Inf2 – Foundations of Data Science 2025

Task: Semester 1 Week 7 Workshop – Linear Regression

FOUNDATIONS OF DATA SCIENCE

October 2025

In the S1 Week 7 workshop, we will explore concepts related to **Simple Linear Regression**. The goal is to try to build intuition, understand the calculations surrounding Linear Regression, and be able to discuss good regression practices with others.

Task 1 – Intuition

You have weather data from continental Europe during the mid-19th century. For this task, you will try to predict the maximum temperature for a given day from the minimum temperature during that day. (The dataset has many other weather-related features, but we'll only use these two for now.) The goal is to create a linear regression model with min temperature as the independent variable and max temperature as the dependent variable.

The model you want to eventually fit will be $y = \beta_0 + \beta_1 x$, where β_0 and β_1 are the parameters (the same as in the lectures).

1. Before seeing any data, can you make a guess about what the optimal linear regression parameters would be? Write it down.
2. You still don't have all the data, but now you're given some extra information: the correlation coefficient is approximately $r = 0.9$. Additionally, you also are told that the average daily difference between the highest and lowest temperature in the UK during 1950-80 is about $8°C$. Knowing this, do you have a better idea what the parameters would be? Write the new ones down.
3. Let's see the raw data first - scroll to Figure 1 on the next page. Looking at the data, does anything jump out at you? If there are any issues, what would you do to fix them? Should you even fix them?
4. The actual optimal parameters for the raw data are (approximately) $\hat{\beta}_0 = 10.27$, $\hat{\beta}_1 = 0.95$. Are they different from your previous guesses? Why do you think they're different?
5. You could use other data (i.e. variables) and combine with minimum temperature (perhaps using Multiple Linear Regression) to get a more accurate prediction. What other variables would be useful to predict max temperature?

If you're interested in getting more info about the dataset, or want to play with it yourself, you can find it here.

Avoid looking at the image on the next page before you've done parts 1 and 2!

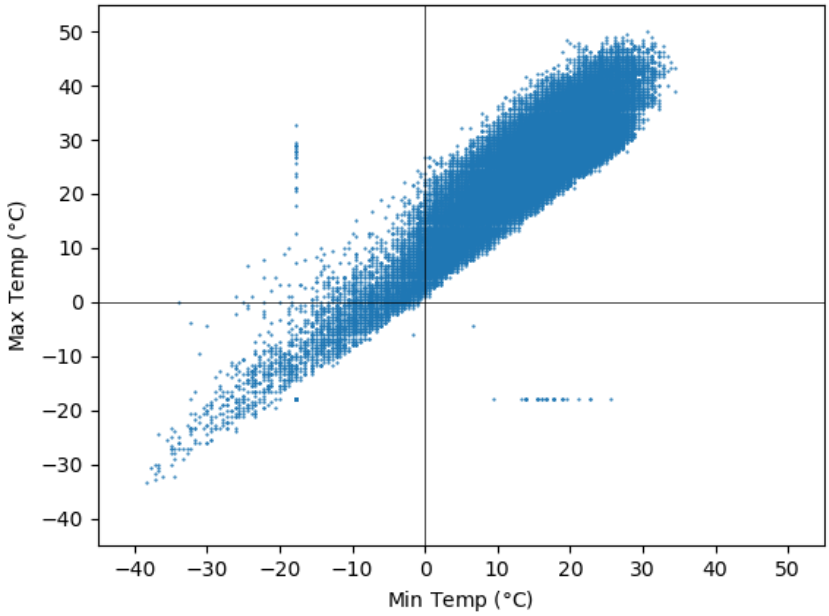

Figure 1: Temperature data for Task 1. Taken from continental Europe during the mid-19th century.

Task 2 – Calculations

This task is about making calculations by hand. In practice, you won't do these by hand – you'll instead use a library method. But it's still useful to do these things once yourself, so you understand what those library methods are doing. As additional motivation, you should know that these types of questions are great to ask in exams as well!

