An introduction to Logistic Regression in R

Developed by Rohan Palmer

Part 1: What is logistic regression? Why and when do we need to use it?

Logistic regression is a special case of linear regression used when the dependent is dichotomous/ordinal (ordered categories) not continuous. There are several statistical issues with dichotomous dependent variables:

- 1. How to interpret a predicted value of 0.74, when the dependent variable is coded as 0's and 1's.
- 2. The error distribution is non-normal.
- 3. Possibility of Heteroscedasticity (unequal residual variance at varying levels of an independent measure).

The solution to modeling this type of data is logistic regression. The goal is to predict a particular category for one or more independent variables which may or may not be continuous. Instead of ordinary least squares regression, maximum likelihood estimation is used in logistic regression. As a result, the statistics of interest are Wald-Chi-square values instead of F or t-values. Fortunately, the traditional method of model comparison and hypothesis testing is unchanged.

Examining the nature of the binary dependent variable

The script and figures that follow are an example of modeling a binary dependent variable using a continuous and contrast coded dependent variable. The script generates a dependent variable, "y" (made up 400 rows of 0's and 1's), a dummy coded variable, "z", and a continuous variable, "x" (made up of random normal numbers). The dependent variable is a True/False measure of whether or not the individual represented by the row of data is male. The dummy coded dependent variable is a Yes/No item for whether or not the child is a "cry baby". The continuous dependent variable is a measure of the number of times the child defecates (poops) in a day.

The figures show that when the traditional optimized least squares regression (OLS) method is applied the model attempts to fit the mean and the variance of the dependent variable. This is inappropriate because the values being fitted are non-existent (making interpretation impossible) and we cannot predict changes in the dependent variable when there is very little or no change. In other [Type text]

words the dependent variable does not vary continuously. Using the "child is male" outcome from above, for 1000 records on this TRUE/FALSE (binary) response what we have are proportions and not means or variances. This is not to say that we could not have used a continuous measure to indicate a child's masculinity. Albeit such an approach would provide more power to detect relationships among covariates, nature is not so obliging. As a proportion the dependent variable is bounded by 0 and 1. The Log transformation of the dependent variable will free up the boundary constraint. However, this becomes the defining line between logistic regression and OLS. By log transforming the data we are essentially taking the natural log of the odds of the dependent occurring or not. Thus, logistic regression estimates the odds of a certain event occurring (in this case the probability of the child being male or female).

```
n <- 200

z <- c(rep(0,132),rep(1,268))

x <- c(rnorm(n), 1+rnorm(n))

y <- c(rep(0,n), rep(1,n))

par(mfrow=c(1,2))

op <- par(lwd=2,cex.main=1.3,cex.lab=1.3,cex.axis=1.3)

plot(y~z,main="Figure 1: Plot of Dichotomous DV & Binary IV",xlab="Cry Baby",ylab="Child is a male: 0=No,1=Yes")

abline(Im(y~z), col='red')

plot(jiggle(y)~x,main=" Figure 2: Plot of Dichotomous DV & Continuos IV",xlab="# of poops per day",ylab="Child is a male: 0=No,1=Yes")

abline(Im(jiggle(y)~jiggle(x)), col='blue')
```

Figure 1: Modeling a binary outcome with dichotomous and continuous predictors

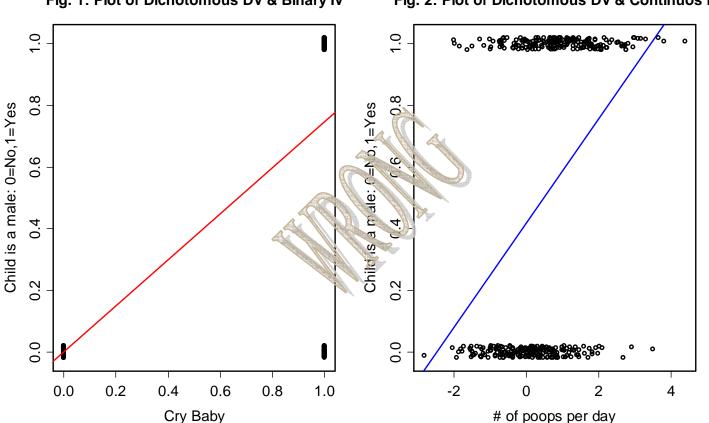


Fig. 1: Plot of Dichotomous DV & Binary IV Fig. 2: Plot of Dichotomous DV & Continuos IV

This equation is the typical setup of a linear model. In logistic regression we model the probability (p) of an outcome in our dependent variable.

Probability (Child is Male) = $\beta 0 + \beta 1 + \dots \beta n$ ----- equation 1.

As stated above, because the above equation is bounded on the left hand side and the right hand side is not we transform it by taking the natural log of the probability.

