Assignment 3

Henrik Olaussen

2023-10-08

Problem 1

a)

Have that

$$E[Y \mid X = x] = \beta_0 + \beta_1 x^{(1)} + ... + \beta_k x^{(k)}$$

where $x^{(k)} = 1$ if $x = k$, and $x^{(k)} = 0$ if $x \neq k$. Furthermore, this can be written as:

$$E[Y \mid X = x] = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$$

with $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_k)$ and $\boldsymbol{x} = (x^{(1)}, x^{(2)}, ..., x^{(k)})^T$. We can think of β_0 as our base case when all entries of $\boldsymbol{x}$ are 0, or as the intercept of our regression. The values of $\boldsymbol{\beta}$ are the slope coefficients in our regression. These entries are also refered to as the regression coefficients, in which we have to estimate. The β_i's hold information about the rate of change per unit of each respective $x^{(k)}$, relative to the other entries of $\boldsymbol{x}$ and the response. However, as we have categorical values, β_i represent the difference between the two cases, $x^{(k)} = 0$ and $x^{(k)} = 1$. The model has the following form for different values of x:

$\underline{x = 0}$:

$$E[Y \mid X = 0] = \beta_0$$

$\underline{x = 2}$:

$$E[Y \mid X = 2] = \beta_0 + \beta_2$$

$\underline{x = k}$:

$$E[Y \mid X = k] = \beta_0 + \beta_k$$

b)

First, we have to convert the data to the desired format:

```
data = chickwts

?chickwts

y = data$weight
data$x = as.numeric(factor(data$feed))-1
x = data$x

for (i in 1:5) {
```

```
  name = paste("x", i, sep="")
  assign(name, ifelse(x == i, 1,0))
}

rm(name)

model1 <- lm(y ~ factor(x1) + factor(x2) + factor(x3) + factor(x4) + factor(x5), data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ factor(x1) + factor(x2) + factor(x3) + factor(x4) +
##     factor(x5), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  323.583     15.834  20.436  < 2e-16 ***
## factor(x1)1 -163.383     23.485  -6.957 2.07e-09 ***
## factor(x2)1 -104.833     22.393  -4.682 1.49e-05 ***
## factor(x3)1  -46.674     22.896  -2.039 0.045567 *
## factor(x4)1  -77.155     21.578  -3.576 0.000665 ***
## factor(x5)1    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

The linear model has been fitted above. The output shows the estimates of $\boldsymbol{\theta}^{*T}$, where $\hat{\beta}_0$ is given by 'intercept', and the estimated sigma, $\hat{\sigma}$, is given by 'Residual standard error'. The estimated values gives the predicted weight of a chicken relative to the base case when they have been fed by one of the feed types. For example, the estimated value is -163.383g relative to the base case if the feed type is $x^{(1)}$. This means that the weight in grams of a chicken that is fed by the feed type corresponding to x1, which is horsebean, has a weight of 163.383g less than the base case (intercept). Furthermore, we want to test independence of X and Y. The following test is performed:

H_0 : Y is independent of X vs H_1 : X and Y are dependent, or equivalently $H_0 : \boldsymbol{\beta}^* = 0$ vs $H_1 : \boldsymbol{\beta}^* \neq 0$. The significance level for this test is $\alpha = 0.01$. A χ^2 test is carried out below:

```
waldtest(model1, test = 'Chisq')
```

```
## Wald test
##
## Model 1: y ~ factor(x1) + factor(x2) + factor(x3) + factor(x4) + factor(x5)
## Model 2: y ~ 1
##   Res.Df Df  Chisq Pr(>Chisq)
## 1     65
```

```
## 2     70 -5 76.824  3.871e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the test, one can see that the p-value is very small (much smaller than 0.01). Consequently, there are evidence against H_0. Hence, the conclusion is that at least one of the entries in $\boldsymbol{\beta}^*$ is not 0, indicating that there is dependence between X and Y.

c)

