Solution 1: Connection between MLE and ERM

(a) We can make use of the "substitution trick" from Exercise Sheet 1, i.e., $z^{(i)} = m(y^{(i)})$. Then, it holds that $z^{(i)} \mid \mathbf{x}$ is distributed as $\mathcal{N}(m(f_{\text{true}}(\mathbf{x}^{(i)})), \sigma^2)$, since $z^{(i)} = m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}$ and $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Note that $(\mathbf{x}^{(1)}, z^{(1)}), \ldots, (\mathbf{x}^{(n)}, z^{(n)})$ are iid, as transforming $y^{(1)}, \ldots, y^{(n)}$ via m to $z^{(1)}, \ldots, z^{(n)}$ preserves the stochastic independence property.

(b) The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \ldots, (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(z^{(i)} \;\middle|\; f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[z^{(i)} - m\left(f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)\right]^2\right). \end{aligned}$$

So, the negative log-likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \ldots, (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned} -\ell(\boldsymbol{\theta}) &= -\log\left(\mathcal{L}(\boldsymbol{\theta})\right) \\ &= -\log\left(\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[z^{(i)} - m\left(f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)\right]^2\right)\right) \\ &\propto \sum_{i=1}^{n} \left[z^{(i)} - m\left(f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)\right]^2 \\ &= \sum_{i=1}^{n} \left[m(y^{(i)}) - m\left(f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)\right]^2. \end{aligned}$$

Thus, the negative log-likelihood for a parameter $\boldsymbol{\theta}$ is proportional to the empirical risk of a hypothesis $f(\cdot \mid \boldsymbol{\theta})$ w.r.t. the generalized L2-loss function of Exercise sheet 1, i.e., $L\left(y, f(\mathbf{x})\right) = \left(m(y) - m(f(\mathbf{x}))\right)^2$.

(c) First, we specify the feature space: $\mathcal{X} = \{1\} \times \mathbb{R}$, i.e., any feature $\mathbf{x} \in \mathcal{X}$ is of the form $\mathbf{x} = (x_1, x_2)^\top = (1, x_2)^\top$ for some $x_2 \in \mathbb{R}$. According to the exercise we use $m(x) = \log(x)$, whose inverse is $m^{-1}(x) = \exp(x)$. Let us rewrite Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ into $y = m^{-1}\left(m(f(\mathbf{x} \mid \boldsymbol{\theta})) + \epsilon\right)$, for some suitable hypothesis $f(\mathbf{x} \mid \boldsymbol{\theta})$:

$$\begin{aligned} y &= \theta_1 \exp(\theta_2 x + \epsilon) \\ &= \exp(\log(\theta_1) \exp(\theta_2 x + \epsilon)) \\ &= \exp(\log(\theta_1) + \theta_2 x + \epsilon) && \text{(Functional equation of exp)} \\ &= \exp(\log(\theta_1) + \log(\exp(\theta_2 x)) + \epsilon) \\ &= \underbrace{\exp}_{=m^{-1}}\left(\underbrace{\log}_{=m}(\theta_1 \exp(\theta_2 x)) + \epsilon\right) && \text{(Functional equation of log)} \\ &= m^{-1}(m(\theta_1 \exp(\theta_2 x)) + \epsilon). \end{aligned}$$

With this, we see that $f(\mathbf{x} \mid \boldsymbol{\theta}) = \theta_1 x_1 \exp(\theta_2 x_2) = \theta_1 \exp(\theta_2 x_2)$ is a suitable functional form for the hypotheses. Thus, we use as our parameter space $\Theta = \mathbb{R}_+ \times \mathbb{R}$ which gives rise to the hypothesis space

$$\mathcal{H} = \{f(\mathbf{x} \mid \boldsymbol{\theta}) = \theta_1 x_1 \exp(\theta_2 x_2) \mid \boldsymbol{\theta} \in \Theta\}.$$