We can alternatively represent equation 2 as this

Logit (probability of child is male) = Log probability of child is male ----- equation 3

The quotient of the two probabilities is typically referred to as the odds-ratio of the event occurring. [Type text]

The following script is an example of fitting a line to logits.

```
x <- seq(0,1, length=100)
x <- x[2:(length(x)-1)]
logit <- function (t) {
log( t / (1-t) )
}
op <- par(lwd=2,cex.main=1.3,cex.lab=1.3,cex.axis=1.3)
plot(logit(x) ~ x, type='l',col="tomato3",main='Figure 3:Plot of Logits against observed values')</pre>
```

Figure 3:Plot of Logits against observed values

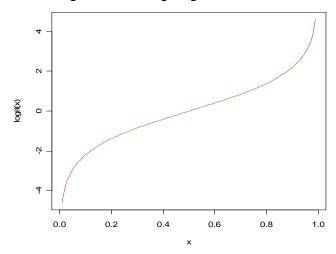
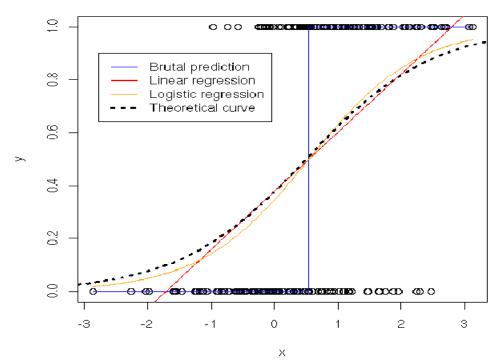


Figure 3 to the left denotes the difference between the values of x and the values of the logit(x). When we plot the two against each other figure 3 indicates the log-odds of an observed value of x occurring. This increased level of interpretation is the reason why logistic regression is the preferred method for modeling dichotomous/ordinal dependent measures.



To further illustrate the difference between logistic regression and OLS figure 4 was incorporated from another WIKI (http://zoonek2.free.fr/UNIX/48_R/12 html). Notice that the

R/12.html). Notice that the model closest to the theoretical curve (similar to the plot of logit(x) against x above) is the logit regression model and OLS regression model. In further detail, the figure compares varied forms of regression for modeling binary data. The brutal prediction method uses least squares regression to predict

the observed values however, because the initial data are 0 and 1 it fails, and spits out infinite values in the output. We can think of the theoretical curve as the probabilities of each of the observed

responses. Logistic regression, using maximum likelihood estimation, models these data points with more precision.

Basic interpretation of odds-ratios and log-transformed odds-ratios (aka logits)

Although logits can take on negative or positive values, odds-ratios must be greater than zero. This can also be explained in terms of probabilities. Considering equation 3 above, as the probability of the numerator approaches zero, the odds ratio approaches zero. Likewise, as the probability of the denominator approaches zero, the odds ratio approaches positive infinity.

An odds-ratio of 1 indicates that the condition or event is equally likely in both groups or for all values of an independent variable (Logit value would NOT be significantly different from zero). An odds ratio greater than 1 indicates that the condition or event is more likely in one group than another or that the odds-ratio increases as the values of the independent variable increases (Logit value would be significantly GREATER THAN zero). An odds-ratio less than 1 indicates that the condition or event is less likely in one group and more likely in another or that the odds-ratio increases for larger values of your independent variable (Logit value would be significantly LESS THAN zero).

A SIMPLE EXAMPLE (EX. 1)

The fictional Metropolitan Board of Health requested a study on one hundred and twenty students (ignoring gender) from a random institution. Students were asked whether or not they have had unsafe sex in the past month (defined as sexual practices that do not safeguard against sexually transmitted diseases (STDs)). Their reply was either a yes or no. Results are presented in Contingency Table 1.

Table 1: Contingency table of sexual practices among students

	Safe Sex	Unsafe Sex	Total
Respondents	90	110	200

Given these results there are several deductions that could be made.

The probability that a person had unsafe sex in this population: 110/200 = 0.55

The probability of safe sex = 1 - probability of unsafe sex: 1-0.55 = 0.45 (i.e., 90/200)

[Type text]

The odds-ratio that a person had unsafe sex as opposed to safe sex in this population = 110/90 = 1.2:1. Likewise the odds of safe sex is the inverse = 90/110 = 0.81:1

A MORE COMPLEX EXAMPLE (EX.2)

The results from the initial study were taken to the Board of Health and they were surprised that the odds-ratio was relatively low so they insisted that we explore further. The board was most interested in determining if male and female students were equally at risk. We return to the university and collect a sample of 200 students (100 males and females). Table 2 is a summary of the results from the new sample.

Table 2: Contingency table of gender specific sexual practices

	Safe Sex	Unsafe Sex	Total
Males	10	90	100
Female	80	20	100
Total	90	110	200

Again, we can ask the same questions as above, but now we can also test if the risk is significantly different between males and females.

The odds of males having unsafe sex in the past month is 90/10 = 9:1

The odds of females having unsafe sex in the past month is 20/80 = 0.25:1

The odds-ratio of males to females for having had unsafe sex in the past month is 9/0.25 = 36.

Males are 36 times more likely to have had unsafe sex in the past month than females.

Using the formula (p/(1-p))/(q/(1-q)) = (p(1-q))/(q(1-p)), where p is the probability of the event occurring in the first group, and q is the probability of the event occurring in the second group:

((90/100) / (10/100)) / ((20/100) / (80/100)) = (0.9*0.8) / (0.1*0.2) = 0.72/0.02 = 36.

Part 2: Generating contingency tables of your data and simple Chi-Square testing

Let's start with how to create contingency tables of our data in R. If you are familiar with SAS, STATA, or SPSS you typically just call up a cross-tabulated frequency table command. In R these tables are called "FLAT TABLES" and the concept behind their creation is the same.

We will start be creating a dataframe for each example that contains the information in a "person-level" format. The resultant "risky sex" variable in the dataframes will be coded as a binary variable where a 0 indicates the absence of risky sexual practices by that individual and a 1 indicates the admission of unsafe sex by that person in the past month.

For example 1:

```
ex1.unsafe <- (rep(1,times=110)) #Prevalence of unsafe sex ex1.safe <- (rep(0,times=90)) #Prevalence of safe sex ex1 <- data.frame(risky_sex=c(ex1.safe,ex1.unsafe)) ex1[,1] <- as.ordered(ex1[,1])
```

NOTE: You must set up you dependent variable as an ordered factor. This is important because (1) you cannot order "Yes" or "No" categories and (2) it will facilitate the identification of your predicted outcome level (i.e. Yes versus No) when fitting logit models.

