IE30301 - Data Mining Assignment 2 (70 Points)

Eldor Fozilov

March 30, 2022

Exercise 1

1.1

1. **Supervised learning** refers a task aimed at predicting a variable of interest (target) based on a set of other variables (predictors). There are many supervised learning techniques such as linear regression, logistic regression, support vector machines and much more, which all share the following characteristic: they require data that has labels (observed targets) for each sample of (training) data in order to "learn" from it and offer a prediction in the end when new data fed into the model. For example, using supervised learning, we can get predictions on whether a particular email is spam or not based on some labeled data of spam and non-spam emails.
2. **Unsupervised learning** refers a task aimed at finding patterns and the underlying structure in the data set itself, which are easy to interpret for humans. This definition can tell us that the goal is completely different to that of supervised learning and there is no need for labeling the data set in unsupervised learning. Unsupervised learning techniques are commonly used to do the following four things: data clustering, anomaly detection, association (think of recommendation systems), and dimensionality reduction.
3. **Regression** is referred to a set of statistical methods whose purpose is to identify cause-and-effect relationships between different variables and/or to predict a variable of interest based on some other variables. In regression models, the target variable should be a continuous variable, while predictor variables can be of any type. For example, using regression analysis, we can predict the income of a person based on his or her education, age and gender (there might be a lot of other relevant variables).
4. **Classification** is similar to regression in that classification models also try to find inherent relationships between different variables and/or predict a target variable based on other variables. However, in classification, as the name might suggest, the target variable is not a continuous variable, but a categorical variable. For example, predicting whether an image contains a cat or not is considered a classification problem, and can be solved using various classification models such as logistic regression, naive bayes, and support vector machines.
5. **Clustering** is considered an unsupervised learning task, in which a clustering algorithm is applied to a dataset to identify groupings between data points that are "similar" to each other. Similarity might mean Euclidean distance between data points if variables

in a dataset are continuous, but it can also mean other things depending on specific problem measures. A classical application of clustering algorithms was in the field of marketing, in which marketers try to find customers with similar buying behavior and create more personalized campaigns for those specific groups. Clustering techniques help them accomplish that task in a data-driven way.

1.2

1. **Nominal** variables represent values that don't have any numeric meaning and thus they cannot be ordered. For example, variable such as gender (male | female), color of a car (blue | red), and types of movies (romance | thriller) can be considered as nominal variables.
2. **Binary** variables can only be in two states, thus is why they are called binary. For example, the result of a coin toss can be either tail or head, but not both, thus it can be considered as a binary variable.
3. **Continuous** variables can take any numerical value within a specified range, and that range can also be infinite. For example, the price of stock can be considered as a continuous variable since it can have value between 0 and a very large positive number, say 10000$.
4. **Numeric** variables have clear numeric meaning and thus can be ranked. For example, the revenue amount of a company in various periods can be considered a numerical variable.
5. **Ordinal** variables might not be represented as numbers per se, however they can be ordered. For example, answers to survey questions, which are related to how much a person agrees with some statement, such as "don't agree", "somewhat agree", "totally agree" can be considered as an ordinal variable.

1.3

1. **Models** help us understand or predict the real world. A good analogy for a model is a paper map, which is an extremely simplified version of the real world, but still can help a lot. A phenomenon that we are trying to understand or predict its future behavior, in reality, can be influenced by sum of a large number of factors, where the individual influences may vary based on those factors and our target. Using a model, we try to estimate those individual influences, which are commonly called as "**parameters**" among people and in textbook definitions. We try to make sure that after estimating those parameters using data, our model results can be trusted and be applicable on new data. In other words, it can **generalize** to new data and give us useful information about the real world.

Exercise 2

Since the given set of data $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ is assumed to be i.i.d and and follow a distribution of $\mathcal{N}(\mu, \sigma^2)$, the likelihood function, which is a parameterized density $p(\boldsymbol{x} \mid \mu, \sigma^2)$, will look like the following:

$$p(\boldsymbol{x} \mid \mu, \sigma^2) = p(X = x_1 \mid \mu, \sigma^2) \cdot p(X = x_2 \mid \mu, \sigma^2) \cdot ... \, p(X = x_n \mid \mu, \sigma^2) = \prod_{k=1}^{n} p(X = x_k \mid \mu, \sigma^2) \quad (2.1)$$

