 is given for

```
(exp(g2)/(1+exp(g2)))
```

[1] 0.0004941278

this is,

$$\hat{\pi}(2) = \frac{e^{\hat{g}(2)}}{1 + e^{\hat{g}(2)}} = 0.00049. \tag{5}$$

The estimator of the logit and its confidence interval provide the basis for the logistic probability [1], and its associated confidence interval.

The estimator of the variance of the estimator of the logit for a subject with SPHEQ = 2 is given, from Table 2, in equation 6

$$\widehat{Var}[g(2)] = \widehat{Var}(\hat{\beta}_0) + 2^2 \times \widehat{Var}(\hat{\beta}_1) + 2 \times 2 \times \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1), \tag{6}$$

this is,

```
(test$cov.scaled[1,1]+2^2*test$cov.scaled[2,2]+2*2*test$cov.scaled[1,2])
```

[1] 0.4893668

and the estimator for respective standard error is given for

```
(se = sqrt(test$cov.scaled[1,1]+2^2*test$cov.scaled[2,2]+2*2*test$cov.scaled[1,2]))
```

[1] 0.6995475

this is,

$$\widehat{SE}[\hat{g}(2)] = \sqrt{\widehat{Var}[\hat{g}(2)]} = 0.70 \tag{7}$$

So the end points of a 95 percent confidence interval for the logit is given in equation 8

```
(g22 = g2+c(-1,1)*abs(qnorm(.975,lower.tail=F))*se)
```

[1] -8.983310 -6.241134

this is,

$$-7.61 \pm 1.96 \times (0.70) = (-8.983310, -6.241134) \tag{8}$$

The end points of a 95 percent confidence interval for the logistic probability is given in equation 9

```
(exp(g22)/(1+exp(g22)))
```

[1] 0.0001254711 0.0019438596

this is,

$$\hat{\pi}(2) = \frac{e^{\hat{g}(2) \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}[\hat{g}(2)]}}{1 + e^{e^{\hat{g}(2) \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}[\hat{g}(2)]}}} = (0.0001254711, 0.0019438596). \tag{9}$$

Therefore, we can be 95 % confident that the probability of having myopia at SPHEQ = 2 could be as low as 0.0125% and as high as 0.1944%.

2 Exercise Two

For this exercise, we need the ICU dataset. The data is available by Professor Lemeshow's course on Canvas Network and it has been studied in Week One, Exercise 2, Part (d).

Complete the following:

1. Using the results from Week One, Exercise 2, Part (d), compute 95 percent confidence intervals for the slope and constant term. Write a sentence interpreting the confidence interval for the slope.

2. Obtain the estimated covariance matrix for the model fit from Week One, Exercise 2, Part (d). Then compute the logit and estimated logistic probability for a 60-year old subject. Then compute a 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence or two interpreting the estimated probability and its confidence interval.

Solving:

Codes for ICU Data:

```
filename = "icu.csv"
if (!file.exists(filename)) {
fileURL = "https://learn.canvas.net/courses/1179/files/461760/download?download_frd=1"
        download.file(fileURL, filename) }
data = read.csv("icu.csv", header = T, sep = ",")
```

1. Codes for model:

```
mod = glm(STA ~  AGE, data = data, family = binomial(link = "logit"))
mod1 = xtable(mod,
caption = "Results of Fitting the Logistic Regression Model to the ICU Data, n = 200.", label = "ta
print(mod1, caption.placement = 'top')
```

Table 3: Results of Fitting the Logistic Regression Model to the ICU Data, n = 200.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0585	0.6961	-4.39	0.0000
AGE	0.0275	0.0106	2.61	0.0091

The sentence for 95 percent confidence intervals for the intercept $\hat{\beta_0}$ is given from:

```
test = summary(mod)
(test$coefficients[1]+c(-1,1)*abs(qnorm(.975,lower.tail=F))*test$coefficients[3])
```

[1] -4.422807 -1.694219

for

$$\hat{\beta_0} \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}(\hat{\beta_0}) \tag{10}$$

and we can be 95 % confident that true value of β_0 is between -4.42 and -1.70.

The sentence for 95 percent confidence intervals for the slope $\hat{\beta_1}$ of coefficient AGE is given from

```
(test$coefficients[2]+c(-1,1)*abs(qnorm(.975,lower.tail=F))*test$coefficients[4])
```

[1] 0.00683723 0.04824799

for

$$\hat{\beta_1} \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}(\hat{\beta_1}) \tag{11}$$

and we can be 95 % confident that increase in the log-odds per one-year increase in AGE is between between 0.0068 and -0.048.