You can use a calculator to help with calculations, but don't use programming/scripts to automate the work for you. Ideally, you should pretend you're in an exam, where you would have to show your work.

In this task, we'll look at some data showing student attendance for FDS tutorials during the semester. Note that the data is made up to make the computations easier, but still somewhat follows actual trends for the course. The data is the following 4 data points – (3, 65), (5, 50), (7, 30), (9, 15) – where the first number denotes the week of the semester, and the second number is attendance in percent. In other words, there was a 65% attendance rate for the workshop in week 3, etc.

1. Take these parameters: $\beta_0 = 75$, $\beta_1 = -5$. Compute the squared loss $f(\beta_0, \beta_1)$ for those parameters on the data using the standard Linear Regression model $y = \beta_0 + \beta_1 x$.
2. With the same parameters, compute the coefficient of determination R^2.
3. Take a look at the residual plot for the parameters (Figure 2, next page). What does it tell you?

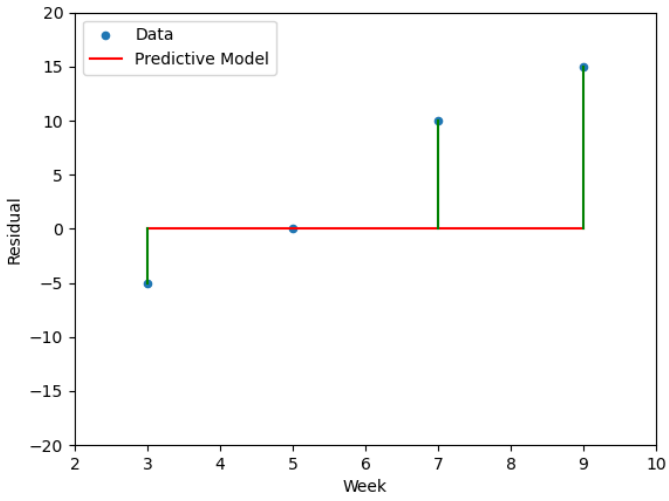

Figure 2: Residual plot for the parameters $\beta_0 = 75$, $\beta_1 = -5$ given in step 1 of Task 2.

4. With all the above info, are the parameters we used good?
5. Compute the actual optimal parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ using the formula from the lectures.
6. Compute the loss and R^2 with these optimal parameters.
7. In Figure 3 you can find a fit to the data and the residual plot with the optimal parameters. Does the fit look better than before?
8. If the semester had a week 11 and we had a workshop that week, what would your model predict for that workshop's attendance? Does it make sense? Should you try to fix it, and if yes - how?
9. If we changed the initial data from percent to fractions (so 0.65, 0.5, 0.4, 0.15), how would that change all the above calculations? (You don't have to recompute anything, but consider how it would change the answers.)

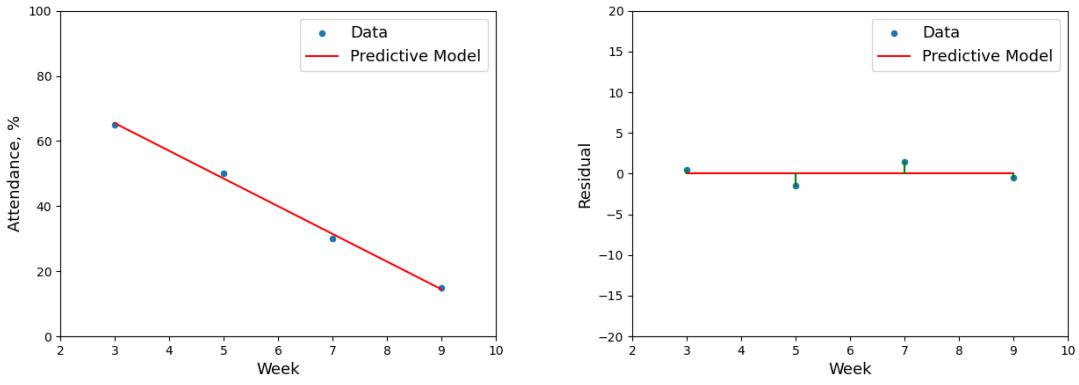

Figure 3: Model fit (left) and residuals (right) for the optimal parameters computed in step 5 of Task 2.