For example 2:

```
Ladies <- "Female"

Gentlemen <- "Male"
gender <- (c(Ladies,Gentlemen)) #Later we will have to convert the gender variable into a contrast code.
fem.safe <- rep(0,times=80) #Prevalence of safe sex in females
fem.unsafe <- rep(1,times=20) #Prevalence of unsafe sex in females
male.safe <- rep(0,times=10) #Prevalence of safe sex in males
male.unsafe <- rep(1,times=90) #Prevalence of unsafe sex in males
ex2 <- data.frame(risky_sex=c(fem.safe,fem.unsafe,male.safe,male.unsafe),gender=rep(gender,each=100))
ex2[,1] <- as.ordered(ex2[,1])
```

Now that we have the data we can proceed to create a frequency table of each of the ordered categories. In each example we are interested in the prevalence of "Risky Sex Behavior". Frequency tables can be generated using the **table()** and **ftable** commands in R. Both functions produce the same results, but the **ftable()** command offers greater control over the structure of your table by allowing you to choose the independent variables you are interested in, and how they will create levels in your dependent variable. It is also important to note that the **ftable()** command is sensitive to (1) the number of dimensions of your table (at least two variables are required), and (2) the class of the object which you are attempting to represent as a table. The tables below demonstrate the use of the **table()** and **ftable()** commands using dataframes. You can also save your tables by equating your table command to an available object name.

```
?table
table(ex1)
ex1
90
        110
table(ex2)
                    Gender
                Female Male
risky_sex
    0
                         10
                 80
                         90
                #OR WE CAN USE ftable()
?ftable
table.1 <- ftable(ex1[c("risky_sex")])
                                                  #THIS WILL NOT WORK
table.2 <- ftable(ex2[c("gender","risky_sex")])
table.2
        risky_sex 0
                          1
gender
                         20
Female
Male
                         90
#Chi-Square Testing
?chisq.test
chisq.test(table(ex1))
    Chi-squared test for given probabilities
data: table(ex1)
X-squared = 2, df = 1, p-value = 0.1573
chisq.test(table.2)
    Pearson's Chi-squared test with Yates' continuity correction
data: table.2
X-squared = 96.1818, df = 1, p-value < 2.2e-16
```

We can test for homogeneity of the proportions (equal cell sizes) in our contingency table using the **chisq.test()** command. The command executes a Pearson Chi-square test along with a Yates continuity correction that makes the Pearson Chi-square statistic to have better agreement with Fisher's Exact test when the sample size is small.

It is very important to note that the complexity of your contingency tables will depend on the levels of your independent variables. This will be demonstrated later.

Part 3: Conducting Logistic Regression Analyses in R

The approach to logistic regression in R is the general linear model **(glm)** function, while specifying the type/family of analysis that is specific to the nature of your dependent measure. The family function, **family ()**, indicates whether we want to do a logit, probit, or other forms of non-linear regression.

How to execute logistic regression in R

In the first example we could not tell if we had significant differences in the prevalence of risky sex without the chi-square test. In the second example we could tell that males had risky sex more often than females by looking at the data but we could only test it using the Chi-square test. As our models [Type text]

become more complex with the use of continuous, ordinal, dichotomous or combinations of independent variables, the logistic regression method becomes more applicable. We will get to those complexities in the next example, but for now you should note that the logistic regression model generates the same results from the Chi-square tests of the first and second examples. The script will also serve as the basis for executing logistic models using the **glm()** function.

```
#For example 1
ex1.test <- glm (risky_sex ~ 1,family=binomial(logit), data=ex1)

#For example 2
gender.code <- (ex2$gender==0.5)*1 - 0.5  #Create a contrast code for dichotomous IV's
ex2.test <- glm(risky_sex ~ gender.code, family=binomial(logit), data=ex2)
summary(ex1.test);summary (ex2.test)
```

Note that the specification of the model is unchanged, with the exception of the family function being set to binomial, and the method of binomial analysis being set to logit. Another important aspect of the analysis is the coding of your dichotomized or leveled factors. This is very important when it comes to interpreting the results of the models.

First we must recode gender in a manner that makes its parameter estimate interpretable. Using -0.5 for females and 0.5 for males, the interpretation for the gender parameter estimate would be, "the amount by which the log of the odds of having risky sex in the past month is greater/lesser in males than females".

The next set of scripts uses the **summary()** command to view the results of the fitted model.

```
summary(ex1.test)
glm(formula = risky_sex ~ 1, family = binomial(logit), data = ex1)
Deviance Residuals:
         1Q
                Median 3Q
-1.264 -1.264 1.093 1.093 1.093
Coefficients:
                                Std. Error
                                                z value Pr(>|z|)
                Estimate
(Intercept)
                0.2007
                                0.1421
                                                1.412 0.158
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 275.26 on 199 degrees of freedom
Residual deviance: 275.26 on 199 degrees of freedom
AIC: 277.26
Number of Fisher Scoring iterations: 3
summary(ex2.test)
Call:
glm(formula = risky_sex ~ gender.code, family = binomial(logit),
  data = ex2
Deviance Residuals:
        1Q
                Median 3Q
 Min
                               Max
-2.146 -0.668 0.459 0.459 1.794
Coefficients:
                Estimate
                                Std. Error
                                                z value
                                                                Pr (>|z|)
(Intercept)
                0.4055
                                0.2083
                                                1.946
                                                                0.0516
                                                                <2e-16 ***
gender.code
               3.5835
                                0.4166
                                                8.601
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 275.26 on 199 degrees of freedom
Residual deviance: 165.10 on 198 degrees of freedom
AIC: 169.10
Number of Fisher Scoring iterations: 4
```

