Machine Learning 1
Exercise 2

Group: BSSBCH

October 30, 2017

Matthias Bigalke, 339547, maku@win.tu-berlin.de
Tolga Buz, 346836, buz_tolga@yahoo.de
Alejandro Hernandez, 395678, alejandrohernandezmunuera@gmail.com
Aitor Palacios Cuesta, 396276, aitor.palacioscuesta@campus.tu-berlin.de
Christof Schubert, 344450, christof.schubert@campus.tu-berlin.de
Daniel Steinhaus, 342563, dany.steinhaus@googlemail.com

1 Maximum-Likelihood Estimation

We consider the problem of estimating using the maximum-likelihood approach the parameters $\lambda, \eta > 0$ of the probability distribution: $p(x,y) = \lambda\eta e^{-\lambda x - \eta y}$ supported on $\mathbb{R}_+^2$. We consider a dataset $D = ((x_1, y_1), ..., (x_N, y_N))$ composed of N independent draws from this distribution.

Task a)

Show that x and y are independent.

A joint distribution of x and y can be expressed as:

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \tag{1}$$

For x and y being independent, $p(x|y) = p(x)$ and $p(y|x) = p(y)$. This leads to:

$$p(x,y) = p(x) \cdot p(y) \tag{2}$$

... and considering the given probability distribution:

$$p(x) \cdot p(y) = \lambda\eta e^{-\lambda x - \eta y} \tag{3}$$

Now the single probabilities $p(x)$ and $p(y)$ have to be computed:

$$\begin{aligned} p(x) &= \int_y p(x,y)dy = \int_y \lambda\eta e^{-\lambda x-\eta y}dy = \lim_{C\to\infty}[-\lambda e^{-\lambda x-\eta y}]_0^C \\ &= \lim_{C\to\infty}(-\lambda e^{-\lambda x-\eta C}) + \lambda e^{-\lambda x} = 0 + \lambda e^{-\lambda x} \\ &= \lambda e^{-\lambda x} \end{aligned} \tag{4}$$

$$\begin{aligned} p(y) &= \int_x p(x,y)dx = \int_x \lambda\eta e^{-\lambda x-\eta y}dx = \lim_{C\to\infty}[-\eta e^{-\lambda x-\eta y}]_0^C \\ &= \lim_{C\to\infty}(-\eta e^{-\lambda x-\eta y}) + \eta e^{-\eta y} = 0 + \eta e^{-\eta y} \\ &= \eta e^{-\eta y} \end{aligned} \tag{5}$$

Multiplying the single probabilities (equations 4 and 5) leads to:

$$p(x)\cdot p(y) = \lambda e^{-\lambda x}\cdot \eta e^{-\eta y} = \lambda\eta e^{-\lambda x-\eta y} = p(x,y) \tag{6}$$

This shows that x and y are independent.

Task b)

Derive a maximum likelihood estimator of the parameter λ based on D.

Considering the given dataset D, the following function can be interpreted as likelihood:

$$p(D|\omega,\lambda) = p(D|\lambda) = \prod_{i=1}^{N} p(x_i|\lambda) \tag{7}$$

The maximum likelihood principle in this case selects the parameter λ which is maximally likely given the dataset D. But the log-likelihood will be maximized since it finds the same maximum:

$$l(\lambda) = ln(p(D|\lambda)) = ln(\prod_{i=1}^{N} p(x_i|\lambda)) = \sum_{i=1}^{N} ln(p((x_i,y_i)|\lambda)) \tag{8}$$

That leads to a maximum likelihood $\hat{\lambda}$ of:

$$\begin{aligned} \hat{\lambda} &= arg\,\underset{\lambda}{max}(l(\lambda)) = arg\,\underset{\lambda}{min}(-l(\lambda)) \\ &= arg\,\underset{\lambda}{min}(-l(\lambda)) = arg\,\underset{\lambda}{min}(-\sum_{i=1}^{N} ln(p((x_i,y_i)|\lambda))) \end{aligned} \tag{9}$$

If $p(D|\lambda)$ is a well-behaved, differentiable function of λ, $\hat{\lambda}$ can be found with using ∇_λ ($\nabla_l(\lambda) = (\frac{\partial l(\lambda)}{\partial \lambda 1}, ..., \frac{\partial l(\lambda)}{\partial \lambda_N})^T$) as gradient operator:

$$\nabla_\lambda l(\lambda) = -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} ln(p((x_i, y_i)|\lambda)) \overset{!}{=} 0 \tag{10}$$

