Lab Assignment 2

Charlie Lock

2023-05-01

Set Up

```
library(leaps)
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

Question 1.

```
toothpaste.data <- read.csv("/Users/charlielock/Documents/R/Datasets/ToothpasteSales.CSV")
```

Question 1.a.

```
lm.toothpaste <- lm(sales ~ budget + ratio + personal_disposal_income, data = toothpa
ste.data)
summary(lm.toothpaste)
```

```
## 
## Call:
## lm(formula = sales ~ budget + ratio + personal_disposal_income, 
##     data = toothpaste.data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -24088  -2568   1021   3836  10100 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               34104.559  17654.144   1.932 0.082187 .  
## budget                        3.746      1.976   1.896 0.087243 .  
## ratio                    -30046.343  22859.674  -1.314 0.218066    
## personal_disposal_income     85.926     17.911   4.797 0.000727 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9574 on 10 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.9598 
## F-statistic: 104.5 on 3 and 10 DF,  p-value: 7.537e-08
```

The p-value of the F-statistic is less than 0.05 so the model can be seen as significant

Question 1.b.

Using the summary() function from the previous question it can be seen that personal_disposal_income has a p-value of less than 0.05 which means it can be considered as important whilst the budget and ratio variables have p-values of greater than 0.05 so they are not considered important

Question 1.c.

```
ncvTest(lm.toothpaste)
```

```
## Non-constant Variance Score Test 
## Variance formula: ~ fitted.values 
## Chisquare = 0.07424735, Df = 1, p = 0.78525
```

H0: Errors have constant variance H1: Errors do not have constant variance

The p-value is just > 0.05 so H0 is rejected which implies that the assumption of constant error variance is not violated.

```
shapiro.test(lm.toothpaste$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.toothpaste$residuals
## W = 0.83777, p-value = 0.01522
```

H0: Errors are normally distributed H1: Errors are not normally distributed

The p-value is < 0.05 so H0 is rejected which implies that the assumption of normally distributed errors is violated.

```
durbinWatsonTest(lm.toothpaste)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1129573      2.211283   0.892
##  Alternative hypothesis: rho != 0
```

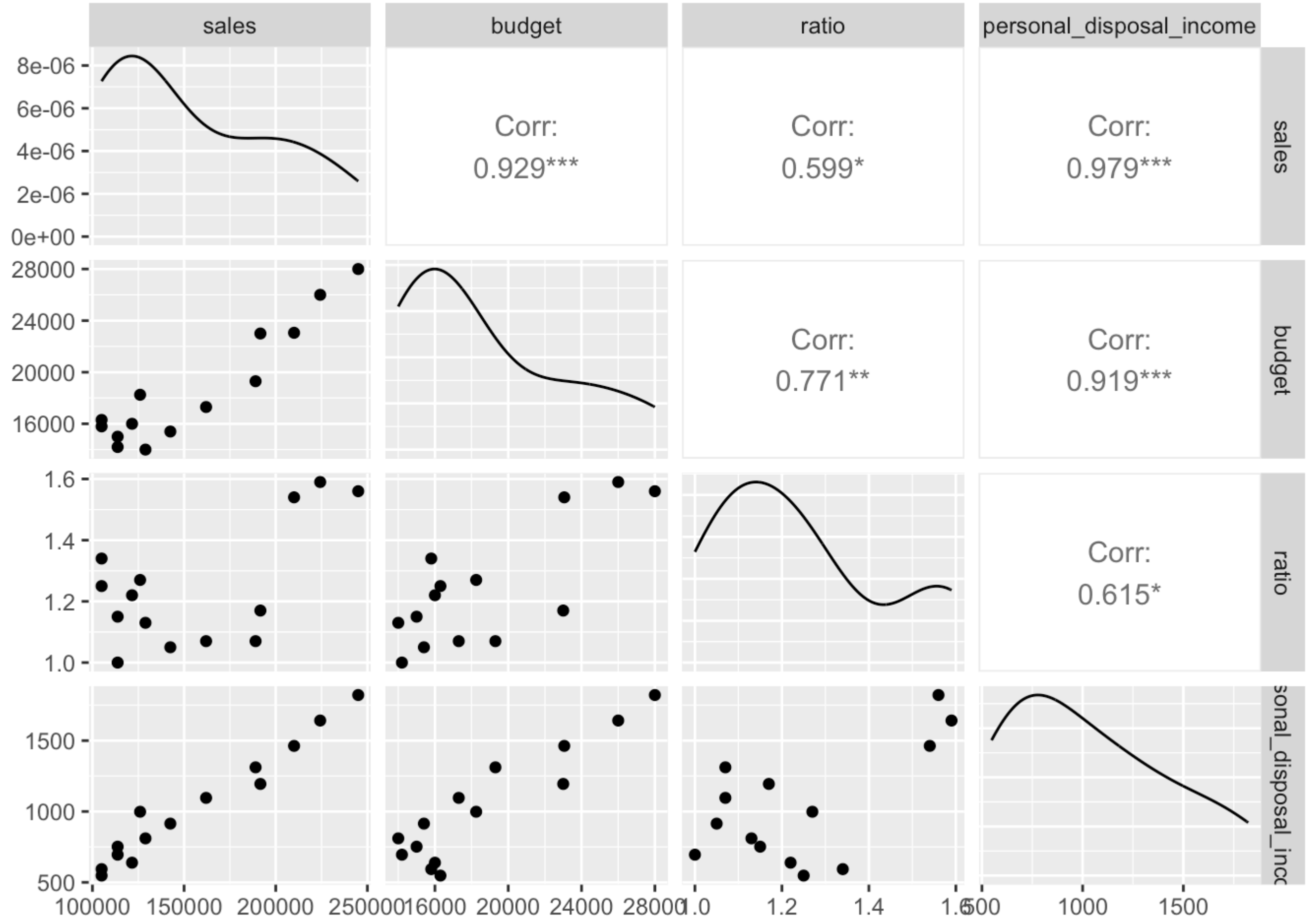
H0: Errors are not correlated H1: Errors are correlated

The p-value is > 0.05 so H0 cannot be rejected and the assumption of errors being correlated is not violated.

Question 1.d.

