

# Homework 4

SAAS DF Spring 2025

Due Monday, March 24th, at 11:59 PM

As always, you are only expected to complete problems marked with a (\*). However, if you want to really grow and flourish as a data scientist, we recommend you attempt the unmarked challenge questions, as well.

## 1 The Wild, Wild West of Regularization (\*)

*Estimated completion time: 60 mins*

- 
- (a) Head over to `regularization.ipynb` and complete the exercises! (\*)
  - (b) **DM/DC Interview Prep:** Without fail, the DM and DC interviews will ask you what overfitting and underfitting are and how to prevent them, and they expect you to come up with a very thorough, clear answer. In 5 to 10 sentences, answer that question in a written paragraph. For some tips on how to structure a "good" interview answer, see the hints section! (\*)
- 

## 2 We Have Time (\*)

*Estimated completion time: 15 mins*

- 
- (a) Head over to `timeseries.ipynb` and complete the written exercises! (\*)
  - (b) Imagine you sell ice cream over the summer, and you would like a way to forecast future ice cream sales volume. That way, you know when you need to stock up on the most popular flavors to meet customer demand.

You notice that in years where you have really bad pollen allergies in the spring, ice cream sales are always higher. You fit a regression model of summer ice cream sales on spring pollen volume and find you are able to forecast future sales within 1% prediction error!

You tell your classmate about your very accurate time series forecasting model when you get back to school, but your classmate scoffs at you, saying, "That's not a good model. Temperature is the confounding variable. When it's hotter outside, more flowers are pollinating, causing more pollen and worse allergies, and when it's hotter in spring, it's probably even hotter in summer, causing more ice cream sales. Your model is meaningless; pollen doesn't cause ice cream sales!"

Is your classmate right? Is your model bad because it's not causal? Why or why not? (\*)

---

## 3 Penguins (\*)

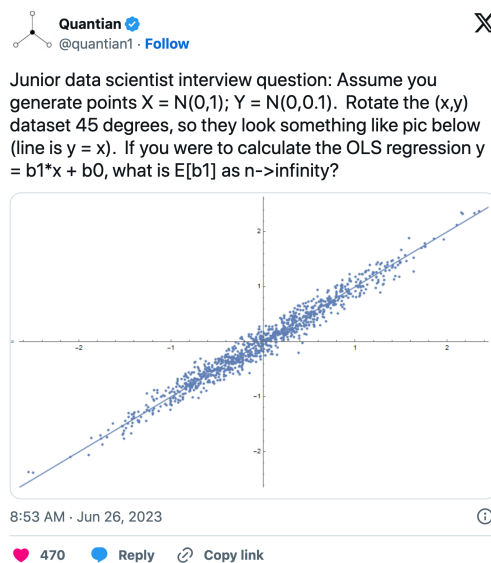
*Estimated completion time: 60 mins*

- 
- (a) Head over to `penguins.ipynb` and complete the exercises! (\*)
- (b) **Reflection:** What did you learn from the assignment? What are you most confident on in your understanding of regression? What are you least confident on in your understanding of regression? (\*)
- 

## 4 A Real Life Data Science Interview: Part 1 (\*)

*Estimated completion time: 30 mins*

Two years ago, the Twitter user @Quantian1 posted the following interview question that they ask to prospective junior data scientists:



This question stirred up a lot of controversy on Twitter, with people going back and forth on whether this is a fair question to ask during a junior-level interview. (I think Twitter user @ryxcommar hit the nail on the head with his take on the matter over at his blog.)

Regardless of your thoughts on whether this is a fair question or not, these are the kinds of technical questions data science interviewers may expect you to be able to answer, and more importantly, the question reflects the importance of actually *understanding* the models we work with rather than just plugging and chugging with scikit-learn. So, it's good to get some practice building up the intuition on how to solve problems like this.

**Note:** Before proceeding to this section's questions, I would recommend reading through the questions in the next section even if you do not attempt any of them. You will need to know what variance and covariance are for this section.

**Note:** I would recommend (c)-(f) in Part 2 of this question as a really worthwhile challenge if you like math and want to develop your technical rigor as a data scientist!