Now we have a misspecified model for some $\boldsymbol{\theta}^*$. However, the estimate of $\boldsymbol{\theta}^*$ will still be the same.

```
waldtest(model1, test = 'Chisq', vcov = sandwich)
```

```
## Wald test
##
## Model 1: y ~ factor(x1) + factor(x2) + factor(x3) + factor(x4) + factor(x5)
## Model 2: y ~ 1
##   Res.Df Df  Chisq Pr(>Chisq)
## 1     65
## 2     70 -5 117.83  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is still very small. Hence, there is still evidence that X and Y are dependent.

d)

Now, we have a different model involving both categorical variables U and a real random variable V, meaning $\boldsymbol{X}^T = (U, V)$. This model also includes an interaction term between U and V, given by $\sum_{k=1}^{\kappa} \delta_k u^{(k)} v$ for $U = u$ and $V = v$. The value δ_k tells us how big of an influence the interaction between $u^{(k)}$ and v has on the response Y. The relationship is modeled as:

$E[Y \mid \boldsymbol{X} = (u, v)] = \beta_0 + \sum_{k=1}^{\kappa} \beta_k u^{(k)} + \gamma v + \sum_{k=1}^{\kappa} \delta_k u^{(k)} v.$

$\underline{(u, v) = (0, v)}$:

$$E[Y \mid \boldsymbol{X} = (0, v)] = \beta_0 + \gamma v$$

$\underline{(u, v) = (1, v)}$:

$$E[Y \mid \boldsymbol{X} = (1, v)] = \beta_0 + \beta_1 u^{(1)} + \gamma v + \delta_1 u^{(1)} v$$

$\underline{(u, v) = (k, v)}$, for some $k \in \{0, 1, ..., \kappa\}$:

$$E[Y \mid \boldsymbol{X} = (k, v)] = \beta_0 + \beta_k u^{(k)} + \gamma v + \delta_k u^{(k)} v$$

e)

Firstly, we fit the new linear model to the babies data set:

```r
data2 = na.omit(babies) #delete rows with NA values

model2 = lm(bwt ~ as.factor(smoke) + gestation + as.factor(smoke):gestation, data = data2)

summary(model2)
```

```
##
## Call:
## lm(formula = bwt ~ as.factor(smoke) + gestation + as.factor(smoke):gestation,
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.023 -11.078  -0.084   9.995  50.499
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  19.63964   10.29098   1.908 0.056580 .
## as.factor(smoke)1           -72.68713   17.23243  -4.218 2.65e-05 ***
## gestation                     0.36962    0.03671  10.069  < 2e-16 ***
## as.factor(smoke)1:gestation   0.23085    0.06176   3.738 0.000194 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.16 on 1170 degrees of freedom
## Multiple R-squared:  0.2249, Adjusted R-squared:  0.2229
## F-statistic: 113.2 on 3 and 1170 DF,  p-value: < 2.2e-16
```

The estimated value of δ_1^*, $\hat{\delta_1}$, is 0.23085. Secondly, we want to test if Y is dependent on U. This is done through the following test:

H_0: Y is independent of U (δ_1^* and β_1^* are equal to 0)
vs H_1 : Y and U are dependent (At least one of δ_1^* and β_1^* is not 0).
By passing in the models $\{Y \mid \boldsymbol{X} = (u, v)\} = \beta_0 + \beta_1 u^{(1)} + \gamma v + \delta_1 u^{(1)} v$ and $\{Y \mid X = v\} = \beta_0 + \gamma v$ to waldtest(), the mentioned test is performed. In addition, to correct for misspecification, the argument vcov = sandwich is passed in as well.

```r
model2_test = lm(bwt ~ gestation, data = data2)

waldtest(model2_test, model2, vcov = sandwich, test = 'Chisq')
```

```
## Wald test
##
## Model 1: bwt ~ gestation
## Model 2: bwt ~ as.factor(smoke) + gestation + as.factor(smoke):gestation
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   1172
## 2   1170  2 82.877  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of the test is shown above. The p-value of the test is much smaller than the significance level of 0.01. Hence there is evidence that H_0 should be discarded. In other words, the test gives evidence that the weight of a baby is dependent on whether the mother smokes or not.