Alternative: Note that we could alternatively rephrase the learning problem by applying the logarithm on both sides of Forbes' model:

$$y = \theta_1 \exp(\theta_2 x + \epsilon) \quad \Leftrightarrow \quad \log(y) = \log(\theta_1) + \theta_2 x + \epsilon,$$

so that we work with the logarithm of the original labels, i.e., we consider $z^{(1)} = \log(y^{(1)}), \ldots, z^{(n)} = \log(y^{(n)})$ instead of $y^{(1)}, \ldots, y^{(n)}$. A suitable hypothesis space is then

$$\mathcal{H} = \{f(\mathbf{x} \mid \boldsymbol{\theta}) = \log(\theta_1)x_1 + \theta_2 x_2 \mid \boldsymbol{\theta} \in \Theta\},$$

which are the linear functions[1] $\mathbf{x}^\top \boldsymbol{\theta}$ of features in $\mathcal{X}$. The empirical risk minimizer in this case is specified by the parameter

$$(\log(\hat{\theta}_1), \hat{\theta}_2)^\top = \hat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{z}, \qquad \boldsymbol{z} = (\log y^{(1)}, \ldots, \log y^{(n)})^\top,$$

(see Chapter 02.02 of I2ML) which for this simple case is:

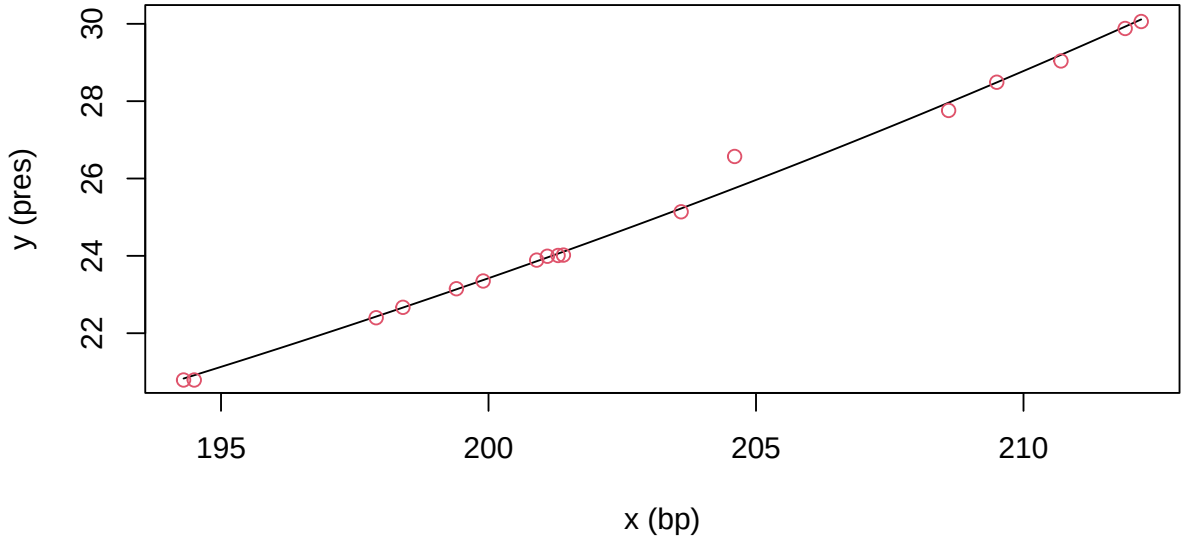
$$\hat{\theta}_2 = \frac{\sum_{i=1}^n (\mathbf{x}_2^{(i)} - \bar{\mathbf{x}}_2)(\log(y^{(i)}) - \overline{\log(y)})}{\sum_{i=1}^n (\mathbf{x}_2^{(i)} - \bar{\mathbf{x}}_2)^2},$$
$$\hat{\theta}_1 = \exp\left(\overline{\log(y)} - \hat{\theta}_2 \bar{\mathbf{x}}_2\right),$$

where $\bar{\mathbf{x}}_2 = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_2^{(i)}$ and $\overline{\log(y)} = \frac{1}{n}\sum_{i=1}^n \log(y^{(i)})$.

