Assignment 13

Sachin Shubham

4/23/2021

Task 1:

In a study for determining the effect of weaning conditions on the weight of 9-week-old pigs, data on weaning (WWT) and 9-week (FWT) weights were recorded for pigs from three litters. One of these litters was weaned at approximately 21 days (EARLY), the second at about 28 days (MEDIUM), and the third at about 35 days (LATE). The data are given in Table 1. Perform an analysis of covariance using FWT as the response, weaning time as the factor, and WWT as the covariate. Comment on the results. Is there a problem with assumptions?

Solution:

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.0.4
```

```
## Loading required package: Matrix
```

```
library(ggplot2)
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.0.4
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
Early_WWT = c(9,9,12,11,15,15,14)
Early_FWT = c(37,28,40,45,44,50,45)

Med_WWT = c(16,16,15,14,14,12,10)
```

```
Med_FWT = c(48,45,47,46,40,36,33)

Late_WWT = c(18,17,16,15,14,14,13)
Late_FWT = c(45,38,35,38,34,37,37)

EARLY_Time=c("EARLY","EARLY","EARLY","EARLY","EARLY","EARLY","EARLY")

MEDIUM_TIME=c("MEDIUM","MEDIUM","MEDIUM","MEDIUM","MEDIUM","MEDIUM","MEDIUM")

LATE_TIME=c("LATE","LATE","LATE","LATE","LATE","LATE","LATE")

WWT<-c(Early_WWT,Med_WWT,Late_WWT)
FWT<-c(Early_FWT,Med_FWT,Late_FWT)
TIME<-c(EARLY_Time,MEDIUM_TIME,LATE_TIME)

dd<-data.frame(WWT,FWT,TIME)

head(dd,5)
```

```
##   WWT FWT  TIME
## 1   9  37 EARLY
## 2   9  28 EARLY
## 3  12  40 EARLY
## 4  11  45 EARLY
## 5  15  44 EARLY
```

```
#Analysis of Covariance for Stages
model_STAGE <- aov(FWT ~ WWT+TIME)
summary(model_STAGE)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## WWT          1  181.5  181.50   15.77 0.000987 ***
## TIME         2  289.8  144.91   12.59 0.000442 ***
## Residuals   17  195.6   11.51                     
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(model_STAGE)
```

```
## Analysis of Variance Table
## 
## Response: FWT
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## WWT        1 181.50 181.500  15.772 0.0009865 ***
## TIME       2 289.82 144.909  12.592 0.0004416 ***
## Residuals 17 195.63  11.508                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#Analysis of Covariance for WWT
model_WWT <- aov(FWT ~ TIME+WWT)
summary(model_WWT)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## TIME          2   77.2    38.6   3.356   0.0591 .
## WWT           1  394.1   394.1  34.244 1.92e-05 ***
## Residuals    17  195.6    11.5
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(model_WWT)
```

```
## Analysis of Variance Table
##
## Response: FWT
##           Df Sum Sq Mean Sq F value    Pr(>F)
## TIME       2  77.24   38.62  3.3559    0.0591 .
## WWT        1 394.08  394.08 34.2445 1.923e-05 ***
## Residuals 17 195.63   11.51
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Conclusion:

1.If the P value is less than 0.05 so it is highly significant.So, there is significant evidence that the weaning time is related to FWT, even after adjusting for WWT.

2.The LATE FWT is significantly less than either EARLY or MEDIUM, therefore the pigs within the same litter are probably not independent because of maternal and genetic effect

Task 2

Skidding is a major contributor to highway accidents. The following experiment was conducted to estimate the effect of pavement and tire tread depth on spinout speed, which is the speed (in mph) at which the rear wheels lose friction when negotiating a specific curve. There are two asphalt (ASPHALT1 and ASPHALT2) pavements and one concrete pavement and three tire tread depths (1-, 2-, and six-sixteenths of an inch). This is a factorial experiment, but the number of observations per cell is not the same. The data are given in Table2.

(a) Perform the analysis of variance using both the dummy variable and ”standard” approaches. Note that the results are not the same although the differences are not very large.

(b) The tread depth is really a measured variable. Perform any additional or alternative analysis to account for this situation.

(c) It is also known that the pavement types can be characterized by their coefficient of friction at 40 mph as follows:

ASPHALT1: 0.35

ASPHALT2: 0.24

CONCRETE: 0.48

Again, perform an alternative analysis suggested by this information. Which of the three analyses is most useful?

