Homework 10 - SOLUTION Multiple Logistic Regression Stats 230 - Titanic Survival & Leukemia

Stat 230 - Homework #11 - Not Collected - Solution

PROBLEMS ASSIGNED: #10.3, #10.6, #10.10, #10.23, #10.24, #10.28

A solution will be posted to aid your studying for the final exam.

For both 10.23 and 10.24, note that randomness and independence are not met as conditions, as this is Titanic data. However, if we are NOT interested in generalizing our results, and just want to find out what happened in regards to the actual Titanic incident, you can still assess that with "inference" procedures, although the issues with independence are worrisome for the probability model.

Exercise 10.3 - Provide the model SOLUTION: The suggested logistic regression model should have an intercept, main effect for x, and main effect for Group. No interaction is necessary because the plot shows roughly parallel lines (if we drew them in).

Exercise 10.6 - Provide the model SOLUTION: The suggested logistic regression model should have an intercept, main effect for x, main effect for Group, and an interaction between x and Group. The interaction is necessary because the plot does not show parallel lines.

Exercise 10.10 - Provide the model SOLUTION: The suggested logistic regression model should have an intercept, main effect for x, main effect for A, and an interaction between x and A. An interaction between x and A is necessary because the difference between response groups in terms of x values at the different levels of A differs. In one case, when A = Low, Y=0 has lower values of X than Y=1, but this reverses when A = High.

```
data("Titanic")
```

Exercise 10.23

part a: Fit a logistic regression with both age and sexcode in the model. Provide the logit and probability forms of the model.

SOLUTION:

```
logm <- glm(Survived ~ Age + SexCode, data = Titanic, family = binomial(logit))
msummary(logm)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.15984    0.21965   -5.28  1.3e-07 ***
## Age         -0.00635    0.00619   -1.03      0.3
## SexCode      2.46600    0.17846   13.82  < 2e-16 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
```

```
## Residual deviance:  795.59  on 753  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 801.6
##
## Number of Fisher Scoring iterations: 4
```

The logit form of the model is: predicted logit(P(Survived)) = -1.16 -0.00635(Age) + 2.466(SexCode).

The probability form of the model would be:

predicted P(Survived) = exp(-1.16 -0.00635(Age) + 2.466(SexCode))/(1 + exp(-1.16 -0.00635(Age) + 2.466(SexCode)))

part b: Comment on the effectiveness of each predictor in the model.

SOLUTION: We can assess via the Wald z-statistics whether each predictor is necessary in the model GIVEN the presence of the other predictor. It appears that SexCode IS useful given Age is in the model (very small p-value), but that Age is NOT useful given SexCode is in the model (p-value of 0.3).

(This could indicate you could refit the model without Age, but here they want to account for it so they leave it in.)

part c: Estimate the probability and odds that an 18 year old man would survive.

SOLUTION:

From the help menu, we see that SexCode was coded as 1=female or 0=male. So, we can find the logit and then work back to odds and probability by plugging in.

```
resp <- -1.16 -0.00635*(18) + 2.466*(0); resp
```

```
## [1] -1.2743
```

```
exp(resp)
```

```
## [1] 0.279627
```

```
prob <- exp(resp) / (1 + exp(resp)); prob
```

```
## [1] 0.218522
```

The odds of survival for the male at age 18 is 0.28, and the associated probability is 0.219.

part d: Repeat c for an 18 year old woman, AND find the odds ratio compared to a man of the same age.

SOLUTION:

```
resp2 <- -1.16 -0.00635*(18) + 2.466*(1); resp2
```

```
## [1] 1.1917
```

```
exp(resp2)
```

```
## [1] 3.29267
```

```
prob2 <- exp(resp2) / (1 + exp(resp2)); prob2
```

```
## [1] 0.767045
```

The odds of survival for the female at age 18 is 3.29, and the associated probability is 0.767. Next, we compute the odds ratio female (numerator) to male (denominator).

```
3.29/0.28
```

```
## [1] 11.75
```

The odds ratio for female vs. male at 18 years of age (female/male) is 11.75.

part e: Redo c and d for a man and woman of age 50.

