BTRY 6020: Module 1

Your Name

Question 1 (3 pts)

We discussed the conceptual framework of a population process of interest, the data that is generated from that process, and a statistic which is calculated from the data. Briefly describe what that might look like in your field of interest. (i.e., 1 to 3 sentences total). Describe a population/process that you are interested in learning about, what type of data might be gathered from that system, and what statistic you might calculate to summarize your data.

Answer

One example would be learning about the distribution of scores in a particular statistics class at Cornell for all time. We could obtain all of the overall scores for the class from as many years as we have access to. We could calculate the sample mean, median and mode to gain an idea of the average mark in the class. We could also calculate the sample variance to have an understanding of the spread of the scores.

Question 2 (2 pts)

In Lecture 1, we discussed various tradeoffs in statistical analysis: bias vs variance, model complexity vs interpretation, false positives vs false negatives. Briefly discuss (i.e., 1 to 3 sentences total) a situation in your field of interest where one of those trade-offs might occur.

Answer

One might be interested in the real estate market and studying how the price of a home is associated with certain characteristics of the home. If we fit a model for predicting the price of a home, we can include 1000s of covariates on the home and get very precise estimates. However, that model may be difficult to explain to a home buyer who wants to know why their offer price should be at a specific value or why a house they are interested in is so expensive. Thus, fitting a less complicated model may be easier to explain and there is a trade-off between complexity and interpretation.

Question 3 (1 pt)

Go to the website https://guessthecorrelation.com/, and guess the correlation for at least 10 plots. Record your guess and the true correlation for the last plot below.

Answer

Guess:

Truth:

Question 4 (2 pt)

The following data is on the spring snow depth on Mt. Rainier from 1920 - 2013.

- year: year of snow depth measurement (1920-2013)

- snow: maximum snow depth in cm measured in April at 1500m elevation at Paradise Ranger Station on Mt. Rainier
- temp: mean temperature the previous winter (Nov. - Apr. mean, in C)
- prec: mean precipitation the previous winter (liquid water equivalent, in cm)
- nino: Index of El Nino Southern Oscillation (ENSO), (sea surface temperature anomaly (C) from the long term seasonal mean observed in the Nino 3.4 region of the Central Pacific (basically the middle of the Pacific on the equator)

Fit a linear model with snow depth as the dependent variable and the El Nino index as the independent variable. What are the estimated intercept and slope?

```r
snow_data <-
  read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/snow_data.csv")

names(snow_data)
```

```
## [1] "year" "prec" "temp" "snow" "nino"
```

Answer

```r
mod1 <- lm(snow ~ nino, data = snow_data)
mod1
```

```
##
## Call:
## lm(formula = snow ~ nino, data = snow_data)
##
## Coefficients:
## (Intercept)         nino
##      484.23       -71.62
```

Question 5 (2 pt)

Give an interpretation for the estimated slope parameter which you could explain to a colleague.

Answer

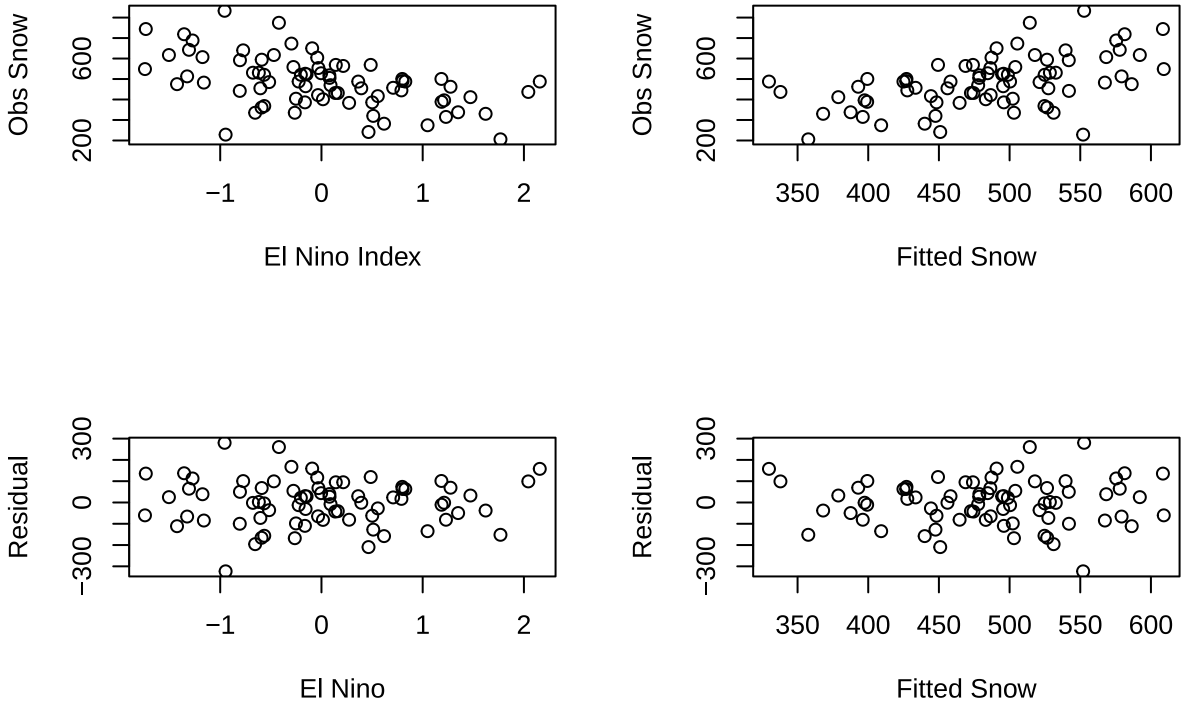
Given two years which differ in the El Nino index by 1 unit, we would expect the year with a higher El Nino index to have a snow depth which is 71.62 cm lower than the year with the lower El Nino index.

