Lab 6 - Overdispersion and Poisson Regression

Lab Goals

In this lab we will explore the overdispersion parameter, and poisson regression models for count data in R. In particular, we will examine:

1. Calculating the overdispersion parameter
2. Goodness of fit testing
3. Poisson Regression

Skin Cancer Data

The data set *minn.csv* includes Binomial data for the prevalence of skin cancer by age for a random sample of women from Minnesota. Here we will fit a logistic regression model to look at the relationship between `age` and the probability of developing skin cancer. `age` was originally a categorical variable. However, to look at the relationship between increasing age and skin cancer, women in each category are assigned the average age for that category. The variables in the data are summarized here.

Variable	Description
`age`	average age of women in the age group
`Cases`	number of women with skin cancer
`Pop`	total number of women from the age group

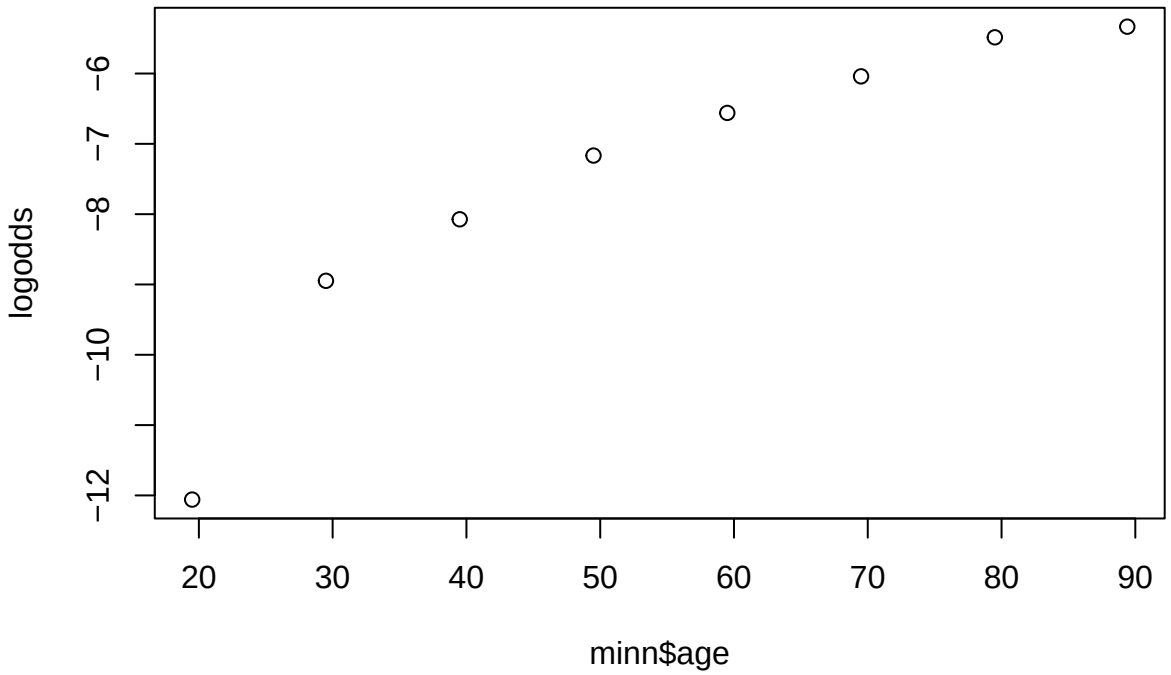
Load the data into the console and into this R Markdown document using the code chunk below.

```
minn <- read.csv("minn.csv")
```

Model Selection

First we will look at the data by plotting the log odds of developing skin cancer by `age`.

```
prob = minn$Cases/minn$Pop
logodds = log(prob/(1-prob))
plot(minn$age,logodds)
```



1. Does the relationship between the log odds of developing skin cancer and age seem to be linear? *No, there appears to be a quadratic relationship between the log odds of developing skin cancer and age.*

Next we will fit a logistic regression model for this study without a quadratic term.

```
cancer.glm=glm(cbind(Cases,Pop-Cases)~age,family=binomial,data=minn)
summary(cancer.glm)
```

```
##
## Call:
## glm(formula = cbind(Cases, Pop - Cases) ~ age, family = binomial,
##     data = minn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8644  -1.6687  -0.0714   1.2002   1.9857
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.55629    0.16901  -62.46   <2e-16 ***
## age           0.06374    0.00248   25.70   <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: 846.631  on 7  degrees of freedom
## Residual deviance:  44.102  on 6  degrees of freedom
## AIC: 91.488
##
## Number of Fisher Scoring iterations: 4
```

2. Based on the model fit, what is the estimated multiplicative effect on the odds of developing cancer of each additional 5 years of age? *The additive effect on the log odds for each additional 5 years of age*

is $5 \times 0.064 = 0.32$. *Thus the multiplicative effect on the odds of developing cancer of an additional 5 years of age is* $e^{0.32} = 1.38$. Determine a 95% confidence interval for this multiplicative effect. *Using the Wald statistic an approximate confidence interval for the additive effect on the log odds is* $(0.32 - 1.96 \times \sqrt{(25 * 0.0025^2)}, 0.32 + 1.96 \times \sqrt{(25 * 0.0025^2)}) = (0.2955, 0.3445)$. *Exponentiating each endpoint, a 95% confidence interval for the multiplicative effect on the odds is (1.344,1.411).*

```
0.32-1.96*sqrt(25*0.0025^2)
```

```
## [1] 0.2955
```

```
0.32+1.96*sqrt(25*0.0025^2)
```

```
## [1] 0.3445
```

```
exp(0.2955)
```

```
## [1] 1.343798
```

```
exp(0.3445)
```

```
## [1] 1.411284
```