SOLUTION:

```
resp<- -1.16 -0.00635*(50) + 2.466*(0); resp
```

```
## [1] -1.4775
```

```
exp(resp)
```

```
## [1] 0.228207
```

```
prob <- exp(resp) / (1 + exp(resp)); prob
```

```
## [1] 0.185805
```

```
resp2<- -1.16 -0.00635*(50) + 2.466*(1); resp2
```

```
## [1] 0.9885
```

```
exp(resp2)
```

```
## [1] 2.6872
```

```
prob2<- exp(resp2) / (1 + exp(resp2)); prob2
```

```
## [1] 0.728792
```

The odds of survival for the male at age 50 is 0.228, and the associated probability is 0.186. The odds of survival for the female at age 50 is 2.69, and the associated probability is 0.729.

```
2.69/0.228
```

```
## [1] 11.7982
```

The odds ratio (female/male) at 50 years of age is 11.8.

part f: Compare the odds ratios from d and e. What happens? Will this always be the case?

SOLUTION: The odds ratios are identical, up to round off errors. This will happen at every value of age, as this is a property of logistic regression models. The probabilities will change but the odds ratios are constant.

Exercise 10.24 Add an interaction term to the model from 10.23

part a: Explain how the coefficients related to separate models for males and females. It may be easiest to just provide the two different logit models that would result.

SOLUTION:

```
logm2 <- glm(Survived ~ Age * SexCode, data = Titanic, family = binomial(logit))
msummary(logm2)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.29875    0.27770   -1.08     0.28
## Age         -0.03637    0.00926   -3.93  8.6e-05 ***
## SexCode      0.59986    0.40805    1.47     0.14
## Age:SexCode  0.06572    0.01369    4.80  1.6e-06 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
```

```
## Residual deviance:  770.56  on 752  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 778.6
##
## Number of Fisher Scoring iterations: 4
```

The coefficients can be described in many ways, but it may just be easiest to give the two different logit models - one for males and one for females.

For males, predicted logit(P(Survived)) = -0.29875 - 0.03637(Age)
For females, predicted logit(P(Survived))= -0.29875 + 0.59986 - 0.03637(Age) + 0.06572(Age)
which when appropriate terms are combined yields:
predicted logit(P(Survived)) = 0.301 + 0.0294(Age)

The intercept and slope coefficient of Age have different signs - both negative for male and both positive for female.

> part b: Use a nested LRT to determine if adding Age and the interaction with Sex improves over a model using just SexCode.

SOLUTION:

```
logm3 <- glm(Survived ~ SexCode, data = Titanic, family = binomial(logit))
msummary(logm3)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6080     0.0919   -17.5   <2e-16 ***
## SexCode       2.3012     0.1349    17.1   <2e-16 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1688.1  on 1312  degrees of freedom
## Residual deviance: 1355.5  on 1311  degrees of freedom
## AIC: 1360
##
## Number of Fisher Scoring iterations: 4
```

To perform a nested drop-in deviance test, we need the reduced model specified. We already have the full model. When we fit the reduced model, a quick inspection shows that this model has MANY more observations than logm2 (our full model), because a number of Age values were not known for logm2. In order to make a proper comparison, we need to take that into account, and fit the model with just SexCode on the SAME data points as logm2, in order for the residual deviances to be comparable. This adjustment is important - it is NECESSARY to have comparable quantities for the drop-in-deviance test to work correctly. It can be done several ways, including using augment on logm2, or by filtering out missing values for Age.

```
Titanic2 <- Titanic[which(Titanic$Age != "NA"), ] #filter can also be used
TitanicAug <- augment(logm2)
```

The first command takes out the observations with Age missing and makes a new data set Titanic2 that could be used. TitanicAug also has those points removed, as it is the augmented data set from the full model.

```
logm2 <- glm(Survived ~ Age*SexCode, data = TitanicAug, family = binomial(logit))
msummary(logm2)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.29875    0.27770   -1.08     0.28
## Age         -0.03637    0.00926   -3.93  8.6e-05 ***
```

```
## SexCode      0.59986    0.40805    1.47     0.14
## Age:SexCode  0.06572    0.01369    4.80  1.6e-06 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  770.56  on 752  degrees of freedom
## AIC: 778.6
##
## Number of Fisher Scoring iterations: 4
```

```
logm3 <- glm(Survived ~ SexCode, data = TitanicAug, family = binomial(logit))
msummary(logm3)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.355      0.114   -11.8   <2e-16 ***
## SexCode        2.472      0.178    13.9   <2e-16 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  796.64  on 754  degrees of freedom
## AIC: 800.6
##
## Number of Fisher Scoring iterations: 4
```