Its corresponding log-likelihood function is given by

$$l(\mu, \sigma^2) = \sum_{k=1}^{n} \ln p(X = x_k \mid \mu, \sigma^2) \quad (2.2)$$

Since we know the distribution of each data point, we can represent the log-likelihood as

$$l(\mu, \sigma^2) = \sum_{k=1}^{n} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \, e^{\frac{-(x_k - \mu)^2}{2\sigma^2}} \right) \quad (2.3)$$

By utilizing logarithmic function properties such as $\ln \boldsymbol{a} \cdot \boldsymbol{b} = \ln \boldsymbol{a} + \ln \boldsymbol{b}$, $\ln \boldsymbol{a}^{\boldsymbol{b}} = \boldsymbol{b} \ln \boldsymbol{a}$ and $\ln \boldsymbol{e}^{\boldsymbol{r}} = \boldsymbol{r}$, we can write the above expression as

$$\begin{aligned} l(\mu, \sigma^2) &= \sum_{k=1}^{n} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \, e^{\frac{-(x_k - \mu)^2}{2\sigma^2}} \right) = \sum_{k=1}^{n} \left[\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \left(\frac{(x_k - \mu)^2}{2\sigma^2} \right) \right] = \\ &= n \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{k=1}^{n} \left(\frac{(x_k - \mu)^2}{2\sigma^2} \right) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2 \end{aligned} \quad (2.4)$$

Now, in order to find $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$, we will take the partial derivatives of the log-likelihood function with respect to μ and σ^2, and then equal them to zero.

$$\begin{aligned} \frac{\partial l}{\partial \mu} = -\frac{2}{2\sigma^2} \cdot (-1) \sum_{k=1}^{n} (x_k - \mu) = \frac{1}{\sigma^2} \left(\sum_{k=1}^{n} x_k - n\mu \right) = 0 \\ \Longrightarrow \sum_{k=1}^{n} x_k = n\mu \end{aligned} \quad (2.5)$$

Based on equation 2.5, we can conclude that indeed $\hat{\mu}_{ML} = \frac{1}{n} \sum_{k=1}^{n} x_k$

We substitute that $\hat{\mu}_{ML}$ into the following equation instead of μ:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} (x_k - \mu)^2 = 0 \quad (2.6)$$

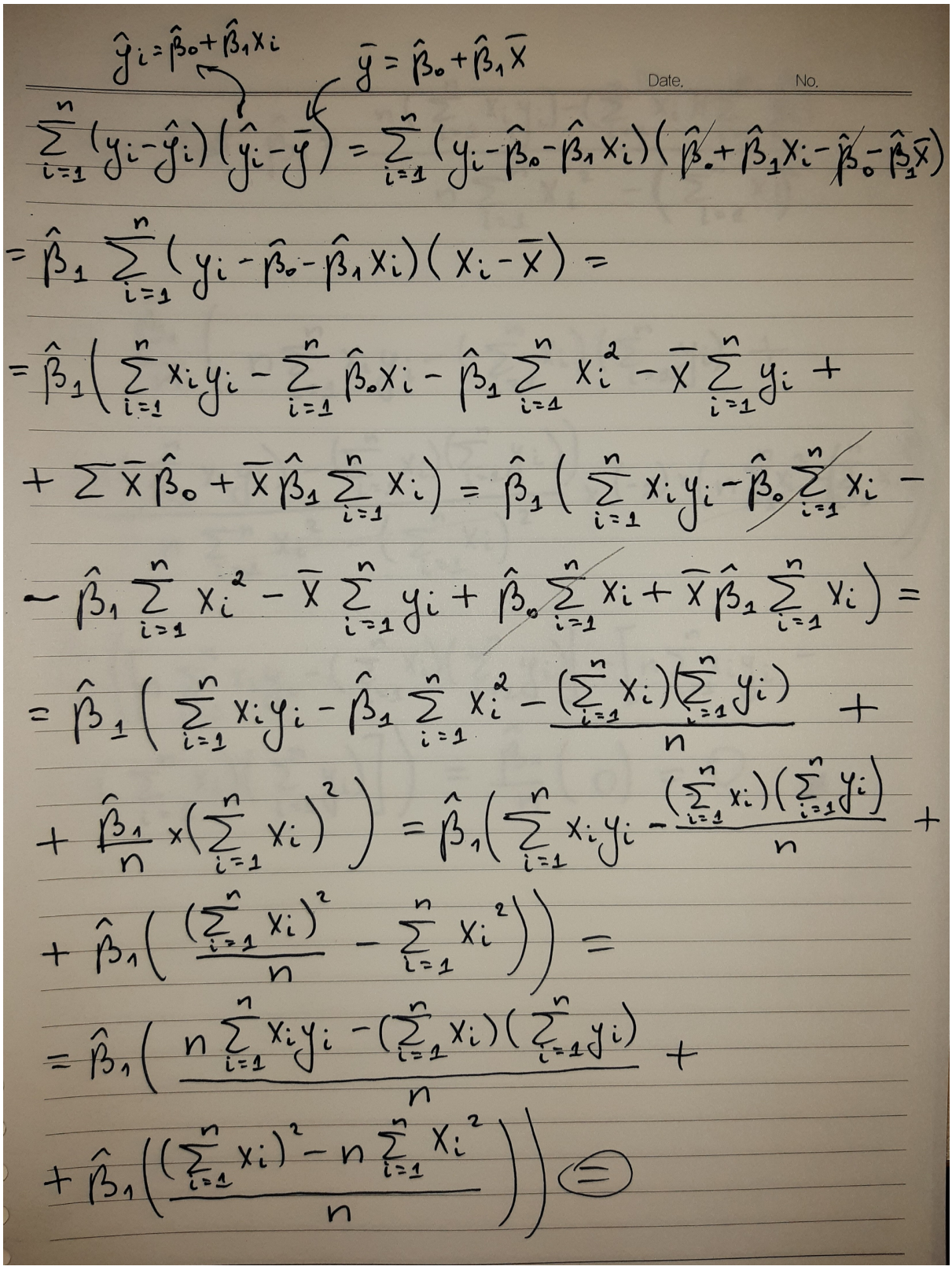
$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} (x_k - \hat{\mu}_{ML})^2 = 0 \quad (2.7)$$