Solution:

```
OBS<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26)
PAVE<-c("ASPHALT1","ASPHALT1","ASPHALT1","ASPHALT1","ASPHALT1","ASPHALT1","ASPHALT1","CONCRETE","CONCRE
        "CONCRETE","CONCRETE","CONCRETE","CONCRETE","CONCRETE","CONCRETE","ASPHALT2","ASPHALT2","ASPHAL
        "ASPHALT2","ASPHALT2","ASPHALT2","ASPHALT2","ASPHALT2")
PAVE_NUM<-c(0.35,0.35,0.35,0.35,0.35,0.35,0.35 ,0.48,0.48,0.48,0.48,0.48,
        0.48,0.48,0.48,0.48,0.48,0.24,0.24,0.24,0.24,0.24,0.24,0.24,0.24,
        0.24)

TREAD<-c(1,1,2,2,2,6,6,1,1,1,1,2,2,2,6,6,6,1,1,1,2,2,2,6,6,6)
SPEED<-c(36.5, 34.9, 40.2, 38.2, 38.2, 43.7,43.0, 40.2,41.6,42.6,41.6,40.9,42.3,
         45.0,47.1,51.2,51.2,33.4,38.2,34.9, 36.8,35.4,35.4,40.2,40.9,43.0)
levels(PAVE)
```

```
## NULL
```

```
TH<-as.factor(TREAD)
PV<-as.factor(PAVE)
df_task2<-data.frame(OBS,PAVE,TH,SPEED)
head(df_task2,5)
```

```
##   OBS     PAVE TH SPEED
## 1   1 ASPHALT1  1  36.5
## 2   2 ASPHALT1  1  34.9
## 3   3 ASPHALT1  2  40.2
## 4   4 ASPHALT1  2  38.2
## 5   5 ASPHALT1  2  38.2
```

```
#(a)
#Standard
model.task2_standard = aov(SPEED ~  PAVE*TH,data=df_task2)
summary(model.task2_standard)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## PAVE         2 237.19  118.59  45.185 1.57e-07 ***
## TH           2 244.36  122.18  46.551 1.27e-07 ***
## PAVE:TH      4  10.24    2.56   0.975    0.447
## Residuals   17  44.62    2.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model.task2_standard)
```

```
## Analysis of Variance Table
##
## Response: SPEED
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## PAVE       2 237.188 118.594 45.1854 1.571e-07 ***
## TH         2 244.356 122.178 46.5510 1.269e-07 ***
## PAVE:TH    4  10.239   2.560  0.9753    0.4468
## Residuals 17  44.618   2.625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Dummy
model.task2_dummy = glm(SPEED ~  PAVE*TH,data=df_task2)
summary(model.task2_dummy)
```

```
##
## Call:
## glm(formula = SPEED ~ PAVE * TH, data = df_task2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7333  -0.6667  -0.3917   1.0583   2.7000
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        35.7000     1.1456  31.164  < 2e-16 ***
## PAVEASPHALT2       -0.2000     1.4789  -0.135 0.894015
## PAVECONCRETE        5.8000     1.4030   4.134 0.000694 ***
## TH2                 3.1667     1.4789   2.141 0.047039 *
## TH6                 7.6500     1.6201   4.722 0.000197 ***
## PAVEASPHALT2:TH2   -2.8000     1.9842  -1.411 0.176233
## PAVECONCRETE:TH2   -1.9333     1.9283  -1.003 0.330098
## PAVEASPHALT2:TH6   -1.7833     2.0915  -0.853 0.405701
## PAVECONCRETE:TH6    0.6833     2.0385   0.335 0.741569
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 2.624608)
##
##     Null deviance: 536.402  on 25  degrees of freedom
## Residual deviance:  44.618  on 17  degrees of freedom
## AIC: 107.83
##
## Number of Fisher Scoring iterations: 2
```

```
anova(model.task2_dummy)
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: SPEED
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev
## NULL                          25     536.40
## PAVE     2  237.188          23     299.21
## TH       2  244.356          21      54.86
## PAVE:TH  4   10.239          17      44.62
```

```
Anova(model.task2_dummy, test="F")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: SPEED
## Error estimate based on Pearson residuals
##
##             Sum Sq Df F value    Pr(>F)
## PAVE      253.912  2 48.3715 9.626e-08 ***
## TH        244.356  2 46.5510 1.269e-07 ***
## PAVE:TH    10.239  4  0.9753    0.4468
## Residuals  44.618 17
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#(b)
df_task2_alt<--data.frame(OBS,PV,TREAD,SPEED)
```

```
## Warning in Ops.factor(left): ’-’ not meaningful for factors
```

```
head(df_task2_alt,5)
```

```
##   OBS PV TREAD SPEED
## 1  -1 NA    -1 -36.5
## 2  -2 NA    -1 -34.9
## 3  -3 NA    -2 -40.2
## 4  -4 NA    -2 -38.2
## 5  -5 NA    -2 -38.2
```

```
model.task2_alternative = aov(SPEED ~  TREAD+PV,data=df_task2)
summary(model.task2_alternative)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## TREAD          1 226.69  226.69   90.86 2.86e-09 ***
## PV             2 254.83  127.42   51.07 5.41e-09 ***
## Residuals     22  54.88    2.49
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(model.task2_alternative)
```

```
## Analysis of Variance Table
##
## Response: SPEED
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## TREAD      1 226.686 226.686  90.865 2.865e-09 ***
## PV         2 254.832 127.416  51.074 5.412e-09 ***
## Residuals 22  54.884   2.495
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Anova(model.task2_alternative, test="F")
```

```
## Anova Table (Type II tests)
##
## Response: SPEED
##           Sum Sq Df F value    Pr(>F)
## TREAD     244.329  1  97.938 1.459e-09 ***
## PV        254.832  2  51.074 5.412e-09 ***
## Residuals  54.884 22
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#(c)
df_task2_c<-data.frame(OBS,PAVE_NUM,TREAD,SPEED)
head(df_task2_c,5)
```

```
##   OBS PAVE_NUM TREAD SPEED
## 1   1     0.35     1  36.5
## 2   2     0.35     1  34.9
## 3   3     0.35     2  40.2
## 4   4     0.35     2  38.2
## 5   5     0.35     2  38.2
```

```
model.task2_c = glm(SPEED ~  PAVE_NUM*TREAD,data=df_task2_c)
summary(model.task2_c)
```

```
##
## Call:
## glm(formula = SPEED ~ PAVE_NUM * TREAD, data = df_task2_c)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7015  -1.0001  -0.4649   1.0950   3.5036
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     27.5766     2.0135  13.696 3.02e-12 ***
## PAVE_NUM        24.5548     5.3032   4.630 0.000129 ***
## TREAD            0.7815     0.5526   1.414 0.171299
## PAVE_NUM:TREAD   1.8548     1.4700   1.262 0.220248
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 2.733577)
##
##     Null deviance: 536.402  on 25  degrees of freedom
## Residual deviance:  60.139  on 22  degrees of freedom
## AIC: 105.59
##
## Number of Fisher Scoring iterations: 2
```

```
anova(model.task2_c)
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: SPEED
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev
## NULL                              25     536.40
## PAVE_NUM        1  226.515        24     309.89
## TREAD           1  245.396        23      64.49
## PAVE_NUM:TREAD  1    4.352        22      60.14
```