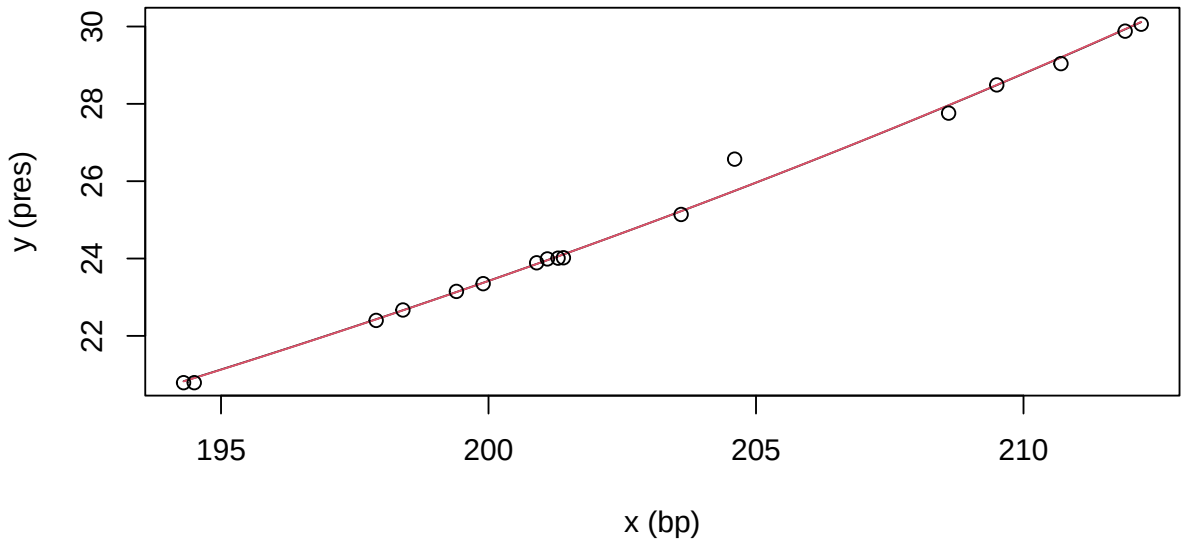
```r
#' @param X the feature input matrix X
#' @param y the outcome vector y
#' @param theta coefficient vector for the model (2-dimensional)

# Load MASS and data set forbes
library(MASS)
data(forbes)
attach(forbes)

# initialize the data set
X = cbind(rep(1,17),bp)
y = pres

#' function to represent your models via the parameter vector theta = c(theta_1, theta_2)
#' @return a predicted label y_hat for x
f <- function(x, theta){

  return((exp(theta[2]*x[2])*theta[1]*x[1]))

}

#' @return a vector consisting of the optimal parameter vector
optim_coeff <- function(X,y){

  #' @return the empirical risk of a parameter vector theta
  emp_risk <- function(theta){
    sum( (log(y) - log(apply(X,1,f,theta)))^2   )
  }


  return(
    optim(c(0.4,0.5),
          emp_risk,
          method = "L-BFGS-B",
          lower=c(0,-Inf),
          upper=c(Inf,Inf))$par)
  # note that c(0.4,0.5) can be replaced by any other theta vector
  # satisfying the constraint theta[1]>0
}
```

[1] Note that $\log(\theta_1)$ can be any value in $\mathbb{R}$.

```r
# optimal coefficients
hat_theta = optim_coeff(X,y)
print(hat_theta)

## [1] 0.38050968 0.02059961

# Checking Forbes' model visually

f_x <- function(x, theta){

  return((exp(theta[2]*x)*theta[1]))

}

curve(f_x(x,theta = hat_theta),min(bp),max(bp),xlab="x (bp)",ylab="y (pres)")
points(pres~bp,col=2)
```




```r
# Alternative solution

hat_theta_2 = cov(bp,log(pres))/(var(bp))
hat_theta_1 = exp(mean(log(pres))-hat_theta_2*mean(bp))

curve(f_x(x,theta = hat_theta),min(bp),max(bp),xlab="x (bp)",ylab="y (pres)")
curve(f_x(x,theta = c(hat_theta_1,hat_theta_2)),min(bp),max(bp),add=T,col=2)
points(pres~bp)
```


y (pres)
22
24
26
28
30
195
200
205
210
x (bp)