Lab 4: Logistic Regression

put the date, 2017

### Names: Daniel Meza, Adolfo Huerta

##### What each person contributed to the lab:

**Please make sure to show all R code and output after each question so that I can see your work.** Write a sentence for each numerical value produced describing its meaning in context with the proper units.

## Problem 1

An additive to interior paint has been recently developed that may greatly increase the ability of the paint to resist staining. An investigation was conducted to determine whether the additive is safe when exposed to children. Various amounts of the additive were fed to test animals and the number of animals developing liver tumors was recorded in the file paint.csv.

### 1.a.

State the statistical analysis research question for your analysis.

Is the additive to interior paint safe when exposed to children?

### 1.b.

Name the response variable of interest below and state the type (quantitative, qualitative)

The response variable is the animals with tumors and it is qualitative.

### 1.c.

Name the explanatory variables of interest below and state the types (quantitative, qualitative)

The explanatory variables of interest are ppm which is quantitative and the number of animals in the study which is quantitative.

### 1.d.

Determine the coefficients for each variable in the model and provide an interpretation of each parameter estimate in the context of the problem. Interpret the intercept using the Probability Model. Does the intercept make sense practically? Interpret the slope using both the Odds and the Probability Model.

#Provide R code and output/graphs   
  
library(readr)  
paint <- read\_csv("~/Desktop/ucsc/paint.csv")

## Parsed with column specification:  
## cols(  
## ppm = col\_integer(),  
## NumAnimals = col\_integer(),  
## AnimalswTumors = col\_integer()  
## )

names(paint)

## [1] "ppm" "NumAnimals" "AnimalswTumors"

attach(paint)  
  
ppm<-c(rep(0,0), rep(0,30),   
 rep(10,2), rep(10,18),   
 rep(25,2), rep(25,18),   
 rep(50,7), rep(50,27),   
 rep(100,25), rep(100,5),  
 rep(200,30), rep(200,0)  
)  
length(ppm)

## [1] 164

tumor<-c(rep(1,0), rep(0,30),   
 rep(1,2), rep(0,18),   
 rep(1,2), rep(0,18),   
 rep(1,7), rep(0,27),   
 rep(1,25), rep(0,5),  
 rep(1,30), rep(0,0)   
)  
length(tumor)

## [1] 164

indTumors <- data.frame(ppm, tumor)  
write.csv(indTumors, "/home/math/Documents/paint\_me.csv")  
#   
# write.csv(indTumors, "C:/Users/danie/Documents/R/paint\_me.csv")  
View(indTumors)  
  
length(indTumors$ppm)

## [1] 164

tumor1<-glm(formula = tumor~ppm,family = binomial())  
  
summary(tumor1)

##   
## Call:  
## glm(formula = tumor ~ ppm, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.85218 -0.40952 -0.21516 0.04746 2.55579   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.754260 0.562704 -6.672 2.53e-11 \*\*\*  
## ppm 0.052712 0.008325 6.332 2.42e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 221.068 on 163 degrees of freedom  
## Residual deviance: 90.977 on 162 degrees of freedom  
## AIC: 94.977  
##   
## Number of Fisher Scoring iterations: 6

exp(-3.754260)

## [1] 0.02341777

exp(0.052712)

## [1] 1.054126

exp(-3.754260)/(1+exp(-3.754260))

## [1] 0.02288193

exp(0.052712)/(1+exp(0.052712))

## [1] 0.5131749

When the additive is zero ppm the odds that an animal develops a liver tumor is .02342. When the ppm increases to the next level tested ppm the odds of an animal developing a tumor is 1.054 times the odds of an animla getting a tumor at ppm level x.

When the ppm level is zero, the probability that an animal develops a tumor is .0229.

The interpretation of the intercept is practical becuase there is very little chance that an animal develops a tumor when no additive is added.

Since the value of the slope is possitve(.513), then the probability of an animal developing a tumor increases as ppm increases.

### 1.e.

State the hypotheses (null and alternative) for the Logistic Regression Analysis mathematically and in a complete sentence. Here are some symbols you may find useful to adapt:

The probability of an animal developing a tumor is not associated with additive ppm.

The probability of an animal developing a tumor is associated with additive ppm.

### 1.f.

Conduct the appropriate analysis using the Wald Z-test and provide a written interpretation of the results in the context of the problem. Include your vital statistics (test statistic, significance level, and p-value) appropriately reported.