This leads to:

$$\begin{aligned}
\nabla_\lambda l(\lambda) &= -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} ln(p((x_i, y_i)|\lambda)) \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} ln(\lambda \eta e^{-\lambda x_i - \eta y_i}) \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} (ln(\lambda) + ln(\eta) + (-\lambda x_i - \eta y_i)) \\
&= -\sum_{i=1}^{N} (\frac{1}{\lambda} - x_i) \\
&= -\sum_{i=1}^{N} (\frac{1}{\lambda}) + \sum_{i=1}^{N} (x_i) \\
&= -(\frac{n}{\lambda}) + \sum_{i=1}^{N} (x_i)
\end{aligned} \tag{11}$$

Considering the the minimum condition from equation 10, this leads to:

$$\begin{aligned}
&-(\frac{n}{\lambda}) + \sum_{i=1}^{N} (x_i) = 0 \\
\Leftrightarrow &\lambda = \frac{n}{\sum_{i=1}^{N} (x_i)}
\end{aligned} \tag{12}$$

Task c)

Derive a maximum likelihood estimator of the parameter λ based on D under the constraint $\eta = \frac{1}{\lambda}$.

The solution can be found analogously to Task b) except the adjusted function $p(x, y) = \lambda \frac{1}{\lambda} e^{-\lambda x - \frac{1}{\lambda} y}$. Starting point for the following computation is equation 10:

$$
\begin{aligned}
\nabla_\lambda l(\lambda) &= -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} ln(p((x_i, y_i)|\lambda)) \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} ln(\lambda \frac{1}{\lambda} e^{-\lambda x_i - \frac{1}{\lambda} y_i}) \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial \lambda} (ln(\lambda) + ln(\frac{1}{\lambda}) - \lambda x_i - \frac{1}{\lambda} y_i) \\
&= -\sum_{i=1}^{N} \frac{1}{\lambda} - \frac{1}{\lambda} - x_i + \frac{1}{\lambda^2} y_i \\
&= -\sum_{i=1}^{N} -x_i + \frac{1}{\lambda^2} y_i \\
&= \sum_{i=1}^{N} x_i - \frac{1}{\lambda^2} \sum_{i=1}^{N} y_i
\end{aligned}
\tag{13}
$$

Now considering the condition $\nabla_\lambda l(\lambda) \overset{!}{=} 0$:

$$
\begin{aligned}
&\sum_{i=1}^{N} x_i - \frac{1}{\lambda^2} \sum_{i=1}^{N} y_i = 0 \\
\Leftrightarrow &\sum_{i=1}^{N} x_i = \frac{1}{\lambda^2} \sum_{i=1}^{N} y_i \\
\Leftrightarrow &\lambda^2 = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} \\
\Leftrightarrow &\lambda = \sqrt{\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}}
\end{aligned}
\tag{14}
$$

Task d)

Derive a maximum likelihood estimator of the parameter λ based on D under the constraint $\eta = 1 - \lambda$.

The solution can be found analogously to Task b) except the adjusted function $p(x, y) = \lambda(1 - \lambda)e^{-\lambda x - (1-\lambda)y}$. Starting point for the following computation is equation 10:

$$
\begin{aligned}
\nabla_\lambda l(\lambda) &= -\sum_{i=1}^{N} \frac{\partial}{\partial\lambda} ln(p((x_i, y_i)|\lambda)) \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial\lambda} ln(\lambda(1-\lambda)e^{-\lambda x_i-(1-\lambda)y_i}) \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial\lambda} (ln(\lambda) + ln(1-\lambda) - \lambda x_i - (1-\lambda)y_i) \\
&= -\sum_{i=1}^{N} (\frac{1}{\lambda} - \frac{1}{1-\lambda} - x_i + y_i)
\end{aligned}
\tag{15}
$$

