

Solutions to extra exercises STK2100

Vinnie Ko

April 19, 2018

Exercise 1

(a)

i) Least squares

The *least squares* method: find parameter values that minimizes the residual sum of squares (RSS)

$$\text{RSS} = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2.$$

So, the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2.$$

Or, by using matrix algebra,

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{where } \mathbf{X} \in \mathbb{R}^{n \times (p+1)} \text{ and } \mathbf{y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^{p+1} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

which leads to:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

ii) Maximum likelihood

The error term in linear regression is defined as

$$\varepsilon_1, \dots, \varepsilon_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2).$$

By adding $\mathbf{x}_i^T \boldsymbol{\beta}$ to each ε_i , we obtain

$$\mathbf{x}_1^T \boldsymbol{\beta} + \varepsilon_1, \dots, \mathbf{x}_n^T \boldsymbol{\beta} + \varepsilon_n \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$$

We can now write the likelihood function by using independence

$$L = \prod_{i=1}^n f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2} \right].$$

Then, the log-likelihood is

$$\ell = \log(L) = \sum_{i=1}^n \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2} \right] = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2. \quad (1)$$

To find the maximum likelihood estimator of $\boldsymbol{\beta}$, we have to maximize the equation (1) with respect to $\boldsymbol{\beta}$. And maximizing (1) with respect to $\boldsymbol{\beta}$ is equal to minimizing $\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$ with respect to $\boldsymbol{\beta}$.

Therefore, the maximum likelihood estimator is same as least squares estimator this case.

(b)

From (a), we have

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiate RSS this with respect to $\boldsymbol{\beta}$.

$$\begin{aligned} \text{RSS} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}^T - \boldsymbol{\beta}^T \mathbf{X}^T) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} \\ \frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} &= \frac{\partial (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= 0 - \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X} + (\mathbf{X}\mathbf{X}^T)^T) \boldsymbol{\beta} \\ &= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} \end{aligned} \quad (2)$$

This first derivative should equal to 0. So,

$$\begin{aligned} -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} &= 0 \\ \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T \mathbf{y} \\ \boldsymbol{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \end{aligned}$$

Therefore, the maximum likelihood estimate for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

which is also the least squares estimator.

For further career in statistics, it is handy to know the following matrix differentiation rules:

Let the scalar α be defined by $\alpha = \mathbf{b}^T \mathbf{A} \mathbf{x}$ where $\mathbf{b}$ and $\mathbf{A}$ are not a function of $\mathbf{x}$, then

$$\frac{\partial \mathbf{b}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T \mathbf{b}, \quad \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{b}}{\partial \mathbf{x}} = \mathbf{A} \mathbf{b}, \quad \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x}.$$

These three rules are actually a special case of a more general rule:

Let the scalar α be defined by $\alpha = \mathbf{u}^T \mathbf{A} \mathbf{v}$ where $\mathbf{u} = \mathbf{u}(\mathbf{x})$, $\mathbf{v} = \mathbf{v}(\mathbf{x})$ and $\mathbf{u} : \mathbb{R}^m \rightarrow \mathbb{R}^m$, $\mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, then

$$\frac{\partial \mathbf{u}^T \mathbf{A} \mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{A} \mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \mathbf{A}^T \mathbf{u}.$$

Note that there are several conventions in matrix calculus. In this solution, we stick to the *denominator* layout (a.k.a. *Hessian* formulation).

(c)

In the previous exercise, we obtained the log-likelihood function

$$\ell = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

Differentiate the log-likelihood function with respect to σ^2

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

This first derivative should be equal to 0. So,

$$\begin{aligned} \frac{n}{2\sigma^2} &= \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \\ \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \\ &= \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

Therefore, the maximum likelihood estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Note that $\hat{\sigma}^2$ is a *biased* estimator of σ^2 . The unbiased estimator can be obtained by replacing n with $n - p$.

(d)

i)

We prove more general case $E[\mathbf{X}\mathbf{Y}] = E[\mathbf{X}]E[\mathbf{Y}]$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{p \times m}$, $\mathbf{X} \perp \mathbf{Y}$ and $1 \leq i \leq n$,

$1 \leq k \leq p, 1 \leq j \leq m.$

$$\begin{aligned} \mathbb{E}[\mathbf{XY}] &= \mathbb{E} \left[\begin{bmatrix} x_{1,1} & \cdots & x_{1,k} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,k} & \cdots & x_{i,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,k} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} y_{1,1} & \cdots & y_{1,j} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{k,1} & \cdots & y_{k,j} & \cdots & y_{k,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{p,1} & \cdots & y_{p,j} & \cdots & y_{p,m} \end{bmatrix} \right] \\ &= \mathbb{E} \left[\begin{bmatrix} \sum_{k=1}^p x_{1,k}y_{k,1} & \cdots & \sum_{k=1}^p x_{1,k}y_{k,j} & \cdots & \sum_{k=1}^p x_{1,k}y_{k,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{k=1}^p x_{i,k}y_{k,1} & \cdots & \sum_{k=1}^p x_{i,k}y_{k,j} & \cdots & \sum_{k=1}^p x_{i,k}y_{k,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{k=1}^p x_{n,k}y_{k,1} & \cdots & \sum_{k=1}^p x_{n,k}y_{k,j} & \cdots & \sum_{k=1}^p x_{n,k}y_{k,m} \end{bmatrix} \right] \end{aligned}$$

We can see that the (i, j) entry of $\mathbf{XY}$ is defined as $(\mathbf{XY})_{i,j} = \sum_{k=1}^p x_{i,k}y_{k,j}$. For arbitrary i and j , we have

$$\begin{aligned} \mathbb{E}[(\mathbf{XY})_{i,j}] &= \mathbb{E} \left[\sum_{k=1}^p x_{i,k}y_{k,j} \right] = \sum_{k=1}^p \mathbb{E}[x_{i,k}y_{k,j}] \\ &= \sum_{k=1}^p \mathbb{E}[x_{i,k}] \mathbb{E}[y_{k,j}] \\ &= (\mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{Y}])_{i,j} \end{aligned}$$

That is, $\mathbb{E}[\mathbf{XY}] = \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{Y}]$.

Now, we prove $\mathbb{E}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}]$ where $\mathbf{X} \in \mathbb{R}^{n \times m}$, $\mathbf{Y} \in \mathbb{R}^{n \times m}$ and $1 \leq i \leq n, 1 \leq j \leq m$. Note that $\mathbf{X}$ and $\mathbf{Y}$ don't have to be independent.

The (i, j) entry of $\mathbf{X} + \mathbf{Y}$ is defined as: $(\mathbf{X} + \mathbf{Y})_{i,j} = x_{i,j} + y_{i,j}$.

$$\begin{aligned} \mathbb{E}[(\mathbf{X} + \mathbf{Y})_{i,j}] &= \mathbb{E}[x_{i,j} + y_{i,j}] = \mathbb{E}[x_{i,j}] + \mathbb{E}[y_{i,j}] \\ &= \mathbb{E}[X_{i,j}] + \mathbb{E}[Y_{i,j}] \end{aligned}$$

That is, $\mathbb{E}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}]$.

Finally, by combining the two properties that we just proved, we obtain

$$\mathbb{E}[\mathbf{AZ} + \mathbf{b}] = \mathbb{E}[\mathbf{A}] \mathbb{E}[\mathbf{Z}] + \mathbb{E}[\mathbf{b}] = \mathbf{A} \mathbb{E}[\mathbf{Z}] + \mathbf{b}.$$

ii)

Consider $\mathbf{b}, \mathbf{d}, \mathbf{X}, \mathbf{Y} \in \mathbb{R}^n$ and $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{m \times n}$.

The scalar-version of covariance is defined as:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Consider a matrix $(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T$. The (i, j) entry of this matrix is $(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])$. Thus, the (i, j) entry of the matrix $\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T]$ is $\text{Cov}(X_i, Y_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])]$. That is,

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T].$$

We have

$$\begin{aligned} \text{Cov}(\mathbf{A}\mathbf{X} + \mathbf{b}, \mathbf{C}\mathbf{Y} + \mathbf{d}) &= \mathbb{E}[(\mathbf{A}\mathbf{X} + \mathbf{b} - \mathbb{E}[\mathbf{A}\mathbf{X} + \mathbf{b}])(\mathbf{C}\mathbf{Y} + \mathbf{d} - \mathbb{E}[\mathbf{C}\mathbf{Y} + \mathbf{d}])^T] \\ &= \mathbb{E}[(\mathbf{A}\mathbf{X} + \mathbf{b} - \mathbf{A}\mathbb{E}[\mathbf{X}] - \mathbf{b})(\mathbf{C}\mathbf{Y} + \mathbf{d} - \mathbf{C}\mathbb{E}[\mathbf{Y}] - \mathbf{d})^T] \\ &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbf{A}\mathbb{E}[\mathbf{X}])(\mathbf{C}\mathbf{Y} - \mathbf{C}\mathbb{E}[\mathbf{Y}])^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{C}(\mathbf{Y} - \mathbb{E}[\mathbf{Y}]))^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T \mathbf{C}^T] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T] \mathbf{C}^T \\ &= \mathbf{A}\text{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{C}^T. \end{aligned}$$

That is,

$$\text{Cov}(\mathbf{A}\mathbf{X} + \mathbf{b}, \mathbf{C}\mathbf{Y} + \mathbf{d}) = \mathbf{A}\text{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{C}^T. \quad (3)$$

When $\mathbf{A}\mathbf{X} + \mathbf{b} = \mathbf{C}\mathbf{Y} + \mathbf{d}$, (3) has a special case

$$\text{Var}(\mathbf{A}\mathbf{X} + \mathbf{b}, \mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}\text{Var}(\mathbf{X})\mathbf{A}^T.$$

Note that I assumed that $\mathbf{A}$ and $\mathbf{b}$ are not random matrices/vectors and that their dimensions are well defined such that matrix operations $(+, -, \times, \dots)$ are possible.

(e)

Consider two arbitrary vectors $\mathbf{a}, \mathbf{X} \in \mathbb{R}^n$.

By using (3) from the previous exercise, we have

$$\text{Var}(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \text{Var}(\mathbf{X}) \mathbf{a}.$$

Notice that $\mathbf{a}^T \mathbf{X}$ is a scalar. For convenience we call it α . So, $\alpha = \mathbf{a}^T \mathbf{X}$.

By definition, variance is a non-negative real number, which implies

$$\mathbf{a}^T \text{Var}(\mathbf{X}) \mathbf{a} = \text{Var}(\alpha) \geq 0.$$

That is, a covariance matrix is always positive semi-definite.