#Provide R code and output/graphs  
summary(tumor1)

##   
## Call:  
## glm(formula = tumor ~ ppm, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.85218 -0.40952 -0.21516 0.04746 2.55579   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.754260 0.562704 -6.672 2.53e-11 \*\*\*  
## ppm 0.052712 0.008325 6.332 2.42e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 221.068 on 163 degrees of freedom  
## Residual deviance: 90.977 on 162 degrees of freedom  
## AIC: 94.977  
##   
## Number of Fisher Scoring iterations: 6

(z=6.332,p-value=2.42e-10,sig level=.05) Based on our sample using a significance level of .05, we will reject the null hypothesis that the probability that an anial develops a tumor is not associated with ppm level and support the alternative that the probability of an animal developing a tumor is associated with additive ppm level.

### 1.g.

Conduct the appropriate analysis using the Likelihood Ratio Test and provide a written interpretation of the results in the context of the problem. Include your vital statistics (test statistic, degrees of freedom, significance level, and p-value) appropriately reported.

#Provide R code and output/graphs  
G= 221.068-90.977  
1-pchisq(G,1)

## [1] 0

anova(tumor1,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: tumor  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 163 221.068   
## ppm 1 130.09 162 90.977 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(G=130.09, p-value< 2.2e-16,df=1,sig level=.05) Based on our sample using a significance level of .05, we will reject the null hypothesis and support the alternative that there is significant evidence to suggest that ppm level is a significant predictor of an animal developing a tumor.

### 1.h.

Do your conclusions to the hypothesis tests in parts g and f agree or disagree? Are the results identical? Why or why not?

The conclusions in both parts g and f agree since the p-value in both was less than the significance level. The p-values are almost the same with justa small difference. The difference can be due to the fact that the methods use different distributions to caluculate the test statistics and p-value. The first method uses the wald z test while the second method uses the a comparison between the full model and the reduced model. The different distributions are normal distribution and the chisquare distribution.

## Problem 2

What happens when someone who is under-aged needs a ride home after a night of drinking? The concern of the media today focuses on those who drive home in this situation. However, what about those who ride with someone who has been drinking? We will examine what factors indicate a risk for riding home with a drinking driver (ride.alc.driver). We will examine risk factors such as gender (female; female=1, 0=male), grade (grade; factor 9, 10, 11, 12), age (age4), ever tried to smoke (smoke; 1=yes, 0=no) (we will ignore driver's license).

### 2.a.

State the statistical analysis research question for your analysis.

What factors indicate a risk for riding home with a drinking driver.

### 2.b.

Name the response variable of interest below and state the type (quantitative, qualitative)

The response variable is riding home with a drinking driver, which is the variable ride.alc.driver in our dataset and it is qualitative.

### 2.c.

Name the explanatory variables of interest below and state the types (quantitative, qualitative)

The explanatory variables are female which is qualitative, grade which is also qualitative since there are only four options (i.e 9,10,11,12). The other explanatory variables are age4 which is quantitative since it is numeric, but can also be seen as qualitative since the only options range from 14 to 18 and finally smoke which is binary so it is qualitative.

### 2.d.

Researchers decide to consider separate models for age and grade.

Model 1:

Model 2:

1. What would motivate that decision?
2. Which model would you choose - the one including age or the one including grade? Justify your selection using relevant statistics from the model outputs.

#Provide R code and output/graphs  
  
library(readr)  
YouthRisk2007 <- read\_csv("~/Desktop/ucsc/YouthRisk2007.csv")

## Parsed with column specification:  
## cols(  
## ride.alc.driver = col\_integer(),  
## female = col\_integer(),  
## grade = col\_integer(),  
## age4 = col\_integer(),  
## smoke = col\_integer(),  
## DriverLicense = col\_integer()  
## )

names(YouthRisk2007)

## [1] "ride.alc.driver" "female" "grade" "age4"   
## [5] "smoke" "DriverLicense"

attach(YouthRisk2007)  
  
grade\_rdd <- glm(formula = ride.alc.driver ~ grade, family = binomial())  
summary(grade\_rdd)