Thirdly, if the mother is a smoker, the resulting estimated formula is:

$$\{Y \mid \boldsymbol{X} = (1, v)\} = 19.63964 - 72.68713 + 0.36962v + 0.23085v$$

Problem 2

a)

Estimate the parameter:

```
data3 = GAGurine
model3 = lm(GAG ~., data = data3)

summary(model3)
```

```
##
## Call:
## lm(formula = GAG ~ ., data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.950 -4.217 -1.596  2.477 36.470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.89381    0.52553   37.85   <2e-16 ***
## Age         -1.27253    0.07242  -17.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.386 on 312 degrees of freedom
## Multiple R-squared:  0.4974, Adjusted R-squared:  0.4958
## F-statistic: 308.7 on 1 and 312 DF,  p-value: < 2.2e-16
```

Next, we test if there is a decreasing relationship between GAG and Age, using the following hypothesizes at significance level 0.05:

H_0 : There is no decreasing relationship between GAG and Age ($\beta_1 = 0$)
vs H_1 : There is a decreasing relationship ($\beta_1 < 0$).
Moreover, this is an one-sided hypothesis test, meaning we have to divide the p-value by 2 since the waldtest() returns a two-sided test.

```
waldtest(model3, vcov = sandwich)
```

```
## Wald test
##
## Model 1: GAG ~ Age
```

```
## Model 2: GAG ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    312
## 2    313 -1 290.38 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is very small (dividing by 2 will not change the conclusion), H_0 will be discarded at the 0.05 level. Consequently, there is evidence that there is a negative relationship between GAG and Age.

b) Moreover, one can test non-linearity. The following test is performed:

$$H_0 : \{Y \mid X = x\} = \beta_0^* + \beta_1^* x + E$$

vs

$$H_1 : \{Y \mid X = x\} = \beta_0^* + \beta_1^* x + \beta_2^{*2} x + \beta_3^{*3} x + \beta_4^{*4} x$$

```
model3_nonlin = lm(GAG ~ Age + I(Age^2) + I(Age^3) + I(Age^4), data = data3)

waldtest(model3, model3_nonlin, test = 'Chisq', vcov = sandwich)
```

```
## Wald test
##
## Model 1: GAG ~ Age
## Model 2: GAG ~ Age + I(Age^2) + I(Age^3) + I(Age^4)
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    312
## 2    309  3 216.95  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value is very small, we reject H_0. This means that the conclusion is that the non-linear model is better.

c)

There are some assumptions made such that the test statistic in b) has the necessary regularity to be asymptotically normal. Firstly, some conditions are needed for the sample risk to converge, uniformly almost-surly, to the risk. The risk is the expected loss (wrt some loss function) of approximating Y. In this case, Y has been approximated as $\beta_0 + \beta_1 x + \beta_2^2 x + \beta_3^3 + \beta_4^4$. For example, the betas that are used in linear regression, $\beta^* \in \theta^*$, are the betas that minimize this risk (called the least square estimator). More general, the parameters that minimize the risk are denoted θ^*. However, in real life, the parameter values, such as $\beta^* \in \theta^*$ must be obtained from the observed data. In other words,the estimated parameters $\hat{\theta_n}$ are obtained by minimizing the sample risk instead. The sample risk is a sum over the observed losses (loss using the observed data). Moreover, the following assumptions are needed:

1. The data $\mathcal{D}_n$ consists of IID samples.

2. The set $\mathbb{T} \in \mathbb{R}^l$ is compact. Here, $\mathbb{T}$ is some parameter space.

3. The loss ℓ is Caratheodory, meaning $\ell(\cdot, \cdot; \boldsymbol{\theta})$ is measurable for each fixed $\boldsymbol{\theta} \in \mathbb{T}$, and $\ell(\boldsymbol{x}, y; \cdot)$ is continuous for each fixed $\boldsymbol{x} \in \mathbb{R}^m$ and $y \in \mathbb{R}$

4. The loss is dominated in the sense that

$$|\ell(\boldsymbol{X}, Y; \boldsymbol{\theta})| \leq \Delta(\boldsymbol{X}, Y)$$

for all $\boldsymbol{\theta} \in \mathbb{T}$, some dominating function Δ, in the sense that

$$E[\Delta(\boldsymbol{X}, Y)] < \infty$$

If in addition, the risk has a unique minimizer on $\mathbb{T}$, and the conditions above holds, then the sample risk converges almost surly to the risk. Furthermore, to assure asymptotic normality, we have the additional assumptions:

1. The data consists of IID samples.

2. For fixed $\boldsymbol{x}$ and y, the loss has all of its first three partial derivatives (i.e. can be differentiated three times) on some open set $\mathbb{S} \in \mathbb{T}$ and $\boldsymbol{\theta}^* \in \mathbb{T}$, satisfying

$$E[\partial_{\boldsymbol{\theta}} \ell(\boldsymbol{X}, Y, \cdot)(\boldsymbol{\theta}^*)] = 0$$

3. In a neighborhood of $\boldsymbol{\theta}^*$, there exists a dominating function $\Delta(\boldsymbol{X}, Y)$, such that

$$\sum_{i,j,k} \left| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \ell(\boldsymbol{X}, Y; \theta) \right| \leq \Delta(\boldsymbol{X}, Y)$$

in the sense that $E[\Delta(\boldsymbol{X}, Y)] < \infty$.

4. The matrix

$$\boldsymbol{A}(\boldsymbol{\theta}^*) = E\left[-\frac{\partial^2 \ell(\boldsymbol{X}, Y; \cdot)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]$$

exists and is non-singular.

5. The matrix

$$\boldsymbol{B}(\boldsymbol{\theta^*}) = E\left[\{\partial_{\boldsymbol{\theta}} \ell(\boldsymbol{X}, Y; \cdot)(\boldsymbol{\theta^*})\} \{\partial_{\boldsymbol{\theta}} \ell(\boldsymbol{X}, Y; \cdot)(\boldsymbol{\theta^*})\} \right]$$

exists.

6. The sequence $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, ...$ satisfies the consistency condition

$$\hat{\boldsymbol{\theta}}_n \xrightarrow[n \to \infty]{a.s} \boldsymbol{\theta}^*$$

and

$$\frac{1}{\sqrt{n}} \partial_{\boldsymbol{\theta}} r_n(\cdot; \mathcal{D}_n)(\hat{\boldsymbol{\theta}}_n) \xrightarrow[n \to \infty]{a.s} \mathbf{0}$$

d)

One can perform another test than the test performed in b). For, example one can perform the likelihood ratio test using the 'lrtest' function:

```
lrtest(model3, model3_nonlin)
```

```
## Likelihood ratio test
##
## Model 1: GAG ~ Age
## Model 2: GAG ~ Age + I(Age^2) + I(Age^3) + I(Age^4)
##   #Df   LogLik Df  Chisq Pr(>Chisq)
## 1   3 -1026.73
## 2   6  -933.62  3 186.21  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One can see that the p-value is very small, meaning we reject H_0. In other words, the test give the same result as in b), namely that the non-linear model is better.

Problem 3

a)

First, we fit the model:

```
data4 = satgpa
model4 = lm(fy_gpa ~ sat_sum + hs_gpa, data = data4)

summary(model4)
```

```
##
## Call:
## lm(formula = fy_gpa ~ sat_sum + hs_gpa, data = data4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11889 -0.34126  0.02434  0.40953  1.62368
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.873433   0.149028  -5.861 6.25e-09 ***
## sat_sum      0.014402   0.001458   9.878  < 2e-16 ***
## hs_gpa       0.579489   0.038456  15.069  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5941 on 997 degrees of freedom
## Multiple R-squared:  0.3581, Adjusted R-squared:  0.3568
## F-statistic: 278.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

And then we use a built-in function to find the CI for each parameter at the $1 - \alpha = 0.9$ level.

```
alpha = 0.1
coefci(model4, vcov = sandwich, level = 1-alpha)
```

```
##                    5 %        95 %
## (Intercept) -1.1071354 -0.63973057
## sat_sum      0.0120289  0.01677462
## hs_gpa       0.5111611  0.64781646
```

b)