Question 6 (3 pt)

Does the linearity assumption seem to hold for the data? Explain why or why not and include a plot in your explanation

Answer

In this case, we have only 1 covariate, so we can simply look at a scatter plot of the data and see that a linear trend seems to hold. We also see that the predicted snow and the observed snow values seem roughly linear. Finally, we see from the bottom two plots that there is not a clear association between the residuals and either the El Nino index or the predicted snow values. Thus, we conclude that the linearity assumption is reasonable for this data.

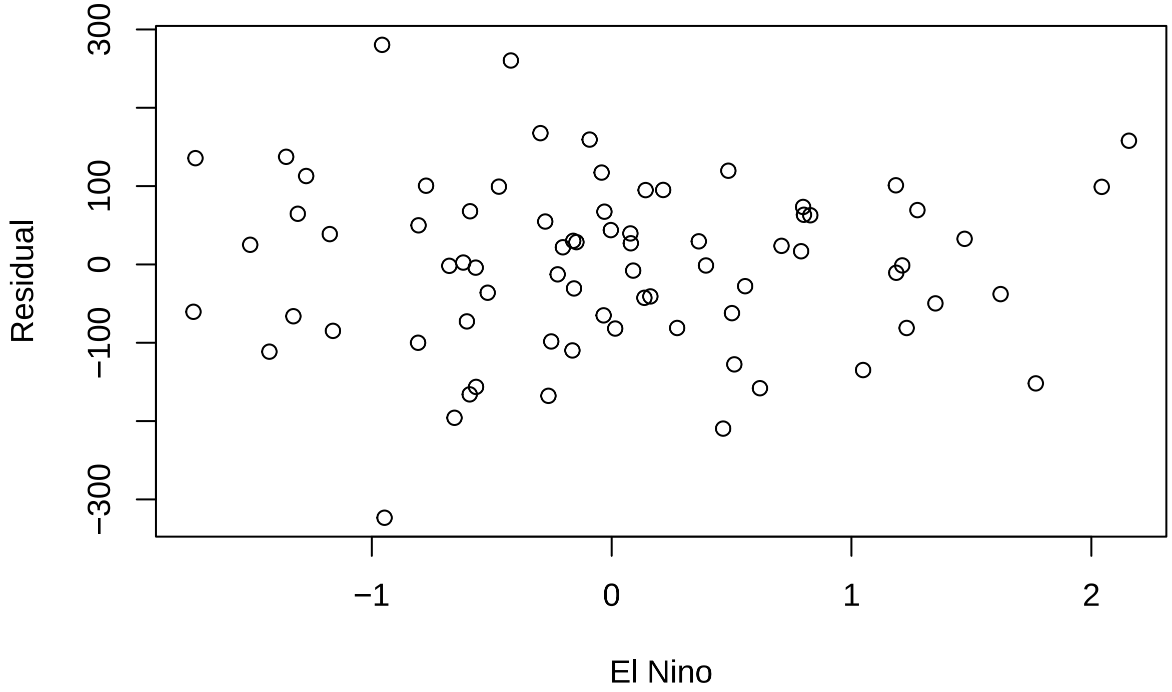


Question 7 (3 pt)