Now considering the condition $\nabla_\lambda l(\lambda) \overset{!}{=} 0$:

$$
\begin{aligned}
&-\sum_{i=1}^{N} (\frac{1}{\lambda} - \frac{1}{1-\lambda} - x_i + y_i) = 0 \\
\Leftrightarrow &-\sum_{i=1}^{N} (y_i - x_i) - \frac{n}{\lambda} + \frac{n}{1-\lambda} = 0 \\
\Leftrightarrow &\frac{n}{1-\lambda} - \frac{n}{\lambda} = \sum_{i=1}^{N} (y_i - x_i) \\
\Leftrightarrow &n\lambda - n(1-\lambda) = \sum_{i=1}^{N} (y_i - x_i)\lambda(1-\lambda) \\
\Leftrightarrow &n\lambda - n + n\lambda = \sum_{i=1}^{N} (y_i - x_i)(\lambda - \lambda^2) \\
\Leftrightarrow &(\sum_{i=1}^{N} (y_i - x_i))\lambda^2 + (2n - \sum_{i=1}^{N} (y_i - x_i))\lambda - n = 0 \\
\Leftrightarrow &\lambda^2 + \frac{2n - \sum_{i=1}^{N} (y_i - x_i)}{\sum_{i=1}^{N} (y_i - x_i)}\lambda - \frac{n}{\sum_{i=1}^{N} (y_i - x_i)} = 0 \\
\Leftrightarrow &\lambda_{1,2} = -\frac{2n - \sum_{i=1}^{N} (y_i - x_i)}{2\sum_{i=1}^{N} (y_i - x_i)} \\
&\qquad \pm \sqrt{(\frac{2n - \sum_{i=1}^{N} (y_i - x_i)}{2\sum_{i=1}^{N} (y_i - x_i)})^2 + \frac{n}{\sum_{i=1}^{N} (y_i - x_i)}}
\end{aligned}
\tag{16}
$$

$$
\begin{aligned}
\Leftrightarrow \lambda_{1,2} &= -\frac{2n - \sum_{i=1}^{N}(y_i - x_i)}{2\sum_{i=1}^{N}(y_i - x_i)} \\
&\quad \pm \sqrt{(\frac{2n - \sum_{i=1}^{N}(y_i - x_i)}{2\sum_{i=1}^{N}(y_i - x_i)})^2 + \frac{n}{\sum_{i=1}^{N}(y_i - x_i)}} \\
\Leftrightarrow \lambda_{1,2} &= -(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2}) \\
&\quad \pm \sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2})^2 + \frac{n}{\sum_{i=1}^{N}(y_i - x_i)}} \\
\Leftrightarrow \lambda_{1,2} &= -(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2}) \\
&\quad \pm \sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)})^2 - \frac{n}{\sum_{i=1}^{N}(y_i - x_i)} + \frac{1}{4} + \frac{n}{\sum_{i=1}^{N}(y_i - x_i)}} \\
\Leftrightarrow \lambda_{1,2} &= -(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2}) \\
&\quad \pm \sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)})^2 + \frac{1}{4}} \\
\Leftrightarrow \lambda_{1,2} &= -(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2}) \\
&\quad \pm \sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} + \frac{1}{2})} \cdot \sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2})} \\
\Leftrightarrow &\sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2})} \\
&\quad \cdot \left(-\sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} - \frac{1}{2})} \pm \sqrt{(\frac{n}{\sum_{i=1}^{N}(y_i - x_i)} + \frac{1}{2})} \right)
\end{aligned}
\tag{17}
$$

2 Linear Regression

Consider the linear regression model $y = x^T\beta + \epsilon$, where $x \in \mathbb{R}^d$ are the predictor variables, $y \in \mathbb{R}$ is the response variable, $\beta \in \mathbb{R}^d$ are the linear regression coefficients, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is random i.i.d. noise. This model can be understood as defining a conditional probability distribution $p(y|x; \beta)$, parameterized by the linear regression coefficients β.

We collect a dataset $\mathcal{D} = ((x_1, y_1), ..., (x_N, y_N))$ of N independent draws of pairs (x_i, y_i). We summarize data into the vector $y = (y_1, ..., y_N) \in \mathbb{R}^N$ and the matrix $X = (x_1, ..., x_N) \in \mathbb{R}^{N\times d}$. We would like to learn for this data a good model parameter β.

The maximum-likelihood solution for β is the parameter for which observed outputs $y_1, ..., y_N$ are the most likely under the model's output distribution. It is obtained by solving the optimization problem:

$$\underset{\beta}{max} \prod_{i=1}^{N} Pr(y = y_i | x = x_i; \beta) \tag{18}$$

and can be shown to have the closed form:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{19}$$

Task a)

Show that $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$, i.e., $\hat{\beta}$ is Gaussian distributed with mean β and covariance matrix $\sigma^2(X^T X)^{-1}$.