Now, the goal is to produce an asymptotic $100(1-\alpha)\%$ confidence set $\mathbb{C}_{\alpha,n}$, such that

$$P((\beta_1^*, \beta_2^*) \in \mathbb{C}_{\alpha,n}) \geq 1 - \alpha$$

at the $1 - \alpha = 0.9$ level. Firstly, we have

$$P((\beta_1^*, \beta_2^*) \in \mathbb{C}_{\alpha,n}) = P(\beta_1^* \in \mathbb{I}^1_{\alpha,n} \cap \beta_2^* \in \mathbb{I}^2_{\alpha,n}) \geq 1 - (1 - P(\beta_1^* \in \mathbb{I}^1_{\alpha,n}) + (1 - P(\beta_2^* \in \mathbb{I}^2_{\alpha,n})))$$

where $\mathbb{I}^k_{\alpha.n}$ is the $(1-\alpha)100\%$ confidence interval for the k'th parameter β_k^*. In the last inequality, we have applied the Bonferroni inequality. Moreover, by evaluating each confidence interval at the $(1-\alpha)100\%$ level, $P(\beta_k^* \in \mathbb{I}^k_{\alpha,n}) \geq 1 - \alpha$, we have that:

$$1 - (1 - P(\beta_1^* \in \mathbb{I}^1_{\alpha,n}) + (1 - P(\beta_2^* \in \mathbb{I}^2_{\alpha,n}))) \geq 1 - 2\alpha$$

Hence, by evaluating each β_k^* at a $(1-\alpha/2)100\%$ confidence interval, we evaluate (β_0^*, β_1^*) at the $(1-\alpha)100\%$. Consequently, we have that

$$\mathbb{C}_{\alpha,n} = \mathbb{I}^1_{\alpha/2,n} \times \mathbb{I}^2_{\alpha/2,n} \implies \mathbb{C}_{0.1,n} = \mathbb{I}^1_{0.05,n} \times \mathbb{I}^2_{0.05,n}$$

Where $\mathbb{I}^1_{0.05,n}$ and $\mathbb{I}^2_{0.05,n}$ is computed in the code chunk below.

```
coefci(model4, vcov = sandwich, level = 1-0.05)
```

```
##                    2.5 %      97.5 %
## (Intercept) -1.15198596 -0.59487998
## sat_sum      0.01157352  0.01723001
## hs_gpa       0.49804814  0.66092944
```

As a result, we have:

$$\mathbb{C}_{0.1,n} = \mathbb{I}^1_{0.05,n} \times \mathbb{I}^2_{0.05,n} = [0.01157352, 0.01723001] \times [0.49804814, 0.66092944]$$

c)

Furthermore, we can construct an asymptotic $(1-\alpha)100\%$ confidence ellipse for the parameters (β_0^*, β_1^*). From the lecture notes, we have that

$$[\hat{\beta}_1, \hat{\beta}_2]^T \overset{A}{\sim} N([\beta_1^*, \beta_2^*]^T, \frac{1}{n}\hat{C}(\hat{\beta}_1, \hat{\beta}_2))$$

Where $\hat{C}(\beta_1^*, \beta_2^*)$ is the sandwich estimator. Moreover, we get that:

$$n([\hat{\beta}_1, \hat{\beta}_2]^T - [\beta_1^*, \beta_2^*]^T)^T \hat{C}^{-1}(\hat{\beta}_1, \hat{\beta}_2)([\hat{\beta}_1, \hat{\beta}_2]^T - [\beta_1^*, \beta_2^*]^T) \overset{A}{\sim} Q = \chi^2(2)$$

We then get that:

$$P(Q \leq q_{2,1-\alpha}) \geq 1 - \alpha$$

Meaning that our confidence ellipse for $[\beta_0^*, \beta_1^*]^T$ is:

$$\mathbb{E}_{\alpha,n} = \{[\beta_1, \beta_2] \in \mathbb{R}^2 : n([\hat{\beta}_1, \hat{\beta}_2]^T - [\beta_1, \beta_2]^T)^T \hat{C}^{-1}(\hat{\beta}_1, \hat{\beta}_2)([\hat{\beta}_1, \hat{\beta}_2]^T - [\beta_1, \beta_2]^T) \leq q_{2,1-\alpha}\}$$

Moreover, we have

$$n \begin{bmatrix} \hat{\beta}_1 - \beta_1 & \hat{\beta}_2 - \beta_2 \end{bmatrix} \begin{bmatrix} \hat{C}_{11} & \hat{C}_{12} \\ \hat{C}_{21} & \hat{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{bmatrix} \leq q_{2,1-\alpha}$$