(f)

i)

In linear regression, we have $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. By using the fact that $\mathbf{X}$ and $\boldsymbol{\beta}$ are not a random matrix/vector, we have

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{X}]\mathbb{E}[\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta} + 0 = \mathbf{X}\boldsymbol{\beta}.$$

ii)

By using the results from the previous exercises, we can write

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \beta + \varepsilon) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \\ &= \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon.\end{aligned}$$

Take the expectation of the obtained expression of $\hat{\beta}$:

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon] \\ &= \beta + \mathbb{E}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \mathbb{E}[\varepsilon] \\ &= \beta.\end{aligned}$$

(g)

By using the results from the previous exercises, we can write:

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \text{Var}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var}(\mathbf{Y}) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\end{aligned}$$

(h)

```
# For reproducibility.
set.seed(1)

# Set parameter values.
n.vec = seq(10, 100, by = 5)
p = 5
sigma.val = 1

# Make a frame to write down results.
var.beta.0 = as.data.frame(matrix(NA, ncol = 2, nrow = length(n.vec)))
colnames(var.beta.0) = c("n", "var.beta.0")

for (i in 1:length(n.vec)) {
  # Select the value of n.
  n = n.vec[i]
  # Make a frame.
  X = matrix(NA, nrow = n, ncol = p)
  # 1st column contains only 1.
  X[,1] = 1
  # Generate random values from standard normal distribution.
  for (j in 2:p) {
    X[,j] = rnorm(n, mean = 0, sd = 1)
```

```

}
# Create the covariance matrix of beta.
cov.mat.beta = sigma.val*solve(t(X)%*%X)
# Write down the result.
var.beta.0[i,1] = n
var.beta.0[i,2] = cov.mat.beta[1,1]
}

# Plot the result.
plot(x = var.beta.0[,1], y = var.beta.0[,2],
     xlab = "n", ylab = expression(paste("Var(", hat(beta)[0], ")")),
     main = "", font.main = 1)

```

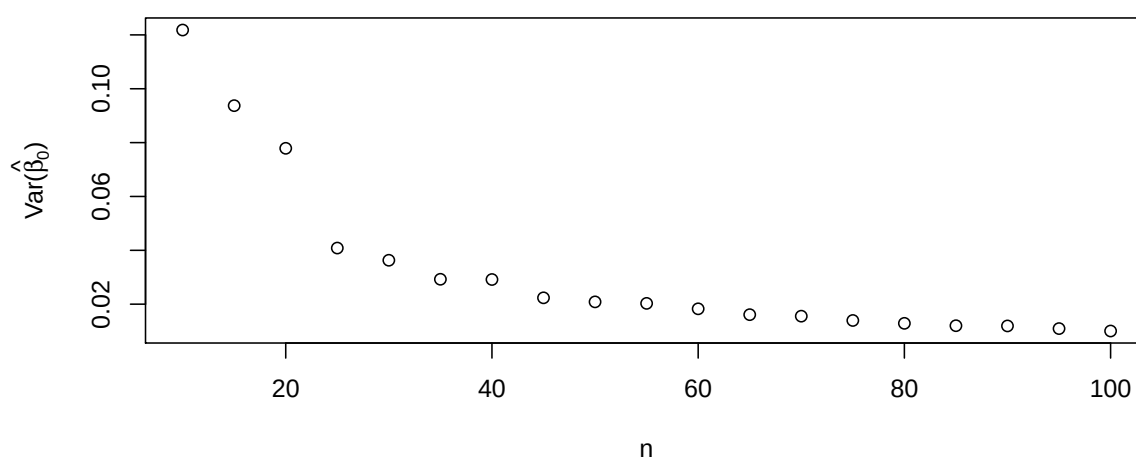


Figure 1: Result of exercise 1 (h).

(i)

```

# For reproducibility.
set.seed(1)

# Set parameter values.
p.vec = seq(20, 32, by = 1)
n = 31
sigma.val = 1

# Make a frame to write down results.
var.beta.0 = as.data.frame(matrix(NA, ncol = 2, nrow = length(p.vec)))
colnames(var.beta.0) = c("p", "var.beta.0")

for (i in 1:length(p.vec)) {
  # Select the value of p.
  p = p.vec[i]
  # Make a frame.

```

```

X = matrix(NA, nrow = n, ncol = p)
# 1st column contains only 1.
X[,1] = 1
# Generate random values from standard normal distribution.
for (j in 2:p) {
  X[,j] = rnorm(n, mean = 0, sd = 1)
}
# Create the covariance matrix of beta.
cov.mat.beta = sigma.val*solve(t(X)%*%X)
# Write down the result.
var.beta.0[i,1] = p
var.beta.0[i,2] = cov.mat.beta[1,1]
}

# Plot the result.
plot(x = var.beta.0[,1], y = var.beta.0[,2],
     xlab = "p", ylab = expression(paste("Var(", hat(beta)[0], ")")),
     main = "", font.main = 1)

```

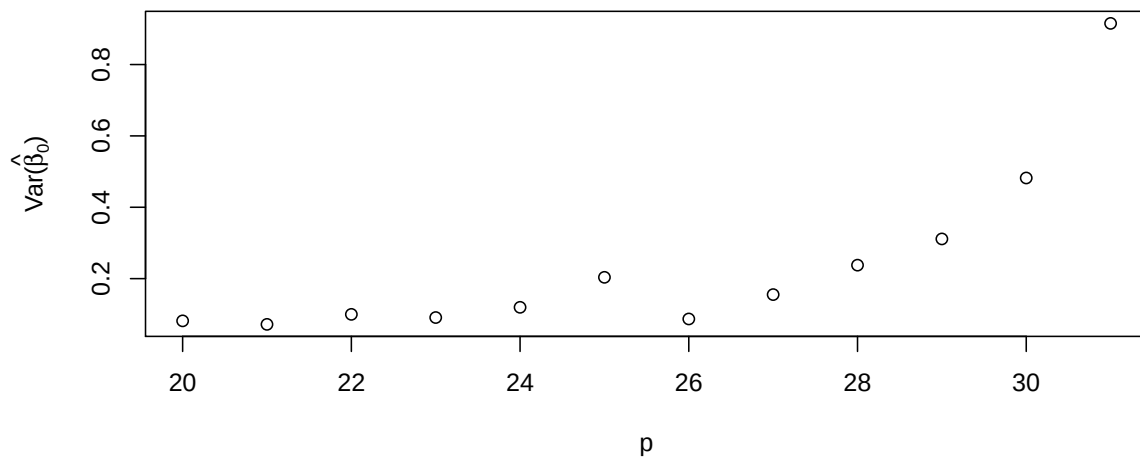


Figure 2: Result of exercise 1 (i).

(j)

Consider a linear regression setting with n data points and p predictors. In this situation, we have to estimate $p + 1$ parameters $(\beta_0, \dots, \beta_p)$ based on n observations.

When n is small (relative to p), $\hat{\beta}_j$ is easily affected by the randomness of an individual data point. But, when n is large (relative to p), this individual effect on $\hat{\beta}_j$ becomes smaller. Therefore, as n increases (relative to p), $\text{Var}(\beta_0)$ decreases.

When $n < p$, there is no unique solution to least squares method and we will get an error in R.

The relationship between $\frac{p}{n}$ and $\text{Var}(\beta_0)$ might be difficult to see in the plots above because n and p are relatively small. So, we generate the same plots again with larger n and p :

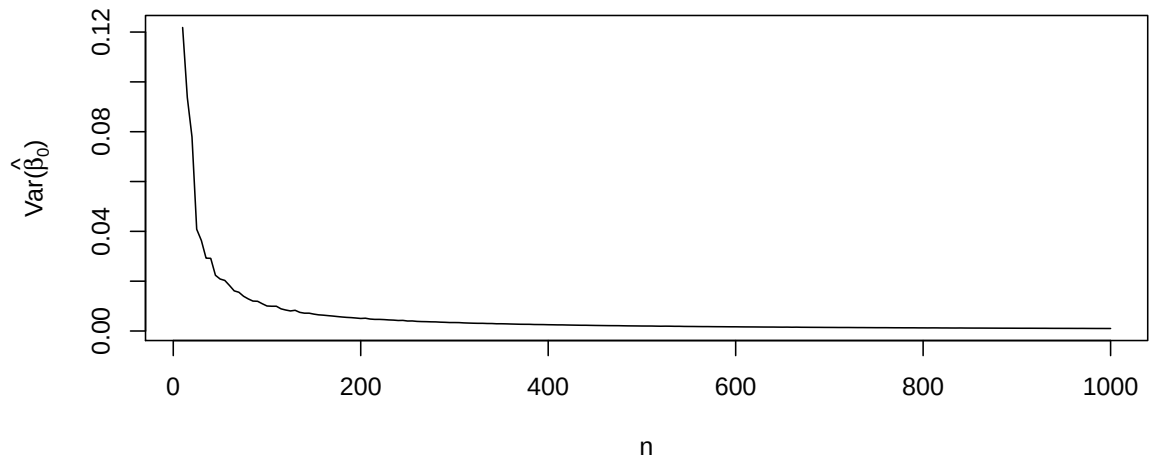


Figure 3: Exercise 1 (h) with: $p = 5$ and $n = 10, 15, \dots, 995, 1000$

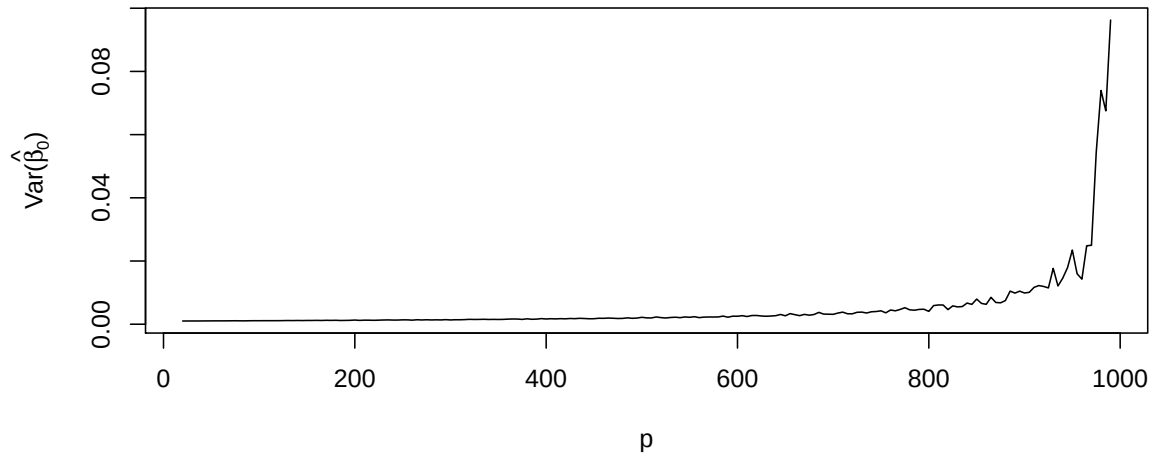


Figure 4: Exercise 1 (i) with: $n = 1000$ and $p = 20, 25, \dots, 985, 990$

Exercise 2

(a)

$$\begin{aligned}
 \text{EPE}(f) &= \mathbb{E}[L(Y, f(\mathbf{X}))] = \mathbb{E}[(Y - f(\mathbf{X}))^2] \\
 &= \int_{\mathbf{x}} \int_y (y - f(\mathbf{x}))^2 p(\mathbf{x}, y) dy d\mathbf{x} \\
 &= \int_{\mathbf{x}} \int_y (y - f(\mathbf{x}))^2 p(\mathbf{x}) p(y|\mathbf{x}) dy d\mathbf{x} \\
 &= \int_{\mathbf{x}} \left(\int_y (y - f(\mathbf{x}))^2 p(y|\mathbf{x}) dy \right) p(\mathbf{x}) d\mathbf{x} \\
 &= \int_{\mathbf{x}} (\mathbb{E}_{Y|\mathbf{X}} [(Y - f(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x}]) p(\mathbf{x}) d\mathbf{x} \\
 &= \mathbb{E}_{\mathbf{X}} [\mathbb{E}_{Y|\mathbf{X}} [(Y - f(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x}]]
 \end{aligned}$$

We are looking for a function f that minimizes $\text{EPE}(f)$ given the data (i.e. $\mathbf{X} = \mathbf{x}$). $\text{EPE}(f)$ becomes

$$\text{EPE}(f) = \mathbb{E}_{\mathbf{X}} [\mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [(Y - f(\mathbf{x}))^2 | \mathbf{X} = \mathbf{x}]]$$

Since all $\mathbf{X}$ are replaced by the given data $\mathbf{x}$, we can ignore $\mathbb{E}_{\mathbf{X}}[\cdot]$. So,

$$\text{EPE}(f) = \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [(Y - f(\mathbf{x}))^2 | \mathbf{X} = \mathbf{x}]$$

We are looking for a function f that minimizes this expression, which is by definition

$$f(\mathbf{x}) = \arg \min_c \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [(Y - c)^2 | \mathbf{X} = \mathbf{x}].$$

(b)

We want the value c that minimizes L .

$$\begin{aligned}
 L &= \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [(Y - c)^2 | \mathbf{X} = \mathbf{x}] \\
 &= \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y^2 - 2Yc + c^2 | \mathbf{X} = \mathbf{x}] \\
 &= \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y^2 | \mathbf{X} = \mathbf{x}] - 2c \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y | \mathbf{X} = \mathbf{x}] + c^2
 \end{aligned}$$

Take the first derivative.

$$\frac{\partial L}{\partial c} = -2 \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y | \mathbf{X} = \mathbf{x}] + 2c$$

This first derivative should equal to 0.

$$\begin{aligned}
 -2 \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y | \mathbf{X} = \mathbf{x}] + 2c &= 0 \\
 c &= \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y | \mathbf{X} = \mathbf{x}]
 \end{aligned}$$

Take the second derivative.

$$\frac{\partial^2 L}{\partial c^2} = 2 > 0$$

Therefore, $c = \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}} [Y | \mathbf{X} = \mathbf{x}]$ is the minimizer of L .

