Assignment 2 - Group 9

Lukas Unruh, Teije Langelaan, Gidon Quint

11 maart, 2024

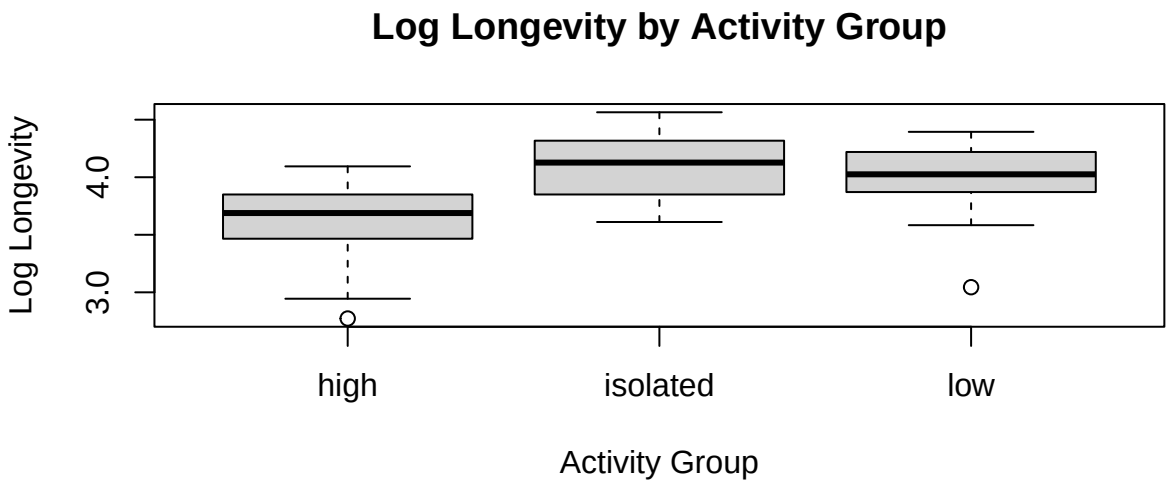
We used the library lme4, glmnet, Matrix, car and ggplot2 with options(digits=3) and a seed of 333*

Exercise 1

```
fruitfly_data=read.table(file="fruitflies.txt",header=TRUE)
fruitfly_data$loglongevity <- log(fruitfly_data$longevity)
```

a)

```
boxplot(loglongevity ~ activity, data = fruitfly_data,
        xlab = "Activity Group", ylab = "Log Longevity",
        main = "Log Longevity by Activity Group")
```



We perform ANOVA to test whether sexual activity influences longevity. ANOVA does assume normality, however, our TA mentioned that we do not need to check for normality.

- H_0: The mean loglongevity is the same across all activity groups.

- H_1: The mean loglongevity is not the same across all activity groups.

```
longaovA <- lm(loglongevity ~ activity, data = fruitfly_data); anova(longaovA)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## activity   2   3.67   1.833    19.4 1.8e-07 ***
## Residuals 72   6.80   0.094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since we get a p-value of 1.8e-7, we reject H_0 and conclude that sexual activity influences longevity.

The estimated longevities for the three conditions are (asked the TA and we assume normality, therefore checking the means):

```
tapply(fruitfly_data$longevity, fruitfly_data$activity, mean)
```

```
##     high isolated      low
##     38.7     63.6     56.8
```

From the values you can see that the mean of "high" is a lot lower than the mean of the two other groups. This suggests that increased sexual activity may be associated with decreased longevity in fruit flies.

b)

- H_0: The mean loglongevity is the same across all activity groups after accounting for the effect of thorax.
- H_1: The mean loglongevity is not the same across all activity groups after accounting for the effect of thorax.

```
longaovB <- lm(loglongevity ~ thorax + activity, data = fruitfly_data)
anova(longaovB)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value Pr(>F)
## thorax     1   5.43    5.43   132.2 <2e-16 ***
## activity   2   2.11    1.06    25.7  4e-09 ***
## Residuals 71   2.92    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since we get a p-value of 4e-9, we reject H_0 and conclude that sexual activity influences longevity.

```r
summary(longaovB)$coefficients
```