In the next derivation, the hats on the entries to the sandwich estimator has been dropped due to simplicity.

$$\frac{n}{C_{11}C_{22} - C_{12}C_{21}}((\hat{\beta}_1 - \beta_1)^2 C_{22} + (\hat{\beta}_2 - \beta_2)(\hat{\beta}_1 - \beta_1)(-C_{21} - C_{12}) + (\hat{\beta}_2 - \beta_2)^2 C_{11}) \leq q_{2,1-\alpha}$$

$$\implies \frac{n}{C_{11}C_{22} - C_{12}C_{21}} \cdot \frac{1}{q_{2,1-\alpha}}((\hat{\beta}_1 - \beta_1)^2 C_{22} + (\hat{\beta}_2 - \beta_2)(\hat{\beta}_1 - \beta_1)(-C_{21} - C_{12}) + (\hat{\beta}_2 - \beta_2)^2 C_{11}) \leq 1$$

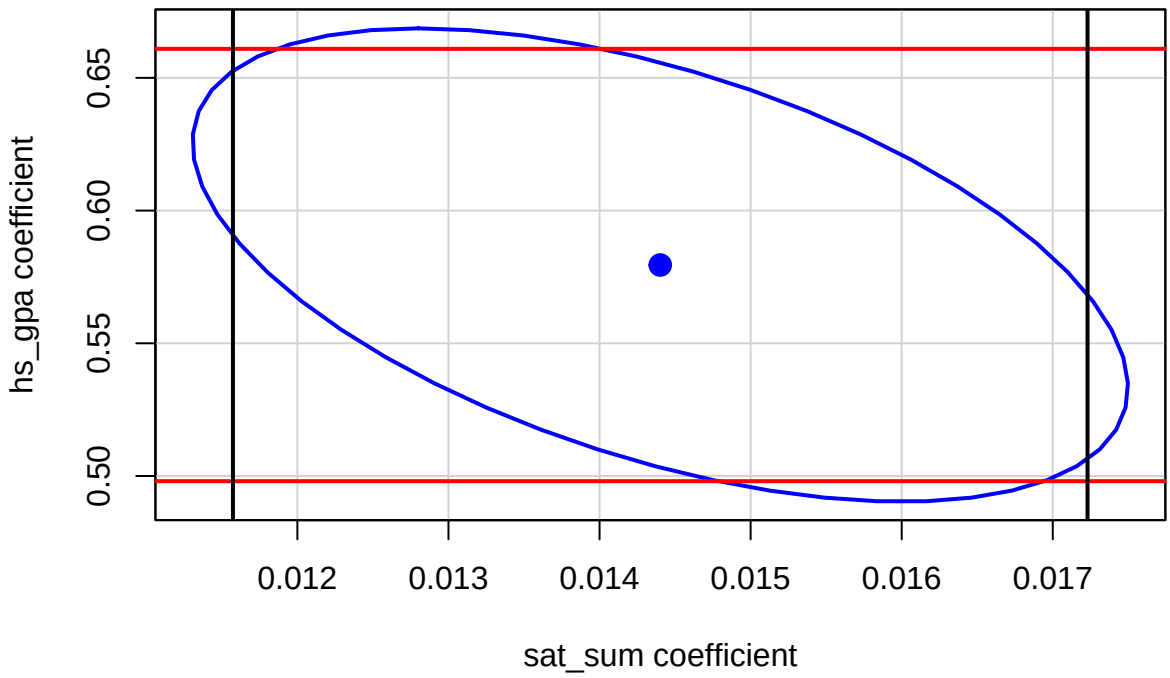
We find the values of the sandwich estimator by the following function:

```
sandwich(model4)
```

```
##                (Intercept)       sat_sum        hs_gpa
## (Intercept)  0.0201495046 -1.177396e-04 -2.316233e-03
## sat_sum     -0.0001177396  2.077218e-06 -3.093475e-05
## hs_gpa      -0.0023162332 -3.093475e-05  1.722389e-03
```

