Inf2 – Foundations of Data Science 2025

Task: Semester 1 Week 7 Workshop – Linear Regression

October 2025

Task 1 – Intuition

In this task most of the questions involve discussion but. It should hopefully not take too long.

1. The very first basic guess would likely involve β_1 being 1. If you remember the "regression to the mean" part of the lecture, you should be able to guess that β_1 is likely going to be at least a bit less than 1, but possibly not exactly by much. For β_0, you should be able to come up with a reasonable guess by, e.g., imagining what the average difference between the coldest and hottest time of a day may be.
2. One of the results you should be aware of is that the optimal β_1 should be $\hat{\beta}_1 = r\sigma_y/\sigma_x$. So β_1 should be proportional to r. If you assume $\sigma_y \approx \sigma_x$ (which is not unreasonable) then $\hat{\beta}_1 \approx 0.9$.
   The second piece of info about the $8°C$ difference between the highest and lowest temperature in the UK is a bit of a trick. The dataset is continental Europe, and it would be risky to assume that has the same temperature statistics as the UK.
3. The obvious thing here is the two "spokes" to the side of the large clump of data (one at around $x = -17$ and one at $y = -17$) – these clearly seem erroneous. We can think of no reason why those are there, so it would not be unreasonable to remove them.
   Also, there are a few other points that have the min temp higher than the max temp (everything below the $x = y$ line). This also doesn't really make sense for any reasonable definition of the data.
   Still, the problematic points number around 50 from the dataset of around 10000 data points, so they do not make much of a dent in the calculations (but it's still worth removing them to not have obvious issues with the visualisation).
4. —
5. There's a lot of potential data that can be used to help the prediction, e.g. humidity or wind speed. You can also take a look at the full dataset to maybe get some ideas.

Task 2 – Calculations

1. The squared loss is $f(\beta_0, \beta_1) = \sum_{i=1}^{4}(y_i - \hat{y}_i)^2$ where $\hat{y}_i$ are the predictions $\hat{y}_i = \beta_0 + \beta_1 x$ given the provided $\beta_0 = 75$ and $\beta_1 = -5$. The predictions are:

$$\hat{y}_1 = 75 - 5 \times 3 = 60$$

$$\hat{y}_2 = 75 - 5 \times 5 = 50$$
$$\hat{y}_3 = 75 - 5 \times 7 = 40$$
$$\hat{y}_4 = 75 - 5 \times 9 = 30$$

Therefore, the loss is:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{4} (y_i - \hat{y}_i)^2 = (65-60)^2 + (50-50)^2 + (30-40)^2 + (15-30)^2 = 350$$

2. In the lectures, we defined the coefficient of determination as $R^2 = 1 - SSE/SST$ where SSE is the *sum of squared errors* and SST is the *total sum of squares*. SSE in this case is the loss we already computed, while $SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = n\sigma_y^2$. Calculating this gives $R^2 = 0.759$ up to 3 significant digits. (Note you can get $R^2 = 0.678$ instead if you used the formula for the lecture $SST = (n-1)\sigma_y^2$. This result is also fine. For the small amount of data here, the difference between the two possible results is big, but for large n the difference would be negligible.)
3. The main takeaway from the residual plot is that the residuals are clearly correlated with the independent variable – as x (the week) increases, so does the residual. For a good model fit, the residuals should be uncorrelated with x.
4. While the number for the loss is big, it doesn't really mean much without a frame of reference. However, R^2 is somewhat low for a dataset this simple (one could expect it to be close to 1), and the residual plot clearly shows there is more to desire from the residuals.
5. There are multiple ways to compute the optimal parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. For example, you can use $\hat{\beta}_1 = r\sigma_y/\sigma_x$ but this requires computing both the variation in x and in y and also the correlation coefficient. The most straightforward way would be to use the direct formulas from the second-to-last field of the first Linear Regression Lecture. Either way, the result should be:

$$\hat{\beta}_0 = 91.0$$
$$\hat{\beta}_1 = -8.5$$

6. The process is the same as before, just with the new parameters:

$$f(\hat{\beta}_0, \hat{\beta}_1) = 5$$
$$R^2 = 0.997 \ (or\ 0.995)$$

7. The residuals now don't look correlated with x at all, which implies the fit is good.
8. The prediction for Week 11 would be $91 - 11 \times 8.5 = -2.5$. Of course, a negative attendance doesn't make sense. But, you can argue that the model has only seen data in the range of Week 3 to Week 9, so it would become inaccurate with data outside that range. Also, it's the data scientist's responsibility to input reasonable values to the model – inputting "Week −10" into the model would also give a nonsensical result, but that's not a reasonable thing to do. Still, there are things you could do to not have predictions outside of the range of 0% to 100% attendance (the simplest would be to hard limit predictions within that range).

9. Nothing important would really change. The optimal parameters and the residuals would be both 100 times smaller, and the loss would be 10000 times smaller (since it's squared). However, the R^2 values will remain the exact same. If you change the y-axes of the plots to be 100 times smaller, they would look the same too.
   You could view this as a simple (and not particularly useful) transformation of the data.

Task 3 – Simpson's Paradox

Steps 1, 2, and 4 should be mostly self-explanatory - running a simple linear regression works when you assume all the data points are a single group, but once you have the four different groups as shown in Figure 5 it wouldn't be able to capture each individual group's behaviour. Note here that the group is a confounding variable here.

For step 3, you can split the dataset into four parts - one for each group - and create a different model for each of them (so just have four completely different models in the end). Or you can use multiple linear regression and have the group be an additional independent variable. (Though for the latter it would make more sense to add three binary variables instead - one indicating if a data point belongs to group 1, one indicating group 2, and one indicating group 3 - you don't need a fourth one because if the first three are all 0 (off), then that can indicate group 4 already).

For step 5, we refer you to the paper's own explanation in Part II on page 2 (link to paper for reference: von Kügelgen et al. (2021)).

Task 4 – Reading and understanding a data science article

1. Question: What is/are the specific question(s) that the article tries to answer?

   Generally, the best place to look to see what question(s) a scientific article is trying to answer are going to be the abstract and the introduction (typically near the end of the introduction).

   This article is trying to predict Olympic medals at the 2016 Summer Olympic Games. More specifically, it is investigating if there is a correlation between socio-economic variables and Olympic success.

2. Context: What dataset(s) does it use? How were the datasets collected? Which methods were used to analyse the problem?

   The article is primarily using Olympic Games data from games held between 1996 and 2008 (this can be found in the first paragraph of the section "*Predicting Olympic medals*"). In addition, their "sophisticated model" uses socio-economic data about countries such as GDP and population.

   The article doesn't concretely reference any of the specific datasets that it's using - it might have been better to do so (so that an interested reader might be able to replicate the results, for example). However, all the data they use is quite public, so

it should be readily accessible for anyone who wants to follow-up on the study.

In terms of methods, the article uses multiple linear regression - they describe the exact models they use in the box titled "*The estimated models*".

3. Correctness: Do the assumptions appear to be valid?

   The assumptions here are largely to do with what can affect Olympic performance (e.g. GDP, a country's political system, etc). These overall make sense - nothing stood out to me as being erroneous. It could be that some of the factors are less important that others, but a linear regression model can implicitly learn that by placing a close-to-zero value for the respective coefficient. It's also likely that there are other variables that haven't been considered by the article that would be beneficial to consider.

4. Model: What model(s) is the paper using? What variables does the model(s) use?

   As said above, everything important about the models is written in the "*The estimated models*" box. The article uses two models - a simple one with only 2 variables (3 coefficients), and a sophisticated one with 8 variables (9 coefficients). Each variable is described and justified there as well - please refer to the paper for details.

   Note also that the article also uses Mean Absolute Error as a measure of model fit (see the final box titled "*Mean absolute error and mean forecast error*" in the article), but this is beyond the scope of the course.

5. Clarity: Is the paper well written? Is there anything you would do to change it?

   The article is fairly well written. What the authors are doing is clear, and their interpretations are relatively easy to understand. The separate box for "*The estimated models*" helps with identifying how the model works (which is generally one of the most important aspects of a data science article).