Identifying and using the statistics of a logistic regression output

There are several statistics obtained from the summary of the logistic regression model. The first statistic to examine is the fit statistics. **Null deviance** is the negative twice the log-likelihood (-2LL) fit estimate for the intercept only model. **Residual Deviance** is the difference in the -2LL for the null model and the model being tested. This difference is distributed as a Chi-square and allows us to compare nested models. The next thing we want to examine is the coefficients of our parameters and how well they predict changes in our dependent variable (we will examine interpretation later). The summary also provides **Standard error estimates** and the ratio of the coefficients to the standard

error which are $Wald-\chi^2$ statistics for testing whether the regression coefficients are equal to zero (but the table presents **z-values** which are the square root of the $Wald-\chi^2$). The summary of the model also produces **p-values** that indicate the size of the effect of the parameter. Finally, the **Akaike** Information Criterion (AIC) statistic is a fit index that takes into account the parsimony of the model by penalizing for the number of parameters (i.e. degrees of freedom, df). The AIC can be used to compare models with varying numbers of parameters; AIC methodology attempts to find the model that best explains the data with a minimum of free parameters.

The **anova()** command can also be used to conduct a likelihood ratio test that compares the fit of the different regression models, as long as they are based on the same dataset. In the box below **anova(ex2.test2)** compares the fit of the NULL (Intercept only) model and the model predicting the difference in logits between males and females. The table shows that for 1df difference the Chisquare's reduced by 110.16 points. Thus, there is a significant difference between genders.

anova (ex2.test)
Analysis of Deviance Table
Model: binomial, link: logit
Response: risky_sex
Terms added sequentially (first to last)

Df DevianceResid. Df Resid. Dev
NULL 199 275.26
gender.code 1 110.16 198 165.10

This comparison is always present in the summary of the model; however if you would like to compare other models that you have tested and saved as an object you can do so with this command.

How to interpret the findings of a logistic regression model

For the first example we tested if the likelihood of a student having had risky sex was equal to that of them not having had risky sex. In other words, are the probabilities equal? Which translates into: Are the odds equal to one? Or, Are the log of the odds (logit) equal to zero?

Model 1: Log (Odds that Risky Sex = 1) = $\beta_0 + \epsilon$

H0: $\beta_0 = 0$

Results: Model 1: Log (Odds that Risky Sex = 1) = $0.20 + \varepsilon$

[Type text]

How to interpret the coefficient of the intercept only model: The log of the odds of having had risky sex in the past month is 0.20. You can also choose to convert the logit (log of the odds) back to an odds-ratio by taking the exponent of the intercept ($\exp(0.20) = 1.22:1$). You can also convert the odds back to probabilities:

```
Odds = p/(1-p).

1.22 = p/(1-p)

1.22(1-p) = p

1.22 - 1.22p = p

1.22 = 2.22p

p = 1.22/2.22 = 0.55
```

Coupled with the $Wald-\chi^2$ and p-value we know that the 0.22 is not significantly different from 0.Thi can also be interpreted as, the probabilities 0.55 and 0.45 are not significantly different. In terms of odds, the odds are 1:1. Notice that the odds and probability statistics are the same as when done by hand.

For the second example we sought to determine the extent to which the likelihood of reporting unsafe sex was different in males and females. Thus we are testing a model in which the coefficient for the gender parameter (i.e. the difference between males and females) is zero.

```
Model 1: Log (Odds that Risky Sex = 1) = \beta_0 + \epsilon
Model 2: Log (Odds that Risky Sex = 1) = \beta_0 + \beta_1(Gender) + \epsilon
H0: \beta_1 = 0
Results: Model 2: Log (Odds that Risky Sex = 1) = 0.41 + 3.58(Gender) + \epsilon
```

How to interpret the intercept in an augmented model: When Gender is equal to zero, the log of the odds of having had risky sex in the past month is 0.41. You can also choose to convert the logit (log of the odds) back to odds by taking the exponent of the intercept ($\exp(0.41) = 1.5$). You can also convert the odds back to probabilities:

```
1.5 = p/(1/p)

1.5(1-p) = p

1.5 - 1.5p = p

1.5 = 2.5p

p = 1.5/2.5 = 0.6
```

More importantly what does Gender = 0 mean if gender is coded as -0.5 and 0.5? The correct way of describing the intercept would be: "On average across males and females the logit of having had

risky sex in the past month is 0.41". This is why it is important to code your categorical independent variable to target the question being asked.

How to interpret the Slope in an augmented model: For each unit increase in Gender, the log of the odds of having had risky sex in the past month increases by 3.58. In terms of odds, you would say: "the odds of having had risky sex increases by a factor of 36" (i.e., $\exp(3.58) = 36$). In this example there is only a one unit increase in Gender that represents the difference between males and females, thus we are getting an odds-ratio of 36:1. Notice that the odds-ratio is the same as when we did it by hand in Part 1.

Part 4: Multivariate Logistic Regression

After taking our latest findings back to Board of Health, they ask that we go back to and gather more data. This time they have asked that we gather the following information:

- 1. The age of the participants
- 2. Their academic year in college (freshman, sophomore, junior, senior).
- 3. Their annual income.
- 4. Their sexual preferences (i.e. male or female).