(c)

In the previous exercise, we showed that $c = E_{Y|\mathbf{X}=\mathbf{x}}[Y|\mathbf{X}=\mathbf{x}]$ is the minimizer of $EPE(f)$. We plug in the given expression of Y into this solution.

$$\begin{aligned} c &= E_{Y|\mathbf{X}=\mathbf{x}}[Y|\mathbf{X}=\mathbf{x}] \\ &= E_{Y|\mathbf{X}=\mathbf{x}}[g(\mathbf{x}) + \varepsilon|\mathbf{X}=\mathbf{x}] \\ &= g(\mathbf{x}) + E_{Y|\mathbf{X}=\mathbf{x}}[\varepsilon|\mathbf{X}=\mathbf{x}] \\ &= g(\mathbf{x}) \end{aligned}$$

So, $f(\cdot)$ is the optimal predictor when $f(\cdot) = g(\cdot)$.

(d)

$$\begin{aligned} EPE(f) &= E[(Y - f(\mathbf{X}))^2] \\ &= E[(Y - E[Y] + E[Y] - f(\mathbf{X}))^2] \\ &= E[(Y - E[Y])^2 + (E[Y] - f(\mathbf{X}))^2 + 2(Y - E[Y])(E[Y] - f(\mathbf{X}))] \\ &= E[(Y - E[Y])^2] + E[(E[Y] - f(\mathbf{X}))^2] + 2E[(Y - E[Y])(E[Y] - f(\mathbf{X}))] \\ &= E[(Y - E[Y])^2] + E[(E[Y] - f(\mathbf{X}))^2] + 2E_X[E[(Y - E[Y])(E[Y] - f(\mathbf{X}))|\mathbf{X}]] \\ &\quad \text{(Law of total expectation)} \\ &= E[(Y - E[Y])^2] + E[(E[Y] - f(\mathbf{X}))^2] + 2E_X[(E[Y] - f(\mathbf{X}))E[(Y - E[Y])|\mathbf{X}]] \\ &= E[(Y - E[Y])^2] + E[(E[Y] - f(\mathbf{X}))^2] \\ &= \text{Var}(Y) + E[(E[Y] - f(\mathbf{X}))^2] \\ &= \text{Var}(f(\mathbf{X}) + \varepsilon) + E[(E[Y] - f(\mathbf{X}))^2] \\ &= \text{Var}(f(\mathbf{X})) + \text{Var}(\varepsilon) + E[(E[Y] - f(\mathbf{X}))^2] \\ &= \text{Var}(f(\mathbf{X})) + \sigma^2 + E[(E[Y] - f(\mathbf{X}))^2] \end{aligned}$$

The last term will be 0 when $E[Y] = f(\mathbf{X})$. So, the lower bound is $\text{Var}(f(\mathbf{X})) + \sigma^2$.

Exercise 3

(a)

This is quite straightforward.

$$\begin{aligned}
 \text{EPE}(f) &= \mathbb{E}[L(Y, f(\mathbf{X}))] = \mathbb{E}[1 - I_{\{f(\mathbf{x})\}}(y)] \\
 &= \int_{\mathbf{x}} \int_y (1 - I_{\{f(\mathbf{x})\}}(y)) p(\mathbf{x}, y) dy d\mathbf{x} \\
 &= \int_{\mathbf{x}} \int_y (1 - I_{\{f(\mathbf{x})\}}(y)) p(\mathbf{x}) p(y|\mathbf{x}) dy d\mathbf{x} \\
 &= \int_{\mathbf{x}} \left(\int_y (1 - I_{\{f(\mathbf{x})\}}(y)) p(y|\mathbf{x}) dy \right) p(\mathbf{x}) d\mathbf{x} \\
 &= \int_{\mathbf{x}} (1 - \Pr(Y = f(\mathbf{x})|\mathbf{X} = \mathbf{x})) p(\mathbf{x}) d\mathbf{x}
 \end{aligned}$$

(b)

$$\text{EPE}(f) = \int_{\mathbf{x}} \{1 - \Pr(Y = f(\mathbf{x})|\mathbf{X} = \mathbf{x})\} p(\mathbf{x}) d\mathbf{x}$$

We are looking for a function f that minimizes this expression, which is by definition

$$f(\mathbf{x}) = \arg \min_k [1 - \Pr(Y = k|\mathbf{X} = \mathbf{x})] = \arg \max_k [\Pr(Y = k|\mathbf{X} = \mathbf{x})] \text{ where } k \in 0, 1.$$

Since $f(\mathbf{x})$ is a binary predictor, we have only 2 options for the value of $f(\mathbf{x})$: 0 and 1.

We are maximizing $\Pr(Y = k|\mathbf{X} = \mathbf{x})$. So, if $\Pr(Y = 0|\mathbf{X} = \mathbf{x}) < \Pr(Y = 1|\mathbf{X} = \mathbf{x})$, $k = 1$. And if $\Pr(Y = 0|\mathbf{X} = \mathbf{x}) > \Pr(Y = 1|\mathbf{X} = \mathbf{x})$, $k = 0$.

Notice that $\Pr(Y = 0|\mathbf{X} = \mathbf{x}) + \Pr(Y = 1|\mathbf{X} = \mathbf{x}) = 1$. So, the decision boundary is at $\Pr(Y = 0|\mathbf{X} = \mathbf{x}) = \Pr(Y = 1|\mathbf{X} = \mathbf{x}) = 0.5$

Therefore,

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \Pr(Y = 1|\mathbf{X} = \mathbf{x}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

(c)

Intuitively,

$$f(\mathbf{x}) = \begin{cases} K-1 & \text{if } K-1 = \arg \max_k [\Pr(Y = k|\mathbf{X} = \mathbf{x})] \\ K-2 & \text{if } K-2 = \arg \max_k [\Pr(Y = k|\mathbf{X} = \mathbf{x})] \\ \vdots & \vdots \\ 1 & \text{if } 1 = \arg \max_k [\Pr(Y = k|\mathbf{X} = \mathbf{x})] \\ 0 & \text{otherwise} \end{cases}$$

(d)

Assume $k_{opt} : \arg \max_k [\Pr(Y = k | \mathbf{X} = \mathbf{x})]$. We get an error when $Y \neq k_{opt}$. The probability that this happens is $1 - \Pr(Y = k_{opt} | \mathbf{X} = \mathbf{x})$ which corresponds to $1 - \max_k \Pr(Y = k | \mathbf{x})$.

Exercise 4

(a)

See `extra4.r` on the course webpage.

(b)

It's given that

$$X \sim N(0, 1), \eta \sim N(0, 1) \text{ and } X \perp \eta.$$

Now we assume

$$Z = 0.9X + \sqrt{1 - 0.9^2}\eta,$$

then

$$\begin{aligned}\text{Var}(Z) &= \text{Var}(0.9X + \sqrt{1 - 0.9^2}\eta) \\ &= 0.9^2\text{Var}(X) + (1 - 0.9^2)\text{Var}(\eta) \\ &= 0.9^2 \cdot 1 + (1 - 0.9^2) \cdot 1 \\ &= 1.\end{aligned}$$

By using the rule $\text{Cov}(aX + bY, cU + dV) = ac\text{Cov}(X, U) + ad\text{Cov}(X, V) + bc\text{Cov}(Y, U) + bd\text{Cov}(Y, V)$, we have

$$\begin{aligned}\text{Cov}(X, Z) &= \text{Cov}(X, 0.9X + \sqrt{1 - 0.9^2}\eta) \\ &= \text{Cov}(X, 0.9X) + \text{Cov}(X, \sqrt{1 - 0.9^2}\eta) \\ &= 0.9\text{Cov}(X, X) + \sqrt{1 - 0.9^2}\text{Cov}(X, \eta) \\ &= 0.9 \cdot 1 + \sqrt{1 - 0.9^2} \cdot 0 \\ &= 0.9\end{aligned}$$

and

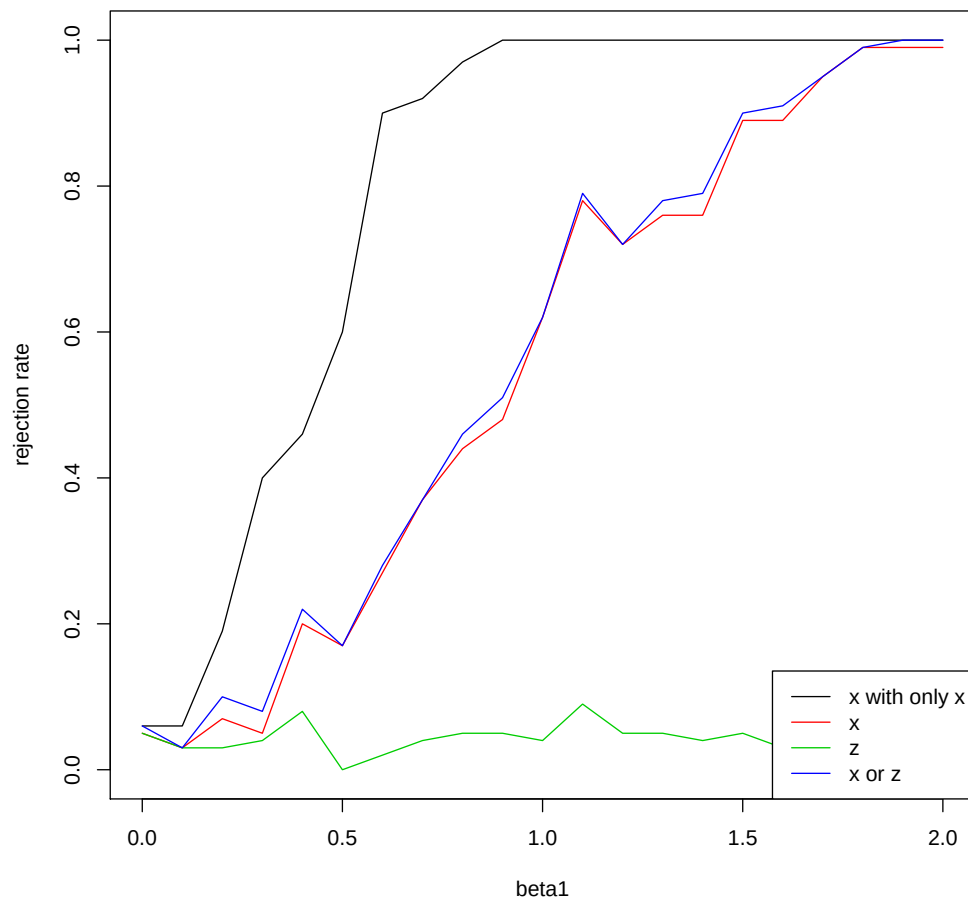
$$\text{Cor}(X, Z) = \frac{\text{Cov}(X, Z)}{\sqrt{\text{Var}(X)\text{Var}(Z)}} = 0.9.$$

So, defining Z with $\{Z = 0.9X + \sqrt{1 - 0.9^2}\eta \text{ and } \eta \sim N(0, 1)\}$ is same as defining Z with $\{Z \sim N(0, 1) \text{ and } \text{Cor}(X, Z) = 0.9\}$.