For our own interest first $\hat{\beta}$ and $\hat{\sigma}^2$ are derived. The actual answer to this task's position starts in the subsection after the next subsection (in *Derivation of the distribution of $\hat{\beta}$ and $\hat{\sigma}^2$*).

Derivation of $\hat{\beta}$ and $\hat{\sigma}^2$

Starting from a conditional probability density function of the dependent variable

$$p(y_i|X) = \frac{1}{\sqrt{2\pi\sigma}} \cdot exp(-\frac{1}{2} \cdot \frac{(y_i - x_i\beta)^2}{\sigma_0^2}) \tag{20}$$

the likelihood function is

$$\begin{aligned} p(D|\beta,\sigma^2) &= \prod_{i=1}^{N} p(y_i|X;\beta,\sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \cdot exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2) \end{aligned} \tag{21}$$

This leads to a log likelihood function of

$$\begin{aligned} l(\beta,\sigma^2) &= ln(\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \cdot exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2)) \\ &= -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2 \end{aligned} \tag{22}$$

Maximizing the log likelihood function $\underset{\beta,\sigma^2}{max} l(\beta,\sigma^2)$ for β and σ^2 finds $\hat{\beta}$ and $\hat{\sigma}^2$. The first-order conditions are

$$\begin{aligned} \nabla_\beta l(\beta,\sigma^2) = 0 \\ \frac{\partial}{\partial\sigma^2} l(\beta,\sigma^2) = 0 \end{aligned} \tag{23}$$

For ∇_β, it is calculated

$$\begin{aligned} \nabla_\beta l(\beta,\sigma^2) &= \nabla_\beta(-\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2) \\ &= \frac{1}{\sigma^2}(\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i\beta) \end{aligned} \tag{24}$$

which is only equal to zero if

$$(\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i\beta) = 0 \tag{25}$$

This leads to

$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i^T y_i}{\sum_{i=1}^{N} x_i^T x_i)} = (X^T X)^{-1} X^T y \tag{26}$$

The partial derivative of the log-likelihood with respect to σ^2 is

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}l(\beta,\sigma^2) &= \frac{\partial}{\partial\sigma^2}(-\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2) \\ &= \frac{1}{2\sigma^2}(\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2)\end{aligned} \tag{27}$$

which (assuming $\sigma^2 \neq 0$) is equal to zero only if

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\hat{\beta})^2 \tag{28}$$

$\hat{\beta}$ does not depend on $\hat{\sigma}^2$, the solution is explicit. This proves that $\hat{\beta} = (X^TX)^{-1}X^Ty$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\hat{\beta})^2$.

Derivation of the distribution of $\hat{\beta}$ and $\hat{\sigma}^2$

Now it has to be shown that the parameter vector $\left[\hat{\beta}, \hat{\sigma}^2\right]^T$ is (asymptotically) normally distributed. Starting from the original conditional probability densitiy function, we consider the score vector $\nabla_\beta ln(p(y_i|X))$ (which indicates the sensetivity of a likelihood function) for the first d entries:

$$\begin{aligned}\nabla_\beta ln(p(y_i|X)) &= \nabla_\beta(ln(\frac{1}{\sqrt{2\pi\sigma}} \cdot exp(-\frac{1}{2} \cdot \frac{(y_i - x_i\beta)^2}{\sigma_0^2}))) \\ &= \nabla_\beta(-\frac{1}{2}ln((2\pi) - \frac{1}{2}ln(\sigma^2) - \frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}) \\ &= \frac{x_i^T(y_i - x_i\beta)}{\sigma^2}\end{aligned} \tag{29}$$

The (d+1) entry then comes from:

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}(ln(p(y_i|X))) &= \frac{\partial}{\partial\sigma^2}(-\frac{1}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) - \frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2})) \\ &= -\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(y_i - x_i\beta)^2}{(\sigma^2)^2}\end{aligned} \tag{30}$$

Further the Hessian matrix (matrix of second derivates) can be expressed with:

$$\nabla^2_{\beta,\sigma^2}ln(p(y_i|X))) = \begin{bmatrix}\nabla_\beta(\nabla_\beta ln(p(y_i|X))) & \nabla_\beta(\frac{\partial}{\partial\sigma^2}ln(p(y_i|X))) \\ \nabla_\beta(\frac{\partial}{\partial\sigma^2}ln(p(y_i|X))) & \frac{\partial}{\partial\sigma^2}(\frac{\partial}{\partial\sigma^2}ln(p(y_i|X)))\end{bmatrix} \tag{31}$$