Conclusion:

The tread depth and pavement types both appeared to be extremely significant («0.05 p-value) when related to the speed a car spins out using ANOVA from (b) and therefore I would say this analysis is the most useful.

Task 3

Cochran and Chamlin (2006) used data from the National Opinion Research Council - General Social Survey (NORC-GSS) to compare whites' and blacks' opinions of the death penalty. The data consisted of responses from 32,937 participants collected between 1972 and 1996. (The question was not asked every year.) The outcome variable was whether the respondent did or did not support the death penalty. Their hypotheses concerned both the possible difference between blacks and whites, and the possible change in that difference over time. The authors provided a table of the percentage of whites and blacks each year that supported the death penalty, shown in Table.

(a) Convert the percentages given in Table 3 to the Ln(odds) within each race and year, and plot the ln(odds) versus year. Comment on any patterns you see. If there is a trend in time, does it appear linear or quadratic?

(b) Use logistic regression to model the probability a person will support the death penalty, as a function of race and year. Is there significant evidence that a quadratic term in year improves the model? Assume that in each year's sample there were 1100 whites and 400 blacks.

(c) Attempt to improve your model by adding interactions of race with the linear and quadratic variables in time. Do the interactions significantly improve the model?

(d) The authors of the study refer to the gap between white and black support as ”enduring”. Are your results in part (c) consistent with this?