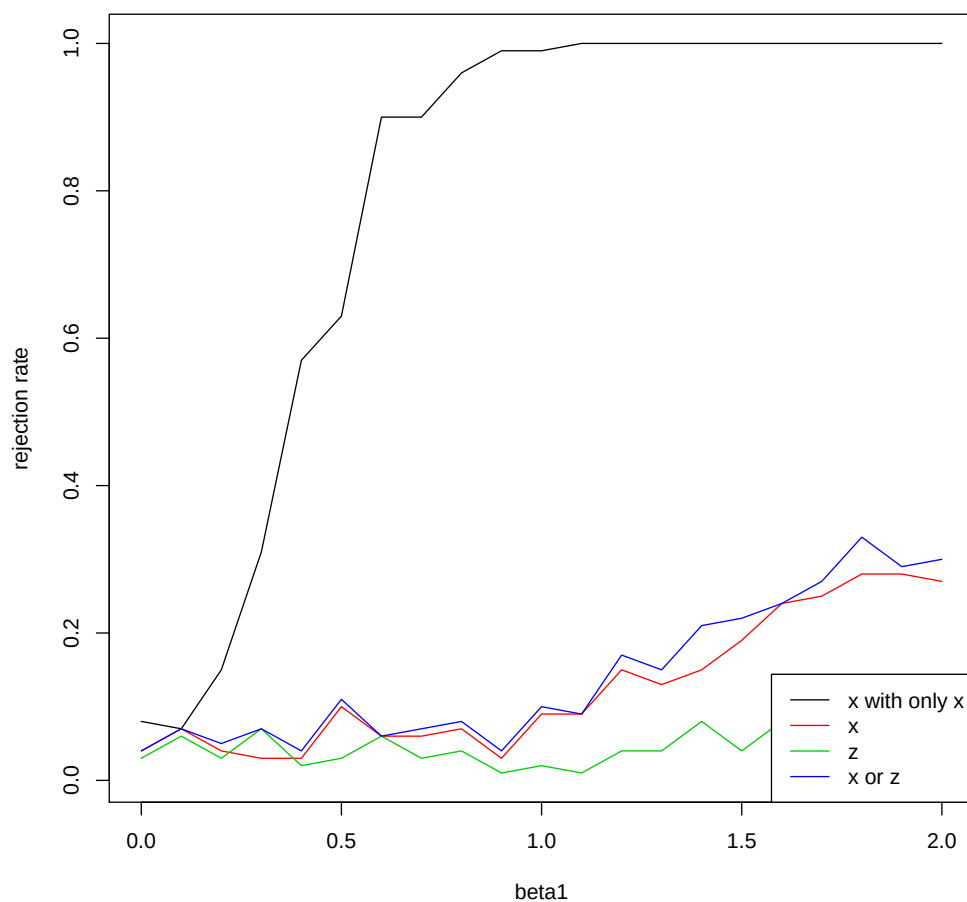
When we simulate (X, Z) in `R`, both generating algorithms will give the same result, except for the differences created by the random number generator.

(c)

See `extra4_extended.r` on the course webpage.



(d)



(e)

z has a high correlation with x . When z is added to the model, it takes over a part of variance y that is previously explained by x . So, the rejection rate of β_x decreases.

Exercise 5

(a)

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[(\mathbf{x}^*)^T \hat{\boldsymbol{\beta}}] = (\mathbf{x}^*)^T \mathbb{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{x}^*)^T \boldsymbol{\beta} = \theta \quad \text{where} \quad \mathbf{x}^* = \begin{bmatrix} 1 \\ x_1^* \\ \vdots \\ x_p^* \end{bmatrix}$$

So, $\hat{\theta}$ is an unbiased estimator of θ .

(b)

$$\sigma_{\hat{\theta}}^2 = \text{Var}(\hat{\theta}) = \text{Var}((\mathbf{x}^*)^T \hat{\boldsymbol{\beta}}) = (\mathbf{x}^*)^T \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}^* = (\mathbf{x}^*)^T \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* = \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*$$

Here, $\mathbf{X}$ is the design matrix that is used to fit the model (i.e. to estimate $\boldsymbol{\beta}$) and $\sigma^2 = \text{Var}(\varepsilon)$. $\mathbf{x}^*$ is the new data point for prediction.

(c)

i)

From the previous exercise, we have: $\sigma_{\hat{\theta}}^2 = \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*$. So, $s_{\hat{\theta}}^2 = \hat{\sigma}_{\hat{\theta}}^2 = \hat{\sigma}^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*$

$$\begin{aligned} T &= \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \\ &= \frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}} \\ &= \frac{\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}}}{\sqrt{\frac{\hat{\sigma}^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}{\sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}}} \\ &= \frac{\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \\ &= \frac{\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}}}{\sqrt{\frac{\hat{\sigma}^2 (n-p-1)}{\sigma^2 (n-p-1)}}} \\ &= \frac{\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*}}}{\sqrt{\frac{\hat{\sigma}^2 (n-p-1)}{\sigma^2} \cdot \frac{1}{n-p-1}}} \\ &= \frac{Z}{\sqrt{\frac{X}{n-p-1}}} \end{aligned}$$

Now we need to show that $Z \sim N(0, 1)$, $X \sim \chi_{n-p-1}^2$ and $Z \perp\!\!\!\perp X$.

We know that $\hat{\theta} \sim N(\theta, \sigma^2(\mathbf{x}^*)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*)$. So, $Z = \frac{\hat{\theta} - \theta}{\sqrt{\sigma^2(\mathbf{x}^*)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}} \sim N(0, 1)$.

As a direct result of the given property, we obtain: $X = \frac{\hat{\sigma}^2}{\sigma^2}(n - p - 1) \sim \chi_{n-p-1}^2$.

It's given that $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$. So, $\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2(\mathbf{x}^*)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}} \perp\!\!\!\perp \frac{\hat{\sigma}^2}{\sigma^2}(n - p - 1)$.

Therefore,

$$T = \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} = \frac{Z}{\sqrt{\frac{X}{n-p-1}}} \sim t_{n-p-1}.$$

ii)

$$T = \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \sim t_{n-p-1}$$

So,

$$\begin{aligned} & P\left(t_{\frac{\alpha}{2}, n-p-1} \leq \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \leq t_{1-\frac{\alpha}{2}, n-p-1}\right) \\ &= P\left(\hat{\theta} - t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{\hat{\theta}} \leq \theta \leq \hat{\theta} + t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{\hat{\theta}}\right) \\ &= P\left(\hat{\theta} - t_{1-\frac{\alpha}{2}, n-p-1} \cdot \hat{\sigma} \sqrt{(\mathbf{x}^*)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*} \leq \theta \leq \hat{\theta} + t_{1-\frac{\alpha}{2}, n-p-1} \cdot \hat{\sigma} \sqrt{(\mathbf{x}^*)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}\right) \\ &= 1 - \alpha \end{aligned}$$

To sum up, the $100(1 - \alpha)\%$ confidence interval for θ : $\left[\hat{\theta} - t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{\hat{\theta}}, \hat{\theta} + t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{\hat{\theta}}\right]$

(d)

$$\begin{aligned} \mathbb{E}[Y^* - \hat{\theta}] &= \mathbb{E}\left[(\mathbf{x}^*)^T\boldsymbol{\beta} + \varepsilon^* - (\mathbf{x}^*)^T\hat{\boldsymbol{\beta}}\right] \\ &= (\mathbf{x}^*)^T\boldsymbol{\beta} - (\mathbf{x}^*)^T\mathbb{E}[\hat{\boldsymbol{\beta}}] + \mathbb{E}[\varepsilon^*] \\ &= (\mathbf{x}^*)^T\boldsymbol{\beta} - (\mathbf{x}^*)^T\boldsymbol{\beta} + 0 \\ &= 0 \end{aligned}$$

The result that we obtain here is $\mathbb{E}[\hat{\theta}] = \mathbb{E}[Y^*]$ and not $\mathbb{E}[\hat{\theta}] = Y^*$.

$$\mathbb{E}[\hat{\theta}] - Y^* = (\mathbf{x}^*)^T\boldsymbol{\beta} - (\mathbf{x}^*)^T\boldsymbol{\beta} - \varepsilon^* = -\varepsilon^* \neq 0$$

(e)

$$\begin{aligned}\sigma_{Y^* - \hat{\theta}}^2 &= \text{Var}((\mathbf{x}^*)^T \boldsymbol{\beta} - (\mathbf{x}^*)^T \hat{\boldsymbol{\beta}} + \varepsilon^*) \\ &= \text{Var}((\mathbf{x}^*)^T \hat{\boldsymbol{\beta}}) + \text{Var}(\varepsilon^*) \\ &= \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + \sigma^2 \\ &= \sigma^2 ((\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)\end{aligned}$$

(f)

i)

First, show that $Y^* - \hat{\theta}$ follows a normal distribution.

$$\begin{aligned}Y^* - \hat{\theta} &= (\mathbf{x}^*)^T \boldsymbol{\beta} - (\mathbf{x}^*)^T \hat{\boldsymbol{\beta}} + \varepsilon^* \\ &= (\mathbf{x}^*)^T \boldsymbol{\beta} - (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} + \varepsilon^* \\ &= (\mathbf{x}^*)^T \boldsymbol{\beta} - (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \varepsilon^* \\ &= (\mathbf{x}^*)^T \boldsymbol{\beta} - (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} - (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\varepsilon} + \varepsilon^* \\ &= -(\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\varepsilon} + \varepsilon^* \\ &\sim N(0, (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 ((\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T) + N(0, \sigma^2) \\ &\quad \text{(because } -(\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\varepsilon} \perp \varepsilon^*) \\ &= N(0, \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T ((\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T) + N(0, \sigma^2) \\ &= N(0, \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*) + N(0, \sigma^2) \\ &= N(0, \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^*) + N(0, \sigma^2) \\ &= N(0, \sigma^2 (\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + \sigma^2) \quad \text{(by the additivity of independent normal distributions)} \\ &= N(0, \sigma^2 ((\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1))\end{aligned}$$

We have: $\sigma_{Y^* - \hat{\theta}}^2 = \sigma^2 ((\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)$. So, $s_{Y^* - \hat{\theta}}^2 = \hat{\sigma}_{Y^* - \hat{\theta}}^2 = \hat{\sigma}^2 ((\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)$

$$\begin{aligned}
T &= \frac{Y^* - \hat{\theta}}{s_{Y^* - \hat{\theta}}} \\
&= \frac{Y^* - \hat{\theta}}{\sqrt{\hat{\sigma}^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}} \\
&= \frac{\frac{Y^* - \hat{\theta}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}}}{\frac{\sqrt{\hat{\sigma}^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}}} \\
&= \frac{\frac{Y^* - \hat{\theta}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \\
&= \frac{\frac{Y^* - \hat{\theta}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}}}{\sqrt{\frac{\hat{\sigma}^2(n-p-1)}{\sigma^2(n-p-1)}}} \\
&= \frac{\frac{Y^* - \hat{\theta}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}}}{\sqrt{\frac{\hat{\sigma}^2(n-p-1)}{\sigma^2} \cdot \frac{1}{n-p-1}}} \\
&= \frac{Z}{\sqrt{\frac{X}{n-p-1}}}
\end{aligned}$$

Now we need to show that $Z \sim N(0, 1)$, $X \sim \chi_{n-p-1}^2$ and $Z \perp\!\!\!\perp X$.

We know that $Y^* - \hat{\theta} \sim N(0, \sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1))$. So, $Z = \frac{Y^* - \hat{\theta}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}} \sim N(0, 1)$.

As a direct result of the given property, we obtain: $X = \frac{\hat{\sigma}^2}{\sigma^2}(n-p-1) \sim \chi_{n-p-1}^2$.

It's given that $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$. So, $\frac{Y^* - \hat{\theta}}{\sqrt{\sigma^2((\mathbf{x}^*)^T(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1)}} \perp\!\!\!\perp \frac{\hat{\sigma}^2}{\sigma^2}(n-p-1)$.