##   
## Call:  
## glm(formula = ride.alc.driver ~ grade, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8911 -0.8729 -0.8549 1.4937 1.5612   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.31729 0.17759 -7.417 1.19e-13 \*\*\*  
## grade 0.04990 0.01675 2.979 0.00289 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16531 on 13319 degrees of freedom  
## Residual deviance: 16522 on 13318 degrees of freedom  
## (67 observations deleted due to missingness)  
## AIC: 16526  
##   
## Number of Fisher Scoring iterations: 4

G\_grade <- grade\_rdd$null.deviance - grade\_rdd$deviance  
p\_val\_grade <- 1 - pchisq(G\_grade, 1)  
anova(grade\_rdd, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: ride.alc.driver  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 13319 16531   
## grade 1 8.8821 13318 16522 0.00288 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

age\_rdd <- glm(formula = ride.alc.driver ~ age4, family = binomial())  
summary(age\_rdd)

##   
## Call:  
## glm(formula = ride.alc.driver ~ age4, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9224 -0.8905 -0.8287 1.4946 1.6108   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.18115 0.25182 -8.661 < 2e-16 \*\*\*  
## age4 0.08593 0.01550 5.545 2.94e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16552 on 13332 degrees of freedom  
## Residual deviance: 16521 on 13331 degrees of freedom  
## (54 observations deleted due to missingness)  
## AIC: 16525  
##   
## Number of Fisher Scoring iterations: 4

G\_age <- age\_rdd$null.deviance - age\_rdd$deviance  
p\_val\_age <- 1 - pchisq(G\_age, 1)  
anova(age\_rdd, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: ride.alc.driver  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 13332 16552   
## age4 1 30.864 13331 16521 2.767e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. This decision from the researchers would be motivated by suspecting that there is collinearity between grade and age since from common knowledge we would expect grade nine to be about fourteen to fifteen, grade ten to be fifteen to sixteen, grade eleven to be sixteen to seventeen and so on. Also another possibility is that they are testing which would be the best model with best prediction ability between these two models.
2. I would choose the model containing age after doing some analysis on both models that were considered by the researchers. This is because for our age model we get a G-statistic of 30.86 and for the grade model we get a G-statistic of 8.88. Since the G-statistic for the age model is larger, this tells us our evidence is stronger using age as an explanatory variable as compared to using grade. Finally, our p-value for our age model is 2.767e-08, which is significantly smaller than the p-value for the grade model being 0.00288. We know that a small p-value indicates strong evidence against the null hypothesis and in our case since the p-value of the age model is smaller, we have stronger evidence against our null hypothesis that beta one is equal to zero.

### 2.e.

Use the selected model and state the SPECIFIC hypotheses (null and alternative) for the Logistic Regression Analysis mathematically and in a complete sentence. Here are some symbols you may find useful to adapt:

Probability of riding with a drunk driver is not associated with age.

Probability of riding with a drunk driver is associated with age.

### 2.f.

Conduct the appropriate analysis for part 2.e. and provide a written interpretation of the results. Include your vital statistics (test statistic, degrees of freedom if applicable, alpha, and p-value) appropriately reported in parenthesis at the end of your statement.

#Provide R code and output/graphs  
summary(age\_rdd)

##   
## Call:  
## glm(formula = ride.alc.driver ~ age4, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9224 -0.8905 -0.8287 1.4946 1.6108   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.18115 0.25182 -8.661 < 2e-16 \*\*\*  
## age4 0.08593 0.01550 5.545 2.94e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16552 on 13332 degrees of freedom  
## Residual deviance: 16521 on 13331 degrees of freedom  
## (54 observations deleted due to missingness)  
## AIC: 16525  
##   
## Number of Fisher Scoring iterations: 4

(z=5.545,p-value=2.94e-08,sig level=.05) Based on our sample using a significance level of .05, we will reject the null hypothesis that the probability of riding with a drunk driver is not associated with age and support the alternative that the probability of riding with a drunk driver is associated with age.

### 2.g.

Determine the coefficients for each variable in the model and provide an interpretation of each slope in the context of the problem. Include answers to the following questions: 1. Is there a difference between male and female youth and their choice to ride with a driver who has been drinking? 2. Is there evidence that older students make better decisions about driving with a drinking driver? 3. Are smokers more likely to ride with drivers who have been drinking?