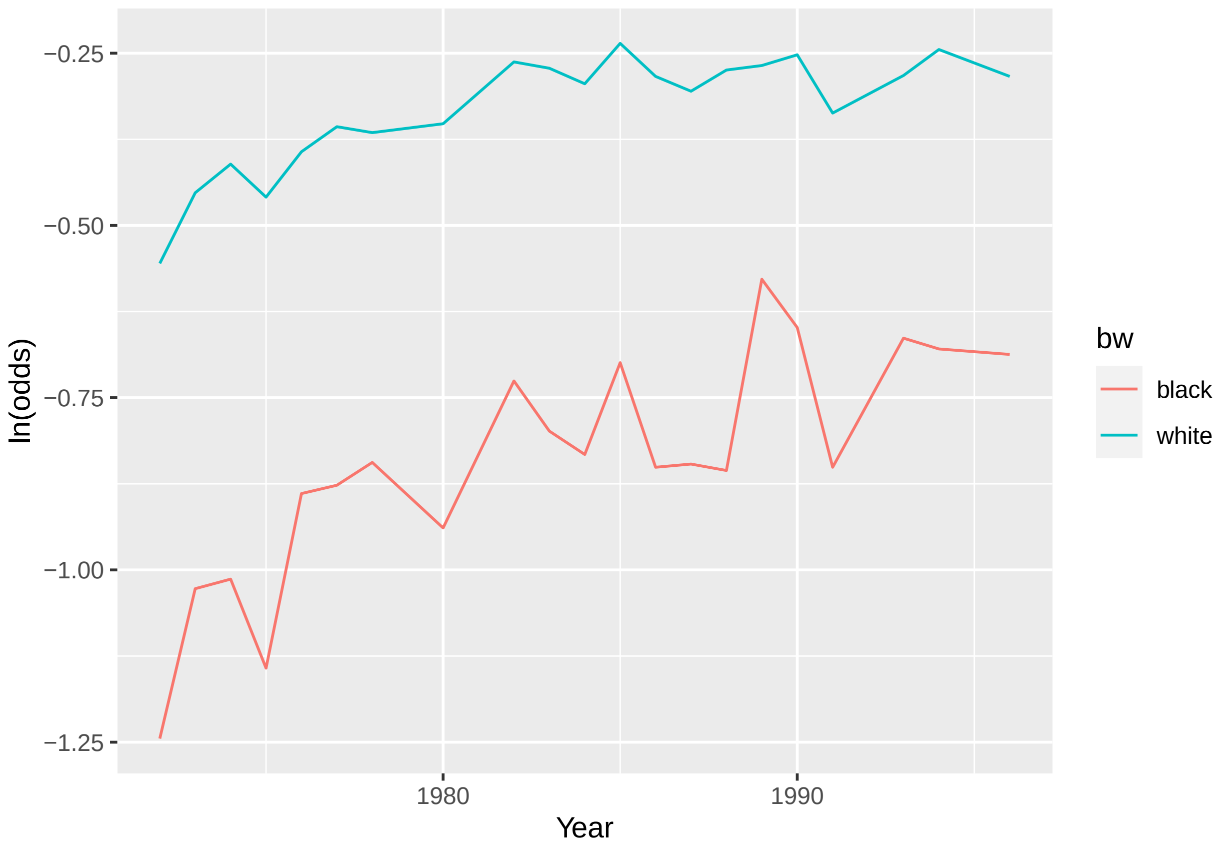
```
race_white_perc <- c(57.4, 63.6, 66.3, 63.2, 67.5, 70.0, 69.4, 70.3, 76.9,
                    76.2, 74.5, 79.0, 75.3, 73.7, 76.0, 76.5, 77.7, 71.4,
                    75.4, 78.3, 75.3)
race_black_perc <- c(28.8, 35.8, 36.3, 31.9, 41.1, 41.6, 43.0, 39.1, 48.4,
                     45.0, 43.5, 49.7, 42.7, 42.9, 42.5, 56.1, 52.3, 42.7,
                     51.5, 50.7, 50.3)
year <- c(1972, 1973, 1974, 1975, 1976, 1977, 1978, 1980, 1982,
          1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991,
          1993, 1994, 1996)


# Convert %s to ln(odds) within each race and year


death_pen <- data.frame(year, race_white_perc, race_black_perc )
head(death_pen,5)
```

```
##   year race_white_perc race_black_perc
## 1 1972            57.4            28.8
## 2 1973            63.6            35.8
## 3 1974            66.3            36.3
## 4 1975            63.2            31.9
## 5 1976            67.5            41.1
```

```
ln_oddsB <- log(race_black_perc/100)
ln_oddsW <- log(race_white_perc/100)

# a)


ln_odds<-c(ln_oddsW,ln_oddsB)

bw<-c("white","white","white","white","white","white","white","white","white",        "white","white","w

dt_task3_a<-data.frame(ln_odds, year,bw)
head(dt_task3_a,5)
```

```
##      ln_odds year    bw
## 1 -0.5551259 1972 white
## 2 -0.4525567 1973 white
## 3 -0.4109803 1974 white
## 4 -0.4588659 1975 white
## 5 -0.3930426 1976 white
```

```
p <- ggplot(data = dt_task3_a, aes(x = year,
                                   y = ln_odds, group = bw,
                                   shape = bw, color=bw))

p + geom_line()+xlab("Year")+ylab("ln(odds)")
```



```
#b

library(reshape2)


dt<-data.frame(race_white_perc,race_black_perc,year)

y_white<-c(floor(dt$race_white_perc*1100/100))
y_black<-c(floor(dt$race_black_perc*400/100))
n_white<-c(1100-y_white)
n_black<-c(400-y_black)


year_sub = year-1972

nbt<-data.frame(year_sub,y_white,y_black,n_white,n_black)
#melt
nbtMelt <- melt(nbt, na.rm = FALSE, value.name = "odds", variable='race', id.vars='year_sub')

nbtMelt$odds2 <- nbtMelt$race
nbtMelt$odds2 <- as.character(nbtMelt$odds2)

#calculate number of white and black for yes and no
nbtMelt$odds2[nbtMelt$odds2 =="y_white"] <- 1
nbtMelt$odds2[nbtMelt$odds2 =="y_black"] <- 1
nbtMelt$odds2[nbtMelt$odds2 =="n_white"] <- 0
```

```
nbtMelt$odds2[nbtMelt$odds2 =="n_black"] <- 0

# Expand rows based on number of votes
nbtExpand <- nbtMelt[rep(row.names(nbtMelt), nbtMelt$odds), 1:4]

nbtExpand$race2 <- as.character(nbtExpand$race)
nbtExpand$race2[nbtExpand$race2 == "y_white"] <- 0
nbtExpand$race2[nbtExpand$race2 == "y_black"] <- 1
nbtExpand$race2[nbtExpand$race2 == "n_white"] <- 0
nbtExpand$race2[nbtExpand$race2 == "n_black"] <- 1

# Quad year
nbtExpand$yearsSquared <- nbtExpand$year_sub^2
# make odds2 & race2 numeric
nbtExpand$odds2 <- as.numeric(nbtExpand$odds2)
nbtExpand$race2 <- as.numeric(nbtExpand$race2)
#logistic model
nbtExpand$race2 <- as.factor(nbtExpand$race2)
#dataframe
head(nbtExpand,5)
```

```
##     year_sub    race odds odds2 race2 yearsSquared
## 1          0 y_white  631     1     0            0
## 1.1        0 y_white  631     1     0            0
## 1.2        0 y_white  631     1     0            0
## 1.3        0 y_white  631     1     0            0
## 1.4        0 y_white  631     1     0            0
```

```
#logistic regression model
mod_b = glm(odds2 ~ race2 + year_sub, family = binomial("logit"), data = nbtExpand)
summary(mod_b)
```

```
## 
## Call:
## glm(formula = odds2 ~ race2 + year_sub, family = binomial("logit"), 
##     data = nbtExpand)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7797  -1.1134   0.7473   0.8782   1.4522  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  0.596073   0.023717   25.13   <2e-16 ***
## race21      -1.222291   0.026647  -45.87   <2e-16 ***
## year_sub     0.031579   0.001726   18.30   <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: 41000  on 31499  degrees of freedom
## Residual deviance: 38535  on 31497  degrees of freedom
## AIC: 38541
##
## Number of Fisher Scoring iterations: 4
```

```
#c
mod_c = glm(odds2 ~ race2 + year_sub + yearsSquared, family = binomial("logit"), data=nbtExpand)
summary(mod_c)
```

```
##
## Call:
## glm(formula = odds2 ~ race2 + year_sub + yearsSquared, family = binomial("logit"),
##     data = nbtExpand)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7021  -1.1473   0.7402   0.8688   1.5291
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4280617  0.0310892  13.769   <2e-16 ***
## race21       -1.2251418  0.0266822 -45.916   <2e-16 ***
## year_sub      0.0813135  0.0062589  12.992   <2e-16 ***
## yearsSquared -0.0021946  0.0002649  -8.284   <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: 41000  on 31499  degrees of freedom
## Residual deviance: 38466  on 31496  degrees of freedom
## AIC: 38474
##
## Number of Fisher Scoring iterations: 4
```

```
anova(mod_b,mod_c)
```

```
## Analysis of Deviance Table
##
## Model 1: odds2 ~ race2 + year_sub
## Model 2: odds2 ~ race2 + year_sub + yearsSquared
##   Resid. Df Resid. Dev Df Deviance
## 1     31497      38535
## 2     31496      38466  1   68.265
```

Conclusion:

(a) The Yes answers were much higher in whites rather than blacks starting in 1972 but over the years the two demographics have started to converge with a relatively steady white population supporting the death penalty and blacks increasingly supporting the death penalty over the years.

