PS6

BSS Stat 20

2022-08-01

NOTICE

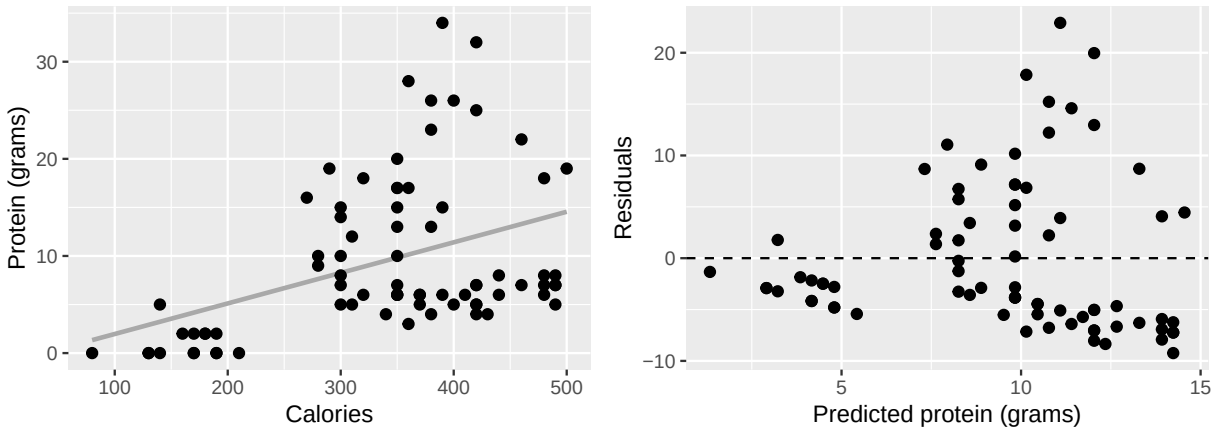
This problem set is worth zero points; you do not have to turn it in. However, if you want to assess yourself fairly, go ahead and type your answers in an RMarkdown file, knit and save to an html or pdf. Feel free to collaborate or use any kind of help while completing this assignment.

This notice will be pasted at the beginning of every problem set.

Questions

Question 1

The scatterplot below shows the relationship between the number of calories and amount of protein (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we might be interested in predicting the amount of protein a menu item has based on its calorie content.



part a

Describe the relationship between number of calories and amount of protein (in grams) that Starbucks food menu items contain.

part b

In this scenario, what are the predictor and outcome variables?

part c

Why might we want to fit a regression line to these data?

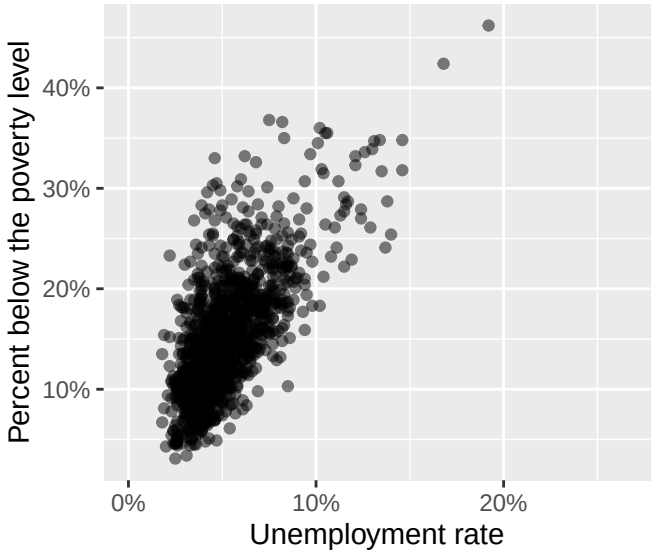
part d

What does the residuals vs. predicted plot tell us about the variability in our prediction errors based on this model for items with lower vs. higher predicted protein?

Question 2

The following scatterplot shows the relationship between percent of population below the poverty level (`poverty`) from unemployment rate among those ages 20-64 (`unemployment_rate`) in counties in the US, as provided by data from the 2019 American Community Survey.

The regression output for the model for predicting `poverty` from `unemployment_rate` is also provided.



term	estimate	std.error	statistic	p.value
`(Intercept)`	4.604	0.349	13.182	<0.0001
`unemployment_rate`	2.054	0.062	33.110	<0.0001

part a

Write out the linear model.

part b

Interpret the intercept.

part c

Interpret the slope.

part d

For this model R^2 is 46%. Interpret this value.

part e

Calculate the correlation coefficient.

Question 3 - True or False

No credit will be given without an explanation.

part a

It's always helpful to add more predictors to your linear model, since R^2 will always increase.

part b

Consider using a linear model to predict Y using the variable X veruss using a linear model to predict Y using variables X *and* Z. The interpretations of the coefficient of X in the two cases should not be the same because of the inclusion of Z in the latter case.

part c

There is the potential for two different people fitting the same k-nearest neighbors algorithm to a set of data to come out with different predictions.

part d

I am considering modeling the heights of students at UC Berkeley using linear regression and the heights of those students' parents as a predictor. Because the heights of students at UC Berkeley are likely normally distributed, I should transform them first before using them as an outcome variable in the regression.

part e

Say we are fitting k-nearest neighbors to a set of data with an outcome variable (Class A or Class B) and two predictors. When we decrease k, we should expect that it will become easier to eyeball which combinations of the two variables will result in a prediction for either Class A or Class B.

Question 4

Now, consider the `penguins` data-set included the `palmerpenguins` package in R.

All of your steps here should be reproducible (when applicable).

part a

What is the unit of observation in this dataset? (Consult the help file if you need to).

part b

Using `sample()`, generate a random number between 1 and the number of rows in the dataset and save it into an object.

part c

Print out the row of the dataset given by the number you drew in **part b**, and save the flipper length value (`flipper_length_mm`) into an object. Identify the `Species` value.

part d

Using *only* `dplyr` *commands*, predict the `Species` value in the first row with `flipper_length_mm` as a predictor variable using k-nearest neighbors with $k = 10$. (You do not have to use the commands in a single pipeline).

If there are ties between classes, code up a way to randomly break them.

part e

Was your row classified correctly?