We see that the logm2 output doesn't change when run on TitanicAug (or Titanic2), because the missing observations were removed. We now have comparable output in logm3 to compare to logm2. We pull out the residual deviances from the output and perform the LRT to show the details, as requested.

```
G <- 796.64 - 770.56; G
```

```
## [1] 26.08
```

```
dfdiff <- 754 - 752
pchisq(G, dfdiff, lower.tail = FALSE)
```

```
## [1] 2.1717e-06
```

Our LRT test statistic is 26.08, and our p-value is found using a chi-square distribution with 2 df. The associated p-value is 2.17 times 10^(-6), which is tiny. This is a very small p-value and provides strong evidence that at least one of the terms involving Age (either the main effect or the interaction) is important in this model.

This can be done using the anova command as well:

```
anova(logm3, logm2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ SexCode
## Model 2: Survived ~ Age * SexCode
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       754      796.6
## 2       752      770.6  2    26.09 2.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data("Leukemia")
```

Exercise 10.28

part a: Which predictors appear to be useful?

SOLUTION:

To answer this, we must fit the specified model in the description. Then we can look at the model summary.

```
logleuk <- glm(Resp ~ Age + Smear + Infil + Index + Blasts + Temp, data = Leukemia, family = binomial(l
msummary(logleuk)
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 108.33115   41.84379    2.59   0.0096 **
## Age          -0.06231    0.02746   -2.27   0.0233 *
## Smear        -0.00469    0.04005   -0.12   0.9068
## Infil         0.03104    0.03789    0.82   0.4126
## Index         0.37281    0.13247    2.81   0.0049 **
## Blasts        0.03267    0.04605    0.71   0.4780
## Temp         -0.11162    0.04263   -2.62   0.0088 **
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 39.275  on 44  degrees of freedom
## AIC: 53.28
##
## Number of Fisher Scoring iterations: 6
```

In the model summary, we identify 3 predictors that appear to be useful to the model in the presence of the other predictors (i.e., using their Z tests and associated p-values). From the model summary, at alpha = 0.05, Age, Index, and Temp have significant p-values.

part b: Interpret relationship between Age and Resp and then between Temp and Resp.

SOLUTION:

```
exp(-0.06231) #Age
```

```
## [1] 0.939592
```

```
exp(-0.11162) #Temp
```

```
## [1] 0.894384
```

Be sure your interpretations reference the other predictors in the model being accounted for.

Age is an effective predictor in this model (p-value = 0.023) and has a negative coefficient, so the probability of responding to treatment appears to decrease for older patients, accounting for the effects of the other terms. According to the odds ratio, the odds of responding go down by a factor of about 0.94 for every extra year of age (decrease of 6 percent), after accounting for the other variables in the model.

Temp is also an effective term in the model (p-value =0.009) with a negative coefficient. Higher temperatures appear to be associated with lower probabilities of responding, with the odds going down by about a factor of 0.89 for every extra tenth of a degree, after accounting for the other variables in the model.

part c: Is it reasonable to include insignificant predictors in a final model?

SOLUTION:

Yes. Book solution: Yes, a predictor that is "insignificant" in one model might actually be important to include in a final model. For example, we might have two predictors that are strongly related to each other and the response. When both are in the model, their individual tests may show large P-values since each is not needed if the other is in the model. However, dropping one might cause the P-value for the other to decrease dramatically when the similar predictor is no longer in the model, making it important to keep at least one of the two predictors in the final model.

Generally speaking, you can definitely encounter situations where you want to account for a variable's effects but not have a real interest in the size of effect or significance, but need to demonstrate it was accounted for.

> part d: Compare model with 6 preds to model with the 3 significant preds. Comment on coefficient stability.