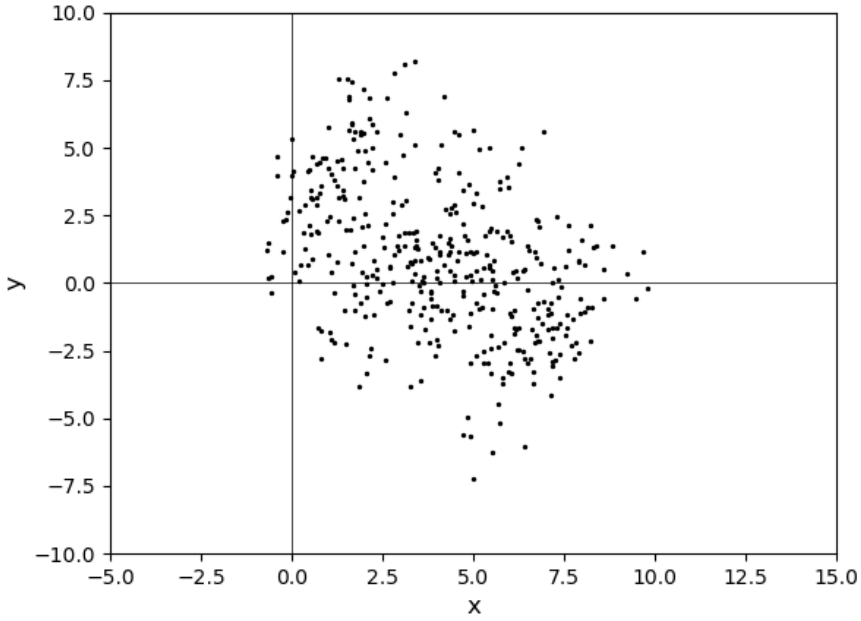

Figure 4: Data with arbitrary labels x and y.

Task 3 – Simpson's Paradox

In this task, we'll talk about a phenomenon called Simpson's paradox that appears in real data. We'll look at a toy example of it first, and then consider how it can explain a real-world data science example.

1. Assume we have some two-dimensional (or bivariate) data (x_i, y_i), $i = 1...n$. The data is shown in Figure 4 above. What prediction do you think a linear regression model with independent variable x and dependent variable y would make? In other words, what would the model fit look like?
2. Now look at Figure 5 on the next page (and try to avoid looking at Figure 6 for now). It turns out that there is an extra factor within the data - each data point comes from one of 4 different groups, which are coloured differently in Figure 5. Looking at this, does your proposed fit from the previous step still make sense?
3. How would you use linear regression as taught in the lectures to capture the data better? There are multiple possible ways to do it, so try to come up with alternatives.
4. Look at Figure 6 - it shows separate linear regression fits for each group of data (and also one fit for the full aggregated data). Simpson's paradox is a phenomenon in which a trend appears in several groups of data but disappears or reverses when the groups are combined. Generally, it happens when there are confounding variables in play (in this case, the groups are the confounding variables).