The board is interested in whether or not these variables are relevant to risky sex behavior at the college level, either separately or combined. Unlike the previous examples this problem is more complex. In order to truly test each of these parameters we have to go beyond our simple contingency table chi-square analyses and utilize logistic regression to its full extent. The following hypotheses were tested with the newly gathered information.

- 1. The odds-ratio of risky sex for persons of average income in the population is equal to 1 (i.e., the logit is not significantly different from zero).
- 2. Males are at a greater risk for risky sex behaviors regardless of the effects of income.
- 3. The likelihood of risky sex increases as age increases, controlling for the effects of income.
- 4. The odds of risky sex increases as academic year in college increases regardless of income.
- 5. Persons who are heterosexual are at the same level of risk for risky sex behaviors as homosexuals regardless of income.
- 6. The difference in likelihood of risky sex between males and females is not altered by (1) age, and (2) academic year in college, regardless of income.

[Type text]

The script and output below demonstrate how to answer these questions.

```
#Multivariate Logistic Regression
gender <- (c(-0.5,0.5))
sexes <- rep(gender,each=100)
fem.safe <- rep(0,times=80)
fage.safe <- rep(c(18,19,20,21,22,23),c(25,15,12,10,10,8))
fem.unsafe <- rep(1,times=20)
                                                          #AS females get older the incidence of risky sex increases
fage.unsafe <- rep(c(18,19,20,21,22,23),c(0,1,3,3,4,9))
male.safe <- rep(0,times=10)
mage.safe <- rep(c(18,19,20,21,22,23),c(0,0,1,2,3,4))
male.unsafe <- rep(1,times=90)
mage.unsafe <- rep(c(18,19,20,21,22,23),c(25,22,18,9,8,8)) #AS males get older the incidence of risky sex decreases
set.seed(1234)
lover <- rnorm(200,1,1)
lover[runif(100,0,1)<.40] <- 0
sex.partner <- (lover==0)*1
                                 #about 40% of lovers are females and they are randomly distributed throughout the
                                 #So we have same sex and opposite sex couples.
#Creating the College Student Annual Income (per thousand) Variable
set.seed(1234)
income <- rnorm(n=200,mean=15,sd=5) #No effect - randomly distributed throughout the data
#We must also mean deviate this variable before we ue it in our analyses thus to make interpretation easy.
cincome <- round(income-mean(income),digits=2)
#Creating the college year variable
#By randomizing this variable we expect little or no relationship between risky sex and academic year.
#Academic ear ranges from 1(Freshmen) to 5 (super seniors).
#We must mean center academic year, so that the intercept is interpretable.
set.seed(1234)
vear <- sample(1:5,size=200,replace=TRUE)</pre>
cyear <- round(year-(mean(year)),digits=2)</pre>
cyear
#Remember that we need to contrast code sex to make it interpretable
                                         #Let's first manually recode gender to be 0.5 if males and -0.5 if female
sex.code <- ((sexes==0.5)*1 - 0.5)
sex.code
                                         #Males are 0.5; females -0.5
#We also have to contrast code partner sex
partner.code <- ((sex.partner==1)*1 - 0.5) #So the interpretation is such that how much the risk increases for male
partners #compared to female partners
partner.code
                                         #Males are 0.5; females -0.5
#Lastly, we must place all this data into a dataframe called "example3".
example3 <-
data.frame(rsex=c(fem.safe,fem.unsafe,male.safe,male.unsafe),gender=sex.code,age=c(fage.safe,fage.unsafe,mage.saf
e,mage.unsafe),year=cyear,income=cincome,partner=partner.code)
example3$rsex <- as.ordered(example3$rsex)
example3 <- example3[order(example3$gender,example3$age,example3$year,example3$partner),]
example3
```

Step 1: Exploring your data to the fullest extent with ftable()

We have partially explored creating levels in your dependent variable by increasing the number of independent variables in our **ftable()** command. The scripts below creates a series of tables that

explore the relationship between the variables in the dataset called **example3** (income is excluded because it is a continuous variable with a very wide range of values).

Looking at partner gender and risky sex we find that there is no relationship between the two.

```
#Risky sex by gender of sexual partner
table.2 <- ftable(example3[c("partner","rsex")])
table.2
    rsex 0 1
partner
-0.5 62 72
0.5 28 38

chisq.test(table.2)
    Pearson's Chi-squared test with Yates' continuity correction

data: table.2
X-squared = 0.1316, df = 1, p-value = 0.7168
```

Looking at gender and sexual partner we find no evidence of a specific preference in either gender.

```
table.3 <- ftable(example3[c("gender","partner")])
table.3
        partner
                         -0.5
                                  0.5
gender
-0.5
                 67
                         33
0.5
                 67
                         33
chisq.test(table.3)
    Pearson's Chi-squared test with Yates' continuity correction
data: table.3
X-squared = 0.0226, df = 1, p-value = 0.8805
```

Age does not appear to be related to the prevalence of past month risky sex practices.

```
table.4 <- ftable(example3[c("age", "rsex")])
table.4
                        1
        rsex
age
18
                25
                        25
                15
                        23
19
                13
                        21
20
21
                12
                        12
22
                13
                        12
                        17
23
                12
chisq.test(table.4)
    Pearson's Chi-squared test
data: table.4
X-squared = 2.4936, df = 5, p-value = 0.7775
```