$$\frac{1}{2\sigma^4}\sum_{k=1}^{n}(x_k - \hat{\mu}_{ML})^2 = \frac{n}{2\sigma^2} \tag{2.8}$$

$$\Longrightarrow \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu}_{ML})^2 \tag{2.9}$$

Exercise 3

3.1



We know that $\hat{\beta}_1 = \dfrac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$.

$\Rightarrow (=) \dfrac{\hat{\beta}_1}{n}\Bigg(n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) + \dfrac{\left(n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \times (-1) \times \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\Bigg)$

$= \dfrac{\hat{\beta}_1}{n}\left(\left[n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right] - \left[n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]\right) = \dfrac{\hat{\beta}_1}{n}(0) = 0$ ■

3.2

3.2

We know that $\hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

and $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

We will now use those facts.

$$\boxed{E(\hat{\beta}_1)} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \times \left(n\sum_{i=1}^{n} x_i E(y_i) - \right.$$

$$\left. - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} E(y_i)\right) \right) =$$ for now, let's label $\dfrac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$ as A to simplify things.

$$= A\left(n\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \beta_0 + \beta_1 x_i\right)\right) =$$

$$= A\left(n\beta_0 \sum_{i=1}^{n} x_i + n\beta_1 \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i\right)\right)$$

$$= A\left(n\beta_0 \sum_{i=1}^{n} x_i + n\beta_1 \sum_{i=1}^{n} x_i^2 - n\beta_0 \sum_{i=1}^{n} x_i - \beta_1 \left(\sum_{i=1}^{n} x_i\right)^2 \right) =$$

$$= A\left(n\beta_1 \sum_{i=1}^{n} x_i^2 - \beta_1 \left(\sum_{i=1}^{n} x_i\right)^2 \right) = \beta_1 A\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \right)$$

$$= \beta_1 \times \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \times \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \right) =$$

$$= \beta_1$$

Date. No.

$\hat{\beta}_1$ can also be represented as $\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

because $\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}\bar{x}y_i}{\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)} =$

$= \dfrac{\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2} =$

$= \dfrac{\sum_{i=1}^{n}x_iy_i-\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}}{\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2} =$

$= \dfrac{\sum_{i=1}^{n}x_iy_i-\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2} = \dfrac{n\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{\frac{n}{\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}}} =$

$= \dfrac{n\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$, which is $\hat{\beta}_1$

Date. No.