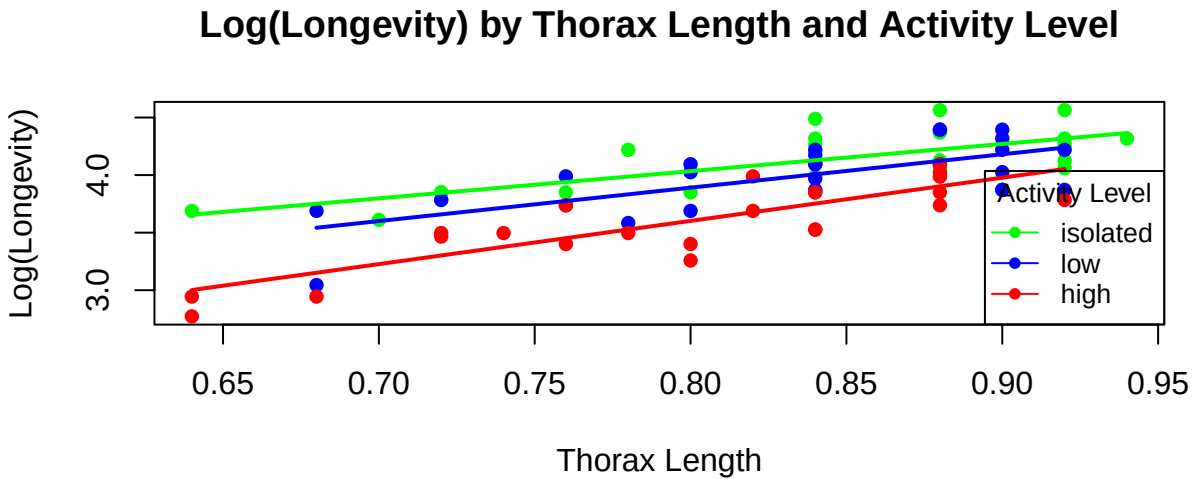
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.219     0.2486    4.90 5.79e-06
## thorax               2.979     0.3067    9.71 1.14e-14
## activityisolated     0.410     0.0584    7.02 1.07e-09
## activitylow          0.286     0.0585    4.88 6.18e-06
```

Activityisolated and activitylow are both positive, with isolated having a larger coefficient than low. Showing that compared to high sexual activity, being isolated is associated with an increase in log-longevity (which translates to an increase in actual longevity), while low sexual activity is also associated with an increase in log-longevity but to a lesser extent than isolation. Therefore, sexual activity appears to decrease longevity compared to no activity, with higher activity decreasing it more.

```r
average_thorax = mean(fruitfly_data$thorax)
new_data = expand.grid(activity = unique(fruitfly_data$activity), thorax = average_thorax)
new_data$predicted_loglongevity = predict(longaovB, newdata = new_data)
new_data$predicted_longevity = exp(new_data$predicted_loglongevity)
print(new_data)
```

```
##   activity thorax predicted_loglongevity predicted_longevity
## 1 isolated  0.825                  4.09                59.5
## 2      low  0.825                  3.96                52.5
## 3     high  0.825                  3.68                39.5
```

c)



The scatter plot with regression lines shows a positive relationship between thorax length and loglongevity, but no clear interaction clear difference in this effect between sexual activity groups.

We perform ANCOVA:

- H_0: There is no interaction effect between activity and thorax on loglongevity.

- H_1: There is an interaction effect between activity and thorax on loglongevity.

```
interaction_model = lm(loglongevity ~ thorax * activity, data = fruitfly_data)
anova(interaction_model)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## thorax           1   5.43    5.43  135.62 < 2e-16 ***
## activity         2   2.11    1.06   26.38 3.1e-09 ***
## thorax:activity  2   0.15    0.08    1.93    0.15
## Residuals       69   2.76    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(interaction_model)$coefficients
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.598      0.419    1.43 1.58e-01
## thorax                      3.755      0.522    7.20 5.78e-10
## activityisolated            1.546      0.584    2.65 1.01e-02
## activitylow                 0.972      0.642    1.51 1.35e-01
## thorax:activityisolated    -1.393      0.712   -1.96 5.45e-02
## thorax:activitylow         -0.854      0.779   -1.10 2.77e-01
```

The interaction between thorax length and sexual activity is non significant p-value of 0.15. The interaction between thorax length and isolated sexual activity is marginally (p-value = 0.055) significant, indicating a possible difference in this effect between flies with no sexual activity and those with high sexual activity

d)

```
model_summary <- summary(longaovA)
adjusted_r_squared <- model_summary$adj.r.squared
cat("R-adjusted model A: ", adjusted_r_squared)
```

```
## R-adjusted model A:  0.332
```

```
model_summary <- summary(longaovB)
adjusted_r_squared <- model_summary$adj.r.squared
cat("\nR-adjusted model B: ", adjusted_r_squared)
```

```
##
## R-adjusted model B:  0.709
```

Model A without thorax interaction has a R-squared = 0.332, model B with thorax interaction has a R-squard = 0.709. Therefore, we should prefer model B. Additionally, model B is model A with an added significant variable namely thorax. However, none of the analyses are wrong.

e)

- H_0: There is no difference in the mean longevity among different activity levels after accounting for thorax length.
- H_1: There is a difference in the mean longevity among different activity levels after accounting for thorax length.