(b) Yes, there is significant evidence that the quadratic term improves the model

(c) The likelihood ratio test has chi square as 4.686 with 2 degree of freedom, which is not significant at alpha equals 5% and it has p value below 0.10. Hence there is no significant evidence of interaction.

(d) (I) Yes, the gap between white and black support as "enduring". The first model shows exceptionally strong evidence for a gap.

(II) Since the interaction was not significant, the gap in the ln(odds) does not appear to be changing much with the time.

(III) Therefore, the plot of the probabilities might shows a minute distance between the Blacks and White.

Task 4

Popkin (1991) presented the data shown in Table 4 for number of auto crashes and number of alcohol-related (A/R) auto crashes for young drivers in North Carolina. You are interested in whether the probability a crash will be A/R is related to age and gender.

(a) Construct a profile plot (similar to those for the two-way ANOVA) for the Ln(odds) that a crash will be alcohol-related, using the age category on the horizontal axis and separate symbols for gender. Discuss the apparent effects. Is there a graphical suggestion of an interaction?

(b) Construct a profile plot in the same way as for part (a), using the empirical probability that a crash will be alcohol-related. Is there a graphical suggestion of an interaction?

(c) Construct a dummy variable system for the age category and gender, and fit a logistic regression that only includes main effects. Interpret the main effects,using the profile plot from part (a).

(d) Fit a logistic regression that includes main effects and interactions.

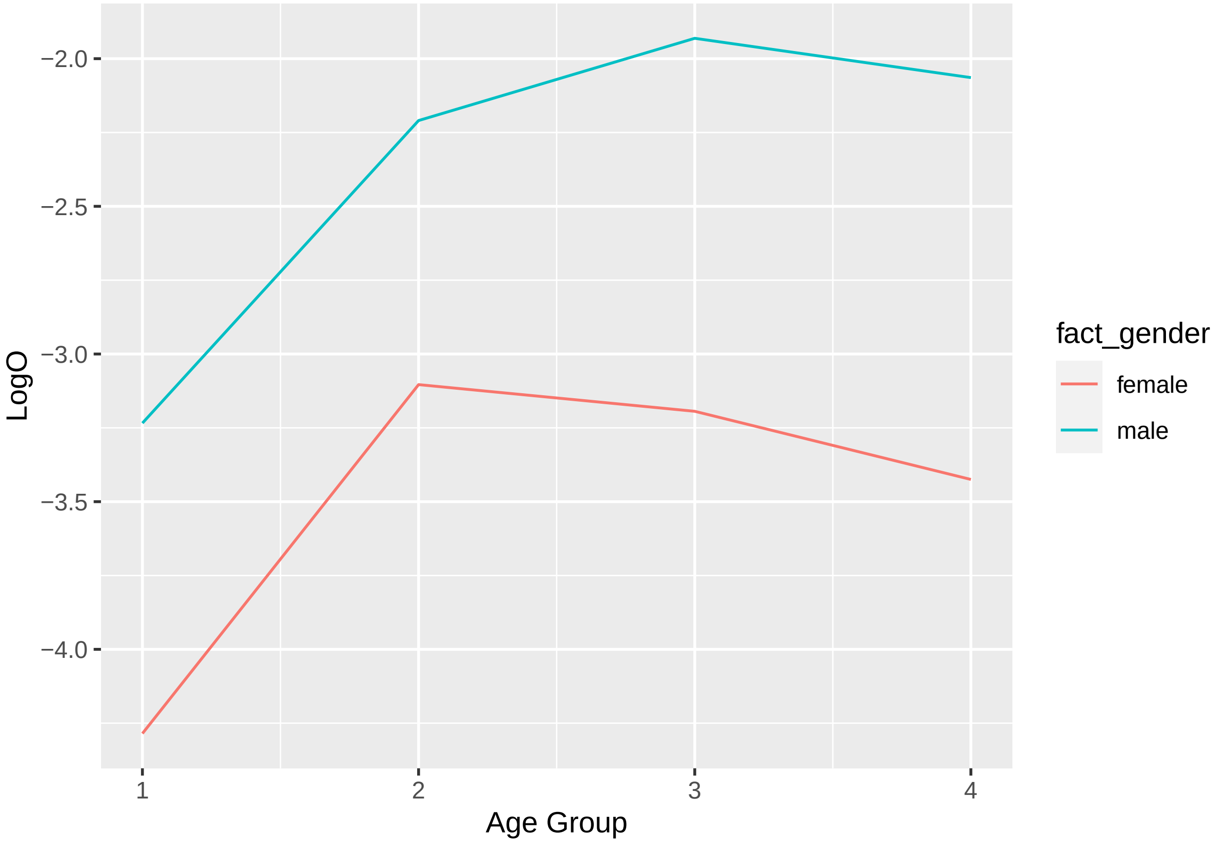
(e) Construct a likelihood ratio test for the null hypothesis that none of the interactions are significant. Interpret the results.