- 
- (a) One of the central aspects of this problem is "rotating the dataset" by 45 degrees. Write down the  $2 \times 2$  rotation matrix in terms of an arbitrary angle  $\theta$ , and then solve for the values of the rotation matrix for when  $\theta = 45$  degrees. (\*)

**Hint:** Refer to the Week 1 lecture slides if you need a refresher on rotation matrices!

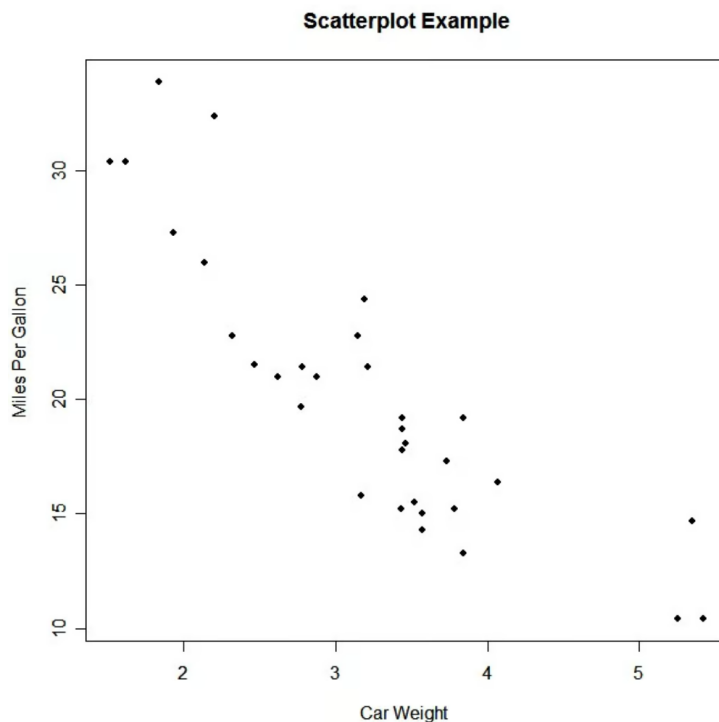
- (b) Head over to `interview.ipynb` and work through the notebook exercises! (\*)
-

## 5 Simple Linear Regression: The Least Squares Estimator

*Estimated completion time: 45 mins*

**Note:** Even if you do not want to attempt this optional section, I would recommend reading the questions at the very least to get an idea of important properties of linear models.

Recall the least squares problem from lecture. Basically, we are trying to find the best **linear** approximation to some given data, like in the scatterplot below.



We usually never find a linear fit to the data that *exactly* passes through all the data in a perfectly-determined linear relationship. Because of that, we define the **least squares problem** that estimates the best intercept and slope parameters that minimize the **mean squared error** between the data points and the line of best fit, i.e.:

$$\text{Find } \hat{\alpha}, \hat{\beta}, \text{ that minimize } \frac{1}{n} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

- (a) Solve the minimization problem above for **just the intercept term**  $\hat{\alpha}$ . Take the partial derivative of the mean squared error with respect to  $\hat{\alpha}$ , set the derivative equal to 0, and solve for  $\hat{\alpha}$ . You may keep your answer in terms of  $\hat{\beta}$  and the sample means of  $x$  and  $y$ , i.e.  $\frac{1}{n} \sum_{i=1}^n y_i = \bar{y}$  and  $\frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$ .
- (b) **Show** that the line of best fit will always contain the mean value of both  $x$  and  $y$  by rearranging your solution for  $\hat{\alpha}$  in part (a).
- (c) A common data transformation is **standardizing** your predictor variable  $x$  and target variable  $y$  by subtracting the mean from each variable and dividing by the standard deviation, i.e.  $x' = \frac{x - \bar{x}}{\text{SD}(x)}$  and  $y' = \frac{y - \bar{y}}{\text{SD}(y)}$ . Standardized  $x', y'$  will have mean 0 and standard deviation 1. **Show** that if you standardize your variables  $x$  and  $y$ , then the intercept term will be at the origin  $(0, 0)$ .
- (d) We define the **covariance** between two variables  $x$  and  $y$  as a measure of how much those two variables "move together." Based on this notion of covariance, we can define the **correlation** between two variables  $x$  and  $y$  as a kind of "standardized" version of covariance bounded between -1 and 1 (covariance, on the other hand, is unbounded). Thus, we can say:

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\text{SD}(x)\text{SD}(y)}$$

One way to define the least squares estimate for slope parameter  $\beta$  is  $\hat{\beta} = \text{Corr}(x, y) \frac{\text{SD}(y)}{\text{SD}(x)}$ . That is to say, for every 1 standard deviation increase in  $x$ , our model estimates that  $y$  will increase by 1 standard deviation, scaled by the correlation coefficient. (When  $x$  and  $y$  are standardized, the correlation coefficient *is* the slope.)

