Report Assignment 2 - Group 4

Filip Muntean, mmi349 | 2663515 Marin Marian, mmn519 | 2698703
Matei Anton, man471 | 2789103

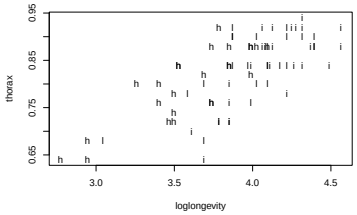
```
library(dplyr)
library(ggplot2)
library(GGally)
library(car)
```

Exercise 1

```
data <- read.delim("fruitflies.txt", sep = "")
data$loglongevity = log(data$longevity)
```

a)

```
plot(thorax~loglongevity, pch=unclass(activity))
```




```
activty = as.factor(data$activity)
ffaov = lm(loglongevity~activity, data=data)
anova(ffaov)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##            Df Sum Sq Mean Sq F value    Pr(>F)
## activity    2 3.6665  1.8333  19.421 1.798e-07 ***
## Residuals 72 6.7966  0.0944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova test, we do have that activity does influence the loglongevity of the specimens, as the p-value is way below the 0.05 threshold.

```r
summary(ffaov)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95531 -0.13338  0.02552  0.20891  0.49222
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.60212    0.06145  58.621  < 2e-16 ***
## activityisolated  0.51722    0.08690   5.952 8.82e-08 ***
## activitylow       0.39771    0.08690   4.577 1.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3072 on 72 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3324
## F-statistic: 19.42 on 2 and 72 DF,  p-value: 1.798e-07
```

```r
print("The estimated loglongevities for isolated, low, and high activity respectively, without taking th
```

```
## [1] "The estimated loglongevities for isolated, low, and high activity respectively, without taking
```

```r
print(c(3.6 + 0.52, 3.6 + 0.4, 3.6))
```

```
## [1] 4.12 4.00 3.60
```

It seems that the more sexually active the specimens are, the smaller the loglongevity.

b)

```r
ffaov = lm(loglongevity~thorax+activity)
anova(ffaov)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax     1 5.4322  5.4322 132.175 <2e-16 ***
## activity   2 2.1129  1.0565  25.705  4e-09 ***
## Residuals 71 2.9180  0.0411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even, when including thorax into the analysis as an explanatory variable, activity still influences the loglongevity.

```r
summary(ffaov)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4858 -0.1612  0.0104  0.1510  0.3574
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.21893    0.24865   4.902 5.79e-06 ***
## thorax             2.97899    0.30665   9.715 1.14e-14 ***
## activityisolated   0.40998    0.05839   7.021 1.07e-09 ***
## activitylow        0.28570    0.05849   4.885 6.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2027 on 71 degrees of freedom
## Multiple R-squared:  0.7211, Adjusted R-squared:  0.7093
## F-statistic:  61.2 on 3 and 71 DF,  p-value: < 2.2e-16
```

As the previous test, the longevity seems to be decreasing as sexual activity increases.

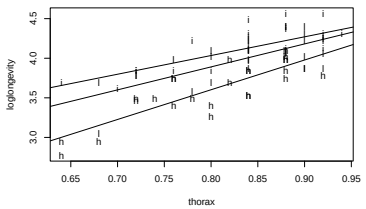
To get the loglongevities for the (global) average thorax length per activity group:

```r
c(predict(ffaov, data.frame(activity="isolated", thorax=mean(thorax))), predict(ffaov, data.frame(activi
```

```
##        1        1        1
## 4.085190 3.960910 3.675209
```

c)

```r
plot(loglongevity~thorax, pch=unclass(activity))
abline(lm(loglongevity~thorax, data=data[data$activity=="isolated",]))
abline(lm(loglongevity~thorax, data=data[data$activity=="low",]))
abline(lm(loglongevity~thorax, data=data[data$activity=="high",]))
```



It seems thorax and loglongevity are very much related as the bigger the thorax length, the higher the longevity. Also, there doesn't seem to be any indication that the lines wouldn't be parallel, so .

```r
ffaov = lm(loglongevity~activity*thorax)
anova(ffaov)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## activity         2 3.6665  1.8332 45.7687 2.228e-13 ***
## thorax           1 3.8786  3.8786 96.8327 9.020e-15 ***
## activity:thorax  2 0.1542  0.0771  1.9251    0.1536
## Residuals       69 2.7638  0.0401
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, from the test, there is no significant interaction between sexual activity and thorax length

d)

Even though non of the analysis is wrong (with or without the thorax length), the one which includes the thorax length is preferable as the thorax length explains much of the variance of loglongevity in our analysis.

##e)