```
fruitfly_data$activity = as.factor(fruitfly_data$activity)
ancova_model = lm(longevity ~ thorax + activity, data = fruitfly_data)
summary(ancova_model)$coefficients
```

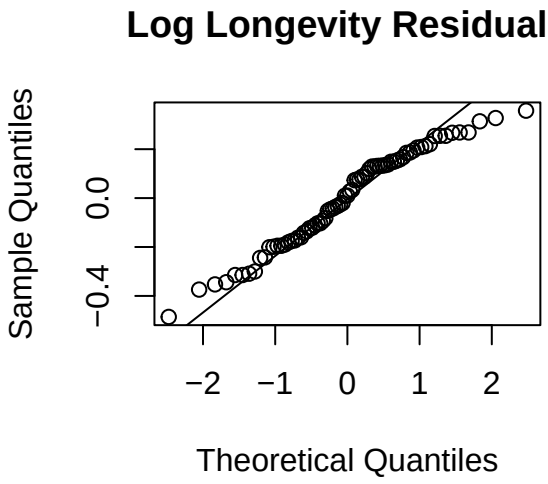
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -67.4      12.75   -5.28 1.33e-06
## thorax              132.6      15.72    8.43 2.62e-12
## activityisolated     20.1       2.99    6.70 4.13e-09
## activitylow          13.1       3.00    4.35 4.43e-05
```

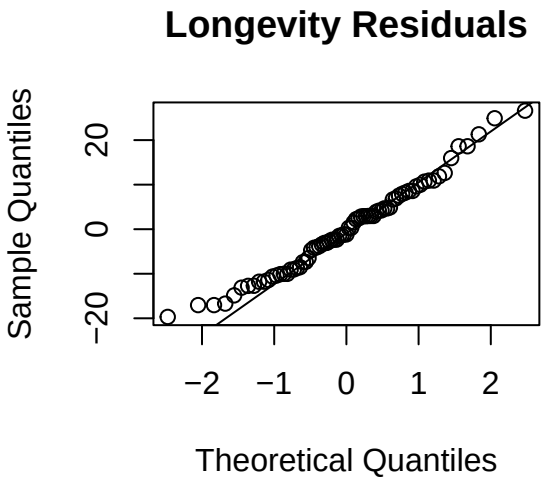
```
anova(ancova_model)
```

```
## Analysis of Variance Table
##
## Response: longevity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## thorax     1  10959   10959     101 2.6e-15 ***
## activity   2   4967    2483      23 2.0e-08 ***
## Residuals 71   7673     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(ancova_model, test = "F")
```

```
## Single term deletions
##
## Model:
## longevity ~ thorax + activity
##          Df Sum of Sq   RSS AIC F value  Pr(>F)
## <none>                 7673 355
## thorax    1      7687 15360 405    71.1 2.6e-12 ***
## activity  2      4967 12640 389    23.0 2.0e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



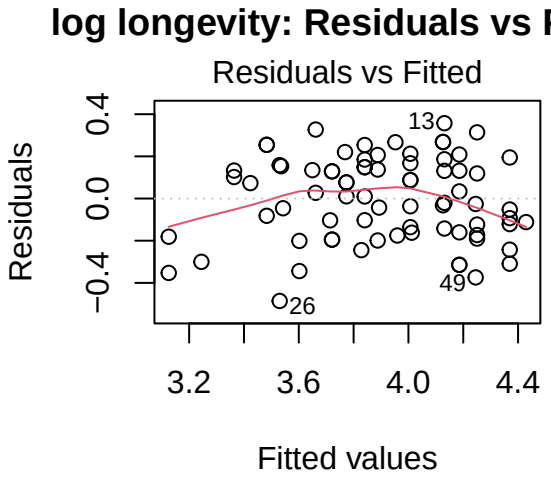


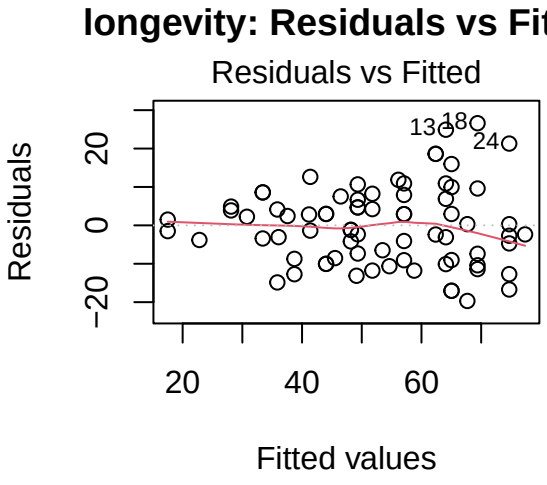