#Provide R code and output/graphs   
gender\_rdd <- glm(formula = ride.alc.driver ~ female, family = binomial())  
summary(gender\_rdd)

##   
## Call:  
## glm(formula = ride.alc.driver ~ female, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0302 -1.0302 -0.6777 1.3321 1.7799   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.35433 0.03206 -42.24 <2e-16 \*\*\*  
## female 0.99777 0.04059 24.58 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 15725 on 12631 degrees of freedom  
## Residual deviance: 15088 on 12630 degrees of freedom  
## (755 observations deleted due to missingness)  
## AIC: 15092  
##   
## Number of Fisher Scoring iterations: 4

exp(-1.25433)

## [1] 0.2852669

exp(0.99777)

## [1] 2.712227

exp(-1.25433) / (1 + exp(0.99777))

## [1] 0.07684523

summary(age\_rdd)

##   
## Call:  
## glm(formula = ride.alc.driver ~ age4, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9224 -0.8905 -0.8287 1.4946 1.6108   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.18115 0.25182 -8.661 < 2e-16 \*\*\*  
## age4 0.08593 0.01550 5.545 2.94e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16552 on 13332 degrees of freedom  
## Residual deviance: 16521 on 13331 degrees of freedom  
## (54 observations deleted due to missingness)  
## AIC: 16525  
##   
## Number of Fisher Scoring iterations: 4

exp(-2.18115)

## [1] 0.1129116

exp(0.08593)

## [1] 1.08973

exp(-2.18115) / (1 + exp(0.08593) )

## [1] 0.05403167

smoker\_rdd <- glm(formula = ride.alc.driver ~ smoke, family = binomial())  
summary(smoker\_rdd)

##   
## Call:  
## glm(formula = ride.alc.driver ~ smoke, family = binomial())  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0442 -1.0442 -0.6375 1.3168 1.8404   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.49031 0.03318 -44.92 <2e-16 \*\*\*  
## smoke 1.16853 0.04113 28.41 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 16099 on 12998 degrees of freedom  
## Residual deviance: 15229 on 12997 degrees of freedom  
## (388 observations deleted due to missingness)  
## AIC: 15233  
##   
## Number of Fisher Scoring iterations: 4

exp(-1.49031)

## [1] 0.2253028

exp(1.16853)

## [1] 3.21726

exp(-1.49031) / (1 + exp(1.16853))

## [1] 0.05342398

Slope Interpretations:

Q1- The slope tells us that when female increases by one, in other words when the person is female, the odds of riding with a drunk driver are 2.712 times the odds of riding with a drunk driver when the person is male. The positive beta one tells us that the probability of riding with a drunk driver increases as female increases, person is female instead of male. In the probability model interpretation, it tells us that when female is zero, the person is male, the probability he will ride with a drunk driver is 0.0768.

Q2 -The slope tells us that when age increases by one year (x + 1), the odds of riding with a drunk driver are 1.08973 times the odds of riding with a drunk driver when the person is x years old. The positive beta one tells us that the probability of riding with a drunk driver increases as the age increases.

Q3 - The slope tells us that when smoker increases by one (x + 1), a person is a smoker, the odds of riding with a drunk driver are 3.21726 times the odds of riding with a drunk driver when the person is not a smoker (x = 0). The positive beta one in this question tells us that the probability of riding with a drunk driver increases as the smoker status increases, in other words the person is a smoker.

Q1 - Coefficients are $\_0 $ = -1.25433 and = 0.9977. We can conclude that there is a difference between male and female youth and their choice to ride with a driver who has been drinking. There is evidence to suggest that females are more likely to ride with a drunk driver. We reject the null hypothesis (Z-statistic = 24.58, p-value = 2e-16, significance level = 0.05) that the probability of riding with a drunk driver is not associated with female.

Q2 - Coefficients are $\_0 $ = -2.18115 and = 0.08593. We can conclude that there is evidence that older students do not make better decisions about driving with a drinking driver. We reject the null hypothesis ( Z-statistic = 5.545, sig.level = 0.05, p-value = 2.94e-08) that the probability of riding with a drunk driver is not associated with female. This is also apparent in our interpretation of the slope above.

Q3 - Coefficients are $\_0 $ = -1.49031 and = 1.16853 . From our data analysis we have evidence to suggest that smokers are more likely to ride with drunk drivers. We reject the null hypothesis (Z-statistic = 28.41, significance level = 0.05, p-value = 2e-16) that the probability of riding with a drunk driver is not associated with smoker status.

The End.