```r
ffaov = lm(longevity~thorax+activity)
anova(ffaov)
```

```
## Analysis of Variance Table
##
## Response: longevity
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax     1 10959.3 10959.3 101.409 2.557e-15 ***
## activity   2  4966.7  2483.4  22.979 2.016e-08 ***
## Residuals 71  7673.0   108.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(ffaov)
```

```
##
## Call:
## lm(formula = longevity ~ thorax + activity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.688  -8.622  -1.176   6.790  26.605
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -67.375     12.750  -5.284 1.33e-06 ***
## thorax            132.618     15.725   8.434 2.62e-12 ***
## activityisolated   20.066      2.994   6.701 4.13e-09 ***
```

```
## activitylow        13.054      2.999   4.352 4.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 71 degrees of freedom
## Multiple R-squared:  0.6749, Adjusted R-squared:  0.6611
## F-statistic: 49.12 on 3 and 71 DF,  p-value: < 2.2e-16
```

Even though the test still holds without taking the log, now the influence from sexual activity is negative which is harder to interpret. Taking the log of the number of days was indeed a wise choice.

Exercise 2

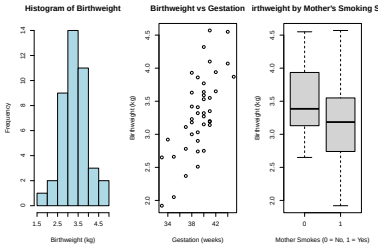
a)

```
birthweight <- read.csv("Birthweight.csv", header = TRUE)
summary(birthweight)
```

```
##        ID             Length       Birthweight       Headcirc      Gestation
##  Min.   :  27.0   Min.   :43.00   Min.   :1.920   Min.   :30.0   Min.   :33.00
##  1st Qu.: 537.2   1st Qu.:50.00   1st Qu.:2.940   1st Qu.:33.0   1st Qu.:38.00
##  Median : 821.0   Median :52.00   Median :3.295   Median :34.0   Median :39.50
##  Mean   : 894.1   Mean   :51.33   Mean   :3.313   Mean   :34.6   Mean   :39.19
##  3rd Qu.:1269.5   3rd Qu.:53.00   3rd Qu.:3.647   3rd Qu.:36.0   3rd Qu.:41.00
##  Max.   :1764.0   Max.   :58.00   Max.   :4.570   Max.   :39.0   Max.   :45.00
##      smoker            mage           mnocig          mheight          mppwt
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.000   Min.   :149.0   Min.   :45.00
##  1st Qu.:0.0000   1st Qu.:20.25   1st Qu.: 0.000   1st Qu.:161.0   1st Qu.:52.25
##  Median :1.0000   Median :24.00   Median : 4.500   Median :164.5   Median :57.00
##  Mean   :0.5238   Mean   :25.55   Mean   : 9.429   Mean   :164.5   Mean   :57.50
##  3rd Qu.:1.0000   3rd Qu.:29.00   3rd Qu.:15.750   3rd Qu.:169.5   3rd Qu.:62.00
##  Max.   :1.0000   Max.   :41.00   Max.   :50.000   Max.   :181.0   Max.   :78.00
##       fage          fedyrs          fnocig         fheight          lowbwt
##  Min.   :19.0   Min.   :10.00   Min.   : 0.00   Min.   :169.0   Min.   :0.0000
##  1st Qu.:23.0   1st Qu.:12.00   1st Qu.: 0.00   1st Qu.:175.2   1st Qu.:0.0000
##  Median :29.5   Median :14.00   Median :18.50   Median :180.5   Median :0.0000
##  Mean   :28.9   Mean   :13.67   Mean   :17.19   Mean   :180.5   Mean   :0.1429
##  3rd Qu.:32.0   3rd Qu.:16.00   3rd Qu.:25.00   3rd Qu.:184.8   3rd Qu.:0.0000
##  Max.   :46.0   Max.   :16.00   Max.   :50.00   Max.   :200.0   Max.   :1.0000
##      mage35
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.09524
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

First, to investigate the problems of potential and influence points, as well as colinearity, we will analyse the data in order to find any apparent relationships between variables or potential outliers.