Therefore,

$$T = \frac{Y^* - \hat{\theta}}{s_{Y^* - \hat{\theta}}} = \frac{Z}{\sqrt{\frac{X}{n-p-1}}} \sim t_{n-p-1}.$$

ii)

$$T = \frac{Y^* - \hat{\theta}}{s_{Y^* - \hat{\theta}}} \sim t_{n-p-1}$$

So,

$$\begin{aligned}
& P \left(t_{\frac{\alpha}{2}, n-p-1} \leq \frac{Y^* - \hat{\theta}}{s_{Y^* - \hat{\theta}}} \leq t_{1-\frac{\alpha}{2}, n-p-1} \right) \\
&= P \left(\hat{\theta} - t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{Y^* - \hat{\theta}} \leq Y^* \leq \hat{\theta} + t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{Y^* - \hat{\theta}} \right) \\
&= P \left(\hat{\theta} - t_{1-\frac{\alpha}{2}, n-p-1} \cdot \hat{\sigma} \sqrt{(\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1} \leq Y^* \leq \hat{\theta} + t_{1-\frac{\alpha}{2}, n-p-1} \cdot \hat{\sigma} \sqrt{(\mathbf{x}^*)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^* + 1} \right) \\
&= 1 - \alpha
\end{aligned}$$

To sum up, the $100(1-\alpha)\%$ *prediction interval* for Y^* : $\left[\hat{\theta} - t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{Y^* - \hat{\theta}}, \hat{\theta} + t_{1-\frac{\alpha}{2}, n-p-1} \cdot s_{Y^* - \hat{\theta}} \right]$

Exercise 7

(a)

(i)

$p = 0$, so the model: $Y_i = \beta_0 + \varepsilon_i$, and $X = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

The least squares estimate is given by:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

which in this case leads to

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

So,

$$\hat{y}_i = \bar{y} \text{ for } 1 \leq i \leq n.$$

(ii)

Same procedure as in (i), but you have to replace $\mathbf{X}$ and $\mathbf{y}$ with $\mathbf{X}_{-i}$ and $\mathbf{y}_{-i}$ by removing the i -th data point.

The resulting prediction:

$$\widehat{y}_i^{-i} = \frac{1}{n-1} \sum_{j \neq i} y_j.$$

(iii)

$$\begin{aligned} \mathbf{H} &= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \cdot \left([1 \cdots 1] \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \cdot [1 \cdots 1] \\ &= n^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \cdot [1 \cdots 1] \\ &= \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \end{aligned}$$

Thus,

$$h_{ii} = \frac{1}{n}.$$

(iv)

$$\begin{aligned}
y_i - \widehat{y}_i^i &= y_i - \frac{\sum_{j \neq i} y_j}{n-1} \\
&= y_i - \frac{\sum_{i'=1}^n y_{i'} - y_i}{n-1} \\
&= y_i - \frac{\frac{\sum_{i'=1}^n y_{i'}}{n} - \frac{y_i}{n}}{\frac{n-1}{n}} \\
&= y_i - \frac{\widehat{y}_i - \frac{y_i}{n}}{1 - \frac{1}{n}} \\
&= \frac{(1 - \frac{1}{n})y_i + \frac{y_i}{n} - \widehat{y}_i}{1 - \frac{1}{n}} \\
&= \frac{y_i - \widehat{y}_i}{1 - \frac{1}{n}} \\
&= \frac{y_i - \widehat{y}_i}{1 - h_i} \quad (\text{by using the result from (iii)})
\end{aligned}$$

(b)

(i)

$$\begin{aligned}
M_n &= \mathbf{X}_n^T \mathbf{X}_n \\
&= \begin{bmatrix} x_{1,1} & \cdots & x_{i,1} & \cdots & x_{n,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1,j} & \cdots & x_{i,j} & \cdots & x_{n,j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1,p} & \cdots & x_{i,p} & \cdots & x_{n,p} \end{bmatrix} \cdot \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,p} \end{bmatrix} \\
&= [\mathbf{x}_1 \cdots \mathbf{x}_n] \cdot \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \\
&= \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T
\end{aligned}$$

(ii)

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

if and only if

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T) \left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}} \right) = \mathbf{I} \text{ and } \left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}} \right) (\mathbf{A} + \mathbf{u}\mathbf{v}^T) = \mathbf{I}$$

For convenience, let's write $c = \frac{1}{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}}$.

First condition:

$$\begin{aligned}
(\mathbf{A} + \mathbf{u}\mathbf{v}^T) \left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1}}{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}} \right) &= (\mathbf{A} + \mathbf{u}\mathbf{v}^T) (\mathbf{A}^{-1} - c \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1}) \\
&= \mathbf{A} \mathbf{A}^{-1} - c \mathbf{A} \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} + \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} - c \mathbf{u} (\mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}) \mathbf{v}^T \mathbf{A}^{-1} \\
&= \mathbf{I} - c \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} + \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} - c (\mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}) \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} \\
&= \mathbf{I} + (-c + 1 - c \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}) \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} \\
&= \mathbf{I} + \left(\frac{-1 + 1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u} - \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}}{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}} \right) \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} \\
&= \mathbf{I}
\end{aligned}$$

Second condition:

$$\begin{aligned}
\left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1}}{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}} \right) (\mathbf{A} + \mathbf{u}\mathbf{v}^T) &= (\mathbf{A}^{-1} - c \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1}) (\mathbf{A} + \mathbf{u}\mathbf{v}^T) \\
&= \mathbf{A}^{-1} \mathbf{A} + \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T - c \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T - c \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \\
&= \mathbf{I} + (1 - c - c \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T) \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \\
&= \mathbf{I} + \left(\frac{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u} - 1 - \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}}{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}} \right) \mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \\
&= \mathbf{I}
\end{aligned}$$

Thus,

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{u} \mathbf{v}^T \mathbf{A}^{-1}}{1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}}.$$

(iii)

Let $\tilde{\mathbf{x}}_n = \mathbf{M}_{n-1}^{-1} \mathbf{x}_n$, then

$$\begin{aligned}
\mathbf{M}_n^{-1} &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \\
&= \left(\sum_{i=1}^{n-1} \mathbf{x}_i \mathbf{x}_i^T + \mathbf{x}_n \mathbf{x}_n^T \right)^{-1} \\
&= (\mathbf{M}_{n-1}^{-1} + \mathbf{x}_n \mathbf{x}_n^T)^{-1} \\
&= \mathbf{M}_{n-1}^{-1} - \frac{\mathbf{M}_{n-1}^{-1} \mathbf{x}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1} \mathbf{x}_n} && \text{(by using the Sherman-Morrison formula)} \\
&= \mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n}
\end{aligned}$$

(iv)

$$\begin{aligned}
\hat{\beta} &= \hat{\beta}_n \\
&= (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{y}_n \\
&= \mathbf{M}_n^{-1} \mathbf{X}_n^T \mathbf{y}_n \\
&= \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \sum_{i=1}^n \mathbf{x}_i y_i \\
&= \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \left(\sum_{i=1}^{n-1} \mathbf{x}_i y_i + \mathbf{x}_n y_n \right) \\
&= \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\mathbf{X}_{n-1}^T \mathbf{y}_{n-1} + \mathbf{x}_n y_n) \\
&= \mathbf{M}_{n-1}^{-1} \mathbf{X}_{n-1}^T \mathbf{y}_{n-1} + \mathbf{M}_{n-1}^{-1} \mathbf{x}_n y_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \mathbf{X}_{n-1}^T \mathbf{y}_{n-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \mathbf{x}_n y_n \\
&= \hat{\beta}_{n-1} + \mathbf{M}_{n-1}^{-1} \mathbf{x}_n y_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \hat{\beta}_{n-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \mathbf{x}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \mathbf{M}_{n-1}^{-1} \mathbf{x}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \tilde{\mathbf{x}}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n)
\end{aligned}$$

Or alternatively,

$$\begin{aligned}
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \tilde{\mathbf{x}}_n y_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n - \frac{\mathbf{x}_n^T \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n y_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \left(1 - \frac{\mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \tilde{\mathbf{x}}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \frac{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n - \mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \tilde{\mathbf{x}}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \frac{1}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \tilde{\mathbf{x}}_n y_n
\end{aligned}$$

(v)

To prove $\left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right)^{-1} = \mathbf{I} + \tilde{\mathbf{x}}_n \mathbf{x}_n^T$, show

$$\left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\mathbf{I} + \tilde{\mathbf{x}}_n \mathbf{x}_n^T) = \mathbf{I} \quad \text{and} \quad (\mathbf{I} + \tilde{\mathbf{x}}_n \mathbf{x}_n^T) \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) = \mathbf{I}$$

or simply use the Sherman-Morison formula.

Use the result from (iv):

$$\begin{aligned}\hat{\beta}_n &= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n) \\ \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right)^{-1} \hat{\beta}_n &= \hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n \\ \hat{\beta}_{n-1} &= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right)^{-1} \hat{\beta}_n - \tilde{\mathbf{x}}_n y_n.\end{aligned}$$

Use the equality that we just proved

$$\hat{\beta}_{n-1} = (\mathbf{I} + \tilde{\mathbf{x}}_n \mathbf{x}_n^T) \hat{\beta}_n - \tilde{\mathbf{x}}_n y_n.$$

(vi)

$$\begin{aligned}y_n - \hat{y}_n^- &= y_n - \mathbf{x}_n^T \hat{\beta}_{n-1} \\ &= y_n - \mathbf{x}_n^T \left((\mathbf{I} + \tilde{\mathbf{x}}_n \mathbf{x}_n^T) \hat{\beta}_n - \tilde{\mathbf{x}}_n y_n \right) \\ &= y_n - \left((\mathbf{x}_n^T + \mathbf{x}_n^T \tilde{\mathbf{x}}_n \mathbf{x}_n^T) \hat{\beta}_n - \mathbf{x}_n^T \tilde{\mathbf{x}}_n y_n \right) \\ &= y_n - \left((1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n) \mathbf{x}_n^T \hat{\beta}_n - \mathbf{x}_n^T \tilde{\mathbf{x}}_n y_n \right) \\ &= y_n + \mathbf{x}_n^T \tilde{\mathbf{x}}_n y_n - (1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n) \mathbf{x}_n^T \hat{\beta}_n \\ &= (1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n) y_n - (1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n) \hat{y}_n \\ &= (1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n) (y_n - \hat{y}_n)\end{aligned}$$

(vii)

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \cdot \mathbf{M}_n^{-1} \cdot [\mathbf{x}_1 \cdots \mathbf{x}_n]$$

From this, we can directly see that $(\mathbf{H})_{i,j} = \mathbf{x}_i^T \mathbf{M}_n^{-1} \mathbf{x}_j$.

$$\begin{aligned}
(\mathbf{H})_{n,n} &= \mathbf{x}_n^T \mathbf{M}_n^{-1} \mathbf{x}_n \\
&= \mathbf{x}_n^T \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \mathbf{x}_n \\
&= \mathbf{x}_n^T \left(\mathbf{M}_{n-1}^{-1} \mathbf{x}_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1} \mathbf{x}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \\
&= \mathbf{x}_n^T \left(\tilde{\mathbf{x}}_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \\
&= \mathbf{x}_n^T \tilde{\mathbf{x}}_n \left(1 - \frac{\mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \\
&= \mathbf{x}_n^T \tilde{\mathbf{x}}_n \left(\frac{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n - \mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \\
&= \frac{\mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n}
\end{aligned}$$

First, we use the result from (vii):

$$\begin{aligned}
h_n &= \frac{\mathbf{x}_n^T \tilde{\mathbf{x}}_n}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \\
h_n + h_n \mathbf{x}_n^T \tilde{\mathbf{x}}_n &= \mathbf{x}_n^T \tilde{\mathbf{x}}_n \\
h_n &= \mathbf{x}_n^T \tilde{\mathbf{x}}_n - h_n \mathbf{x}_n^T \tilde{\mathbf{x}}_n \\
\frac{h_n}{1 - h_n} &= \mathbf{x}_n^T \tilde{\mathbf{x}}_n.
\end{aligned}$$

We plug this result into the equation we just obtained:

$$\begin{aligned}
y_n - \widehat{y}_n^n &= (1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n) (y_n - \widehat{y}_n) \\
&= \left(1 + \frac{h_n}{1 - h_n} \right) (y_n - \widehat{y}_n) \\
&= \frac{y_n - \widehat{y}_n}{1 - h_n}.
\end{aligned}$$

This verifies equation (5.2) in the textbook for $i = n$.

