1. What are the differences of nonparametric and parametric regressions? Give one example for a nonparametric regression and one for a parametric regression. Give an argument for each kind of method. (5p

Nonparametric regressions don’t impose a functional form on the pattern of association between the variables, regressions with parameters do specify an explicit shape for the function. Locally weighted scatterplot smoothing (lowess/loess) is nonparametric, parametric example could be a simple linear regression. Nonparametric regressions are good for finding out what kind of association there is between our variables, however, it can only be used for interpolation, not for extrapolation. In addition, the slope coefficient has to be calculated for each observation even for interpolation, whereas for example simple linear regression gives 1 slope parameter for the association that is easy to interpret.

Graphical user interface, application

Description automatically generated

Y: log hotel price

X: distance from city center

Constant = 5 means that hotels with 0 distance from city center would have a price of log 5 (in whatever unit we have price in). This is not very useful since usually there are no hotels exactly in the city center and log price should be converted back to absolute units for clarity.

Distance = -0.2 means that for 1 km (if km is the unit we have distance in our sample) larger distance from center prices are 20% lower on avg.

The coefficients are statistically significant (CI doesn’t include 0) so they can be generalized to hotels in 2017 weekend night in the specific city within 5km from its center.

R2 = 0.291 means that predicted values are capturing some of the variation of the actual values, but we are far from a perfect fit. If we are not aiming for prediction in this model, this is not a big concern, the model could explain a linear relationship between the variables well.

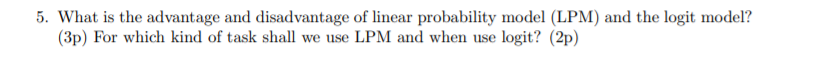
Text

Description automatically generated

95% CI would be [-0.17,-0.23] meaning that in the population hotels with the 1 km larger distance from the center are expected to be between 17 and 23% cheaper with 95% confidence. 80% CI would be narrower as we would produce a less precise generalization with more confidence in it.

A white text with black text

Description automatically generated



N = 100000 Y: log income of families X: country (1 = CAN 0 = USA) beta = 0.07

A: Families that live in Canada are expected to have 7% higher income than families living in USA on avg. in the sample.

B: Z: having children (0 = no 1 =yes)

Beta1 = 0.15 means that families living in Canada are expected to have 15% higher income than families living in US on avg. in the sample when comparing families that have the same status regarding having kids i.e. if we only compare families with no kids or only families with kids.

C: We don’t know. Beta2 = -0.1 means that families with kids earn on avg. 10% less when comparing families living in the US. However, we cannot conclude casual effect based on this, this is only the association in our sample.

D: Beta3 = 0.2 tells us that the gap between avg. earning is opposite in Canada vs US, see below.

E: Beta1

Y: log income of families

X1: country (0 = USA 1 = CAN)

X2: kids (0 = NO 1 = YES)

Beta 1 = 0.08 –- Families with No kids in CAN earn avg. 8% more than families in US .

Beta 2 = -0.1 –- Families with kids earn 10% less than families with no kids in US.

Beta 1 + 3 = 0.28 –- Families with kids in CAN earn avg. 28% more than families in US.

Beta 2 + 3 = 0.1 –- Families with kids earn 10% more than families with no kids in Can.

Text

Description automatically generated

Y: avg. afternoon temp ©

X: log tourists airport

Beta = -0.05 means that on avg. for 1% more tourists landing a day the avg. afternoon temp decreases 0.0005 Celsius. However, this coefficient doesn’t account for seasonality, trend. We know that global avg. temperatures rise while number of tourists landing in Madrid may be decreasing in our sample (especially if COVID years are included at the end) showing a negative trend between the two variables. B, When accounting for seasonality, it’s only 0.0001 Celsius in same month.

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

1 A

2 B

3 C – false positive == rejecting the null incorrectly

4 A

Y: smoking (binary)

X1: age

X2: gender (0 male 1 female)

Interaction

Beta1 = -0.01 1 year is 1% less likely for males

Beta2 = -0.15 females are 15% less likely at age 0

Beta3 = 0.03 1 year is 2% more likely for females and females are 12% less likely

5 D