```
ggpairs(data = toothpaste.data, columns = c(2,3,4,5))
```

```
## Warning in geom_point(): All aesthetics have length 1, but the data has 16 rows.
## ℹ Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



The sales variable is correlated to the budget and personal disposable income variables by 0.929 and 0.979 respectively. This strongly indicates multicollinearity.

The budget variable is correlated to the ratio variable by 0.771 and also to the personal disposable income variable by 0.919. This also strongly indicates multicollinearity between those variables.

Question 1.e.

```
full=lm(sales~., data = toothpaste.data)
null=lm(sales~1, data = toothpaste.data)
```

Stepwise Regression Method

```
step(null, scope = list(upper = full), direction = "both")
```

```
## Start:  AIC=302.64
## sales ~ 1
## 
##                            Df  Sum of Sq        RSS    AIC
## + personal_disposal_income  1 2.8411e+10 1.2467e+09 260.27
## + Year                      1 2.6668e+10 2.9902e+09 272.51
## + budget                    1 2.5611e+10 4.0466e+09 276.75
## + ratio                     1 1.0637e+10 1.9020e+10 298.42
## <none>                                   2.9658e+10 302.63
## 
## Step:  AIC=260.27
## sales ~ personal_disposal_income
## 
##                            Df  Sum of Sq        RSS    AIC
## + budget                    1 1.7162e+08 1.0750e+09 260.19
## <none>                                   1.2467e+09 260.27
## + Year                      1 3.9874e+07 1.2068e+09 261.81
## + ratio                     1 5.7459e+05 1.2461e+09 262.26
## - personal_disposal_income  1 2.8411e+10 2.9658e+10 302.63
## 
## Step:  AIC=260.19
## sales ~ personal_disposal_income + budget
## 
##                            Df  Sum of Sq        RSS    AIC
## + ratio                     1  158364757  916674966 259.96
## <none>                                   1075039723 260.19
## - budget                    1  171624035 1246663758 260.27
## + Year                      1    8275335 1066764387 262.08
## - personal_disposal_income  1 2971530267 4046569990 276.75
## 
## Step:  AIC=259.96
## sales ~ personal_disposal_income + budget + ratio
## 
##                            Df  Sum of Sq        RSS    AIC
## <none>                                    916674966 259.96
## - ratio                     1  158364757 1075039723 260.19
## + Year                      1   14933380  901741586 261.73
## - budget                    1  329414202 1246089168 262.26
## - personal_disposal_income  1 2109684588 3026359554 274.68
```

```
## 
## Call:
## lm(formula = sales ~ personal_disposal_income + budget + ratio, 
##     data = toothpaste.data)
## 
## Coefficients:
##              (Intercept)  personal_disposal_income                    budget  
##                34104.559                    85.926                     3.746  
##                    ratio  
##               -30046.343
```

The Stepwise Regression method suggests that the model with the best fit includes all 3 predictor variables (personal_disposal_income, budget and ratio) as this has the minimum AIC value of 259.96.