Does it seem like the variance of errors are constant across the range of the covariate? Explain why or why not and include a plot in your explanation

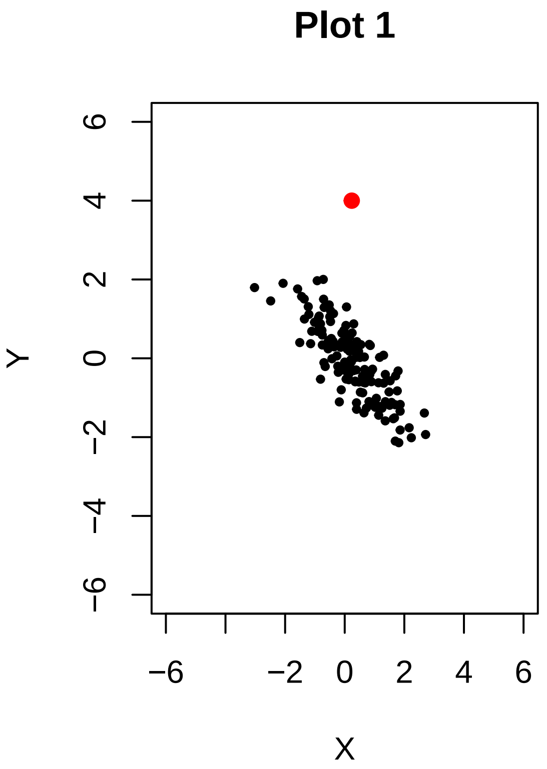
Answer

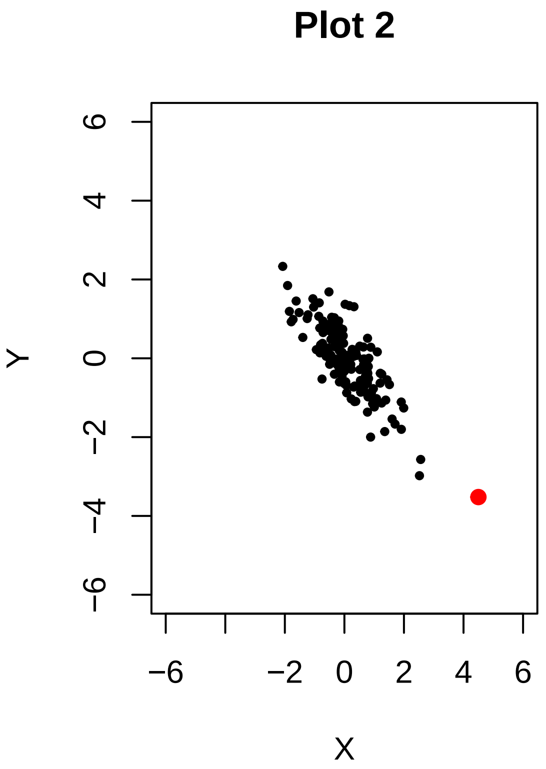
Looking across the range of the covariate, it seems that the variance of the errors is reasonably constant. There is not a huge difference at either end of the range.



Question 8 (2 pt)

In the following plots, say whether the red dot has high/low leverage and high/low influence. Give a brief explanation for why (1 sentence each)





Answer

In Plot 1, the point is a clear outlier in the Y direction, but it is not an outlier in the X direction, so it has low leverage. The point is actuall positioned directly on the mean of X so including/excluding it will not change the estimated slope. Thus, it has low influence.

In Plot 2, the point is an outlier in the X direction. Thus, it has high leverage. However, it seems to be on the general trend expressed by the other points. Thus, including/excluding will not change the estimated slope significantly. Thus, it also has low influence.

Question 9 (3 pt)

In the lecture, we considered data which showed the price of a bottle of wine as well as the rating of the bottle of wine. A plot is shown in the last slide of Lecture 1. There are two points in the upper right hand corner of the plot that might be considered an outlier. Describe a question you might investigate with this data. For that question of interest, explain whether you would choose to include or exclude those points from your analysis.

Answer

Suppose a grocery store is interested in investigating the relationship between the percieved quality of a wine and how they price the wine. In this case, the $\approx$ \$300 wines is likely not representative of the population they are interested in since it would probably not be sold at the grocery store. Then, we should probably exclude those outliers from our analysis