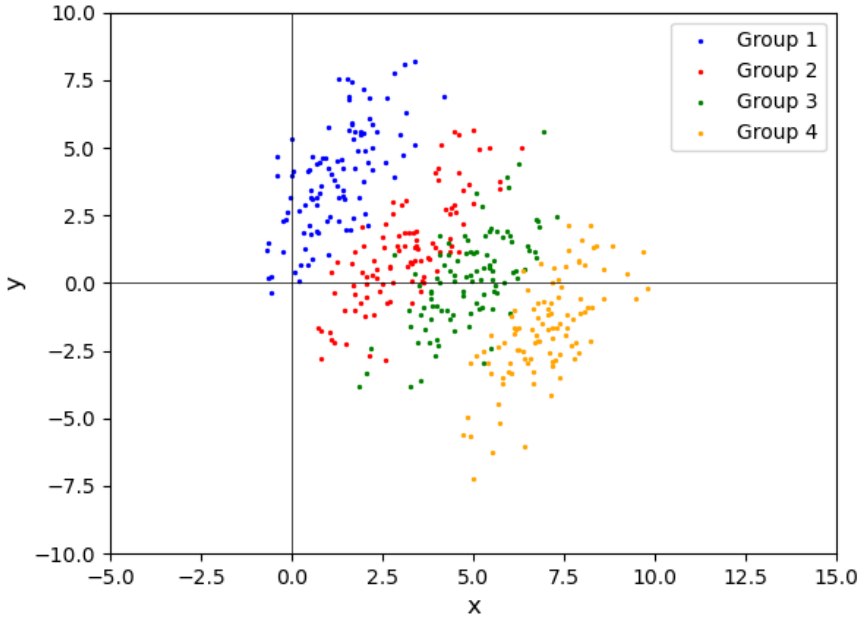

Figure 5: The same data from Figure 4, but this time with the four underlying groups of the data shown explicitly.

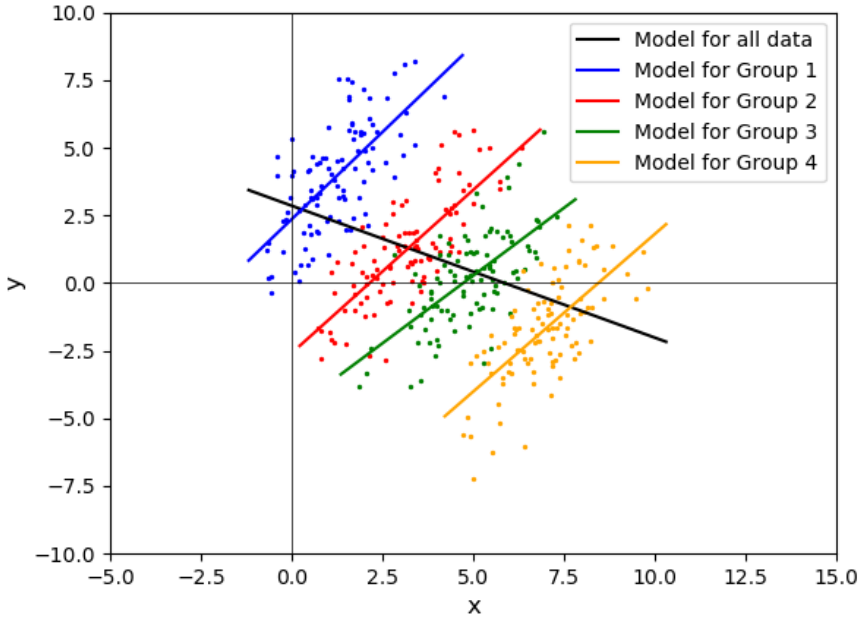

Figure 6: Linear regression fits for the same data from Figure 5. The black line is the model fitted on the full dataset, and the coloured lines are models fitted on the corresponding group of data.

