Data Analysis – BootCamp B4: Regularization

Collaborators:

1. You build a model with two hyperparameters λ and γ. You have 4 good candidate values for λ and 3 possible values for γ, and you are wondering which λ, γ pair will be the best choice. If you were to perform five-fold cross-validation, how many validation errors would you need to calculate?(Exam May 2022)

5 x 4 x 3 = 60 errors

2. In the typical setup of k-fold cross validation, we use a different parameter value on each fold, compute the mean squared error of each fold and choose the parameter whose fold has the lowest loss.

   True or False? FALSE (parameter needs to be the same in each fold)

3. Which of the following are NOT reasonable loss functions? Assume that we can find the optimal parameters for each loss function. Why? (exam April 2022 question)

A $y - \hat{y}$

B $\dfrac{1}{y - \hat{y}}$

C $(y - \hat{y})^2$

D $e^{-(y-\hat{y})^2}$

Since we minimize the loss function to find the optimal model parameters, a loss function must output higher values when the prediction y^ is far away from y and lower values when y^ is close to y.

Option D does not have this property, so we exclude it.

Choice A also is not reasonable, because the optimal y^ is infinite, since we are minimizing the loss function

Choice B is more complicated. The optimal y^ would be very slightly above y, as it would lead to a large negative loss. But if y^ is ever-so-slightly less than y, we have very large positive loss, and if y = y^, which would be a perfect prediction, the loss is undefined, so this is an unreasonable loss function.

4. (Exam April 2023)
   A random sample of 76 apartments is collected near Imaginary University. All of the apartments in the sample have between 1 and 6 bedrooms. The variables recorded for each apartment are Rent (in euros) and the number of Bedrooms. The regression output is:

The dependent variable is Rent
R squared = 62.0% R squared (adjusted) = 61.5%

Variable	Coeff	SE(Coeff)	t-ratio	p-value
Interecpt	357.795	111.6	3.2	0.0020
Bedrooms	400.554	36.42 1	1.0	< 0.0001

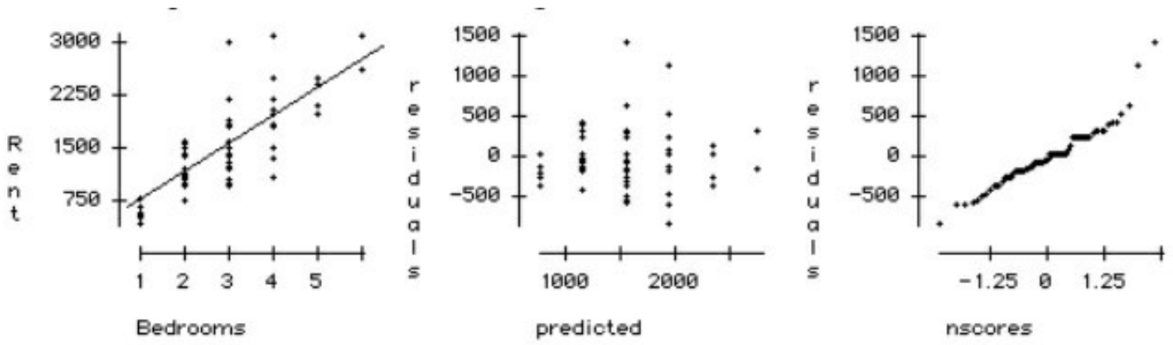
A. The coefficient value for the number of bedrooms (which is approx. 400.6) appears to be too large. A friend of yours suggests that there is something wrong with the model. Do you agree? Justify your answer shortly.

The value of the coefficient depends on the units of the actual variable. In this case #bedrooms is ranging from 1 to 6 and the output variable is rent price (in euros). Therefore, a large coefficient is needed for getting from the independent variable (1-6) to the dependent variable (which is expected to be in the range of hundreds)

B. Explain the meaning of the intercept in the context of this problem and comment on whether you can use it for actual decisions based on the model. Explain shortly your answer.

The intercept, aka 357.795, would be the price of a house with 0 bedrooms (aka a studio). Given that the dataset we are given mentions that the apartments in the sample have 1-6 bedrooms, we actually have no evidence to do the "interpolation", therefore we cannot use it for making any informed decisions

C. On top of the model above, we are also given the following plots (from left to right): (a) the scatterplot with the best line fit, (b) the plot of residuals vs. the predicted values and (c) is the normal probability plot (check online or ask Jerry what does it show for the normality of the residuals) of the residuals. Given the model details (as given in the introduction) and these three plots comment on the fit of the model. Explain your answer shortly.



You have the model fit results and 3 plots. Things you could say:

1)R^2 is 62% which is moderate and the coefficient seems significant.
2) Neither the plot of Rent vs. Bedrooms nor the plot of Residuals vs. Predicted show any evidence of curvature.
3) The Residual vs. Predicted plot appears to thicken in the middle, but this is partly due to two large outliers. Ignoring the outliers, the spread is similar throughout the plot.
4) Random residuals: The problem stated that the apartments were a random sample of apartments around the university. Seems okay.
5) The Normal probability plot of the residuals is fairly straight except for the two outliers.