```
shapB <- shapiro.test(residuals(longaovB))
shapanc <-shapiro.test(residuals(ancova_model))
cat("p-value Log Longevity", shapB$p.value)
```

```
## p-value Log Longevity 0.0575
```

```
cat("\np-value Longevity", shapanc$p.value)
```

```
##
## p-value Longevity 0.318
```





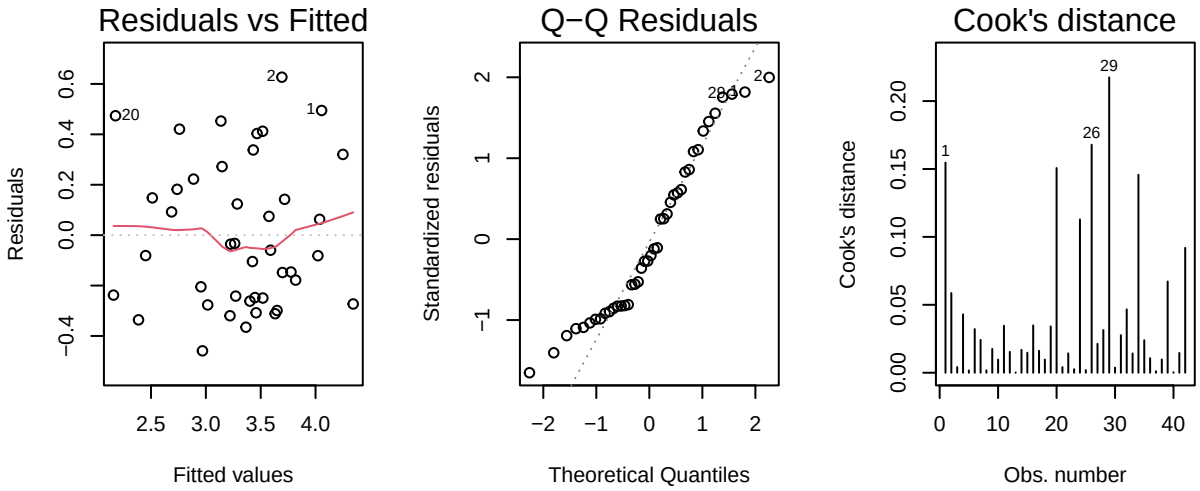
The ANCOVA model using raw longevity data shows a slightly lower adjusted R-squared than the model with log-transformed longevity, hinting that logarithmic transformation may offer a tighter fit. However, this transformation shifts the sample's distribution away from normality, evident from QQ-plots and a

Shapiro-Wilk test p-value of 0.058. Before concluding, it's noteworthy that the log transformation appears to marginally equalize the variance of residuals. Despite this, the deviation from normal distribution in the log-transformed model leads us to prefer the raw longevity model. We believe it is unwise to convert variables before concluding the original variable is problematic.

Exercise 2

```
birthweight_data=read.csv(file="Birthweight.csv",header=TRUE)
```

a)



Upon inspection of plots and Cook's distances, we conclude that there are no influence points in the data set, as no observation has a Cook's distance greater than 1.