So, $\boxed{Var(\hat{\beta}_1)} = Var\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) =$

$= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 \times \sum_{i=1}^{n}\left((x_i-\bar{x})^2 Var(y_i)\right) =$

$= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 \times \sum_{i=1}^{n}\left((x_i-\bar{x})^2 \sigma^2\right) =$

$= \frac{1}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} \times \sigma^2 \times \sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sigma^2}{S_{xx}}$

$\boxed{E(\hat{\beta}_0)} = E(\bar{y}-\hat{\beta}_1\bar{x}) = E(y)\, E\left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1\bar{x}\right) =$

$= \left(\frac{1}{n}\sum_{i=1}^{n} E(y_i)\right) - \beta_1\bar{x} = \left(\frac{1}{n}\sum_{i=1}^{n}\beta_0+\beta_1 x_i\right) -$

$-\beta_1\frac{\sum_{i=1}^{n} x_i}{n} = \beta_0 + \frac{\beta_1}{n}\sum_{i=1}^{n} x_i - \frac{\beta_1}{n}\sum_{i=1}^{n} x_i =$

$= \beta_0$

$\boxed{Var(\hat{\beta}_0)} = Var(\bar{y}-\hat{\beta}_1\bar{x}) = Var(\bar{y}) + \cancel{\beta_1 + \bar{x}^2\beta}$

$+(-\bar{x})^2 Var(\hat{\beta}_1) + 2(-\bar{x})\, Cov(\bar{y}, \hat{\beta}_1) = Var\left(\left(\sum_{i=1}^{n} y_i\right)/n\right)$

$+\bar{x}^2 Var(\hat{\beta}_1) \underset{\text{minus}}{-} 2\bar{x}\, Cov\left(\frac{\sum_{i=1}^{n} y_i}{n}, \hat{\beta}_1\right)$

Date. No.

Let's decompose it.

To find $Var(\hat{\beta}_0)$, we first need to find $Var(\bar{y})$, $Var(\hat{\beta}_1)$ and $Cov(\bar{y}, \hat{\beta}_1)$

We already knew that $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

$$Var(\bar{y}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2} \times \sum_{i=1}^{n} Var(y_i) =$$

$$= \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

this is because y_i and y_j, where $i \neq j$, are considered independent.

Now we only need to find

$Cov(\bar{y}, \hat{\beta}_1)$

To do that, we will use two properties of covariance, which are the following:

① $Cov\left(\sum_{i=1}^{n} A_i, \sum_{j=1}^{n} B_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(A_i, B_j)$

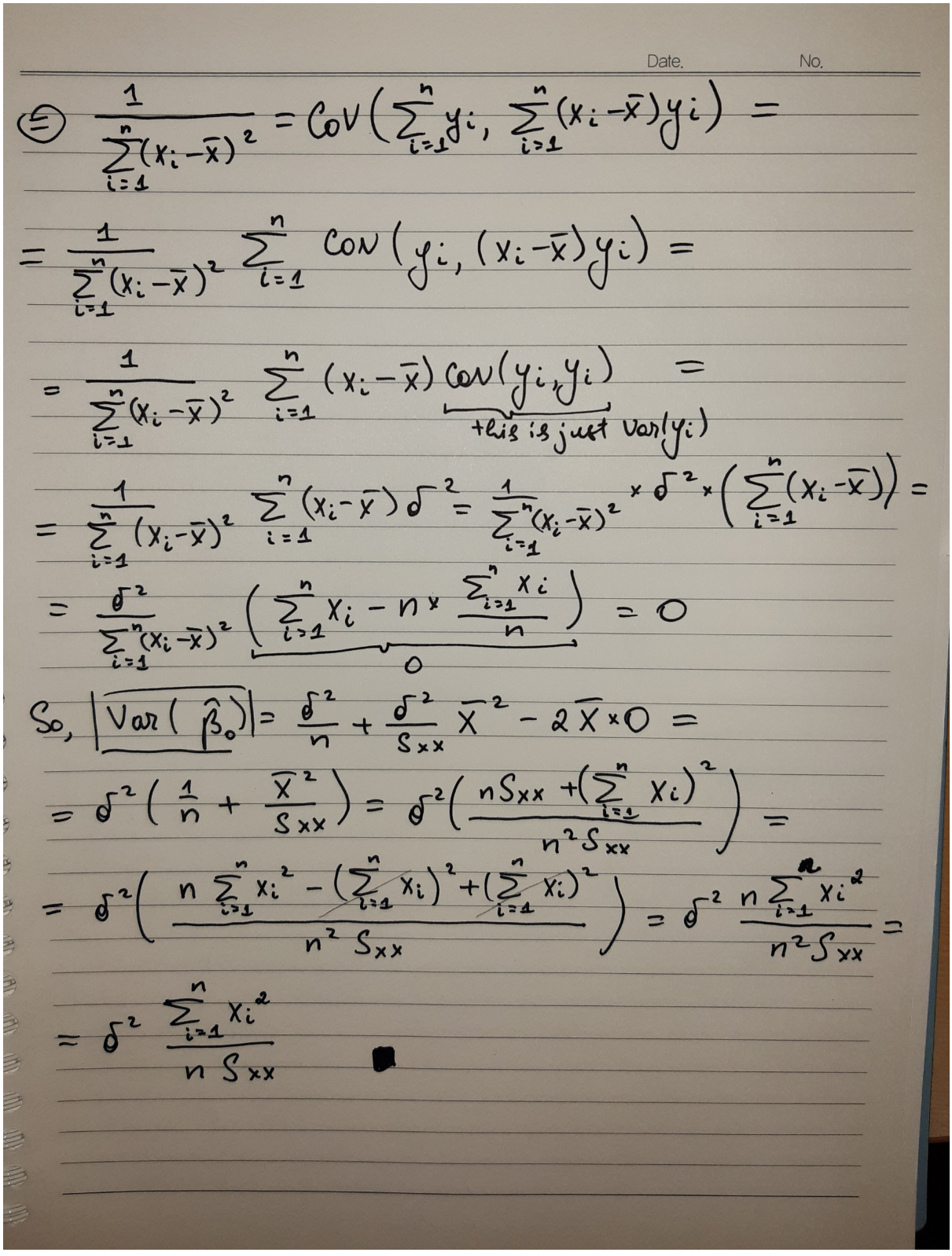
② $Cov(c_1 A, c_2 B) = c_1 c_2 Cov(A, B)$, c_1, c_2 - constants