2. The covariance matrix for model in Table 3 is given in Table 4.

```
A = xtable(test$cov.scaled,
caption = "Estimated Covariance Matrix of the Estimated Coefficients in Table 3.",
label = "tab04")
print(A, caption.placement = 'top')
```

Table 4: Estimated Covariance Matrix of the Estimated Coefficients in Table 3.

	(Intercept)	AGE
(Intercept)	0.48	-0.01
AGE	-0.01	0.00

The estimated logit for a subject with AGE = 60 is given for

```
(g2 = test$coefficients[1]+test$coefficients[2]*60)
```

[1] -1.405957

The estimated probability of dying in the ICU for a 60 year old subject is given for

```
(exp(g2)/(1+exp(g2)))
```

[1] 0.1968726

The estimator of the variance of the estimator of the logit for this subject is given for

```
(test$cov.scaled[1,1]+60^2*test$cov.scaled[2,2]+2*60*test$cov.scaled[1,2])
```

[1] 0.03393532

and the estimator for respective standard error is given for

```
(se = sqrt(test$cov.scaled[1,1]+60^2*test$cov.scaled[2,2]+2*60*test$cov.scaled[1,2]))
```

[1] 0.1842154

So the end points of a 95 percent confidence interval for the logit is given for

```
g22 = g2+c(-1,1)*abs(qnorm(.975,lower.tail=F))*se
(exp(g22)/(1+exp(g22)))
```

[1] 0.1459143 0.2602054

and we are 95 % confident that this probability of dying in the ICU for a 60 year old subject could be as low as 0.14 and as high as 0.26.

3 Exercise Three

First, using the ICU data (given above), consider the multiple logistic regression model of vital status, STA, on age (AGE), cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and race (RACE).

Then, complete the following:

1. The variable RACE is coded at three levels. Prepare a table showing the coding of the two design variables necessary for including this variable in a logistic regression model.
2. Write down the equation for the logistic regression model of STA on AGE, CAN, CPR, INF, and RACE. Write down the equation for the logit transformation of this logistic regression model. How many parameters does this model contain?
3. Write down an expression for the likelihood and log likelihood for the logistic regression model in part 2. How many likelihood equations are there? Write down an expression for a typical likelihood equation for this problem.

Table 5: Coding of the Design Variables for Race

RACE	RACE2	RACE3
White	0	0
Black	1	0
Other	0	1

4. Using a logistic regression package, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part 2. Using these estimates write down the equation for the fitted values, that is, the estimated logistic probabilities.

5. Using the results of the output from the logistic regression package used in part 4, assess the significance of the slope coefficients for the variables in the model using the likelihood ratio test. What assumptions are needed for the p-values computed for this test to be valid? What is the value of the deviance for the fitted model?

6. Use the Wald statistics to obtain an approximation to the significance of the individual slope coefficients for the variables in the model. Fit a reduced model that eliminates those variables with non-significant Wald statistics. Assess the joint (conditional) significance of the variables excluded from the model. Present the results of fitting the reduced model in a table.

7. Using the results from part 6, compute 95 percent confidence intervals for all coefficients in the model. Write a sentence interpreting the confidence intervals for the non-constant covariates.

Solving:

1. In R we just need be sure that race variable is a factor. However the codes of design variables for RACE is given in Table 5.

```
data$RACE = as.factor(data$RACE)
class(data$RACE)
```

[1] "factor"

2. Equation for the logistic regression model of STA on AGE, CAN, CPR, INF, and RACE.

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}, \tag{12}$$

where

$$\hat{g}(x) = \ln\left(\frac{\hat{\pi}(x)}{1 - \hat{\pi}(x)}\right) \tag{13}$$

$$= \beta_0 + \beta_1 \times (AGE) + \beta_2 \times (CAN) + \beta_3 \times (CPR) + \beta_4 \times (INF) + \beta_5 \times (RACE_2) + \beta_6 \times (RACE_3) \tag{14}$$

is the equation for the logit transformation of this logistic regression model and has 7 parameters.

3. An expression for the likelihood is given for

$$l(\beta) = \prod_{i=1}^{200} \left\{ \pi(x_i)^{y_i} \times [1 - \pi(x_i)]^{1-y_i} \right\} \tag{15}$$

and log likelihood is given for

$$\mathcal{L}(\beta) = ln[l(\beta)] = \sum_{i=1}^{200} \{x_i \times ln[\pi(x_i)] + [1 - x_i] \times ln[1 - \pi(x_i)]\} \tag{16}$$

and there are 7 equations.