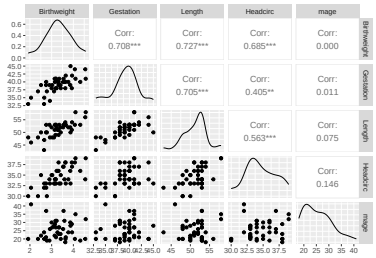
```r
par(mfrow=c(1,3))

hist(birthweight$Birthweight, main="Histogram of Birthweight", xlab="Birthweight (kg)", col="lightblue"
plot(birthweight$Gestation, birthweight$Birthweight, xlab="Gestation (weeks)", ylab="Birthweight (kg)",
boxplot(Birthweight ~ smoker, data=birthweight, main="Birthweight by Mother's Smoking Status", ylab="Bi
```




```r
#reset
par(mfrow=c(1,3))
```

```r
#fancy plot
selected_data <- birthweight[, c("Birthweight", "Gestation", "Length", "Headcirc", "mage")]
ggpairs(selected_data)
```

```
## plot: [1, 1] [==>-------------------------------------------------------------] 4% est: 0s
## plot: [1, 2] [====>-----------------------------------------------------------] 8% est: 1s
## plot: [1, 3] [=======>--------------------------------------------------------] 12% est: 1s
## plot: [1, 4] [========>-------------------------------------------------------] 16% est: 2s
## plot: [1, 5] [============>---------------------------------------------------] 20% est: 2s
## plot: [2, 1] [==============>-----------------------------------------------] 24% est: 1s
## plot: [2, 2] [================>---------------------------------------------] 28% est: 1s
## plot: [2, 3] [==================>-------------------------------------------] 32% est: 1s
## plot: [2, 4] [=====================>----------------------------------------] 36% est: 1s
## plot: [2, 5] [========================>-------------------------------------] 40% est: 1s
## plot: [3, 1] [==========================>-----------------------------------] 44% est: 1s
## plot: [3, 2] [=============================>--------------------------------] 48% est: 1s
## plot: [3, 3] [================================>-----------------------------] 52% est: 1s
## plot: [3, 4] [===================================>--------------------------] 56% est: 1s
## plot: [3, 5] [======================================>---------------------------] 60% est: 1s
## plot: [4, 1] [=========================================>----------------------] 64% est: 1s
## plot: [4, 2] [============================================>-------------------] 68% est: 1s
## plot: [4, 3] [===============================================>----------------] 72% est: 1s
## plot: [4, 4] [==================================================>-------------] 76% est: 0s
## plot: [4, 5] [=====================================================>----------] 80% est: 0s
## plot: [5, 1] [========================================================>-------] 84% est: 0s
## plot: [5, 2] [===========================================================>----] 88% est: 0s
## plot: [5, 3] [==============================================================>-] 92% est: 0s
## plot: [5, 4] [=================================================================>] 96% est: 0s
## plot: [5, 5] [==================================================================]100% est: 0s
```

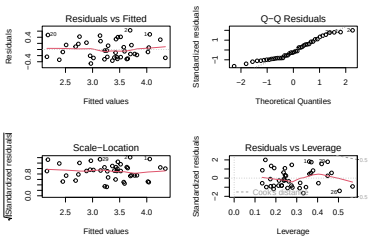



```
par(mfrow=c(1,1))
```

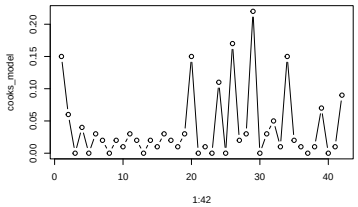
Through this analysis, we observe a birthweight distribution, with a bell-shaped shape, which suggests a normal distribution with potential outliers. Longer gestations may lead to more fetal growth, indicating a positive association between gestation duration and birthweight. Smoking during pregnancy may result in a lower median birthweight in children born to smokers compared to non-smokers.

Furthermore, we investigate the problem of potential and influence points as such:

```
model <- lm(Birthweight ~ Length + Headcirc + Gestation + mage + mnocig + mheight + mppwt + fage + fedy

par(mfrow=c(2,2))
plot(model)
```




```
# reset
par(mfrow=c(1,1))

cooks_model = round(cooks.distance(model),2)
plot(1:42, cooks_model,type="b")
```



Based on the Cooks model above, we can conclude that there are numerous influence points, while most 0.