```
vif_values <- vif(TJ_BW)
print(vif_values)
```

```
##    Length  Headcirc Gestation      mage    mnocig   mheight     mppwt      fage
##      3.12      1.84      2.51      4.03      1.42      3.11      2.38      4.52
##    fedyrs    fnocig   fheight
##      1.61      1.71      1.62
```

```
cor_fage_mage <- cor(birthweight_data$mage, birthweight_data$fage)
print(cor_fage_mage)
```

```
## [1] 0.807
```

Upon inspection of correlation matrix, we note that variables mage and fage appear to be highly correlated (0.807), however their VIF values are under 5, therefore we conclude that the model has no problem of collinearity.

b)

```
final_model = lm(Birthweight ~ Headcirc + Gestation, data=birthweight_data)
summary(final_model)$coefficients
```

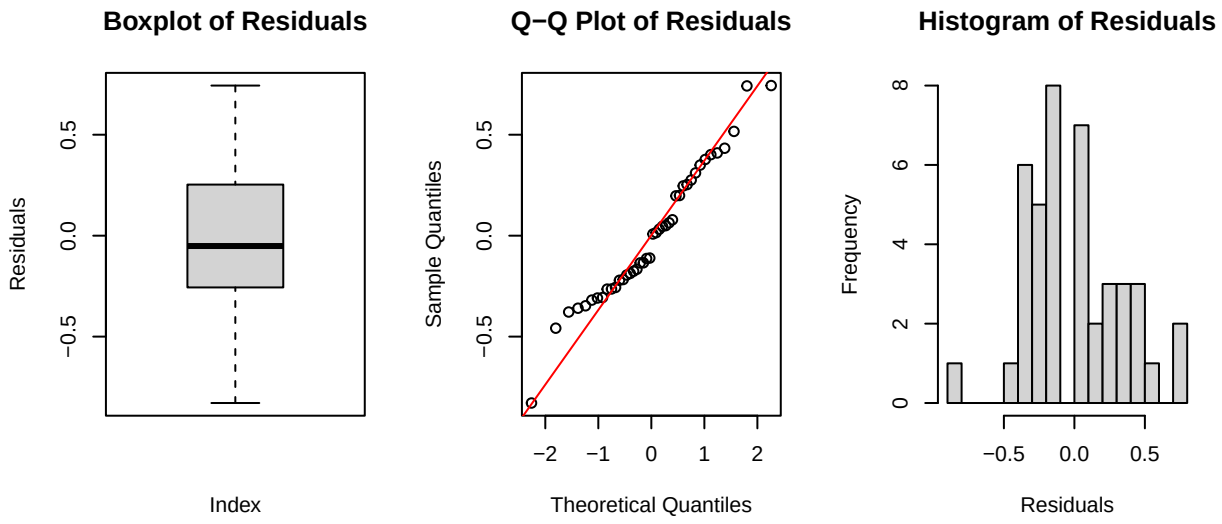
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.448     0.9394   -5.80 9.83e-07
## Headcirc       0.120     0.0245    4.89 1.77e-05
## Gestation      0.118     0.0222    5.30 4.85e-06
```

Order of removal: fage, mheight, fedyrs, fnocig, mnocig, Length, fheight, mage, mppwt

- H_0: The residuals from the linear regression model follow a normal distribution.
- H_1: The residuals from the linear regression model do not follow a normal distribution.

```
residuals = residuals(final_model)
test_results_sw = shapiro.test(residuals)
cat("The p-value of the Shapiro-Wilk test is:", test_results_sw$p.value)
```

```
## The p-value of the Shapiro-Wilk test is: 0.295
```



We do not reject H_0 so we cannot conclude that the residuals are not normally distribtuted, however after taking a good look at the QQ plot and the histoplot we conclude that our normality of residual distribution is questionable.

c)

```
fitted_values = fitted(final_model)
average_values = data.frame(
  Headcirc = mean(birthweight_data$Headcirc, na.rm = TRUE),
```

```
  Gestation = mean(birthweight_data$Gestation, na.rm = TRUE)
)
confidence_intervals = predict(final_model, newdata = average_values, interval =
                               "confidence", level = 0.95)
prediction_intervals = predict(final_model, newdata = average_values, interval =
                               "prediction", level = 0.95)
print("95% Confidence Intervals for the Mean Response:")
```

```
## [1] "95% Confidence Intervals for the Mean Response:"
```

```
print(confidence_intervals)
```

```
##    fit  lwr  upr
## 1 3.31 3.21 3.42
```

```
print("95% Prediction Intervals for a New Observation:")
```

```
## [1] "95% Prediction Intervals for a New Observation:"
```

```
print(prediction_intervals)
```

```
##    fit  lwr  upr
## 1 3.31 2.61 4.02
```

The output indicates that for babies with average head circumference and gestation, the model predicts a mean Birthweight of 3.31 kg, with a 95% confidence interval of 3.21 to 3.42 kg for the mean and a wider 95% prediction interval of 2.61 to 4.02 kg for individual observations, reflecting the increased uncertainty in predicting specific outcomes.