Solution:

```
age_num<-c(1,1,2,2,3,3,4,4)
fact_age<-as.factor(age_num)
gender<-c("male","female","male","female","male","female","male","female")
gender_num<-c(0,1,0,1,0,1,0,1)
fact_gender<-as.factor(gender)
level_gender <- c('male', 'female')
total<-c(14589,8612, 21708, 10941, 25664, 13709, 41304, 25183)
ar<-c(553, 117, 2147, 470, 3250, 540, 4652, 794)




logO = log((ar/total)/(1-ar/total))

odds = (ar/total)/(1-ar/total)


dt_task4_a<-data.frame(fact_age,fact_gender,logO)
dt_task4_a
```

```
##   fact_age fact_gender      logO
## 1        1        male -3.234023
## 2         1      female -4.285059
## 3         2        male -2.209466
## 4         2      female -3.103632
## 5         3        male -1.931031
## 6         3      female -3.194052
## 7         4        male -2.064171
## 8         4      female -3.424804
```

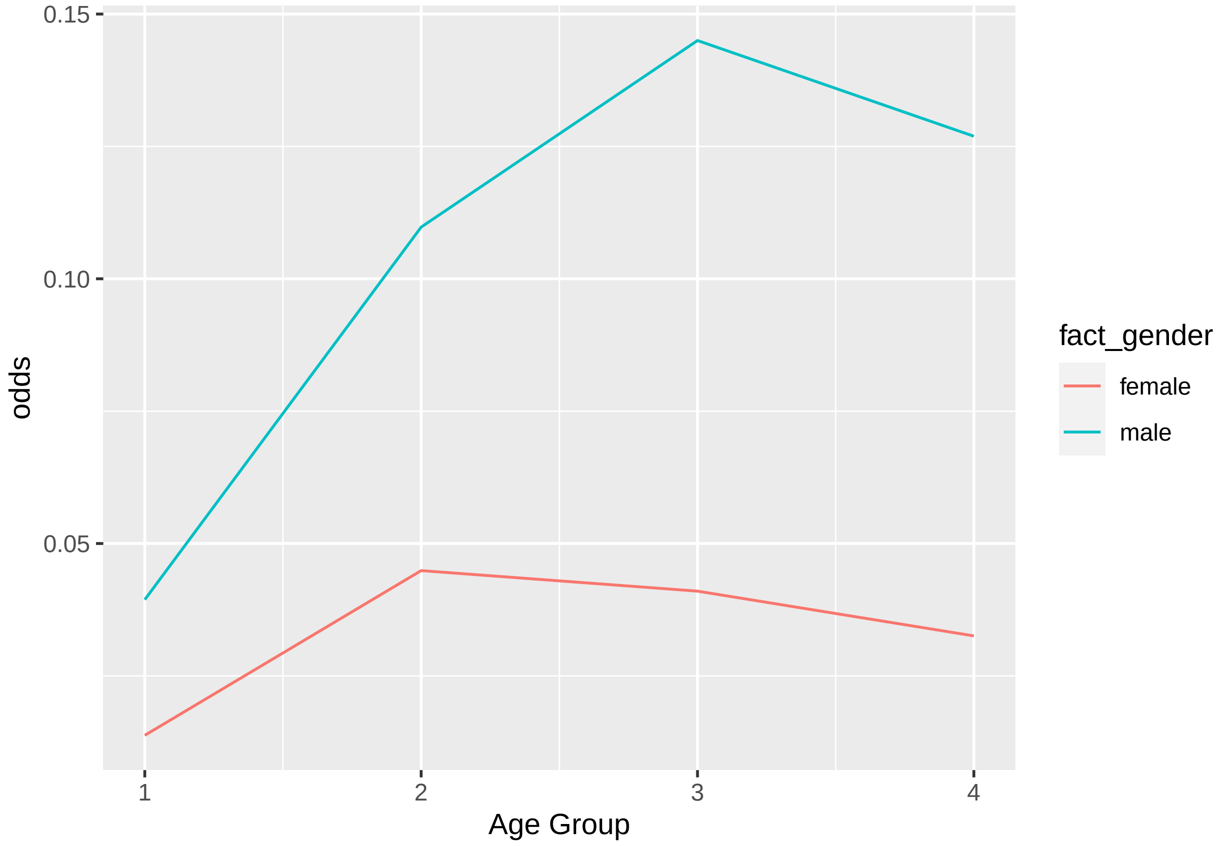
```
#a

p <- ggplot(data = dt_task4_a, aes(x = age_num,
                                   y = logO, group = fact_gender,
                                   shape = fact_gender, color=fact_gender))


p + geom_line()+xlab("Age Group")+ylab("LogO")
```



```
#b

dt_task4_b<-data.frame(fact_age,fact_gender,odds)
dt_task4_b
```

```
##   fact_age fact_gender       odds
## 1        1        male 0.03939869
## 2        1      female 0.01377281
## 3        2        male 0.10975921
## 4        2      female 0.04488588
## 5        3        male 0.14499866
## 6        3      female 0.04100539
## 7        4        male 0.12692350
## 8        4      female 0.03255566
```

```
p <- ggplot(data = dt_task4_b, aes(x = age_num,
                                   y = odds, group = fact_gender,
                                   shape = fact_gender, color=fact_gender))

p + geom_line()+xlab("Age Group")+ylab("odds")
```



```
#c

age_category <- c("< 18","< 18","18-20","18-20","21-24","21-24",">= 25",">= 25")
gender <- c("M","F","M","F","M","F","M","F")
alcohol_odds <- c(553/(14589-553),117/(8612-117),
                  2147/(21708-2147),470/(10941-470),
                  3250/(25664-3250),540/(13709-540),4652/(41304-4652),
                  794/(25183-794))
df_task4c <- data.frame(age_category,gender,alcohol_odds)
df_task4c
```

```
##   age_category gender alcohol_odds
## 1         < 18      M   0.03939869
## 2         < 18      F   0.01377281
## 3        18-20      M   0.10975921
## 4        18-20      F   0.04488588
## 5        21-24      M   0.14499866
## 6        21-24      F   0.04100539
## 7        >= 25      M   0.12692350
## 8        >= 25      F   0.03255566
```

```
df_task4c$age1 <- ifelse(df_task4c$age_category == "< 18", 1, 0)
df_task4c$age2 <- ifelse(df_task4c$age_category == "18-20", 1, 0)
df_task4c$age3 <- ifelse(df_task4c$age_category == "21-24", 1, 0)
df_task4c$age4 <- ifelse(df_task4c$age_category == ">= 25", 1, 0)
df_task4c$gender <- ifelse(df_task4c$gender == "F", 1, 0)
df_task4c$odds <- df_task4c$alcohol_odds
df_task4c$total <- c(14589,8612,21708,10941,25664,13709,41304,25183)

model_4c <- glm( odds ~ gender  + age2 + age3 + age4,
                 weight=total, data = df_task4c,
                 family=binomial("logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(model_4c)
```

```
##
## Call:
## glm(formula = odds ~ gender + age2 + age3 + age4, family = binomial("logit"),
##     data = df_task4c, weights = total)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -0.8775   2.0528  -2.5609   6.0267   0.5651  -1.3456   1.6164  -3.7068
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.15670    0.03886  -81.23   <2e-16 ***
## gender      -1.31033    0.02485  -52.72   <2e-16 ***
## age2         1.11870    0.04343   25.76   <2e-16 ***
## age3         1.37229    0.04202   32.66   <2e-16 ***
## age4         1.20429    0.04102   29.36   <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: 5123.079  on 7  degrees of freedom
## Residual deviance:   66.347  on 3  degrees of freedom
## AIC: 145.4
##
## Number of Fisher Scoring iterations: 4
```

```
#d
model_4d <- glm( odds ~ gender  + age2 + age3 + age4+
                 gender*age2+gender*age3+gender*age4,
                 weight=total,data = df_task4c,
                 family=binomial("logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(model_4d)
```

```