Looking at gender, age, and risky sex together reveals that an age-gender interaction is present.

```
table.5 <- ftable(example3[c("gender","age","rsex")])
table.5
                             0
                                       1
                      rsex
gender
               age
                              25
-0.5
               18
                                       0
                              15
               19
                                       1
               20
                              12
                                       3
               21
                              10
                                       3
               22
                              10
                                       4
                                       9
               23
                              8
0.5
               18
                              0
                                      25
               19
                              0
                                      22
               20
                              1
                                      18
               21
                              2
                                       9
               22
                              3
                                       8
                                       8
chisq.test(table.5)
    Pearson's Chi-squared test
data: table.5
X-squared = 118.5057, df = 11, p-value < 2.2e-16
Warning message:
In chisq.test(table.5): Chi-squared approximation may be incorrect
```

Step 2: Testing the hypotheses using logistic regression

Q1. The odds of risky sex for persons of average income in the population is equal to 1 (i.e., the logit is not significantly different from zero).

```
> Q1 <- glm(rsex ~ income,family=binomial(logit),data=example3)
> summary(Q1)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
              (Intercept)
income
Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' ' 1
AIC: 274.13
> anova(Q1)
Analysis of Deviance Table
Terms added sequentially (first to last)
       Df
               Deviance Resid.
                                             Df
                                                     Resid. Dev
NULL 199
               275.256
                                     198
income 1
               5.127
                                             270.129
```

Take the intercept (which is in logits (log of the odds) and convert it back to odds by taking the exponent of it. Try ?exp; exp(Q1\$coefficients[1]). This gives you the intercept in terms of odds is 1.23. So the first hypothesis is TRUE because 0.21 is not significantly different from zero (**Wald-** χ^2 = (1.44)² = 2.07, **df** = 1, **p-value** = 0.15)

Q2. Males are at a greater risk for risky sex behaviors regardless of the effects of income.

```
Q2 <- glm(rsex ~ gender+income,family=binomial(logit),data=example3)
summary(Q2)
glm(formula = rsex ~ gender + income, family = binomial(logit), data = example3)
Coefficients:
                                               z value Pr(>|z|)
                               Std. Error
               Estimate
(Intercept)
               0.41297
                               0.21015
                                               1.965 0.0494 *
                                               8.497 <2e-16 ***
                3.60176
                               0.42389
gender
income
               0.07021
                               0.03908
                                               1.796
                                                     0.0724 .
AIC: 167.87
> anova(Q2)
Analysis of Deviance Table
                                       Resid. Df
                                                       Resid. Dev
                      Deviance
                          275.256
NULL
             1
                       110.158
                                                        165.097
Gender
                                        198
Income
             1
                       3.228
                                        197
                                                        161.869
```

The hypothesis is correct. Assuming everyone has the same income, males are at a greater risk than females.

Q3. Testing if the likelihood of risky sex increases as age increases, controlling for the effects of income.

```
Q3 <- glm(rsex ~ age+income,family=binomial(logit),data=example3)
glm(formula = rsex ~ age + income, family = binomial(logit), data = example3)
Coefficients:
               Estimate
                               Std. Error
                                               z value Pr(>|z|)
(Intercept)
               0.45401
                               1.66253
                                               0.273 0.7848
age
               -0.01230
                               0.08234
                                               -0.149 0.8813
               0.06524
                               0.02943
                                               2.217 0.0266 *
income
AIC: 276.11
anova(Q3)
Analysis of Deviance Table
               Deviance Resid. Df
                                       Resid. Dev
       Df
NULL
        199
               275.256
               0.012
                                       198
                                               275.244
age
        1
income 1
               5.137
                                        197
                                               270.106
```

The hypothesis was false. Controlling for income, Age has no relationship with the risk for unsafe sexual practices. Income is significant; probably by chance association. If the sample size of example 3 was larger this effect would most likely be absent.

Q4. Testing if the odds of risky sex increase as academic year in college increases regardless of income.

```
Q4 <- glm(rsex ~ year+income,family=binomial(logit),data=example3)
glm(formula = rsex ~ year + income, family = binomial(logit), data = example3)
Coefficients:
               Estimate
                               Std. Error
                                               z value Pr(>|z|)
               0.21534
                               0.14668
                                               1.468 0.1421
(Intercept)
               0.26751
                               0.10548
                                               2.536 0.0112 *
year
               0.06823
                               0.02966
                                               2.301
                                                       0.0214 *
income
AIC: 269.49
anova(Q4)
Analysis of Deviance Table
                       Deviance Resid. Df
                                               Resid. Dev
            Df
NULL
            199
                       275.256
                       6.242
                                               198
                                                       269.014
year
income
             1
                        5.527
                                               197
                                                        263.486
```

The hypothesis was correct. However, as academic year increases the odds in favor of a risky-sex behavior response decreases. Is this association TRUE? (Recall the creation of example3). Like income, year was independently generated and randomly distributed throughout example3. Its effect in this model is also likely due to a chance association resulting from the small sample size.

Q5. Persons who are heterosexual are at the same level of risk for risky sex behaviors as homosexuals regardless of income.

```
Q5 <- glm(rsex ~ income+gender*partner,family=binomial(logit),data=example3)
glm(formula = rsex ~ income + gender * partner, family = binomial(logit), data = example3)
Coefficients:
                Estimate
                                Std. Error
                                                z value Pr(>|z|)
(Intercept)
                0.49285
                                0.24504
                                                2.011
                                                        0.0443 *
income
                0.06682
                                0.03906
                                                1.711
                                                        0.0871.
                                                        2.85e-14 ***
gender
                3.74231
                                0.49210
                                                7.605
                                                0.701
                                                        0.4830
partner
                0.34578
                                0.49298
                                0.98069
                                                0.603
                                                        0.5463
gender:partner 0.59164
AIC: 171.20
> anova(Q5)
Analysis of Deviance Table
                        Deviance Resid. Df
                                              Resid. Dev
                Df
NULL
                199
                        275.256
                        5.127
                                                198 270.129
income
                 1
                        108,260
                                                     161.869
gender
                 1
                                                197
                        0.290
                                                196 161.579
partner
                 1
                        0.382
                                                195 161.197
gender:partner 1
```

The hypothesis was correct. Over and above the income effects, the difference in risk between males and females did not change if a person's partner was of the same sex or not.