The model is E(y) = 34104.559 + (85.926 * personal_disposal_income) + (3.746 * budget) + (-30046.343 * ratio)

Backwards elimination

```
step(full, data = toothpaste.data, direction = "backward")
```

```
## Start:  AIC=261.73
## sales ~ Year + budget + ratio + personal_disposal_income
##
##                            Df Sum of Sq        RSS    AIC
## - Year                      1  14933380  916674966 259.96
## <none>                                   901741586 261.73
## - ratio                     1 165022801 1066764387 262.08
## - budget                    1 249022750 1150764336 263.14
## - personal_disposal_income  1 307969340 1209710927 263.84
##
## Step:  AIC=259.96
## sales ~ budget + ratio + personal_disposal_income
##
##                            Df  Sum of Sq        RSS    AIC
## <none>                                    916674966 259.96
## - ratio                     1  158364757 1075039723 260.19
## - budget                    1  329414202 1246089168 262.26
## - personal_disposal_income  1 2109684588 3026359554 274.68
```

```
##
## Call:
## lm(formula = sales ~ budget + ratio + personal_disposal_income,
##     data = toothpaste.data)
##
## Coefficients:
##              (Intercept)                    budget                     ratio
##                34104.559                     3.746                -30046.343
## personal_disposal_income
##                   85.926
```

The Backwards elimination method suggests that based off the minimum AIC value of 259.96 all 3 predictor variables should be used in the best fitted model.

The model is E(y) = 34104.559 + (85.926 * personal_disposal_income) + (3.746 * budget) + (-30046.343 * ratio)

Forward selection

```
step(null, scope = list(lower = null, upper = full, direction = "forward"))
```

```
## Start:  AIC=302.64
## sales ~ 1
## 
##                             Df  Sum of Sq        RSS    AIC
## + personal_disposal_income  1 2.8411e+10 1.2467e+09 260.27
## + Year                      1 2.6668e+10 2.9902e+09 272.51
## + budget                    1 2.5611e+10 4.0466e+09 276.75
## + ratio                     1 1.0637e+10 1.9020e+10 298.42
## <none>                                   2.9658e+10 302.63
## 
## Step:  AIC=260.27
## sales ~ personal_disposal_income
## 
##                             Df  Sum of Sq        RSS    AIC
## + budget                    1 1.7162e+08 1.0750e+09 260.19
## <none>                                   1.2467e+09 260.27
## + Year                      1 3.9874e+07 1.2068e+09 261.81
## + ratio                     1 5.7459e+05 1.2461e+09 262.26
## - personal_disposal_income  1 2.8411e+10 2.9658e+10 302.63
## 
## Step:  AIC=260.19
## sales ~ personal_disposal_income + budget
## 
##                             Df  Sum of Sq        RSS    AIC
## + ratio                     1  158364757  916674966 259.96
## <none>                                   1075039723 260.19
## - budget                    1  171624035 1246663758 260.27
## + Year                      1    8275335 1066764387 262.08
## - personal_disposal_income  1 2971530267 4046569990 276.75
## 
## Step:  AIC=259.96
## sales ~ personal_disposal_income + budget + ratio
## 
##                             Df  Sum of Sq        RSS    AIC
## <none>                                    916674966 259.96
## - ratio                     1  158364757 1075039723 260.19
## + Year                      1   14933380  901741586 261.73
## - budget                    1  329414202 1246089168 262.26
## - personal_disposal_income  1 2109684588 3026359554 274.68
```

```
## 
## Call:
## lm(formula = sales ~ personal_disposal_income + budget + ratio, 
##     data = toothpaste.data)
## 
## Coefficients:
##              (Intercept)  personal_disposal_income                    budget  
##                34104.559                    85.926                     3.746  
##                    ratio  
##               -30046.343
```

The results of the forward selection method suggests that all 3 predictor variables should be used in the best fitted model as the minimum AIC is 259.96 when all 3 variables are included.

The model is E(y) = 34104.559 + (85.926 * personal_disposal_income) + (3.746 * budget) + (-30046.343 * ratio)

Question 2. Bysinnosis in Cotton Industry Workers

```
byssinosis.data <- read.csv("/Users/charlielock/Documents/R/Datasets/byssinosis-2.cs
v")
head(byssinosis.data)
```

```
##   BysYes BysNo Total Dust Race Sex Smoke Employ
## 1      3    37    40    1    1   1     1      1
## 2      0    74    74    2    1   1     1      1
## 3      2   258   260    3    1   1     1      1
## 4     25   139   164    1    2   1     1      1
## 5      0    88    88    2    2   1     1      1
## 6      3   242   245    3    2   1     1      1
```

xl = dustiness of the workplace (1 = high, 2 = medium, 3 = low) x2 = race (1 = European, 2 = other) x3 = sex (1 = male, 2 = female) x4 = smoking history (1 = smoker, 2 = nonsmoker) x5 = length of employment in the cotton industry (1 = less than 10 years, 2 = between 10 and 20 years, 3 = more than 20 years)