SOLUTION:

```
logleuk2 <- glm(Resp ~ Age + Index + Temp, data = Leukemia, family = binomial(logit))
msummary(logleuk2)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  87.3880    35.4582    2.46   0.0137 *
## Age          -0.0585     0.0256   -2.29   0.0222 *
## Index         0.3849     0.1215    3.17   0.0015 **
## Temp         -0.0890     0.0361   -2.47   0.0136 *
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 43.265  on 47  degrees of freedom
## AIC: 51.27
##
## Number of Fisher Scoring iterations: 6
```

To do the test by hand, we need residual deviances from each model to create G.

```
G <- 43.265 - 39.375; G
```

```
## [1] 3.89
```

```
df <- 47 - 44; df
```

```
## [1] 3
```

```
pchisq(G, df, lower.tail = FALSE)
```

```
## [1] 0.27359
```

We can also get this with the ANOVA command.

```
anova(logleuk2, logleuk, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Resp ~ Age + Index + Temp
## Model 2: Resp ~ Age + Smear + Infil + Index + Blasts + Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        47      43.27
## 2        44      39.28  3     3.99    0.263
```

There are minor differences due to rounding. In either case, the p-value for the nested drop-in deviance test is large. This suggests we can use the reduced model - our evidence does not support that we need to add at least one of Smear, Infil, or Blasts to the model.

Now we check stability of the coefficients: full versus reduced. For Age, we have coefficients of -0.06231 versus -0.0585. For Index, we have coefficients of 0.37281 versus 0.3849. For Temp, we have coefficients of -0.11162 versus -0.089.

In all three cases, the coefficients are similar. Their signs match and their values are fairly similar. For Age and Index the coefs differ by less than 0.01, while for Temp it is closer to 0.03, still small.

part e: Compare coefs for Age and Temp to models where they are single predictors.

SOLUTION:

You will need to fit the models in 9.41 to answer this part.

```
logAge <- glm(Resp ~ Age, data = Leukemia, family = binomial(logit))
msummary(logAge)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1968     1.0055    2.18    0.029 *
## Age          -0.0468     0.0195   -2.39    0.017 *
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 64.004  on 49  degrees of freedom
## AIC: 68
##
## Number of Fisher Scoring iterations: 4
```

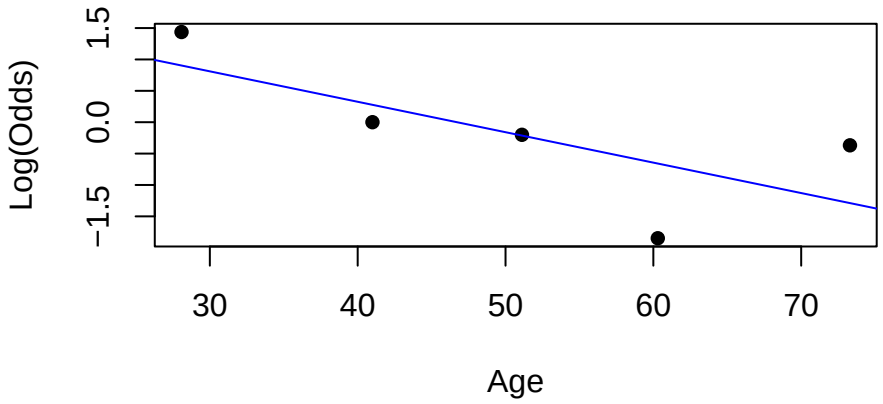
```
logTemp <- glm(Resp ~ Temp, data = Leukemia, family = binomial(logit))
msummary(logTemp)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  38.5507    21.2364    1.82    0.069 .
## Temp         -0.0388     0.0214   -1.82    0.069 .
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 66.766  on 49  degrees of freedom
## AIC: 70.77
##
## Number of Fisher Scoring iterations: 4
```

When Age is by itself, the coefficient is -0.0468 and it is significant. When Temp is by itself, the coefficient is -0.0388, and it is significant at alpha = 0.10, but not at 0.05. The Age coefficient and significance more closely matches the full and reduced models. Temps' coefficient value is more different from the full and reduced models and the significance is different.

We should bear in mind for all of this that we didn't check conditions at any point here, and REALLY should be doing that before looking at interpretations, etc.

```
emplogitplot1(Resp ~ Age, data = Leukemia, ngroups = 5)
```




```
emplogitplot1(Resp ~ Temp, data = Leukemia, ngroups = 5)
```