Next, we investigate the colinearity:

```
vif(model)
```

```
##    Length  Headcirc Gestation      mage    mnocig   mheight     mppwt      fage    fedyrs
##  3.115295  1.835970  2.508132  4.028952  1.416515  3.110390  2.380129  4.517598  1.614641
##    fnocig   fheight
##  1.706549  1.619061
```

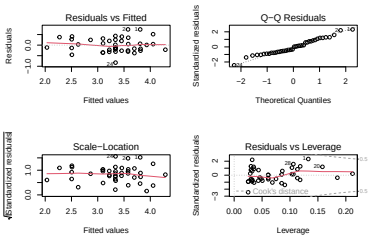
Based on these results, *length*, *mage*, *mheight* (mother's height), and *fage* (father's age) have have some level of colinearity since their VIF values are between 3 and 5, while the rest of the variables have a VIF level of less than 3, which suggests that these variables do not have a strong multicollinearity with the other variables. In conclusion, there does not seem to be a strong problem regarding multicolinearity

b)

```r
full_model <- lm(Birthweight ~ ., data = birthweight)

while (TRUE) {
    model_summary <- summary(full_model)
    p_values <- coef(model_summary)[, 4]  # Extracting p-values
    max_p_value <- max(p_values[p_values < 1])  # Ignoring p-value of the intercept
    if (max_p_value > 0.05) {
        # Remove the least significant variable
        least_significant_var <- names(which.max(p_values))
        formula <- as.formula(paste("Birthweight ~ . -", least_significant_var))
        full_model <- update(full_model, formula)
    } else {
        break
    }
}

par(mfrow=c(2,2))
summary(full_model)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc + Gestation, data = birthweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82889 -0.24763 -0.05136  0.25136  0.74352
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.44799    0.93936  -5.800 9.83e-07 ***
## Headcirc     0.11977    0.02449   4.891 1.77e-05 ***
## Gestation    0.11782    0.02223   5.299 4.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3441 on 39 degrees of freedom
## Multiple R-squared:  0.6911, Adjusted R-squared:  0.6753
## F-statistic: 43.63 on 2 and 39 DF,  p-value: 1.124e-10
```

```r
plot(full_model)
```




```
par(mfrow=c(1,1))
```

Based on the step-down method, the model has resulted in a model with two explanatory variables: Headcirc and Gestation. Both variables are statistically significant with p-values well below the 0.05 threshold, indicating they are strong predictors of Birthweight. The model seems to be a good fit with significant predictors, meeting linearity and homoscedasticity assumptions. However, further investigation is needed to ensure influential points are not unduly affecting the model.

c)

```
avg_headcirc <- mean(birthweight$Headcirc)
avg_gestation <- mean(birthweight$Gestation)

predict_data <- data.frame(Headcirc=avg_headcirc, Gestation=avg_gestation)

predictions <- predict(full_model, newdata=predict_data, interval="prediction", level=0.95)

conf_intervals <- predict(full_model, newdata=predict_data, interval="confidence", level=0.95)

predictions
```

```
##        fit      lwr      upr
## 1 3.312857 2.608563 4.017152
```

```
conf_intervals
```

```
##        fit      lwr      upr
## 1 3.312857 3.205453 3.420261
```

The CI ranges from 2.61 kg to 4.02 kg, indicating uncertainty in the mean birthweight estimate. The PI is narrower, ranging from 3.21 kg to 3.42 kg, specific to predicting an individual's birthweight. If collecting multiple samples, 95% of the intervals would contain the true population mean of birthweight. If predicting individual babies, 95% of their actual birthweights would fall within the CI.