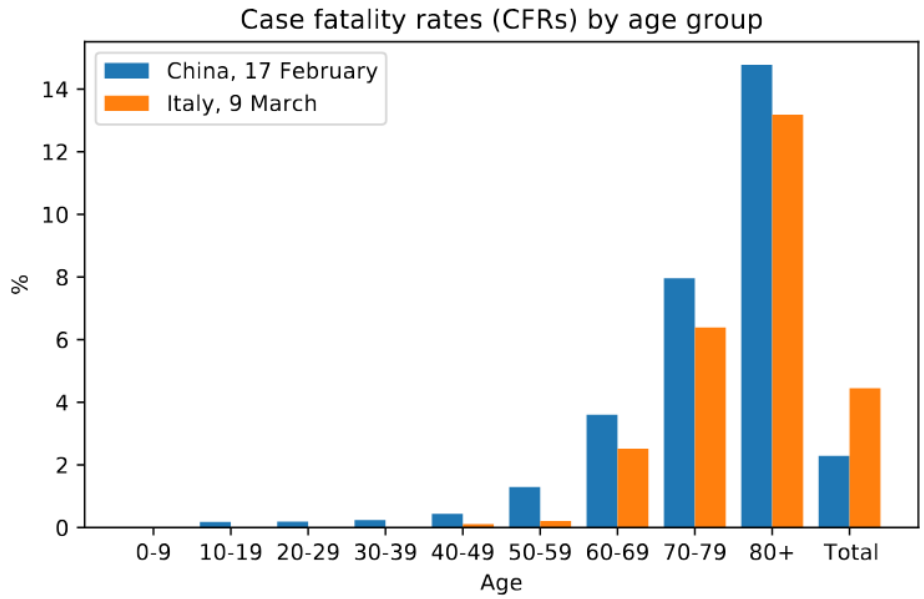

Figure 7: Covid-19 case fatality rates (CFRs) in Italy and China by age group and in aggregated form ("Total"), i.e., incl. all confirmed cases and fatalities up to the time of reporting (see legend). Figure taken from https://arxiv.org/pdf/2005.07180.

5. Now let's see an example of Simpson's Paradox in practice. Take a look at Figure 7. The image shows Covid-19 fatality rates in Italy and China (up to 2021 March 9 and 2021 Feb 17 respectively) as a percentage. It's shown both as a total and for different age groups.

   You can see a counterintuitive pattern in the figure: *for all age groups, CFRs in Italy are lower than those in China, but the total CFR in Italy is higher than that in China.*

   As a group, try to reason why this could've happened. (Hint: it's not because of an issue with the data collection.) You obviously don't have the full context, but you can make suppositions and guesses based on the information provided.

   Important! Do not go to the next step before you've made a guess as to why the data may look the way it does. Ask for hints from the tutors if you have to. In the real world, when given a dataset, you'll likely not just have the answer already prepared for you, so you have to be able to reason about these things yourselves.

6. Take a look at the source of Figure 7, the paper that collated and analysed the data: von Kügelgen, Gresele, and Schölkopf: "Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects.", published in IEEE transactions on artificial intelligence 2.1 (2021): 18-27.

   Find their explanation for the curious pattern. Does it make sense? Does it match what you were thinking about?

Task 4 – Reading and understanding a data science article

In this task, we will look at an existing data science article and try to understand as much as we can from it. This is to help with Learning Outcome 4, which states that at the end of the course you should be able to *"critically evaluate data-driven methods and claims from case studies, in order to identify and discuss a) potential ethical issues and b) the extent to which stated conclusions are warranted given evidence provided."*

The article we will look it is from Significance magazine, and can be found here. The magazine publishes articles that usually have some technical information in them, but they are generally not as technical or complicated as published research papers. We'll refrain from providing more info about the article - you should be able to understand most of what the article is saying.

It's important to note that in practice, even professional researchers don't fully read every paper that they open or end up using. That's not to say the detail in it isn't important, but being able to quickly pick out crucial information from a piece of technical writing is an important and useful skill to develop. As such, you should focus on answering the following questions in your group:

1. Question: What is/are the specific question(s) that the article tries to answer?
2. Context: What dataset(s) does it use? How were the datasets collected? Which methods were used to analyse the problem?
3. Correctness: Do the assumptions appear to be valid?
4. Model: What model(s) is the paper using? What variables does the model(s) use?
5. Clarity: Is the paper well written? Is there anything you would do to change it?