STA 138 Project II

Lam Vu 999119688

March 9, 2017

Problem 1: Logistic Regression

Introduction:

The goal in this analysis of the Car dataset is to determine the relationship between the variable Y (Takes on the value 1 if the customer bought a new car, and 0 otherwise.) and variable X (The age of the customer's current car in months). We examined the dependency of the variables and described any dependency found.

Summary:

Since there is not a steep slope, it does not appear that the age of customer’s current car in months has a significant effect on whether the customer purchases a new car. The estimated regression function we obtained is:logit(pi-hat) = -2.65826 + (0.07635)x. From further test we found there is an effect of car age on purchases of new cars based the Wald Test and likelihood ratio test.

Analysis:

The estimated regression function is: logit(pi-hat) = -2.65826 + (0.07635)x

```
##
## Call:  glm(formula = Y ~ X, family = binomial(logit), data = Car)
##
## Coefficients:
## (Intercept)          X
##    -2.65826     0.07635
##
## Degrees of Freedom: 32 Total (i.e. Null);  31 Residual
## Null Deviance:       45.72
## Residual Deviance: 33.92     AIC: 37.92
```

---

exp(B1):

```
## [1] 1.07934
```

The value of X(age of car) when the probability of purchasing a new car is 0.50

```
## [1] 34.81676
```

The Wald Test Statistic and its corresponding p-value

```
## [1] 2.640003
```

```
## [1] 0.00829053
```

The Likelihood Ratio Test Statistic and its corresponding p-value

```
## [1] 11.79892
```

```
## [1] 0.00059265
```

95% Likelihood Ratio Confidence Interval of exp(B1)

```
## [1] 1.0288 1.1562
```

Interpretation:

The estimated regression function we obtained is:logit(pi-hat) = -2.65826 + (0.07635)x

We observed B1-hat > 0, which suggests the probability of the customers purchasing a new car increases as age of the car does. The odds of purchasing a new car is 1.07934 times of what they were. Since our dataset is of the range 10 to 92 months and a car can't be of age 0, there is no practical interpretation of alpha. The value of X (age of car) when the probability of purchasing a new car is 0.50 is 34.81676 months. The result from the Wald Test statistic is 2.640003 and the likelihood ratio test statistic is 11.79892. We obtained the p-values of 0.00829053 and 0.00059265 respectively. In other words if in reality car age does not affect purchases of new cars then the probability we would observe our data and more extreme is 0.00829053% or 0.00059265%. From these p-values we can reject H0:b1=0 and conclude b1!=0, there is an effect of car age on purchases of new car. From the 95% Likelihood Ratio Confidence Interval of exp(B1) we do not find the interval to include 1 which meant B1 has an effect. In other words we are 95% confident that when age of car increases by 1 month, the odds of purchasing a new car are between 1.0288 and 1.1562 times what they were.

Conclusion:

I conclude that there is an effect of car age on purchases of new cars based on the Wald Test and likelihood ratio test which produced p-values smaller than most conventional alphas= (0.1,0.05,0.01) thus we would reject H0:b1=0 and conclude b1!=0, there is an effect of car age on purchases of new car. From the 95% Likelihood Ratio Confidence Interval of exp(B1) we do not find the interval to include 1 which meant B1 has an effect. In other words we are 95% confident that when age of car increases by 1 month, the odds of purchasing a new car are between 1.0288 and 1.1562 times what they were.

STA 138 Project II

Lam Vu 999119688

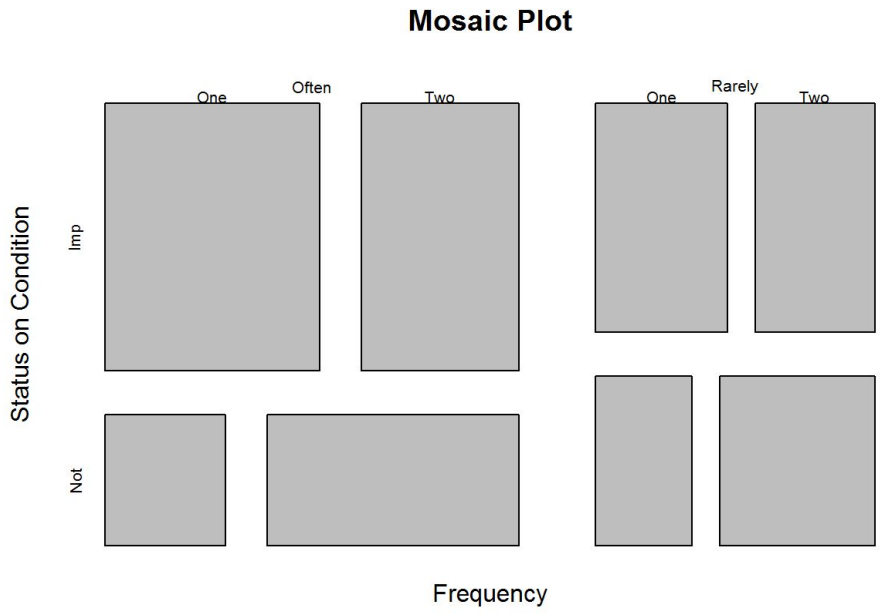
March 9, 2017

Problem 2: Log-Linear Models

Introduction:

The goal in this analysis of the TrailsShort dataset is to determine the relationship between the variable X (The frequency of exercise), variable Y (The status of their condition) and variable Z (What type of drug they were on). We examined the dependencies of the variables and described any dependencies found.