(viii)

Changing the order of data points in the dataset doesn't effect the model. This means that we can set any data point to be $\mathbf{x}_n$. Therefore, equation (5.2) is valid for all $i = 1, \dots, n$.

(c)

Consider a situation where we fitted a model based on n data points. (i.e. We estimated $\widehat{\beta}_n$.) If we get m extra data points (after we already estimated $\widehat{\beta}_n$), we don't have to fit the whole model again, but we can just 'update' our model by using the formulas we obtained. (i.e. We can update $\widehat{\beta}_n$ to $\widehat{\beta}_{n+m}$.)

Exercise 8

(a)

First, realize that natural spline is a cubic spline with a constraint, namely: $g(x)$ is a linear function in the intervals $x \in (-\infty, c_1)$ and $x \in [c_K, \infty)$.

The constraint we want to impose in this exercise (on a cubic spline) is clearly a nested case of this constraint of natural spline. So, this constraint can also be expressed as a natural spline with an extra constraint, namely: The linear function in $x \in (-\infty, c_1)$ and $x \in [c_K, \infty)$ has a slope of 0.

(b)

The constraint requires that $g(x)$ is a constant for $x \in (-\infty, c_1)$ and $x \in [c_K, \infty)$.

Let's look at the first interval. Since $x \in (-\infty, c_1)$, all $(x - c_k)_+^3 = 0$. Which means that all $n_k(x) = 0$. So, $g(x)$ becomes: $g(x) = \theta_0 + \theta_1 x$. For $g(x)$ to be a constant, θ_1 should be 0.

(c)

Cubic spline is a 'stitched' cubic polynomials and the stitching points are called 'knots'. So, each interval created by knots has a single cubic polynomial curve. Therefore, $g(x)$ in the last interval is also a cubic polynomial.

Now, let's write this cubic polynomial in the interval $x \in [c_K, \infty)$ as $g(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$. The constraint from the exercise requires that this polynomial is a constant. So, $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $g(x) = \alpha_0$. Thus, $g'(x) = 0$.

(d)

From (b), we have $\theta_1 = 0$. So $g(x)$ in the interval $x \in [c_K, \infty)$ is $g(x) = \theta_0 + \sum_{k=1}^{K-2} \theta_{k+1} n_k(x)$.

From (c), we know that $g'(x) = 0$ in this interval. So,

$$\begin{aligned}
g'(x) &= \sum_{k=1}^{K-2} \theta_{k+1} n'_k(x) \\
&= \sum_{k=1}^{K-2} \theta_{k+1} (d'_k(x) - d'_{K-1}(x)) \\
&= \sum_{k=1}^{K-2} \theta_{k+1} \left(\left(\frac{(x - c_k)^3 - (x - c_K)^3}{c_K - c_k} \right)' - \left(\frac{(x - c_{K-1})^3 - (x - c_K)^3}{c_K - c_{K-1}} \right)' \right) \\
&= \sum_{k=1}^{K-2} \theta_{k+1} \left(\frac{3(x - c_k)^2 - 3(x - c_K)^2}{c_K - c_k} - \frac{3(x - c_{K-1})^2 - 3(x - c_K)^2}{c_K - c_{K-1}} \right) \\
&= \sum_{k=1}^{K-2} \theta_{k+1} \left(\frac{3(c_K - c_k)(2x - c_k - c_K)}{c_K - c_k} - \frac{3(c_K - c_{K-1})(2x - c_{K-1} - c_K)}{c_K - c_{K-1}} \right) \\
&= \sum_{k=1}^{K-2} \theta_{k+1} (3(2x - c_k - c_K) - 3(2x - c_{K-1} - c_K)) \\
&= 3 \sum_{k=1}^{K-2} \theta_{k+1} (c_{K-1} - c_k) \\
&= 0
\end{aligned}$$

(e)

Now, we reparametrize $g(x) = \theta_0 + \sum_{k=1}^{K-2} \theta_{k+1} n_k(x)$.

Let,

$$\eta_k = \begin{cases} \theta_0 & \text{if } k = 0 \\ \theta_{k+1} & \text{if } k \in \{1, \dots, K-2\} \end{cases}$$

and let the new basis function be:

$$m_k(x) = \begin{cases} 1 & \text{if } k = 0 \\ n_k(x) & \text{if } k \in \{1, \dots, K-2\} \end{cases}.$$

This gives

$$g(x) = \sum_{k=0}^{K-2} \eta_k m_k.$$

Exercise 9

(a)

$$\begin{aligned}
P(Y = k) &= \frac{P(Y = k)}{\sum_{l=0}^{K-1} P(Y = l)} \\
&= \frac{\exp \left[\theta_{k,0} + \sum_{j=1}^p \theta_{k,j} x_j \right]}{\sum_{l=0}^{K-1} \exp \left[\theta_{l,0} + \sum_{j=1}^p \theta_{l,j} x_j \right]} \\
&= \frac{\frac{\exp \left[\theta_{k,0} + \sum_{j=1}^p \theta_{k,j} x_j \right]}{\exp \left[\theta_{0,0} + \sum_{j=1}^p \theta_{0,j} x_j \right]}}{\frac{\sum_{l=0}^{K-1} \exp \left[\theta_{l,0} + \sum_{j=1}^p \theta_{l,j} x_j \right]}{\exp \left[\theta_{0,0} + \sum_{j=1}^p \theta_{0,j} x_j \right]}} \\
&= \frac{\exp \left[\theta_{k,0} - \theta_{0,0} + \sum_{j=1}^p (\theta_{k,j} - \theta_{0,j}) x_j \right]}{\sum_{l=0}^{K-1} \exp \left[\theta_{l,0} - \theta_{0,0} + \sum_{j=1}^p (\theta_{l,j} - \theta_{0,j}) x_j \right]} \\
&= \frac{\exp \left[\theta_{k,0} - \theta_{0,0} + \sum_{j=1}^p (\theta_{k,j} - \theta_{0,j}) x_j \right]}{1 + \sum_{l=1}^{K-1} \exp \left[\theta_{l,0} - \theta_{0,0} + \sum_{j=1}^p (\theta_{l,j} - \theta_{0,j}) x_j \right]}
\end{aligned}$$

by defining $\beta_{l,j} = \theta_{l,j} - \theta_{0,j}$, we get

$$= \frac{\exp \left[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j \right]}{1 + \sum_{l=1}^{K-1} \exp \left[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j \right]}$$

By definition, $\sum_{l=0}^{K-1} P(Y = l) = 1$. So,

$$P(Y = 0) = \sum_{l=0}^{K-1} P(Y = l) - \sum_{l=1}^{K-1} P(Y = l) = 1 - \sum_{l=1}^{K-1} P(Y = l).$$

β -model imposes a restriction that $Y = 0$ is set as the reference case. So, β -model has smaller number of parameters than the θ -model.

(b)

$$\begin{aligned}
P(Z_i^k = 1) &= P(Y_i = k | Y_i \in \{0, k\}) \\
&= \frac{P(Y_i = k)}{P(Y_i = 0) + P(Y_i = k)} \\
&= \frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]} \\
&= \frac{1 - \sum_{l=1}^{K-1} P(Y = l) + \frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}}{1 - \sum_{l'=1}^{K-1} \left(\frac{\exp[\beta_{l',0} + \sum_{j=1}^p \beta_{l',j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]} \right) + \frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}} \\
&= \frac{\frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}}{\frac{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j] - \sum_{l'=1}^{K-1} \exp[\beta_{l',0} + \sum_{j=1}^p \beta_{l',j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]} + \frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}} \\
&= \frac{\frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}}{\frac{1}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]} + \frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}} \\
&= \frac{\frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}}{\frac{1 + \exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \sum_{l=1}^{K-1} \exp[\beta_{l,0} + \sum_{j=1}^p \beta_{l,j} x_j]}} \\
&= \frac{\exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}{1 + \exp[\beta_{k,0} + \sum_{j=1}^p \beta_{k,j} x_j]}
\end{aligned}$$

This is equal to logistic regression. So, we can use the theories from logistic regression to estimate β .

Exercise 14

(a)

Bayes classifier is a classifier that minimizes the probability of misclassification (i.e. error rate).

By using Bayes theorem, we have:

$$\Pr(Y|X) = \frac{\Pr(Y, X)}{\Pr(X)} = \frac{\Pr(X|Y) \Pr(Y)}{\Pr(X)}$$

We are given that

$$\Pr(X|Y = k) = \text{Poisson}(\lambda_k) = \frac{(5 + 5k)^x e^{-(5+5k)}}{x!} \quad \text{and} \quad \pi_k = \Pr(Y = k) = \frac{1}{K}$$

So,

$$\begin{aligned} \Pr(X) &= \sum_{k=1}^3 \Pr(X|Y = k) \Pr(Y = k) \\ &= \frac{1}{3} \sum_{k=1}^3 \Pr(X|Y = k) \\ &= \frac{1}{3} \left(\frac{10^x e^{-10}}{x!} + \frac{15^x e^{-15}}{x!} + \frac{20^x e^{-20}}{x!} \right) \\ &= \frac{e^{-10}}{3(x!)} (10^x + 15^x e^{-5} + 20^x e^{-10}) \end{aligned}$$

$$\begin{aligned} \Pr(Y = k|X) &= \frac{\Pr(X|Y = k) \Pr(Y = k)}{\Pr(X)} \\ &= \frac{\frac{(5+5k)^x e^{-(5+5k)}}{x!} \frac{1}{3}}{\frac{e^{-10}}{3(x!)} (10^x + 15^x e^{-5} + 20^x e^{-10})} \\ &= \frac{(5 + 5k)^x e^{(5-5k)}}{10^x + 15^x e^{-5} + 20^x e^{-10}} \\ &= \frac{(1 + k)^x e^{(5-5k)}}{2^x + 3^x e^{-5} + 4^x e^{-10}} \end{aligned}$$

Minimizing the probability of misclassification is equal to maximizing the probability of correct classification. Thus,

$$\text{Bayes classifier: } \arg \max_k \Pr(Y = k|X) = \arg \max_k \left\{ \frac{(1 + k)^x e^{(5-5k)}}{2^x + 3^x e^{-5} + 4^x e^{-10}} \middle| x \right\} = \arg \max_k \left\{ (1 + k)^x e^{5(1-k)} \middle| x \right\}.$$

(b)

Bayes classifier is a classifier that minimizes the probability of misclassification (i.e. error rate).