NOTE THE CODING OF THE GENDER AND PARTENR VARIABLES AND THE RESPECTIVE CODING OF THEIR INTERACTION. What does the parameter estimate of 0.59 represent? It is ½ the change in the logit if a persons was in a homosexual as opposed to a heterosexual relationship.

Q6. The difference in likelihood of risky sex between males and females is not altered by (1) age, and (2) academic year in college, regardless of income.

```
Q6 <- glm(rsex ~ income+gender*age+gender*year,family=binomial(logit),data=example3)
summary(Q6)
glm(formula = rsex ~ income + gender * age + gender * year, family = binomial(logit), data = example3)
Coefficients:
                                Std. Error
                                                z value Pr(>|z|)
                Estimate
                5.31604
                                4.51564
(Intercept)
                                                1.177
                                                        0.23910
income
                0.03681
                                0.04840
                                                0.760
                                                        0.44698
gender
                                                4.648 3.35e-06 ***
                42.10440
                                9.05790
                -0.22194
                                0.20821
                                                -1.066 0.28645
age
                0.57369
                                0.21866
                                                2.624
                                                       0.00870 **
vear
                -1.81472
                                0.41793
                                                -4.342 1.41e-05 ***
gender:age
                                                2.680
                                                       0.00735 **
gender:year
               1.16582
                                0.43494
AIC: 127.45
anova(Q6)
Analysis of Deviance Table
                                        Resid. Df
                Df
                        Deviance
                                                         Resid. Dev
NULL
                199
                        275.256
                        5.127
                                        198
                                                        270,129
income
                1
gender
                1
                        108.260
                                        197
                                                        161.869
                                        196
                1
                        1.044
                                                        160.825
age
                1
                        2.688
                                        195
                                                        158.138
year
gender:age
                1
                        35,433
                                        194
                                                        122,704
gender:year
                        9.250
                                        193
                                                         113.454
```

The hypothesis is false. Controlling for the effects of income the gender logit is affected by age and academic year (chance association), such that as age increases the logit difference between males and females decreases. This implies that females are more likely to have a risky sex response at younger ages, while males have a higher likelihood at older ages (just as the contingency table depicted). The chance significance between gender and academic year indicates that the logit differences between male and females increases by a factor of 1.17 each additional year a person progresses through college.

Part 5: Assumption Testing for Logistic Regression

Testing for Linearity of Relationship between Dependent and Independent Variables

The most important assumption in logistic regression is linearity, since our goal is to model a log-form of our dichotomous variable and assume when we fit a linear model to those points it will do so appropriately. Nonlinearity can be detected in a plot of the *residuals* versus the *predicted/fitted* values. As we build and test our models the residuals indicate how well we predict the observed data. As such, as our model better predicts the observed data the residuals approach zero. Thus a plot of the residuals against the predicted/fitted values would yield a straight line if our model perfectly predicts the observed data, making the residuals for each data point zero.

In determining the presence of linearity the points should be symmetrically distributed about the dotted horizontal line. If nonlinearity is present you should add a non-linear transformation to your independent variable to your model (for example consider using (age)² along with age, or try to add another independent variable. A large deviation from the horizontal line indicates that the linear model does not fit the data well and you should consider adding another predictor. In the series of assumption plots below, the upper left plot indicates that model 6 is a sufficient linear model. When compared to model 1, it is a significant improvement (not shown).

Testing the Normal Distribution of Residuals Assumption

The next important assumption that we must examine is the normality of the deviance residuals (i.e., the residual in our models). This assumption, though very robust, is important to consider, because non-normality would mean that the significance statistics are biased. Consequently, the validity results would be questionable. We can test this assumption using a normal quantile-quantile plot (qq-plot). As depicted in the upper right quadrant of the series of graphs below, most of the data do fall

closely to the line representing a normal distribution hence the data is consistent with a normal distribution.

Testing the Homogeneity of Variance Assumption (Homoscedasticity)

In logistic regression testing the homogeneity of variance assumption is not required because based on the nature of the dependent variable we have a non-normal error distribution that is heteroskedastic. For continuous outcomes focus your attention on the graph in the lower left corner.

Testing for Outliers and their influence on the regression parameters

Another concern is the presence of "Outliers". The presence of outliers can result in one of three things:

- A. **Increased rate of TYPE I Errors**. This occurs when there is an unusual independent variable value for a usual dependent variable value (for e.g. for a scale of 1-10 on your independent variable, 99% of your sample takes on values between 0 and 5 and single subject has a value of 10; they all score similar values on the dependent variable).
- B. Increased rate of TYPE II Errors/Decrease the power to detect an effect. This occurs when there is an unusual dependent variable value for a common independent variable value (e.g., using the previous scales, it is possible to have an individual with a typical value on the independent variable but unlike other individuals has an unusually large or small value on the dependent variable.
- C. The model and the coefficients are wrong!

The standard method for detecting these abnormalities is the Cook's Distance (cooksD) statistic which incorporates the Lever and the studentized deleted residual. The cooksD vs. Leverage plot identifies these influential points so that we may test our model with or without them. Our expectation is a straight line that follows the dashed lines on the graph. In the graph the labeled points, if any, are those with the largest Cook's distances. The lower right graph amongst the figures below indicates that there are at least two data points with high cooksD values.