Now, express  $\hat{\beta}$  in terms of just  $\text{Cov}(x, y)$ . (You should not have a term for  $\text{Corr}(x, y)$  in your answer.)

- (e) **Challenge:** Rederive your solution above in part (d) by solving the least squares minimization problem for  $\hat{\beta}$ . As a hint, the sample covariance of  $x, y$ , is defined as  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  and the sample variance of a random variable  $V$  is defined as  $\frac{1}{n} \sum_{i=1}^n (v_i - \bar{v})^2$

**Note:** This problem is quite challenging. If you find yourself spending longer than you can afford to on it, feel free to skip it! If you get stuck, you can Google the proof to this problem as well to help you guide your solution. (In reality, the most challenging part about this problem is the algebraic manipulations, which require a bit of creativity.)

## 6 A Real Life Data Science Interview: Part 2

*Estimated completion time: 30 mins*

**Note:** Before proceeding to this section's questions, I would recommend reading through the questions in the previous section even if you do not attempt any of them. You will need to know what variance and covariance are for this section.

- (c) Hopefully the notebook gave you some linear algebra intuition of how to approach the problem. We're now going to reconstruct our computational steps so we can solve the problem mathematically!

From the notebook, you may have noticed that after defining our rotation matrix, we applied the linear transformation to each pair of data points  $(x, y)$ . In fact, each data point can be represented by a vector  $\begin{bmatrix} x \\ y \end{bmatrix}$ . Multiply the  $2 \times 2$  rotation matrix by this arbitrary vector and write down the solution to the matrix-vector product.

- (d) Now that you have a new  $2 \times 1$  vector containing the transformed values of your original data  $(x, y)$ , calculate the **variance** of the transformed  $x$ -values (this is the first element in the vector).
- (e) Calculate the **covariance** between the rotated values of the original  $(x, y)$  data points that you solved for in part (c). Some helpful properties of the covariance function are below in the hints section.
- (f) Lastly, compute the ratio between your answers in part (e) and part (d). That is to say, compute the rotated estimate of the slope coefficient  $b'_1 = \frac{\text{cov}(X', Y')}{\text{var}(X')}$ . Your answer should be around 0.98.

## 7 Hints

**Remark (The Wild, Wild West of Regularization).** Here are some tips on how to structure a "good" interview answer:

- A 3/5 answer would maybe just explain what overfitting and underfitting are, but not offer any concrete solutions to combat them. Or, the only solution offered is "get better training data," which is a cop-out answer (and may also not be feasible in a real-world setting).
- A 4/5 answer would both explain what overfitting and underfitting are, and offer 1 strategy on how to combat them but may not elaborate on how that 1 strategy actually works.
- A 5/5 answer would explain what overfitting and underfitting are, offer multiple strategies on how to combat them, and explain all those strategies in clear and concise detail. (Bonus points if you mention the bias-variance tradeoff. A 6/5 answer might even talk about double descent...)

**Remark (A Real Life Data Science Interview: Part 2).** Here are some hints for part (d):

- **Hint:** For a random variable  $V$  and scalar  $c$ , we define the variance  $\text{var}(cV) = c^2 \text{var}(V)$
- **Hint:** For two random variables  $U$  and  $V$ , we define the variance of their difference as  $\text{var}(U - V) = \text{var}(U) + \text{var}(V) - 2\text{Cov}(U, V)$
- **Hint:** Remember that covariance is an un-standardized version of correlation. When two random variables are generated independently (like  $X$  and  $Y$  in the problem), their correlation is 0 (examine the plot you generated in `interview.ipynb` to confirm this). If two random variables are uncorrelated, then their covariance is also equal to 0.

**Remark (A Real Life Data Science Interview: Part 2).** Here are some hints for part (e):

- **Hint:**  $\text{cov}(aX, bY) = ab * \text{cov}(X, Y)$
- **Hint:**  $\text{cov}(aX+bY, cW+dV) = ac*\text{cov}(X, W)+ad*\text{cov}(X, V)+bc*\text{cov}(Y, W)+bd*\text{cov}(Y, V)$
- **Hint:** Remember that  $\text{cov}(X, X) = \text{var}(X)$  and that  $\text{cov}(X, Y) = \text{cov}(Y, X) = 0$  if  $X, Y$  are independent/uncorrelated.