d)

```
par(mfrow = c(1,2))
x=as.matrix(birthweight_data[,-3]) # remove the response variable
y=as.double(as.matrix(birthweight_data[,3])) # 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')
lambda.min=cv.lasso$lambda.min; lambda.1se=cv.lasso$lambda.1se
coef(lasso.mod,s=cv.lasso$lambda.1se) # beta's for the best lambda
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) -1.2754
## ID           .
## Length       0.0119
## Headcirc     0.0801
```

```
## Gestation    0.0309
## smoker       .
## mage         .
## mnocig       .
## mheight      .
## mppwt        .
## fage         .
## fedyrs       .
## fnocig       .
## fheight      .
## lowbwt      -0.2693
## mage35       .
```

```
y.pred=predict(lasso.mod,s=lambda.1se,newx=x.test) # predict for test
mse.lasso=mean((y.test-y.pred)^2) # mse for the predicted test rows
new_model = lm(Birthweight ~ Length + Headcirc + Gestation + lowbwt, data=birthweight_data)
summary(new_model)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.6516     1.2967   -3.59 0.000963
## Length        0.0369     0.0282    1.31 0.197650
## Headcirc      0.0936     0.0263    3.56 0.001036
## Gestation     0.0733     0.0289    2.54 0.015354
## lowbwt       -0.3114     0.1951   -1.60 0.119007
```

Upon inspection of the summary tables, although the LASSO model has a slightly higher R-squared value, we prefer the step-down model, because the two added variables in the LASSO model are insignificant and it does not make sense to add them in respect to complexity.

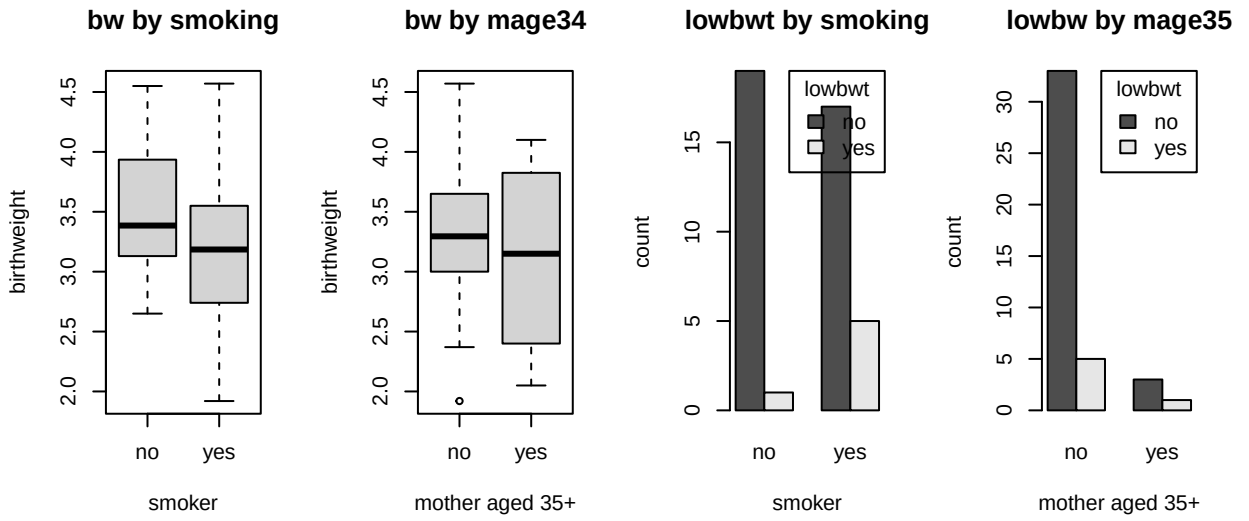
e)

Effect of whether mothers smokes on low birth weight

- H_0: Smoking status of the mother has no effect on the likelihood of the baby being born with low birth weight.
- H_1: Babies born to smoking mothers are more likely to have low birth weight compared to those born to non-smoking mothers.