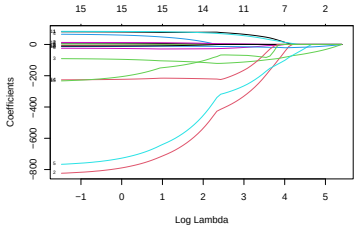
d

```
library(glmnet)

x=as.matrix(birthweight[,-1]) #remove the response variable
y=as.double(as.matrix(birthweight[,1])) #only the response variable

train=sample(1:nrow(x),0.67*nrow(x)) # train by using 2/3 of the data
x.train=x[train,]
```

```
y.train=y[train] # data to train
x.test=x[-train,]
y.test=y[-train] # data to test the prediction quality

lasso.mod=glmnet(x.train,y.train,alpha=1)
cv.lasso=cv.glmnet(x.train,y.train,alpha=1,type.measure='mse')
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per fold
```

```
plot(lasso.mod,label=T,xvar="lambda") #have a look at the lasso path
plot(cv.lasso$glmnet.fit,xvar="lambda",label=T)
```




```
lambda.min=cv.lasso$lambda.min; lambda.1se=cv.lasso$lambda.1se
coef(lasso.mod,s=cv.lasso$lambda.min) #beta's for the best lambda
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 3180.32313
## Length         .
## Birthweight    .
## Headcirc     -51.67671
## Gestation      .
## smoker         .
## mage           .
## mnocig         .
## mheight        .
## mppwt          .
## fage         -15.87569
## fedyrs         .
## fnocig         .
## fheight        .
## lowbwt         .
## mage35         .
```

```
y.pred=predict(lasso.mod,s=lambda.min,newx=x.test) #predict for test
mse.lasso=mean((y.test-y.pred)^2) #mse for the predicted test rows
```

Step-down model predictors: Headcirc, Gestation LASSO model predictors: Length, smoker, mnocig, mage35 The ositive coefficients for Length indicates that as these variables increase, birth weight is expected to increase. Negative coefficients (like for smoker, mnocig, mage35) imply the opposite effect.

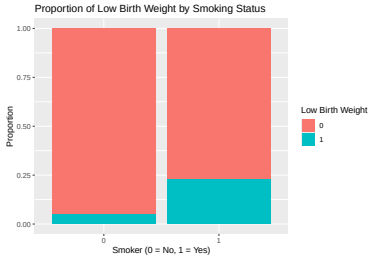
The two methods selected different features as important predictors, Gestation was the only one present in both. The main differences between the two approaches is that step-down prioritizes statistical significance in a classical sense, while LASSO emphasizes predictive power and can handle more correlated predictors.

e

```
data <- birthweight[, c("Gestation", "smoker", "mage35", "lowbwt")]

contingency_table_smoker <- table(data$smoker, data$lowbwt)
print(contingency_table_smoker)
```

```
##
##       0  1
##   0 19  1
##   1 17  5
```

```
ggplot(data, aes(x = factor(smoker), fill = factor(lowbwt))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Low Birth Weight by Smoking Status",
       x = "Smoker (0 = No, 1 = Yes)",
       y = "Proportion",
       fill = "Low Birth Weight")
```

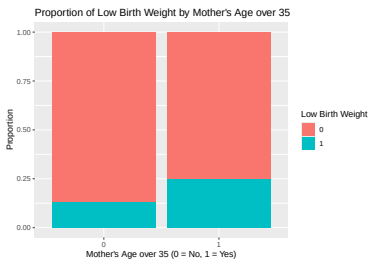


There are far more non-smoking mothers with babies of normal birth weight (19) compared to smoking mothers with babies of normal birth weight (1). A relatively higher number of smoking mothers have babies with low birth weight (5) compared to non-smoking mothers (1). This suggests that smoking might be associated with an increased likelihood of low birth weight.

```
contingency_table_mage35 <- table(data$mage35, data$lowbwt)
print(contingency_table_mage35)
```

```
##
##       0  1
##   0 33  5
##   1  3  1
```

```
ggplot(data, aes(x = factor(mage35), fill = factor(lowbwt))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Low Birth Weight by Mother's Age over 35",
       x = "Mother's Age over 35 (0 = No, 1 = Yes)",
       y = "Proportion",
       fill = "Low Birth Weight")
```



There's a higher number of mothers younger than 35 with babies of normal

birth weight (33) compared to those 35 years or older with babies of normal birth weight (3). But it seems less mothers 35 or older have babies with low birth weight (1) compared to those younger than 35 (5). This suggests that the mother's age (over 35) isn't associated as much as smoking with the risk of having a baby with low birth weight.

f

```
model <- glm(lowbwt ~ Gestation + smoker + mage35, data = data, family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = lowbwt ~ Gestation + smoker + mage35, family = binomial,
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  48.9920    22.5659   2.171   0.0299 *
## Gestation    -1.4633     0.6700  -2.184   0.0290 *
## smoker        5.4495     3.3567   1.623   0.1045
## mage35        0.3223     4.3751   0.074   0.9413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.450  on 41  degrees of freedom
## Residual deviance: 11.803  on 38  degrees of freedom
## AIC: 19.803
##
## Number of Fisher Scoring iterations: 8
```