Question 2.a.

```
bys.model <- glm(cbind(BysYes, BysNo) ~ Dust + Race + Sex + Smoke + Employ, family =
binomial, byssinosis.data)

summary(bys.model)
```

```
##
## Call:
## glm(formula = cbind(BysYes, BysNo) ~ Dust + Race + Sex + Smoke +
##     Employ, family = binomial, data = byssinosis.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4126  -0.7573  -0.2421   0.3688   1.9804
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4852     0.6060  -0.801  0.42331
## Dust         -1.3751     0.1155 -11.901  < 2e-16 ***
## Race          0.2463     0.2061   1.195  0.23203
## Sex          -0.2590     0.2116  -1.224  0.22095
## Smoke        -0.6292     0.1931  -3.259  0.00112 **
## Employ        0.3856     0.1069   3.607  0.00031 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 322.527  on 64  degrees of freedom
## Residual deviance:  69.509  on 59  degrees of freedom
## AIC: 188.19
##
## Number of Fisher Scoring iterations: 5
```

Using the summary of the logistic model it can be seen that the race and sex of a person are not a significant effect on the presence of byssinosis as neither have a p-value of less than the significance level of 0.05. However, the dust, smoke and employ variables all have p-values less than 0.05 and can therefore be considered significant.

Final model: y = 1/(1+e^(-0.4852+(-1.3751 * Dust) + (0.2463 * Race) + (-0.2590 * Sex) + (-0.6292 * Smoke) + (0.3856 * Employ)))

Question 2.b.

H0: Logistic regression model fits is adequate H1: Logistic regression model fits is not adequate

```
deviance(bys.model)
```

```
## [1] 69.50926
```

```
pchisq(bys.model$deviance, df=bys.model$df.residual, lower.tail = FALSE)
```

```
## [1] 0.1645594
```

Using the deviance() function the chi-square statistic is calculated to be 69.509. Using the degrees of freedom of 59 in the pchisq() function outputs a p-value of 0.1646 which is greater than the significant level of 0.05 and therefore the null hypothesis can be accepted and the model can be considered as adequate.

Question 2.c.

```
confint.default(bys.model)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.6729540  0.7025201
## Dust        -1.6015997 -1.1486543
## Race        -0.1576071  0.6501886
## Sex         -0.6737381  0.1557304
## Smoke       -1.0075930 -0.2507521
## Employ       0.1760540  0.5951812
```

Using the confint.default() function, the approximate 95% confidence intervals on the model parameters for the linear logistic regression model.

Question 2.d.i.

xl = dustiness of the workplace (1 = high, 2 = medium, 3 = low) x2 = race (1 = European, 2 = other) x3 = sex (1 = male, 2 = female) x4 = smoking history (1 = smoker, 2 = nonsmoker) x5 = length of employment in the cotton industry (1 = less than 10 years, 2 = between 10 and 20 years, 3 = more than 20 years)

```
(1/(1+exp((-0.4852)+(-1.3751*2)+(0.2463*2)+(-0.2590*1)+(-0.6292*2)+(0.3856*3))))
```

```
## [1] 0.9570328
```

Dustiness of workplace: Medium Race: Non-European Sex: Male Smoking History: Non-smoker Length of employment in cotton industry: More than 20 years

95.70% chance of not having byssinosis

Question 2.d.ii.

```
(1/(1+exp((-0.4852)+(-1.3751*1)+(0.2463*2)+(-0.2590*2)+(-0.6292*1)+(0.3856*3))))
```

```
## [1] 0.7954507
```

Dustiness of workplace: High Race: Non-European Sex: Female Smoking History: Smoker Length of employment in cotton industry: More than 20 years

79.55% chance of not having byssinosis.

Question 2.d.iii.

```
(1/(1+exp((-0.4852)+(-1.3751*2)+(0.2463*1)+(-0.2590*1)+(-0.6292*2)+(0.3856*2))))
```

```
## [1] 0.9766903
```

Dustiness of workplace: Medium Race: European Sex: Male Smoking History: Non-smoker Length of employment in cotton industry: Between 10 and 20 years

97.67% chance of not having byssinosis.

Question 2.d.iv.

```
(1/(1+exp((-0.4852)+(-1.3751*3)+(0.2463*1)+(-0.2590*2)+(-0.6292*2)+(0.3856*1))))
```

```
## [1] 0.9968431
```

Dustiness of workplace: Low Race: European Sex: Female Smoking History: Non-smoker Length of employment in cotton industry: Less than 10 years

99.68% chance of not having byssinosis.