Computing the single entries of the matrix:

$$\begin{aligned}
\nabla_\beta(\nabla_\beta ln(p(y_i|X))) &= \nabla_\beta(\frac{x_i^T(y_i - x_i\beta)}{\sigma^2}) = -\frac{1}{\sigma^2}x_i^T x_i \\
\nabla_\beta(\frac{\partial}{\partial\sigma^2} ln(p(y_i|X))) &= \nabla_\beta(-\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(y_i - x_i\beta)^2}{(\sigma^2)^2}) = -\frac{y_i - x_i\beta}{(\sigma^2)^2}x_i \\
\frac{\partial}{\partial\sigma^2}(\frac{\partial}{\partial\sigma^2} ln(p(y_i|X))) &= \frac{\partial}{\partial\sigma^2}(-\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(y_i - x_i\beta)^2}{(\sigma^2)^2}) = \frac{1}{2(\sigma^2)^2} - \frac{(y_i - x_i\beta)^2}{(\sigma^2)^3}
\end{aligned} \tag{32}$$

So we have the following Hessian matrix:

$$\nabla^2_{\beta,\sigma^2} ln(p(y_i|X))) = \begin{bmatrix} -\frac{1}{\sigma^2}x_i^T x_i & -\frac{y_i - x_i\beta}{(\sigma^2)^2}x_i \\ -\frac{y_i - x_i\beta}{(\sigma^2)^2}x_i & \frac{1}{2(\sigma^2)^2} - \frac{(y_i - x_i\beta)^2}{(\sigma^2)^3} \end{bmatrix} \tag{33}$$

Unsing (Fisher's) information equality, we have that:

$$\begin{aligned}
& Var(\nabla_{\beta,\sigma^2} ln(p(y_i|X)))|_{\beta,\sigma^2} \\
= & -E(\nabla^2_{\beta,\sigma^2} ln(p(y_i|X))) \\
= & \begin{bmatrix} E\left[\frac{1}{\sigma^2}x_i^T x_i\right] & E\left[\frac{y_i - x_i\beta}{(\sigma^2)^2}x_i\right] \\ E\left[\frac{y_i - x_i\beta}{(\sigma^2)^2}x_i\right] & E\left[-\frac{1}{2(\sigma^2)^2} + \frac{(y_i - x_i\beta)^2}{(\sigma^2)^3}\right] \end{bmatrix} \\
= & \begin{bmatrix} \frac{1}{\sigma^2}E\left[x_i^T x_i\right] & \frac{1}{(\sigma^2)^2}E\left[(y_i - x_i\beta)x_i\right] \\ \frac{1}{(\sigma^2)^2}E\left[(y_i - x_i\beta)x_i\right] & \frac{1}{(\sigma^2)^3}E\left[(y_i - x_i\beta)^2\right] - \frac{1}{2(\sigma^2)^2} \end{bmatrix}
\end{aligned} \tag{34}$$

Looking at that, we have first

$$E\left[(y_i - x_i\beta)^2\right] = E\left[\epsilon_i^2\right] = Var\left[\epsilon_i\right] = \sigma^2 \tag{35}$$

and second by the *Law of Iterated Expectations*

$$\begin{aligned}
E\left[(y_i - x_i\beta)x_i^T\right] &= E\left[E\left[(y_i - x_i\beta)x_i^T|X\right]\right] \\
&= E\left[E\left[(y_i - x_i\beta)|X\right]x_i^T\right] \\
&= E\left[E\left[\epsilon_i|X\right]x_i^T\right] \\
&= E\left[0x_i^T\right] \\
&= 0
\end{aligned} \tag{36}$$

This leads to:

$$
\begin{aligned}
&Var(\nabla_{\beta,\sigma^2} ln(p(y_i|X)))|_{\beta,\sigma^2} \\
&= \begin{bmatrix} \frac{1}{\sigma^2} E\left[x_i^T x_i\right] & \frac{1}{(\sigma^2)^2} E\left[(y_i - x_i\beta)x_i\right] \\ \frac{1}{(\sigma^2)^2} E\left[(y_i - x_i\beta)x_i\right] & \frac{1}{(\sigma^2)^3} E\left[(y_i - x_i\beta)^2\right] - \frac{1}{2(\sigma^2)^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2} E\left[x_i^T x_i\right] & 0 \\ 0 & \frac{\sigma^2}{(\sigma^2)^3} - \frac{1}{2(\sigma^2)^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2} E\left[x_i^T x_i\right] & 0 \\ 0 & \frac{1}{2(\sigma^2)^2} \end{bmatrix}
\end{aligned} \tag{37}
$$