Now we will fit the model with the quadratic term.

```r
cancer2.glm=glm(cbind(Cases,Pop-Cases)~age+I(age*age),family=binomial,data=minn)
summary(cancer2.glm)
```

```
##
## Call:
## glm(formula = cbind(Cases, Pop - Cases) ~ age + I(age * age),
##     family = binomial, data = minn)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -2.22316   0.98841   0.39489   0.69945  -0.40323  -0.98403   0.78740  -0.09141
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.329e+01  5.613e-01 -23.679  < 2e-16 ***
## age           1.628e-01  1.858e-02   8.761  < 2e-16 ***
## I(age * age) -8.234e-04  1.499e-04  -5.494 3.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 846.6311  on 7  degrees of freedom
## Residual deviance:   8.3238  on 5  degrees of freedom
## AIC: 57.71
##
## Number of Fisher Scoring iterations: 4
```

3. Determine the likelihood ratio statistic and reference distribution for testing whether the parameter for the quadratic term is significantly different from 0. *From above: Residual deviance (reduced model) - Residual deviance (full model) = 44.102-8.3238 = 35.7782. The reference distribution is a chi-square distribution with 1 degree of freedom.*

4. Perform the likelihood ratio test from (2) using the `anova()` function. What is the p-value for this test? 2.211×10^{-9}

```
anova(cancer.glm,cancer2.glm,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Cases, Pop - Cases) ~ age
## Model 2: cbind(Cases, Pop - Cases) ~ age + I(age * age)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         6     44.102
## 2         5      8.324  1   35.778 2.211e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. What is the p-value from the Wald test for determining whether the parameter for the quadratic term is significantly different from 0? 3.93×10^{-8} Does it match the p-value from the LRT? *No* If not, why? *The LRT and the Wald test are two different tests for the same hypothesis. Both use approximate distributions for the test statistic under the null hypothesis where the approximation gets better as sample size increases. These tests have better agreement as sample size increases.*

Overdispersion

We should also check our final model for overdispersion. Overdispersion is present when the variability in the observations is more than we would expect for binomial data. A goodness-of-fit test can be used to determine whether overdispersion is present in the model (however it may also indicate other problems with the model). One test statistic that can be used for the goodness-of-fit test is the deviance of the model.

1. Using the output from `summary(cancer2.glm)`, what is the goodness-of-fit test statistic? *residual deviance*= 8.3238

The null hypothesis of this test is that the model fits the data well. Large values of the goodness-of-fit test indicate the model does not fit the data well. The reference distribution for this test is a chi-square distribution with degrees of freedom equal to the residual degrees of freedom. Here we will calculate the p-value for this test.

```
1-pchisq(8.3238,(8-3))
```

```
## [1] 0.1392702
```

2. Does there appear to be a problem with overdispersion in the model? *No, the p-value of this test is much larger than any significance level we would normally consider.*

3. If overdispersion is present in the model, one way to adjust the model for overdispersion is to multiply each of the standard errors by the square root of the estimated *overdispersion parameter*. If you set `family=quasibinomial` in the `glm()` function, the standard errors of each estimated parameter will be adjusted for overdispersion. How are the p-values for the Wald tests affected by this adjustment? *The p-values are all higher reflecting the extra variability associated with the estimated parameters due to overdispersion.*

```
cancerquas.glm=glm(cbind(Cases,Pop-Cases)~age+I(age*age),family=quasibinomial,data=minn)
summary(cancerquas.glm)
```

```
##
## Call:
## glm(formula = cbind(Cases, Pop - Cases) ~ age + I(age * age),
##     family = quasibinomial, data = minn)
##
## Deviance Residuals:
```

```
##        1         2         3         4         5         6         7         8
## -2.22316   0.98841   0.39489   0.69945  -0.40323  -0.98403   0.78740  -0.09141
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.329e+01  6.533e-01 -20.344 5.31e-06 ***
## age           1.628e-01  2.163e-02   7.527 0.000655 ***
## I(age * age) -8.234e-04  1.744e-04  -4.720 0.005242 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.354694)
##
##     Null deviance: 846.6311  on 7  degrees of freedom
## Residual deviance:   8.3238  on 5  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Again, you could also run this as...

```
cancerquas2.glm=glm(Cases/Pop~age+I(age*age),weights=Pop,family=quasibinomial,data=minn)
summary(cancerquas2.glm)
```

```
##
## Call:
## glm(formula = Cases/Pop ~ age + I(age * age), family = quasibinomial,
##     data = minn, weights = Pop)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -2.22316   0.98841   0.39489   0.69945  -0.40323  -0.98403   0.78740  -0.09141
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.329e+01  6.533e-01 -20.344 5.31e-06 ***
## age           1.628e-01  2.163e-02   7.527 0.000655 ***
## I(age * age) -8.234e-04  1.744e-04  -4.720 0.005242 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.354694)
##
##     Null deviance: 846.6311  on 7  degrees of freedom
## Residual deviance:   8.3238  on 5  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

4. The summary includes the estimated *overdispersion parameter*. What is this value? *1.355*