##
## Call:
## glm(formula = odds ~ gender + age2 + age3 + age4 + gender * age2 +
##     gender * age3 + gender * age4, family = binomial("logit"),
##     data = df_task4c, weights = total)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.19383    0.04256 -75.048  < 2e-16 ***
## gender      -1.07736    0.10178 -10.585  < 2e-16 ***
## age2         1.10062    0.04778  23.037  < 2e-16 ***
## age3         1.41945    0.04610  30.789  < 2e-16 ***
## age4         1.26539    0.04505  28.088  < 2e-16 ***
## gender:age2  0.11286    0.11386   0.991  0.32157
## gender:age3 -0.30044    0.11193  -2.684  0.00727 **
## gender:age4 -0.38590    0.10881  -3.547  0.00039 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  5.1231e+03  on 7  degrees of freedom
## Residual deviance: -3.4191e-12  on 0  degrees of freedom
## AIC: 85.057
##
## Number of Fisher Scoring iterations: 3
```

```
#e
library(lmtest)
```

```
## Warning: package ’lmtest’ was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package ’zoo’ was built under R version 4.0.4
```

```
##
## Attaching package: ’zoo’
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lrtest(model_4c, model_4d)
```

```
## Likelihood ratio test
##
## Model 1: odds ~ gender + age2 + age3 + age4
## Model 2: odds ~ gender + age2 + age3 + age4 + gender * age2 + gender *
##     age3 + gender * age4
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   5 -67.698
## 2   8 -34.528  3 66.34  2.592e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

a) (I) It appears that men have relatively high probability of alcohol related crashes as compared to females.

   (II) As the men get older the likelihood of getting into an alcohol-related crash increases and peaks when they reach above 25 years of age.

   (III) Females start with a relatively low probability of alcohol-related crashes which spikes at 18-20 years old and decreases after turning 21.

b) (I) Empirically, men are shown to have much alcohol related crashes, peaking at about a 14% rate before they turn 25.

   (II) Females have lower odds of being a part of an alcohol related crash peaking before 21 and decreasing thereafter.

c) (I) All dummy variables were extremely significant with females have significantly lower odds of being in alcohol related crashes.

   (II) Both genders start off with low number of alcohol related crashes and increase up to a certain age (males peak at 25 and females peak at age 20).

d) The interaction affects the size of the gap between the ln(odds) of the men and women:

   (I) In the youngest age group, the ln(odds) for women is less than ln(odds) of men.Also for the oldest age group,

   (II) The ln(odds) for women is less than ln(odds) of men. This is consistent with plot in (a).

e) (I) There is significant evidence that at least one of the interactions of a variable between the two models is different.

   (II) Between the men and women data from the 2 models, at least one of the age groups is significantly different.

Task 5

Van den Bos et. al. (2006) analyzed Y = Outcome Satisfaction for 138 participants in an experiment with two factors: Cognitive Busyness (low to high) and Outcome (equal to others, better than others, worse than others) The mean values for Y within each cell are given below.

(a) Construct a profile plot that will allow you to inspect the apparent effects in the data.

(b) The authors cite the following test statistics from the two-way ANOVA:

Main effect for Outcome: F(2,132) = 236.56, p < 0.001

Main effect for Busyness: F(1,132) = 4.36, p < 0.04

Interaction: F(2,132) = 3.38, p < 0.04

Use this information, together with your profile plot, to write a short paragraph explaining the effects of these factors on Outcome Satisfaction.

(c) The authors carried out ”the least significant difference test for means (p <0.05) with the six cells of our design serving as the independent variable.” how many independent samples t-tests are implied by this statements?

(d) In the table above, cell means with the same letter in parentheses were not significantly different using the method described in part (c). The authors state ”there were no efforts of Cognitive Busyness within the equal-to-other and

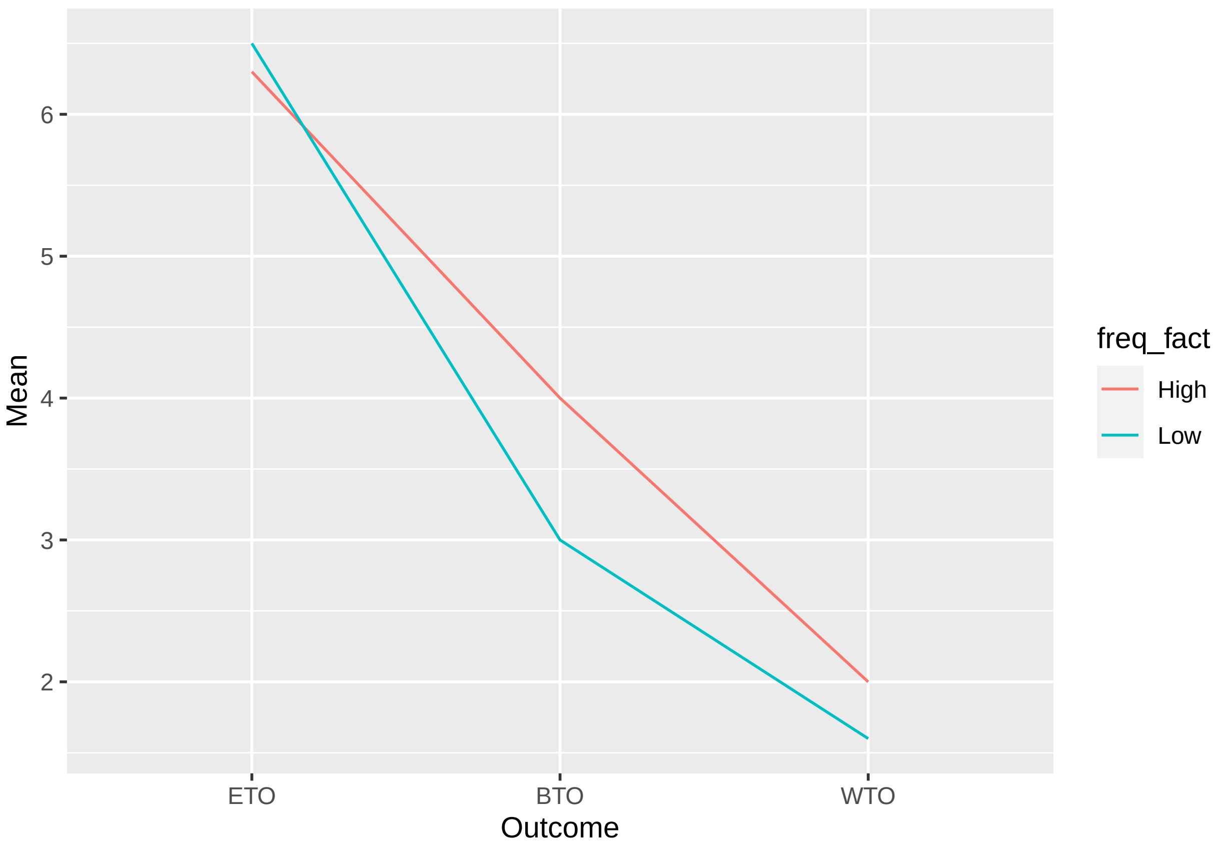
Solution:

(a)