Summary:



From the mosaic plot above we can observe the distribution of status on condition between treatment group and frequency of exercise. We can see that overall more people are having improvement in their status. There is also greater improvement on status in those who exercise often. Those who often exercise in group One seems to be the biggest group of people who improved. We fitted the model ln(μijk)= $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$ which is the best fitting model and it contains two interaction terms. We then obtained the Wald and Likelihood Ratio Confidence Intervals at 5% significance level and found no interaction terms with intervals including 0, this meant that all interaction terms are dependent.

Analysis:

Model Selection:

##	Log-Li	LR	Pearson	df	p-val:LR	p-val:Pear
## F~X+Y+Z	-37.4907	27.5388	27.1055	4	0.0000	0.0000
## F~X+Y+Z+Y*Z	-26.9962	6.5498	6.5710	3	0.0877	0.0869
## F~X+Y+Z+X*Z	-37.2740	27.1055	26.1329	3	0.0000	0.0000
## F~X+Y+Z+X*Y	-34.9878	22.5329	22.2182	3	0.0001	0.0001
## F~X+Y+Z+X*Y+X*Z	-34.7711	22.0996	21.7804	2	0.0000	0.0000
## F~X+Y+Z+X*Y+Y*Z	-24.4933	1.5439	1.5459	2	0.4621	0.4616
## F~X+Y+Z+X*Z+Y*Z	-26.7795	6.1164	6.2010	2	0.0470	0.0450
## F~X+Y+Z+X*Y+X*Z+Y*Z	-24.4701	1.4976	1.4995	1	0.2210	0.2207
## F~X+Y+Z+X*Y+X*Z+Y*Z+X*Y*Z	-23.7213	0.0000	0.0000	0	1.0000	1.0000

##	AIC	BIC
## F~X+Y+Z	82.9814	83.2992
## F~X+Y+Z+Y*Z	63.9924	64.3896
## F~X+Y+Z+X*Z	84.5481	84.9453
## F~X+Y+Z+X*Y	79.9755	80.3727
## F~X+Y+Z+X*Y+X*Z	81.5422	82.0188
## F~X+Y+Z+X*Y+Y*Z	60.9865	61.4632
## F~X+Y+Z+X*Z+Y*Z	65.5590	66.0357
## F~X+Y+Z+X*Y+X*Z+Y*Z	62.9402	63.4963
## F~X+Y+Z+X*Y+X*Z+Y*Z+X*Y*Z	63.4426	64.0781

Best Fit Model:

```
##
## Call:  glm(formula = the.model, family = poisson, data = Trial.S)
##
```

```
## Coefficients:
##  (Intercept)      XRarely        YNot        ZTwo        XRarely:YNot
##       4.7540      -0.5500      -1.1766      -0.2328       0.4137
##    YNot:ZTwo
##       0.8483
##
## Degrees of Freedom: 7 Total (i.e. Null);  2 Residual
## Null Deviance:       84.61
## Residual Deviance: 1.544     AIC: 60.99
```

Test Statistics , P-value, and information criterion

```
##     Log-Li        LR          Pearson        df       p-val:LR    p-val:Pear
##   -24.4933     1.5439     1.5459     2.0000     0.4621     0.4616
##        AIC        BIC
##    60.9865    61.4632
```

Odds Ratios:

```
## [1] 1.512418

## [1] 2.33577
```

95% Wald Confidence Intervals:

```
##                    lower.bound         upper.bound
## (Intercept)       4.58743225          4.92058900
## XRarely          -0.77472528         -0.32536740
## YNot             -1.49698124         -0.85625600
## ZTwo             -0.45066143         -0.01484339
## XRarely:YNot      0.05135062          0.77606917
## YNot:ZTwo         0.47983260          1.21685011
```

95% Likelihood Ratio Confidence Intervals:

```
##                    2.5 %               97.5 %
## (Intercept)       4.58348356          4.91678300
```

## XRarely	-0.77723217	-0.32738782
## YNot	-1.50334269	-0.86179837
## ZTwo	-0.45179857	-0.01554507
## XRarely:YNot	0.05134698	0.77667452
## YNot:ZTwo	0.48266574	1.22039832

Interpretation:

From the model selection process I chose the model with the lowest AIC to be the best fitted model because compared to BIC, AIC is less conservative and will tend to give a model with more predictors which in this case will be helpful in making predictions.

The model chosen was ln(μijk)= $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$. This model produced Pearson Test Statistic and Likelihood Ratio Test Statistic of 1.5459 and 1.5439 respectively when testing whether the current model fits well. The tests produced p-values of 0.4616 and 0.4621 respectively which will fail to reject H0 and conclude the current model fits well. The estimated odds ratio of 1.512418 and 2.33577 were obtained from the interaction terms. This meant that the odds of no improvement if you rarely exercise is 1.512418 times the odds of no improvement if you exercise often. The odds of no improvement if you are in group two is 2.33577 times the odds of no improvement if you are in group one. From both Wald and Likelihood Ratio Confidence Intervals at 5% significance level we found that there are no interaction terms with intervals including 0, this meant that all interaction terms are dependent. X(frequency of exercise) and Y(status of condition) have a dependent relationship. Also Y(status of condition) and Z(Type of drug) too are dependent.

Conclusion:

I conclude that the chosen model ln(μijk)= $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$ fits well. From the model we have two interaction terms. We obtained the Wald and Likelihood Ratio Confidence Intervals at 5% significance level and found no interaction terms with intervals including 0, this meant that all interaction terms are dependent. X(frequency of exercise) and Y(status of condition) are have dependent relationship. Also Y(status of condition) and Z(Type of drug) too are dependent.