Effect of mother's age over 35 on low birth weight

- H_0: There is no difference in the likelihood of having a baby with low birthweight between mothers over the age of 35 and those 35 or younger.
- H_1: Babies born to mothers over the age of 35 are more likely to have low birthweight compared to those born to younger mothers.




```
table(birthweight_data$smoker, birthweight_data$lowbwt)
```

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

```
table(birthweight_data$mage35, birthweight_data$lowbwt)
```

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

The boxplot do not show a clear difference in birthweight resulting from the varibales 'smoker' and 'mage35'. Both the bar graph, and contingency table do suggests a potential influence of smoking on low birthweight. The limited sample sizes <5 for some experimental units, hinder us in getting reliable chi-squared results on these possible effects.

f)

- H_0: The length of gestation (in weeks) has no effect on the likelihood of the baby being born with low birthweight.
- H_1: Shorter gestation periods are associated with a higher likelihood of the baby being born with low birthweight.

```
TJ_LOWBWT <- glm(lowbwt ~ Gestation + mage35 + smoker, data = birthweight_data, family = binomial)
drop1(TJ_LOWBWT, test="Chisq")
```

```
## Single term deletions
##
```

```
## Model:
## lowbwt ~ Gestation + mage35 + smoker
##           Df Deviance  AIC   LRT Pr(>Chi)
## <none>           11.8 19.8
## Gestation  1     31.4 37.4 19.58  9.6e-06 ***
## mage35     1     11.8 17.8  0.01    0.941
## smoker     1     17.6 23.6  5.80    0.016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TJ_LOWBWT2 <- glm(lowbwt ~ Gestation + smoker, data = birthweight_data, family = binomial)
drop1(TJ_LOWBWT2, test="Chisq")
```

```
## Single term deletions
##
## Model:
## lowbwt ~ Gestation + smoker
##           Df Deviance  AIC   LRT Pr(>Chi)
## <none>           11.8 17.8
## Gestation  1     31.5 35.5 19.71    9e-06 ***
## smoker     1     18.0 22.0  6.22    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef(TJ_LOWBWT2))
```

```
## (Intercept)   Gestation      smoker
##    2.24e+21    2.30e-01    2.41e+02
```

In analyzing factors affecting low birthweight (lowbwt), 'mage35' was found not to significantly influence lowbwt (p-value = 0.941), suggesting maternal age over 35 doesn't notably alter the odds of low birthweight. For 'Gestation', each additional week reduces the odds of lowbwt by approximately 77% (odds ratio = 0.23, p-value = 9e-06). 'Smoker' status increases the odds of lowbwt by 241 times (odds ratio = 241, p-value = 0.013), pointing to a substantial risk increase due to smoking. These results highlight the importance of gestation duration and the detrimental impact of smoking on birthweight.

g)

Interaction gestation and smoking - H_0: There is no interaction effect between gestation length and smoking status on the likelihood of low birthweight.

- H_1: There is an interaction effect between gestation length and smoking status on the likelihood of low birthweight.

Interaction gestation and mage35 - H_0: There is no interaction effect between gestation length and maternal age over 35 on the likelihood of low birthweight

- H_1: There is an interaction effect between gestation length and maternal age over 35 on the likelihood of low birthweight

```r
model_gestation_smoker = glm(lowbwt ~ Gestation * smoker, data = birthweight_data, family = binomial)
anova_gestation_smoker <- anova(model_gestation_smoker, test="Chisq")
p_value_gestation_smoker <- anova_gestation_smoker["Gestation:smoker", "Pr(>Chi)"]
model_gestation_mage35 = glm(lowbwt ~ Gestation * mage35, data = birthweight_data, family = binomial)
anova_gestation_mage35 <- anova(model_gestation_mage35, test="Chisq")
p_value_gestation_mage35 <- anova_gestation_mage35["Gestation:mage35", "Pr(>Chi)"]
print(paste("P-value for Gestation:smoker interaction:", p_value_gestation_smoker))
```

```
## [1] "P-value for Gestation:smoker interaction: 0.19804478568693"
```

```r
print(paste("P-value for Gestation:mage35 interaction:", p_value_gestation_mage35))
```

```
## [1] "P-value for Gestation:mage35 interaction: 0.419271591904811"
```

Both interaction terms—'Gestation:smoker' and 'Gestation:mage35'—do not demonstrate statistically significant effects on low birthweight. The p-value for 'Gestation:smoker' is 0.198, and for 'Gestation:mage35', it is 0.42. These p-values indicate that the interactions between gestation length and either smoking status or maternal age over 35 do not significantly alter the risk of low birthweight in these models. We therefore choose the model without interaction terms.

h)

```r
model_g <- glm(lowbwt ~ Gestation + smoker, data = birthweight_data, family="binomial")
new_data <- data.frame(Gestation = c(40, 40), smoker = c(0, 1))
probabilities <- predict(model_g, newdata = new_data, type = "response")
print(probabilities)
```

```
##        1        2
## 6.97e-05 1.65e-02
```