Moreover, the confidence ellipse can be plotted with the function 'confidenceEllipse'. The black lines shows the confidence set (from b) for β_0^* (x-values) and the red lines for β_1^* (y-values):

```
confidenceEllipse(model4, vcov = sandwich, level = 0.9)
abline(a = 0.49804814, b = 0, lwd = 2, col = 'red')
abline(a = 0.66092944, b = 0, lwd = 2, col = 'red')
abline(v = 0.01157352, lwd = 2)
abline(v = 0.01723001, lwd = 2)
```




```
?confidenceEllipse
```

d)

Next, we perform the hypothesis test:

$$H_0 : \{Y \mid \boldsymbol{X} = \boldsymbol{x}\} = \beta_0^* + \beta_1^* u + \beta_2^* v + E$$

vs

$$H_1 : \{Y \mid \boldsymbol{X} = \boldsymbol{x}\} = \beta_0^* + \beta_1^* u + \beta_2^* v + \gamma^* uv + E$$

First we fit the model including the interaction term:

```
model4_alt = lm(fy_gpa ~ sat_sum + hs_gpa + sat_sum:hs_gpa, data = data4)

summary(model4_alt)
```

```
##
## Call:
## lm(formula = fy_gpa ~ sat_sum + hs_gpa + sat_sum:hs_gpa, data = data4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10116 -0.33940  0.01683  0.41035  1.56987
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.875881   0.792732   1.105   0.2695
## sat_sum        -0.002729   0.007763  -0.352   0.7253
## hs_gpa          0.032836   0.246334   0.133   0.8940
## sat_sum:hs_gpa  0.005300   0.002359   2.247   0.0249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5929 on 996 degrees of freedom
## Multiple R-squared:  0.3613, Adjusted R-squared:  0.3594
## F-statistic: 187.8 on 3 and 996 DF,  p-value: < 2.2e-16
```

And then we perform the test.

```
waldtest(model4, model4_alt, vcov = sandwich, test = "Chisq")
```

```
## Wald test
##
## Model 1: fy_gpa ~ sat_sum + hs_gpa
## Model 2: fy_gpa ~ sat_sum + hs_gpa + sat_sum:hs_gpa
##   Res.Df Df Chisq Pr(>Chisq)
## 1    997
## 2    996  1 5.148    0.02327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At 0.1 level, we discard H0, meaning that at this level, the conclusion is that the alternative model (that includes the interaction term) is better.

Problem 4

a)

```
data5 = gpa_study_hours
model5 = lm(gpa ~., data = data5)

summary(model5)
```

```
##
## Call:
## lm(formula = gpa ~ ., data = data5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95130 -0.19456  0.03879  0.21708  0.73872
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.527997   0.037424  94.272   <2e-16 ***
```

```
## study_hours 0.003328   0.001794   1.855   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2837 on 191 degrees of freedom
## Multiple R-squared:  0.01769,    Adjusted R-squared:  0.01255
## F-statistic:  3.44 on 1 and 191 DF,  p-value: 0.06517
```

Interpretation of the value of β_1^* is that for each additional unit(hour) of study, gpa increase by 0.003328.

b)

As the error is $N(0,1)$, we have that $\{Y \mid X = x\} \sim N(\beta_0^* + \beta_1^* x,\ \exp(\gamma_0^* + \gamma_1^* x))$. Moreover, increasing the value of x by one unit gives:

$$\frac{\exp(\gamma_0^* + \gamma_1^*(x+1))}{\exp(\gamma_0^* + \gamma_1^* x)} = \exp(\gamma_1^*)$$

In other words, increasing x by one unit increases the variance with a factor of $\exp(\gamma_1^*)$. Hence, the value $\exp(\gamma_1^*)$ can be interpreted as the variance ratio as x increases with one unit. Larger values of γ_1^* gives a larger ratio. If $\gamma_1^* < 0$, then $\exp(\gamma_1^*) < 1$, meaning the variance decreases with x.

c)

Next, one can estimate the parameter $\boldsymbol{\theta}^* = (\beta_0^*, \beta_1^*, \gamma_0^*, \gamma_1^*)$ by maximum likelihood estimation. The (conditional) pdf of $\{Y \mid X = x\}$ is

$$f(y \mid x) = \frac{1}{\sqrt{2\pi \exp(\gamma_0^* + \gamma_1^* x)}} \exp\{-\frac{1}{2}(\frac{x - (\beta_0^* + \beta_1^* x)}{\sqrt{\exp(\gamma_0^* + \gamma_1^* x)}})^2\}$$