The coefficient for Gestation is -1.4633, with a standard error of 0.6700. This suggests that for each additional week of gestation, the log odds of having a low birth weight baby decrease by 1.4633 units. The p-value (0.0290) indicates that Gestation is statistically significant in predicting low birth weight. Furthermore, the coefficient for smoker is 5.4495, with a standard error of 3.3567. Although the coefficient is positive, suggesting that smoking mothers are associated with increased odds of low birth weight, the p-value (0.1045) is greater than the typical threshold of 0.05 for statistical significance. This suggests that the association between smoking and low birth weight may not be statistically significant in this model at the conventional level. Finally, the coefficient for mage35 is 0.3223, with a standard error of 4.3751. The p-value (0.9413) is much greater than 0.05, indicating that maternal age over 35 is not statistically significant in predicting low birth weight in this model.

Overall, gestation appears to be a significant predictor of low birth weight, with longer gestation associated with lower odds of low birth weight. However, smoking and maternal age over 35 do not appear to have statistically significant associations with low birth weight in this model.

g

```
model_interaction <- glm(lowbwt ~ Gestation*mage35 + Gestation*smoker, data = data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(model_interaction)
```

```
##
## Call:
## glm(formula = lowbwt ~ Gestation * mage35 + Gestation * smoker,
##     family = binomial, data = data)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1244.095 310252.325   0.004    0.997
## Gestation            -37.135   9248.031  -0.004    0.997
## mage35               310.034 943745.834   0.000    1.000
## smoker             -1199.824 310252.326  -0.004    0.997
## Gestation:mage35      -8.239  24325.474   0.000    1.000
## Gestation:smoker      35.939   9248.031   0.004    0.997
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.4498  on 41  degrees of freedom
## Residual deviance:  9.9445  on 36  degrees of freedom
## AIC: 21.944
##
## Number of Fisher Scoring iterations: 23
```

The high p-values for both interaction terms suggest that the effect of gestation on low birth weight risk likely does not differ significantly between mothers over and under 35 or between smoking and non-smoking mothers. The lack of significance with the main effects for the mage351 and smoker1 variables further suggests those predictors may not play a strong role in explaining low birth weight when the interactions are included.

The AIC for this model is 21.944. Therefore, the resulting model would be the logistic regression model without interactions, as it has a lower AIC.

h

```r
new_data <- expand.grid(
  Gestation = 40,
  smoker = c(0, 1),
  mage35 = c(0, 1)
)

probabilities <- predict(model, newdata = new_data, type = "response")

result <- cbind(new_data, Probability = probabilities)
print(result)
```

```
##   Gestation smoker mage35  Probability
## 1        40      0      0 7.199185e-05
## 2        40      1      0 1.647403e-02
## 3        40      0      1 9.936346e-05
## 4        40      1      1 2.259660e-02
```

The probabilities are consistently higher when smoker = 1 (smoking mothers) compared to smoker = 0 (non-smoking mothers), regardless of mother's age. This aligns with the expectation that smoking increases the

risk of low birth weight. The probabilities are slightly lower when mage35 = 1 (mothers over 35) compared to mage35 = 0. This suggests a potential weak effect where older mothers might have a slightly decreased risk of low birth weight in this model. However, the differences are quite small.

i

```
chi_sq_smoker <- chisq.test(contingency_table_smoker)
```

```
## Warning in chisq.test(contingency_table_smoker): Chi-squared approximation may be incorrect
```

```
chi_sq_mage35 <- chisq.test(contingency_table_mage35)
```

```
## Warning in chisq.test(contingency_table_mage35): Chi-squared approximation may be incorrect
```

```
print(chi_sq_smoker)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table_smoker
## X-squared = 1.4358, df = 1, p-value = 0.2308
```

```
print(chi_sq_mage35)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table_mage35
## X-squared = 2.7398e-31, df = 1, p-value = 1
```

Given the high p-values (> 0.05) in both tests, we fail to reject the null hypothesis, indicating that there is insufficient evidence to conclude that there is a significant association between maternal smoking status or maternal age group over 35 and low birth weight based on these tests.