4.
```
mod = glm(STA ~  AGE + CAN + CPR + INF + RACE, data = data,
family = binomial(link = "logit"))
mod1 = xtable(mod,
caption = "Results of Fitting the Logistic Regression Model to the ICU Data, n = 200.",
label = "tab06" )
print(mod1, caption.placement = 'top')
```

Table 6: Results of Fitting the Logistic Regression Model to the ICU Data, n = 200.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.5115	0.8144	-4.31	0.0000
AGE	0.0271	0.0116	2.34	0.0193
CAN	0.2445	0.6168	0.40	0.6918
CPR	1.6465	0.6234	2.64	0.0083
INF	0.6807	0.3804	1.79	0.0736
RACE2	-0.9571	1.0845	-0.88	0.3775
RACE3	0.2597	0.8713	0.30	0.7656

The logit transformation:

$$\hat{g}(x) = -3.51 + 0.03 \times (AGE) + 0.24 \times (CAN) + 1.65 \times (CPR) + 0.68 \times (INF) - 0.96 \times (RACE_2) + 0.26 \times (RACE_3) \tag{17}$$

Equation for the fitted values, that is, the estimated logistic probabilities:

$$\hat{\pi}(x) = \frac{e^{-3.512+0.027\times(AGE)+0.244\times(CAN)+1.646\times(CPR)+0.681\times(INF)+-0.957\times(RACE_2)+0.260\times(RACE_3)}}{1 + e^{-3.512+0.027\times(AGE)+0.244\times(CAN)+1.646\times(CPR)+0.681\times(INF)+-0.957\times(RACE_2)+0.260\times(RACE_3)}} \tag{18}$$

5. The hypothesis for Likelihood Ratio Test (LRT) are

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \tag{19}$$

or

$$H_1 : \exists j \in \{1, \ldots, 6\} : \beta_j \neq 0 \tag{20}$$

and the statistic for LRT is given for:

$$G = D(\text{model without the variable}) - D(\text{model with the variable}), \tag{21}$$

where D (Deviance) is given for

$$D = -2 \times log\left(\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right) \tag{22}$$

With results from code of the function *glm* we can do

$$D = [\text{Null deviance} - \text{Residual deviance}] \tag{23}$$

```
A=summary(mod)
A$null.deviance
```

[1] 200.161

```
A$deviance
```

[1] 179.3007

```
(G = A$null.deviance - A$deviance)
```

[1] 20.86024

```
(p = pchisq(G, 6, lower.tail = F))
```

[1] 0.001943745

As the probability of finding such G assuming H_0 as true is very small ($p < 0.05$) then we reject the null hypotesis and we conclude that at least one coefficient is not null.

6. The results from Table 7 suggest that variables CAN, INF an RACE are probably not significant. So we will use the Likelihood Ratio Test (LRT) to check permanence of these variables in the model.

```
mod0 = glm(STA ~  AGE + CPR, data = data,
family = binomial(link = "logit"))
A = xtable(anova(mod0,mod, test = "LRT"),
caption = "Results Likelihood Ratio Test (LRT) for permanence of CAN, INF an RACE",
label = "tab07")
print(A, caption.placement = 'top')
```

Table 7: Results Likelihood Ratio Test (LRT) for permanence of CAN, INF an RACE

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	197	183.95			
2	193	179.30	4	4.65	0.3249

For statistical reasons we can reduce the model to two variables AGE and CPR. However, since $p < 0.1$ to INF, other motivations to include it in the model should be evaluated.

7.
```
mod0 = glm(STA ~  AGE + CPR, data = data,
family = binomial(link = "logit"))
A = xtable(mod0,
caption = "Results of Fitting the Logistic Regression Model to the ICU Data, n = 200.", label = "ta
print(A, caption.placement = 'top')
```

Table 8: Results of Fitting the Logistic Regression Model to the ICU Data, n = 200.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.3520	0.7455	-4.50	0.0000
AGE	0.0296	0.0111	2.66	0.0079
CPR	1.7841	0.6073	2.94	0.0033

For constant:

```
A=summary(mod0)
(A$coefficients[1]+c(-1,1)*abs(qnorm(.975,lower.tail=F))*A$coefficients[4])
```

[1] -4.813108 -1.890803

For AGE:

```
A=summary(mod0)
(A$coefficients[2]+c(-1,1)*abs(qnorm(.975,lower.tail=F))*A$coefficients[5])
```

[1] 0.007755898 0.051458923

and we can be 95 % confident that increase in the log-odds per one-year increase in AGE is between between 0.0078 and 0.0514.

For CPR:

```
(A$coefficients[3]+c(-1,1)*abs(qnorm(.975,lower.tail=F))*A$coefficients[6])
```

[1] 0.5938116 2.9743726

and we can be 95 % confident that increase in the log-odds to persons who had CPR prior to admission compared with those who had not is between between 0.5938 and 2.9744.

References

[1] David W Hosmer, Stanley Lemeshow, and Rodney Sturdivant. *Applied Logistic regression*. 2013. 3

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. 1