So, error rate of Bayes classifier:

$$\begin{aligned} \Pr(Y \neq \hat{Y}|X) &= 1 - \Pr(Y = \hat{Y}|X) \\ &= 1 - \max \Pr(Y = k|X) \end{aligned}$$

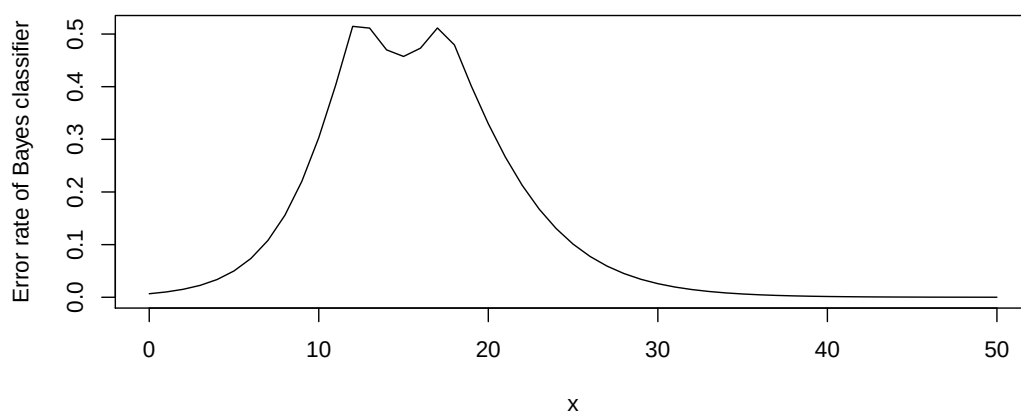
```
# Theoretical error rate of Bayes classifier
theoretical.Bayes.error.rate = function(x,K) {
  prob.mat = data.frame(k = 1:K, prob = NA)

  for (k in 1:K) {
    prob.mat[k, "prob"] = ((1 + k)^x)*exp(5 - 5*k)/(2^x +3^x*exp(-5) + 4^x*exp(-10))
  }
  theo.error.rate = 1 - prob.mat[which.max(prob.mat[, "prob"]), "prob"]

  return(theo.error.rate)
}

theoretical.Bayes.error.rate.vec = Vectorize(theoretical.Bayes.error.rate, vectorize.args = c("x"))

# Plot the theoretical error rate of Bayes classifier.
x.grid = 0:50
y.grid = theoretical.Bayes.error.rate.vec(x.grid, 3)
plot(x = x.grid, y = y.grid, type = "l", xlab = "x", ylab = "Error rate of Bayes classifier")
```



(c)

```
set.seed(1)

# Simulate y
simulated.data = data.frame(y = sample(x = 1:3, size = 1000, replace = T))

# Simulate X
simulated.data[(simulated.data[, "y"] == 1), "x"] = rpois(sum(simulated.data[, "y"] == 1), 10)
simulated.data[(simulated.data[, "y"] == 2), "x"] = rpois(sum(simulated.data[, "y"] == 2), 15)
simulated.data[(simulated.data[, "y"] == 3), "x"] = rpois(sum(simulated.data[, "y"] == 3), 20)

# Bayes classifier
Bayes.classifier = function(x,K) {
  prob.mat = data.frame(k = 1:K, prob = NA)

  for (k in 1:K) {
    prob.mat[k, "prob"] = ((1 + k)^x)*exp(5 - 5*k)/(2^x +3^x*exp(-5) + 4^x*exp(-10))
  }
```

```

    }
    y.hat = prob.mat[which.max(prob.mat[, "prob"]), "k"]

    return(y.hat)
}

# Compute y.hat based on Bayes classifier.
Bayes.classifier.vec = Vectorize(Bayes.classifier, vectorize.args = c("x"))
simulated.data[, "y.hat.Bayes"] = Bayes.classifier.vec(simulated.data[, "x"], 3)
simulated.data[, "is.pred.correct"] =
  as.numeric(simulated.data[, "y.hat.Bayes"] == simulated.data[, "y"])

# Overall error rate
error.rate = 1 - sum(simulated.data[, "y"] == simulated.data[, "y.hat.Bayes"])/nrow(simulated.data)
show(error.rate)

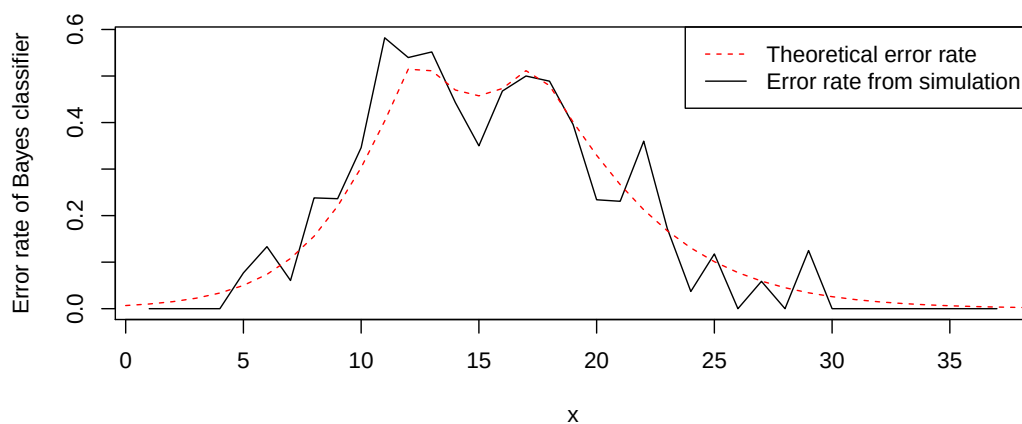
# Error rate per x value
empirical.error.rate.mat = data.frame(
  x = sort(unique(simulated.data[, "x"])),
  n = as.numeric(table(simulated.data[, "x"])),
  n.correct.pred = NA)

for (i in 1:nrow(empirical.error.rate.mat)) {
  x.target = empirical.error.rate.mat[i, "x"]
  empirical.error.rate.mat[i, "n.correct.pred"] =
    sum(simulated.data[(simulated.data[, "x"] == x.target), "is.pred.correct"])
}

empirical.error.rate.mat[, "error.rate"] =
  1 - empirical.error.rate.mat[, "n.correct.pred"]/empirical.error.rate.mat[, "n"]

# Plot the error rate as a function of x.
plot(x = empirical.error.rate.mat[, "x"], y = empirical.error.rate.mat[, "error.rate"],
     type = "l", xlab = "x", ylab = "Error rate of Bayes classifier")
points(x = x.grid, y = y.grid, type = "l", lty = 2, col = "red")
legend("topright", c("Theoretical error rate", "Error rate from simulation"),
      lty = c(2, 1), col = c("red", "black"))

```



Exercise 15

(a)

Use same approach as in exercise 14 (a)

$$\begin{aligned}
 \Pr(X) &= \sum_{k=1}^2 \Pr(X|Y=k) \Pr(Y=k) \\
 &= \frac{1}{2} \sum_{k=1}^2 \Pr(X|Y=k) \\
 &= \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+1)^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \right) \\
 &= \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x+1)^2}{2}} + e^{-\frac{(x-1)^2}{2}} \right)
 \end{aligned}$$

$$\begin{aligned}
 \Pr(Y=k|X) &= \frac{\Pr(X|Y=k) \Pr(Y=k)}{\Pr(X)} \\
 &= \frac{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2}}}{\frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x+1)^2}{2}} + e^{-\frac{(x-1)^2}{2}} \right)} \\
 &= \frac{e^{-\frac{(x-\mu_k)^2}{2}}}{e^{-\frac{(x+1)^2}{2}} + e^{-\frac{(x-1)^2}{2}}}
 \end{aligned}$$

$$\text{Bayes classifier: } \arg \max_k \Pr(Y=k|X) = \arg \max_k \left\{ \frac{e^{-\frac{(x-\mu_k)^2}{2}}}{e^{-\frac{(x+1)^2}{2}} + e^{-\frac{(x-1)^2}{2}}} \middle| x \right\}$$

We can simplify this classifier further. We examine the decision boundary

$$\Pr(Y=1|X) > \Pr(Y=2|X) \iff -(x+1)^2 > -(x-1)^2 \iff x < 0.$$

So, we have

$$\begin{aligned}
 \text{Bayes classifier: } k^{\text{Bayes}} &= \arg \min_k \{1 - \Pr(Y=k|X)\} \\
 &= \arg \max_k \Pr(Y=k|X) \\
 &= \begin{cases} 1 & \text{if } x < 0 \\ 2 & \text{otherwise} \end{cases}
 \end{aligned}$$

(b)

$$\text{We plot } \Pr(Y=1|X) = \frac{e^{-\frac{(x+1)^2}{2}}}{e^{-\frac{(x+1)^2}{2}} + e^{-\frac{(x-1)^2}{2}}}$$

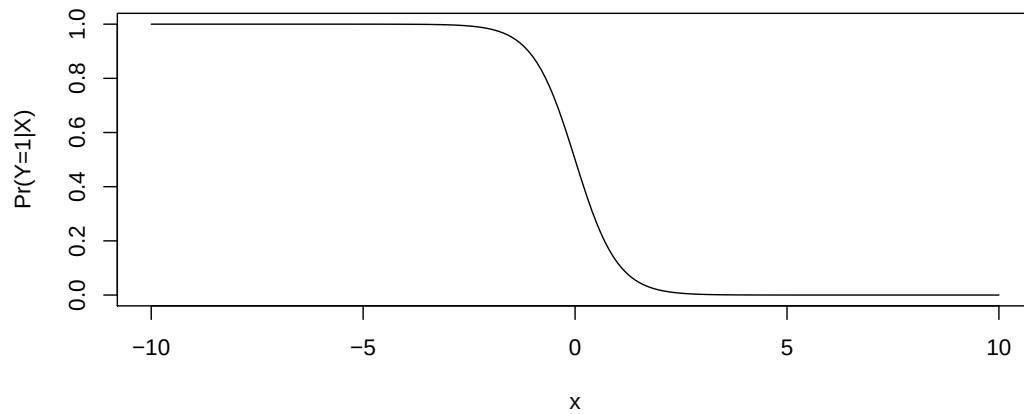
```

prob.y1.cond.x.func = function(x) {
  prob.y1.cond.x = exp(-((x+1)^2)/2) / (exp(-((x+1)^2)/2) + exp(-((x-1)^2)/2))
  return(prob.y1.cond.x)
}

x.grid = seq(from = -10, to = 10, by = 0.1)
y.grid = prob.y1.cond.x.func(x.grid)

plot(x = x.grid, y = y.grid, type = "l", xlab = "x", ylab = "Pr(Y=1|X)")

```



(c)

$$\begin{aligned}
 f_X(x) &= \sum_{k=1}^2 \Pr(X|Y=k) \Pr(Y=k) \\
 &= f(x|y=1)f(y=1) + f(x|y=2)f(y=2) \\
 &= \frac{1}{2}f(x|y=1) + \frac{1}{2}f(x|y=2) \\
 &= \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x+1)^2}{2}} + \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \\
 &= \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x+1)^2}{2}} + e^{-\frac{(x-1)^2}{2}} \right)
 \end{aligned}$$

Null hypothesis testing:

Reject H_0 if $F_X(x) < \frac{\alpha}{2}$ or $F_X(x) > 1 - \frac{\alpha}{2}$ where

$$F_X(x) = \int_{-\infty}^x f_X(u) du = \frac{1}{2}\Phi(x+1) + \frac{1}{2}\Phi(x-1).$$

(d)

When $\alpha = 0$, the confidence interval is $(-\infty, \infty)$ and we will always accept the null hypothesis. In this case, the given classifier is equal to Bayes classifier.

```
Bayes.classifier = function(x) {
  y.hat = as.numeric(x < 0)*1 + as.numeric(x >= 0)*2
  return(y.hat)
}

custom.classifier = function(x, alpha) {
  # Test null hypothesis
  if (
    (1/2*pnorm(x+1) + 1/2*pnorm(x-1) < alpha/2) | (1/2*pnorm(x+1) + 1/2*pnorm(x-1) > 1 - alpha/2)
  ) {
    # Null hypothesis is rejected
    y.hat = c("outlier")
  } else {
    # Null hypothesis is accepted
    y.hat = Bayes.classifier(x)
  }
  return(y.hat)
}

custom.classifier.vec = Vectorize(custom.classifier, vectorize.args = c("x"))
```

