Module 8 Notes

Justin

2018-10-11

Upcoming Assignments/Quizzes

Assignments	Open Time	Due Time
Module 6 Data Quiz	October 12th (1:00 am EST)	October 14th (11:55 pm EST)
Module 6 Conceptual Quiz	October 12th (1:00 am EST)	October 14th (11:55 pm EST)

Notes from Discussion Board/Office Hours

Statistical Power and the p-value

There were a few questions from the previous Module about Power and it related to p-values, α, β, and sample size. Here's a link to a nice explainer from Khan Academy, which is also embedded in the Module 6 notes.

How to tell if assumptions are met?

In the lectures this week, Dr. Baiser demonstrated some graphical approaches for determining whether the assumptions for linear models are met. These approaches are somewhat subjective, we know what patterns we should find if the assumptions are upheld, but unlike the p-value there is no computed statistic and/or threshold for defining if the assumption is met or not. As a result, these techniques are typically much better at tell us when a model assumption **does not** fit than when it does.

How does the amount of data (sample size) related to the assumption of Normality?

In the part of the lecture where Dr. Baiser is discussing assessing the assumption of Normality, he states that it is often difficult to determine if errors are normal distributed if we have only a few data points, and that if we have a lot of data points it doesn't really matter. The first part should make sense, it's difficult to spot a normal distribution with a few data points.

But the second point is less clear. Why doesn't it matter if the errors aren't normal distributed if I have a very large sample size? Does this mean that if I have a very large dataset I can still get valid slope estimates even if my error aren't normally distributed?

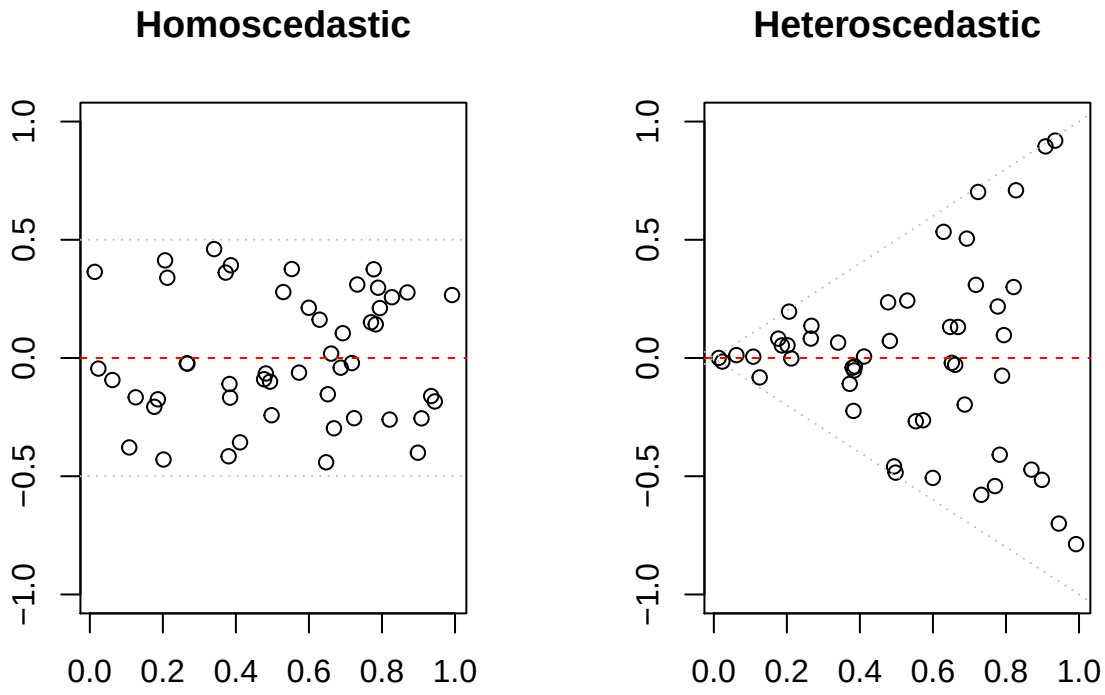
It (very) short answer is probably yes. This has to do with the Central limit theorem, which states that:

> when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

Homoscedasticity vs. Heteroscedasticity

Homoscedasticity is a another word for the homogeneity of variance. This means that the variation we see in our response variable Y is consistent across the range of our predictor variable X or our predicted response $\hat{Y}$. This means that that if we plot the residuals against predicted values, the distance between the points should be about the same over each section of the x-axis (or the *domain* of $\hat{Y}$).

If instead we see some sort of pattern, like higher differences in one section of the graph than in another section, this would be indicative of **heteroscedasticity**:



Again, as we dicussed earlier we're looking for patterns, and it will typically be easier to tell of something is heteroscedastic than so be sure that it is homoscedastic. One real-world example of heteroscedastic data could be height as a function of age. Height typicaly increases as age increase (up to a certain), however there is a larger degree of variance in age-specific height during the pubescent years than outside this age range. I looked all over for some actual data to show this but couldn't find any, if you find some email me and I'll update this page.

Other notes

- Dr. Valle and I try to keep an updated list of stats and programming courses across campus, available here..
- Don't worry, we did skip Module 7, we'll come back to it but the subject of that module is an special type of linear model, so we decided to do this module first.
- For transforming negatively skewed data, try a Fisher transformation.
- Remember that we are checking for normality *in the errors*, not necessarily in the data itself.
- My office hours have moved to Tuesdays from 6:30 to 7:30 pm EST.
- Recall from our earlier module that when we set graphing parameters use `par` R will keep these parameters moving forward. If you've made a multipanel plot and want to make new single panel plots it may be helpful to "reset"" the graphing parameters by running `par(mfrow = c(1,1))`.
- To access the `USairpollution` data for the data exercises, you will have to install and load the `HSAUR2`

```
install.packages("HSAUR2")
library(HSAUR2)
```

	SO2	temp	manu	popul	wind	precip	predays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55.0	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148