Comparing this approach to logistic regression, the chi-squared test is straightforward to implement and interpret, especially when dealing with categorical variables. The disadvantage is that the chi-squared test can only assess associations between categorical variables. If there are continuous predictors or interactions to consider, logistic regression would be more appropriate. Moreover, the chi-squared test does not provide quantification of the strength of the association or estimate the effect size, unlike logistic regression which provides odds ratios.

Exercise 3

a)

```
awards <- read.table("awards.txt", sep = "")

poisson_model <- glm(num_awards ~ factor(prog), family ="poisson", data = awards)

# Summary of the model to check coefficients
summary(poisson_model)
```

```
##
## Call:
## glm(formula = num_awards ~ factor(prog), family = "poisson",
##     data = awards)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.5486     0.1961  -2.797  0.00515 **
## factor(prog)2   0.7068     0.2158   3.275  0.00106 **
## factor(prog)3   0.4432     0.2463   1.799  0.07199 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 216.10  on 197  degrees of freedom
## AIC: 512.42
##
## Number of Fisher Scoring iterations: 5
```

```r
# Predict the number of awards for each program type
predicted_awards <- predict(poisson_model, newdata = data.frame(prog = c(1, 2, 3)), type = "response")

# Display the predicted awards
predicted_awards
```

```
##         1         2         3
## 0.5777778 1.1714286 0.9000000
```

The intercept represents the log count of awards for the baseline group (vocational program), with an estimated coefficient of -0.5486. Program Type 2 (general program) has a significant coefficient of 0.7068, indicating a higher count of awards compared to the vocational program. Program Type 3 (academic program) has a coefficient of 0.4432, suggesting an association with a higher count of awards compared to the vocational program, although less certain than for the general program. Based on the estimated count, it appears that the program type 2 (general) is associated with the highest count of awards, followed by program type 3 (the academic program), and finally program type 1 (the vocational program).

b)

In the situation described in part (a), the number of awards can be considered as the outcome variable, and the type of program (prog) is the independent categorical variable with three levels. Although the number of awards is a count variable, if we are only interested in comparing the medians of the number of awards across different program types without assuming a specific distribution for the counts, the Kruskal-Wallis test could be used.

```r
kruskal.test(num_awards ~ factor(prog), data = awards)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  num_awards by factor(prog)
## Kruskal-Wallis chi-squared = 10.755, df = 2, p-value = 0.00462
```

The Kruskal-Wallis test rejects the null hypothesis (0.0046), indicating a significant difference in award distribution between two program types, even though it doesn't identify specific groups within the dataset.

c)

```
new_model <- glm(num_awards ~ factor(prog) * math, family = "poisson", data = awards)

# Summary to check coefficients
summary(new_model)
```

```
##
## Call:
## glm(formula = num_awards ~ factor(prog) * math, family = "poisson",
##     data = awards)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.578440   1.391426  -1.134    0.257
## factor(prog)2      -1.061226   1.534523  -0.692    0.489
## factor(prog)3       0.962144   1.635965   0.588    0.556
## math                0.020365   0.026950   0.756    0.450
## factor(prog)2:math  0.027437   0.028969   0.947    0.344
## factor(prog)3:math -0.009441   0.032396  -0.291    0.771
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 194.36  on 194  degrees of freedom
## AIC: 496.67
##
## Number of Fisher Scoring iterations: 5
```

```
new_data <- data.frame(prog = factor(c(1, 2, 3)), math = c(56, 56, 56))
predictions <- predict(new_model, newdata = new_data, type = "response")

predictions
```

```
##         1         2         3
## 0.6453249 1.0379350 0.9954753
```

The data reveals that the log count of awards for the baseline program (program 1) is -1.578440 when math score is zero. Factors 2 and 3 represent the log difference in awards between general and vocational programs. The predictions for a math score of 56 for each program type are given at the end of the output:

- Program 1 (vocational): 0.6453249 awards
- Program 2 (general): 1.0379350 awards
- Program 3 (academic): 0.9954753 awards

The predicted number of awards is highest for Program 2 (general) when the math score is 56. Therefore, according to this model, the general program is the best for maximizing the number of awards when a student has a math score of 56. The model predicts that students in general programs earn more awards with a math score of 56 than those in vocational or academic programs, despite no statistically significant interaction effects, possibly due to program type or other unaccounted factors.