And from this, we get an *asymptotic* covariance matrix of

$$
(Var(\nabla_{\beta,\sigma^2} ln(p(y_i|X)))|_{\beta,\sigma^2})^{-1} = \begin{bmatrix} \sigma^2 E\left[x_i^T x_i\right]^{-1} & 0 \\ 0 & 2(\sigma^2)^2 \end{bmatrix} \tag{38}
$$

This shows that $\left[\hat{\beta}, \hat{\sigma}^2\right]^T$ can be approximated by a multivariate normal distribution with β, σ^2 and an asymptotic covariance matrix of

$$
\begin{bmatrix} \sigma^2 E\left[x_i^T x_i\right]^{-1} & 0 \\ 0 & 2(\sigma^2)^2 \end{bmatrix}
$$

whereby the very first entry of the matrix states the covariance matrix for $\hat{\beta}$.

Task b)

Discuss the benefit of knowing the full distribution $\hat{\beta}$ rather than only the estimate itself. What additional statements about β can be made (hint: variable selection)? Assume that σ^2 is known and does not need to be estimated.

With the distribution of the parameter, it is possible to perform a Bayesian Estimation which takes into account the specific distribution. This way, better predictions can be made with less training datapoints. Even if the available datapoints do not represent the underlying distribution sufficiently well, the predictions can get better since the beliefs of how the data should look like, are considered. For dimensional reduction (i.e. reducing the amount of variables while keeping the highest possible accuracy), it is possible to determine highly correlated variables for elimination which will affect the final model the least.

Task c)

Assume we have measured a new datapoint, x_*. We use our regression model to predict the response for $x_* : \hat{y}_* = x_*^T \hat{\beta}$. Derive the distribution of $\hat{y}_*$.

After computing $\hat{\beta} = (X^TX)^{-1}X^Ty$ we can state a vector $\hat{y} = X\hat{\beta}$ with expected value and variance (I is the identity matrix):

$$\begin{aligned}
\hat{y} &= \hat{\beta}X = X(X^TX)^{-1}X^Ty \\
E(\hat{y}) &= X\beta \\
V(\hat{y}) &= X(X^TX)^{-1}X^T\sigma^2I(X(X^TX)^{-1}X^T)^T = \sigma^2X(X^TX)^{-1}X^T
\end{aligned} \tag{39}$$

The residuals with expected value and variance can be found with

$$\begin{aligned}
e &= y - \hat{y} = (I - (X(X^TX)^{-1}X^T))y \\
E(e) &= (I - (X(X^TX)^{-1}X^T))X\beta = 0 \\
V(e) &= (I - (X(X^TX)^{-1}X^T))\sigma^2I(I - (X(X^TX)^{-1}X^T))^T \\
&= \sigma^2(I - (X(X^TX)^{-1}X^T))
\end{aligned} \tag{40}$$

So having the vector of fitted values $\hat{y}$ and the residuals we obtain the predicted value of y when $x = x_*$:

$$\begin{aligned}
\hat{y}_* &= x_*^T\hat{\beta} \\
E(\hat{y}_*) &= x_*^T\hat{\beta} \\
V(\hat{y}_*) &= \sigma^2x_*^T(X^TX)^{-1}x_*
\end{aligned} \tag{41}$$

So y_* is Gaussian distributed.

Task d)

Discuss the benefit of also knowing that distribution in an application of your choice.

In general, it allows the prediction of how frequently several responses will occur, without having information about data points. For example, we can think of a use case for a cinema. It might be interesting to predict the number of viewers who will see a movie, by the characteristics of the movie. This information can be used to choose a cinema hall with the right size or to employ enough staff to serve the viewers. Even the right products can be sold during the movie break. Knowing the distribution of viewers for a certain time can help to predict the number of viewers for a movie without knowing the movie's characteristics.

3 Programming

The programming part is handled in the separate file *sheet02.ipynb*.