For babies born at 40 weeks of gestation, the predicted probability of low birth weight is <0.001% for non-smokers and 0.02% for smokers, highlighting the significant impact of maternal smoking on birth weight outcomes. This underscores the importance of smoking cessation interventions during pregnancy to mitigate the risk of low birth weight.

i)

```r
contingency_table <- table(birthweight_data$lowbwt, birthweight_data$smoker)
chi_squared_result <- chisq.test(contingency_table)
print(chi_squared_result)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 1, df = 1, p-value = 0.2
```

Applying a contingency table test, such as the chi-squared test, to address the questions is a valid approach with certain limitations. A major advantage of this method is its straightforwardness and its capacity to explore associations between categorical variables without assuming a specific relationship form. However, one significant disadvantage is its unreliability with very small sample sizes (e.g., <5 in some groups), which can make chi-squared results misleading, and is this case for this dataset. Additionally, while chi-squared tests are simple and useful for initial explorations, they lack the ability to adjust for other variables or to interpret effects in terms of odds ratios, unlike logistic regression. This simplicity might lead to overlooking complex nuances in the data.

Exercise 3

```
awards_data=read.table(file="awards.txt",header=TRUE)
```

a)

- H_0: The type of program (vocational, general, academic) does not influence the number of awards earned by students.
- H_1: The type of program (vocational, general, academic) has an influence on the number of awards earned by students.

```
awards_data$prog = factor(awards_data$prog)
model = glm(num_awards ~ prog, family = poisson, data = awards_data)
summary(model)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.549      0.196   -2.80  0.00515
## prog2          0.707      0.216    3.27  0.00106
## prog3          0.443      0.246    1.80  0.07199
```

```
new_data = data.frame(prog = factor(c(1, 2, 3)))
predicted_awards = predict(model, newdata = new_data, type = "response")
print(predicted_awards)
```

```
##     1     2     3
## 0.578 1.171 0.900
```

Significant p-values for prog2 (general program) and the intercept in the Poisson regression model suggest that program type affects student awards. Students in the general program are likely to receive more awards than those in vocational ones, shown by significant coefficients. Although academic program students also tend to get more awards, it's not statistically significant. Thus, the general program is the most favorable for receiving awards.

b)

- H_0: There is no difference in the distribution of the number of awards across the different program types (vocational, general, academic).
- H_1: At least one program type has a significantly different distribution of the number of awards compared to the others.

```r
kruskal_test_result = kruskal.test(num_awards ~ prog, data = awards_data)
cat("The p-value of the Shapiro-Wilk test is:", kruskal_test_result$p.value)
```

```
## The p-value of the Shapiro-Wilk test is: 0.00462
```

We reject H0 that everything is the same as the p-value is 0.005. However we can not infer which program type is the best for the number of awards.

c)

```r
awardsinteractionmath = glm(num_awards ~ factor(prog)*math, family="poisson", data=awards_data)
summary(awardsinteractionmath)$coefficients
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.57844     1.3914  -1.134    0.257
## factor(prog)2      -1.06123     1.5345  -0.692    0.489
## factor(prog)3       0.96214     1.6360   0.588    0.556
## math                0.02036     0.0269   0.756    0.450
## factor(prog)2:math  0.02744     0.0290   0.947    0.344
## factor(prog)3:math -0.00944     0.0324  -0.291    0.771
```

```r
awardsmath = glm(num_awards ~ factor(prog)+math, family="poisson", data=awards_data)
summary(awardsmath)$coefficients
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.3726    0.47552   -4.99 6.06e-07
## factor(prog)2   0.4526    0.22475    2.01 4.40e-02
## factor(prog)3   0.5617    0.24748    2.27 2.32e-02
## math            0.0358    0.00834    4.29 1.80e-05
```

```r
new_data = data.frame(prog = factor(c(1, 2, 3), levels = c(1, 2, 3)), math = 56)
predictions = predict(awardsmath, newdata=new_data, type="response")
new_data$predicted_awards = predictions
print(new_data)
```

```
##   prog math predicted_awards
## 1    1   56            0.691
## 2    2   56            1.087
## 3    3   56            1.213
```

No significant interaction results were found therefor we removed the interaction. Looking at the additive model with math included, we conclude that program 3 is the best with a predicted amount of awards of 1.213.