```
Y_vec = matrix(gpa$gpa)
X_vec = matrix(data5$study_hours)

mean_ <- function(beta0, beta1, i) {
  return(beta0 + beta1*X_vec[i])
}

var_ <- function(gamma0, gamma1, i) {
  return(exp(gamma0 + gamma1*X_vec[i]))
}

objective <- function(parameter) {
  beta <- matrix(parameter[1:2], 2, 1)
  gamma <- matrix(parameter[3:4], 2, 1)

  neg_log_like <- 0
  n <- length(Y_vec)

  for (i in 1:n) {
    mu = mean_(beta[1], beta[2], i)
    sigma = sqrt(var_(gamma[1], gamma[2], i))
    neg_log_like <- neg_log_like - (-0.5*log(2*pi)-0.5*log(sigma^2)-1/(2*sigma^2)*(Y_vec[i]-mu)^2)
```

```
  }
  neg_log_like <- neg_log_like / n
  return(neg_log_like)
}

Optimization <- optim(c(1, 1, 1, 1), objective, method = 'BFGS', control = list(maxit = 10000))

print("Optimization Results:")
```

```
## [1] "Optimization Results:"
```

```
print(Optimization)
```

```
## $par
## [1]  3.542099131  0.004117414 -1.516529756 -0.051244276
##
## $value
## [1] 0.2237239
##
## $counts
## function gradient
##       96       32
##
## $convergence
## [1] 0
##
## $message
## NULL
```

The resulting parameters from the optimization are given above under $par in the following order: (beta0, beta1, gamma0, gamma1).

d)

Furthermore, one can test:
H_0 : variance of $Y \mid X = x$ is not a function of x
vs
H_1 : the variance of$Y \mid X = x$ is a function of x.

In other words, we test whether γ_1^* is 0 (indicating that the variance is not a function of x) or not (indicating that the variance is a function of x). We do this by performing a likelihood ratio test (LRT) with $\alpha = 0.01$. The test statistic is:

$$\frac{\mathcal{L}_n(\hat{\beta}_0, \hat{\beta}_1, \gamma_0^*)}{\mathcal{L}_n(\beta_0^*, \beta_1^*, \gamma_0^*, \gamma_1^*)}$$

Where $\mathcal{L}_n(\boldsymbol{\theta}^*)$ is the likelihood function, and $\boldsymbol{\theta}^*$ is MLE of the respective hypothesis. Hence, the MLE of H_0 has to be calculated:

```
objectiveH0 <- function(parameter) {

  neg_log_like <- 0
```

```
  n <- length(Y_vec)

  for (i in 1:n) {
    mu = parameter[1] + parameter[2] * X_vec[i]
    sigma = sqrt(exp(parameter[3]))
    neg_log_like <- neg_log_like - (-0.5*log(2*pi)-0.5*log(sigma^2)-1/(2*sigma^2)*(Y_vec[i]-mu)^2)
  }

  neg_log_like <- neg_log_like / n

  return(neg_log_like)
}

OptimH0 <- optim(c(3,0,1), objectiveH0, method = 'BFGS', control = list(maxit = 10000))

print("Optimization Results:")
```

```
## [1] "Optimization Results:"
```

```
print(OptimH0)
```

```
## $par
## [1]  3.486945886  0.006633234 -2.250774423
##
## $value
## [1] 0.2935502
##
## $counts
## function gradient
##       39       14
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Now, one can perform the LRT:

```
LRT = -2 *log(exp(-OptimH0$value) / exp(-Optimization$value))
LRT
```

```
## [1] 0.1396526
```

The test statistic is given above. The test statistic is distributed as $\chi^2(l_1)$, where $l_1 = 2$. Hence, we have to find the $1-\alpha$-quantile of the $\chi^2(l_1)$ distribution. We reject H_0 if the statistic is larger than the given quantile:

```
alpha = 0.01
qchisq(1-alpha, df = 2)
```

```
## [1] 9.21034
```

As the statistic is smaller than the quantile value, H_0 is not rejected. Moreover, this means that according to this test, there is no evidence that the variance is a function of x.