So, let's find $Cov(\bar{y}, \hat{\beta}_1)$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\sum_{i=1}^{n} y_i, \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \quad (=)$$

$\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is not a random variable (as we assume X to be given).

Thus, we take it to outside



3.3

The reason why we need to use a t-test (instead of z-test) to test the statistical significance of parameters such as $\hat{\beta}_0$ and $\hat{\beta}_1$ is that we don't observe the value of the population variance σ^2 and thus we can only estimate it using the sampled data.

Exercise 4

4.1

If we can prove that $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is equal to $\frac{S_{xy}^2}{S_{xx}}$, which is given as **SSR**, and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is equal to $S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, which is given as **SSE**, then we can be sure that the following expression of statistic $F^\star$ using SSR and SSE is correct:

$$F^\star = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}}$$

Note: in the above equation, **n** is the total number of samples in a data set

So, we will now prove those two things:

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 = \sum_{i=1}^{n}(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2 = \hat{\beta}_1{}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \\ = \left(\frac{S_{xy}}{S_{xx}}\right)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{S_{xy}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{S_{xy}^2}{S_{xx}}\end{aligned} \tag{4.1}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}((y_i - \bar{y}) - (\hat{y}_i - \bar{y}))^2 =$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 =$$

$$S_{yy} - 2\left(\sum_{i=1}^{n} y_i \hat{y}_i - \sum_{i=1}^{n} \bar{y}\hat{y} - \sum_{i=1}^{n} \bar{y} y_i + \sum_{i=1}^{n} \bar{y}^2\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\sum_{i=1}^{n} y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \sum_{i=1}^{n} \bar{y}\hat{y}_i - \sum_{i=1}^{n} \bar{y} y_i + n\bar{y}^2\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$= S_{yy} - 2\left(\hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} \hat{y}_i - \bar{y} \sum_{i=1}^{n} y_i + n\bar{y}^2\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} \hat{y}_i - n\bar{y}^2 + n\bar{y}^2\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} \hat{y}_i\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \bar{y} n \hat{\beta}_0 - \bar{y}\hat{\beta}_1 \sum_{i=1}^{n} x_i\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\hat{\beta}_0 \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \bar{y}\hat{\beta}_1 \sum_{i=1}^{n} x_i\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\left(\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \bar{y}\hat{\beta}_1 \sum_{i=1}^{n} x_i\right) + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\hat{\beta}_1\left(\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i\right) + \frac{S_{xy}^2}{S_{xx}} = S_{yy} - 2\hat{\beta}_1 S_{xy} + \frac{S_{xy}^2}{S_{xx}} =$$