```
#a
freq=c("Low","High","Low","High","Low","High")
freq_fact<-as.factor(freq)
outcome=c("ETO","ETO","BTO","BTO","WTO","WTO")
outcome_fact<-as.factor(outcome)
val=c(6.5,6.3,3.0,4.0,1.6,2.0)
level_order <- c('ETO', 'BTO', 'WTO')
df_task5<-data.frame(outcome_fact,freq_fact,val)
df_task5
```

```
##   outcome_fact freq_fact val
## 1          ETO       Low 6.5
## 2          ETO      High 6.3
## 3          BTO       Low 3.0
## 4          BTO      High 4.0
## 5          WTO       Low 1.6
## 6          WTO      High 2.0
```

```
p <- ggplot(data = df_task5, aes(x = factor(outcome_fact , level = level_order), y = val, group = freq_


p + geom_line()+xlab("Outcome")+ylab("Mean")
```



(b) The plot and statistics show a strong main effect of Outcome with participants having highest satisfaction when perceived as an equal. There is a weak main effect for Busyness where those with low Busyness have less satisfaction.

(c) 6 cells * 5 points / 2 Frequencies = 15 number of independent sampled t-tests

(d) The difference between the High and Low Busyness categories is most pronounced in the better outcome category while there's a low difference in the Equal and Worse categories.