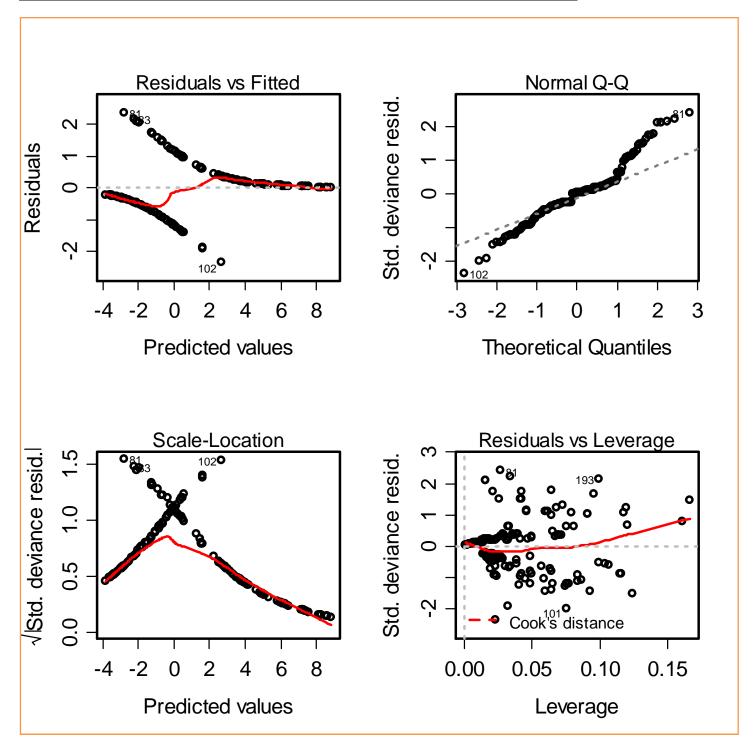
Testing the Independence of Residuals Assumption

A very critical assumption is that there is actually no correlation among the observed dependent variables. In other words knowing something about one observation should not provide any information about another observation. Unfortunately this is a very difficult assumption to test and so we briefly provide mechanisms to avoid dependence among observations in your data. The best protection against violating the independence assumption is to (1) design the experiment in a manner where the likelihood of non-independence in the data is reduced and (2) if you are aware of the cause of the dependence in the data it should be included in the model.

The following R script tests the first four assumptions using the model for question 6 in the multivariate example above.

```
op <- par(mfrow=c(2,2))
op <- par(lwd=2,cex.main=1.3,cex.lab=1.3,cex.axis=1.3)
plot(Q6)
par(op)
```

Figure 5: Plot of the Assumption tests for the logistic model in question 6



Exam Question

Question 1

[Type text]

- A. Use ftable() to create 4 univariate contingency tables of the age(Age), church attendance (Att.church), behavior disinhibition (BD), and early alcohol exposure (Early.alc) variables with lifetime alcohol SUD (L_drink). Describe the relationship between each variable and lifetime alcohol SUD. Provide the necessary statistics to support your claims (i.e. odds-ratios or probabilities, Chi-square, degrees of freedom, and p-values.
- B. Use ftable() to look for an interaction between BD and early alcohol exposure on the prevalence of lifetime alcohol use reports. Does it appear that persons who are exposed under supervision at home endorse having had a use disorder at some point in their lifetime more than persons who are not? Also look at the interaction between BD and church attendance? Does church attendance protect against alcohol use disorders.
- C. Scour the data and determine for each variable how many rows contain missing data for each variable. Try to do this in one line using the apply function. Your output should list each of the variable names and the number of rows with missing data below it.

Question 2

SIDE INFO: Data were created using mean centered forms of Age, SES, and BD so you do not need to center the variables. However, keep in mind that this will affect your interpretation of the model coefficients.

- A. Test the hypothesis that persons from low SES families are at greater risk for a lifetime SUD disorder diagnosis in adolescence than people from high SES families.
- B. Use separate models to test the hypotheses that church attendance and early alcohol exposure are a protective factor for alcohol use, while controlling for SES, and Age. Interpret the intercept and church slope of the church model in terms of odds. For the early alcohol exposure interpret the parameter of the intercept and the alcohol exposure slope in terms of logits.
- C. The literature suggests that BD is the end all. Once you are high on this scale you are very likely to be diagnosed with an alcohol disorder in your lifetime. Test this hypothesis, and based on the model what would the odds in favor of a lifetime alcohol disorder report be if a person, scored 12 on the BD scale, is 16 years old, goes to church, was never exposed to alcohol positively while growing up and whose family earns \$60,000 annually (Note: The dependent

- variable was created the mean deviated forms of Age, BD, and SES). Control for all the other variables in the model. THIS IS ALSO KNWN AS THE KITCHEN SINK MODEL. Have any of the earlier relationships changed once we take BD into account? What is the direction of their effects, does it match up to our expectations?
- D. Test if there is a significant interaction between church attendance and BD, while controlling for Age, SES, and church attendance.

Question 3

- A. Compare the fit of the kitchen sink model to the model with the interaction between BD and church attendance. Which model best fits the data? Justify your answer with the appropriate statistics.
- B. Drop the non-significant parameters from your best fitting model and compare the linearity assumption in your best fitting model and this new model. Which model best fits the data linearly? Are there any other differences in model assumptions that vary between the two models? Use appropriate titles to distinguish the assumptions between the two models.