$$S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

4.2

The null and alternative hypotheses for the F-test in 4.1 are the following:

$$H_0 \ : \ \beta_1 = 0$$

$$H_a \ : \ \beta_1 \neq 0$$

The intuition behind such a formulation of hypotheses can be explained as follows: if we assume that H_0 is true, meaning that $\beta_1 = 0$, this indicates that basically there is no point of constructing a regression line other than the line $y = \overline{y}$ to the data since there is no relationship between the dependent variable **y** and the independent variable **x**, and all the variance in the data is due to the error term ϵ. So, if we run a regression model, estimate the parameters β_0 and β_1, and construct a regression line, then $\frac{SSR}{SSE}$ should be very close to zero, and thus F* statistics, which is $(\boldsymbol{n} - 2)\frac{\boldsymbol{SSR}}{\boldsymbol{SSE}}$ in the case of simple linear regression, should also be very close to zero. However, if H_a is true, meaning that $\beta_1 \neq 0$, then it means fitting a regression line on the data is going to be useful since it can explain some of the variation in the target variable **y**.

4.3

$$\frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2} =$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \frac{2(\sum_{i=1}^{n} x_i)^2}{n} + n\bar{x}^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \frac{2(\sum_{i=1}^{n} x_i)^2}{n} + \frac{(\sum_{i=1}^{n} x_i)^2}{n}} =$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \hat{\beta}_1$$

4.4

Here, we will utilize one of the results we got in 4.1, specifically the fact that $SSR = \frac{S_{xy}^2}{S_{xx}}$, and the equation that we proved in 4.3.

$$F^\star = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{\frac{S_{xy}^2}{S_{xx}}}{s^2} = \frac{\hat{\beta}_1 S_{xy}}{s^2} = \frac{\hat{\beta}_1 \frac{S_{xy}}{S_{xx}} S_{xx}}{s^2} =$$

$$= \frac{(\hat{\beta}_1)^2 S_{xx}}{s^2} = \frac{(\hat{\beta}_1)^2}{\frac{s^2}{S_{xx}}} = \left(\frac{\hat{\beta}_1}{\frac{s}{\sqrt{S_{xx}}}} \right)^2 = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 = (t^\star)^2$$

$$F^\star = (t^\star)^2 \Longrightarrow (F^\star)^{1/2} = |t^\star|$$

Exercise 5

5.1

The formula for finding estimates of parameters β_0, β_1 and β_2 is as follows:

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

However, we should modify the matrix $\boldsymbol{X}$ in order to account for the intercept estimate parameter $\hat{\beta}_0$ in the following way:

$$X = \begin{bmatrix} 1 & 2 & 1 \\ 1 & -2 & -2 \\ 1 & 1 & 0 \\ 1 & 3 & 2 \end{bmatrix}$$

So, now we can represent our regression model as $\hat{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and calculate $\hat{\boldsymbol{\beta}}$, which represents the vector $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$.

$$\hat{\boldsymbol{\beta}} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -2 & 1 & 3 \\ 1 & -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & -2 & -2 \\ 1 & 1 & 0 \\ 1 & 3 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -2 & 1 & 3 \\ 1 & -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} =$$

$$\begin{bmatrix} 4 & 4 & 1 \\ 4 & 18 & 12 \\ 1 & 12 & 9 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -2 & 1 & 3 \\ 1 & -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} =$$

$$\begin{bmatrix} 3 & -4 & 5 \\ -4 & 35/6 & -22/3 \\ 5 & -22/3 & 28/3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -2 & 1 & 3 \\ 1 & -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} =$$

$$\begin{bmatrix} 0 & 1 & -1 & 1 \\ 1/3 & -1 & 11/6 & -7/6 \\ -1/3 & 1 & -7/3 & 5/3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -5/6 \\ 4/3 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -5/6 \\ 4/3 \end{bmatrix}$$

5.2

In [1]:
```
import numpy as np
from sklearn import linear_model
X = np.array([[2 ,1] ,[ -2 , -2] ,[1 ,0] ,[3 ,2]])
y = np.array([0 ,1 ,2 ,3])
```

In [2]:
```
model = linear_model.LinearRegression(fit_intercept = True)
model.fit(X,y)
print(model.coef_, model.intercept_)
```

```
[-0.83333333  1.33333333] 2.0000000000000018
```

We got the same results as in 5.1.