(e)

```
> set.seed(1)
>
> # Simulate y
> simulated.data = data.frame(y = sample(x = 1:2, size = 1000, replace = T))
>
> # Simulate X
> simulated.data[(simulated.data[, "y"] == 1), "x"] = rnorm(sum(simulated.data[, "y"] == 1), mean = -1)
> simulated.data[(simulated.data[, "y"] == 2), "x"] = rnorm(sum(simulated.data[, "y"] == 2), mean = 1,
>
> # Perform classification
> # alpha = 0.05
> y.hat.1 = custom.classifier.vec(x = simulated.data[, "x"], alpha = 0.05)
>
> # alpha = 0.01
> y.hat.2 = custom.classifier.vec(x = simulated.data[, "x"], alpha = 0.01)
>
> # alpha = 0
> y.hat.3 = custom.classifier.vec(x = simulated.data[, "x"], alpha = 0)
>
> # Error rate
> error.rate.1 = 1 - sum(simulated.data[, "y"] == y.hat.1)/nrow(simulated.data)
> error.rate.2 = 1 - sum(simulated.data[, "y"] == y.hat.2)/nrow(simulated.data)
> error.rate.3 = 1 - sum(simulated.data[, "y"] == y.hat.3)/nrow(simulated.data)
>
> cat("Error rate with alpha = 0.05: ", error.rate.1, sep = "", "\n")
Error rate with alpha = 0.05: 0.212
> cat("Error rate with alpha = 0.01: ", error.rate.2, sep = "", "\n")
```

```
Error rate with alpha = 0.01: 0.164
> cat("Error rate with alpha = 0: ", error.rate.3, sep = "", "\n")
Error rate with alpha = 0: 0.148
```

Exercise 16

(a)

$$\sum_{g=1}^G \frac{n_g}{n} \bar{y}_g = \frac{1}{n} \sum_{g=1}^G n_g \bar{y}_g = \frac{1}{n} \sum_{g=1}^G n_g \left(\frac{1}{n_g} \sum_{i \in g} y_i \right) = \frac{1}{n} \sum_{g=1}^G \sum_{i \in g} y_i = \frac{1}{n} \sum_{i=1}^n y_i = \hat{\mu}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_g + \bar{y}_g - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left((y_i - \bar{y}_g)^2 + (\bar{y}_g - \bar{y})^2 + 2 (y_i - \bar{y}_g) (\bar{y}_g - \bar{y}) \right) \\ &= \frac{1}{n} \sum_{g=1}^G \frac{n_g}{n_g} \sum_{i \in g} \left((y_i - \bar{y}_g)^2 + (\bar{y}_g - \bar{y})^2 + 2 (y_i - \bar{y}_g) (\bar{y}_g - \bar{y}) \right) \\ &= \frac{1}{n} \sum_{g=1}^G n_g \left(\frac{1}{n_g} \sum_{i \in g} (y_i - \bar{y}_g)^2 + \frac{1}{n_g} \sum_{i \in g} (\bar{y}_g - \bar{y})^2 \right) + \frac{2}{n} \sum_{g=1}^G \sum_{i \in g} (y_i - \bar{y}_g) (\bar{y}_g - \bar{y}) \\ &= \frac{1}{n} \sum_{g=1}^G n_g \left(\hat{\sigma}_g^2 + (\bar{y}_g - \bar{y})^2 \right) + \frac{2}{n} \sum_{g=1}^G \sum_{i \in g} (y_i - \bar{y}_g) (\bar{y}_g - \bar{y}) \\ &= \frac{1}{n} \sum_{g=1}^G n_g \left(\hat{\sigma}_g^2 + (\bar{y}_g - \bar{y})^2 \right) + \frac{2}{n} \sum_{g=1}^G (\bar{y}_g - \bar{y}) \sum_{i \in g} (y_i - \bar{y}_g) \\ &= \frac{1}{n} \sum_{g=1}^G n_g \left(\hat{\sigma}_g^2 + (\bar{y}_g - \bar{y})^2 \right) + \frac{2}{n} \sum_{g=1}^G (\bar{y}_g - \bar{y}) \cdot 0 \\ &= \frac{1}{n} \sum_{g=1}^G n_g \left(\hat{\sigma}_g^2 + (\bar{y}_g - \bar{y})^2 \right) \end{aligned}$$

(b)

i)

Use the results from (a) with $g_1 = \{1, \dots, n-1\}$ and $g_2 = \{n\}$.

ii)

$$\begin{aligned} \hat{\sigma}_n^2 &= \frac{n_{g_1}}{n} [\hat{\sigma}_{g_1}^2 + (\bar{y}_{g_1} - \bar{y})^2] + \frac{n_{g_2}}{n} [\hat{\sigma}_{g_2}^2 + (\bar{y}_{g_2} - \bar{y})^2] \\ &= \frac{n-1}{n} [\hat{\sigma}_{n-1}^2 + (\bar{y}_{n-1} - \bar{y})^2] + \frac{1}{n} [(y_n - y_n)^2 + (y_n - \bar{y})^2] \\ &= \frac{n-1}{n} \hat{\sigma}_{n-1}^2 + \frac{n-1}{n} (\bar{y}_{n-1} - \bar{y})^2 + \frac{1}{n} (y_n - \bar{y})^2 \\ &= \frac{n-1}{n} \hat{\sigma}_{n-1}^2 + \frac{n-1}{n^2} \left(n(\bar{y}_{n-1} - \bar{y})^2 + \frac{n}{n-1} (y_n - \bar{y})^2 \right) \end{aligned}$$

Use $\bar{y} = \frac{n-1}{n}\bar{y}_{n-1} + \frac{1}{n}y_n$

$$\begin{aligned}
&= \frac{n-1}{n}\hat{\sigma}_{n-1}^2 + \frac{n-1}{n^2} \left(n \left(\bar{y}_{n-1} - \frac{n-1}{n}\bar{y}_{n-1} - \frac{1}{n}y_n \right)^2 + \frac{n}{n-1} \left(y_n - \frac{n-1}{n}\bar{y}_{n-1} - \frac{1}{n}y_n \right)^2 \right) \\
&= \frac{n-1}{n}\hat{\sigma}_{n-1}^2 + \frac{n-1}{n^2} \left(n \left(\frac{1}{n}\bar{y}_{n-1} - \frac{1}{n}y_n \right)^2 + \frac{n}{n-1} \left(\frac{n-1}{n}y_n - \frac{n-1}{n}\bar{y}_{n-1} \right)^2 \right) \\
&= \frac{n-1}{n}\hat{\sigma}_{n-1}^2 + \frac{n-1}{n^2} \left(\frac{1}{n}(\bar{y}_{n-1} - y_n)^2 + \frac{n-1}{n}(y_n - \bar{y}_{n-1})^2 \right) \\
&= \frac{n-1}{n}\hat{\sigma}_{n-1}^2 + \frac{n-1}{n^2}(\bar{y}_{n-1} - y_n)^2 \\
&= \frac{n-1}{n}\hat{\sigma}_{n-1}^2 + \frac{n-1}{n^2}(y_n - \bar{y}_{n-1})^2
\end{aligned}$$

(c)

We already showed this in (iv) of exercise 7 (b). The recap of the result:

$$\begin{aligned}
\hat{\beta}_n &= (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{y}_n \\
&= \mathbf{M}_n^{-1} \mathbf{X}_n^T \mathbf{y}_n \\
&= \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \sum_{i=1}^n \mathbf{x}_i y_i \\
&= \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \left(\sum_{i=1}^{n-1} \mathbf{x}_i y_i + \mathbf{x}_n y_n \right) \\
&= \left(\mathbf{M}_{n-1}^{-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\mathbf{X}_{n-1}^T \mathbf{y}_{n-1} + \mathbf{x}_n y_n) \\
&= \mathbf{M}_{n-1}^{-1} \mathbf{X}_{n-1}^T \mathbf{y}_{n-1} + \mathbf{M}_{n-1}^{-1} \mathbf{x}_n y_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \mathbf{X}_{n-1}^T \mathbf{y}_{n-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \mathbf{x}_n y_n \\
&= \hat{\beta}_{n-1} + \mathbf{M}_{n-1}^{-1} \mathbf{x}_n y_n - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \hat{\beta}_{n-1} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T \mathbf{M}_{n-1}^{-1}}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \mathbf{x}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \mathbf{M}_{n-1}^{-1} \mathbf{x}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \hat{\beta}_{n-1} + \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) \tilde{\mathbf{x}}_n y_n \\
&= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n)
\end{aligned}$$

This equation allows us to update the coefficients (instead of calculating from scratch) when we add or remove a data point.

Instead of doing matrix multiplication with a design matrix of size $n \times p$, we do matrix multiplication between $\left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right)$ and $(\hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n)$ which are of size $p \times p$ and $p \times 1$ respectively. So, when $p < n$, we use less memory.

We can start this algorithm with the ordinary least squares based on the at least $p+1$ data points. This is to avoid the nonsingularity.

(d)

The ‘empty’ linear model (e.g. no predictor) has design matrix: $\mathbf{X} = [1 \cdots 1]^T$ and estimator $\hat{y}_i = \hat{\beta} = \hat{\beta}_0 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n$. We plug this into the result from (c)

$$\mathbf{M}_n = \mathbf{X}_n^T \mathbf{X}_n = n, \quad \tilde{\mathbf{x}}_n = \mathbf{M}_{n-1}^{-1} \mathbf{x}_n = \frac{1}{n-1}$$

$$\begin{aligned} \bar{y}_n = \hat{\beta}_n &= \left(\mathbf{I} - \frac{\tilde{\mathbf{x}}_n \mathbf{x}_n^T}{1 + \mathbf{x}_n^T \tilde{\mathbf{x}}_n} \right) (\hat{\beta}_{n-1} + \tilde{\mathbf{x}}_n y_n) \\ &= \left(1 - \frac{\frac{1}{n-1} \cdot 1}{1 + 1 \cdot \frac{1}{n-1}} \right) \left(\bar{y}_{n-1} + \frac{1}{n-1} y_n \right) \\ &= \frac{n-1}{n} \left(\bar{y}_{n-1} + \frac{1}{n-1} y_n \right) \\ &= \frac{n-1}{n} \bar{y}_{n-1} + \frac{1}{n} y_n \end{aligned}$$

So, (*) is a special case of (**) when $\mathbf{X} = [1 \cdots 1]^T$.

Exercise 17

(a)

β_j follows the normal distribution. So, we expect that $1.000.000 \cdot 0.01 = 10.000$ null hypothesis will be rejected.

(b)

Probability of making at least one error when we use significance level of $\frac{\alpha}{q}$ is $\Pr\left(\bigcup_j \text{reject } H_{0,j}\right)$. We apply Boole's inequality to this

$$\Pr\left(\bigcup_j \text{reject } H_{0,j}\right) \leq \sum_{j=1}^q \Pr(\text{reject } H_{0,j}) = \sum_{j=1}^q \frac{\alpha}{q} = \alpha.$$

Thus, if all $H_{0,j}$'s are true, the probability of making at least one errors is less or equal to α .

$$\text{power} = \Pr(\text{reject } H_{0,j} | H_{1,j} \text{ is true}) = \Pr\left(p_j < \frac{\alpha}{q}\right)$$

Obviously, $\Pr\left(p_j < \frac{\alpha}{1000000}\right) \ll \Pr(p_j < \alpha)$.

(c)

V, S, U, T, R are stochastic.

Without Bonferroni correction and given that all $H_{0,j}$'s are true, the probability of wrongly reject $H_{0,j}$ is α .

$$\text{Type I error rate: } \Pr(V > 0) = 1 - \Pr(V = 0) = 1 - (1 - \alpha)^q$$

If we apply Bonferroni correction,

$$\text{Type I error rate: } \Pr(V > 0) = 1 - \left(1 - \frac{\alpha}{q}\right)^q$$

Since $1 - \frac{\alpha}{q} > 1 - \alpha$, Bonferroni correction decreases the type I error rate. (But, it reduces the power.)

(d)

$q_0 = q$ implies $S = T = 0$. Thus, $E\left[\frac{V}{R}\right] = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{V}\right] = E[1] = 1$

So, we never get $E\left[\frac{V}{R}\right] = 1 \leq \alpha$.

(e)

When $R > 0$, $q_0 = q$ still implies $S = T = 0$. So, we have the same problem as in (d).

(f)

