Outliers and Influence Diagnostics

Simple linear regression – Stat 230

Exploring outliers and influential points

Now, let's consider how outliers and influential points can impact the fitted regression line. To do this, we'll explore a data set containing the bid price and coupon rate of U.S. Treasury bonds. Here's a little background on Treasury bonds:

> US Treasury bonds are among the least risky investments, in terms of the likelihood of your receiving the promised payments. In addition to the primary market auctions by the Treasury, there is an active secondary market in which all outstanding issues can be traded. You would expect to see an increasing relationship between the coupon of the bond, which indicates the size of its periodic payment (twice a year), and the current selling price. The … data set of coupons and bid prices [are] for US Treasury bonds maturing between 1994 and 1998… The bid prices are listed per 'face value' of \$100 to be paid at maturity. Half of the coupon rate is paid every six months. For example, the first one listed pays \$3.50 (half of the 7% coupon rate) every six months until maturity, at which time it pays an additional \$100.[1]

To begin, load the data.

```
bonds <- read.csv("https://aloy.rbind.io/data/bonds.csv")
```

Task 1. Fit a simple linear regression model to predict the bid price given the coupon rate of U.S. Treasury bonds. Write down the fitted model equation.

Task 2. Create a scatterplot of bid price against the coupon rate and superimpose the fitted regression line. Do you see any potential outliers? If so, identify what rows of the data set correspond to these outliers. (Hint: Click on the name of the data set in the "Environment" tab to open a spreadsheet of the data.)

Task 3. Create a plot of the standardized residuals vs. the fitted values. Do you see any potential outliers? If so, identify what rows of the data set correspond to these outliers. Are they the same rows? (Hint: use the `residualPlot()` function from the {car} package with `type = "standard"` to create the plot. To calculate the residuals use `rstandard(fm)`, where `fm` is the name of the fitted regression model.)

Task 4. Create an index plot of the leverage values for each observation. You can do this from the augmented data frame, but a quicker way is to use the `infIndexPlot()` function in the {car} package. `infIndexPlot()` requires that you pass in the fitted regression model as well as the

[1] Siegel A (1997) Practical business statistics (3rd edn). Irwin McGraw-Hill, Boston.

influence diagnostic you want to plot. Use the argument `vars = "hat"` to specify leverage. Identify the high leverage cases (the row numbers are printed for "auto detected" high leverage points).

Task 5. Compare the cases (row numbers) you identified in Task 3 and Task 4. Which, if any, of these points to you expect to have high values of Cook's distance? Why?

Task 6. Create an index plot of the Cook's distance values for each observation. You can do this from the augmented data frame, but a quicker way is to use the `infIndexPlot()` function in the {car} package by setting `vars = "Cook"`. Are there any influential points identified by Cook's distance? If so, identify these points by their row number.

Task 7. Refit the model without the influential points you identified in Task 6. To do this, add a `subset` argument to `lm()` as seen below. Specify the row numbers in the blank (this creates a vector of row numbers to delete). How much did the regression coefficients change?

```
# Fill in the blank with row numbers to delete separated by commas
no_influential <- lm(BidPrice ~ CouponRate, data = bonds, subset = -c(___))
```

Task 8. Are we done? Check the residual plots and recreate the influence index plots you considered in Tasks 4 and 6. Do any additional points cause concern after your have already deleted a few points?

Should we delete the outliers/influential points?

As I discussed in the daily prep videos, it is not always wise to delete outliers and influential points. While this can improve the fit of your model, you might be deleting important information about your population. Of course, if a data entry error was made, then it should be fixed and the analysis should be rerun. In other situations, I recommend following the advice from the flow chart that Ramsey and Schafer include in *The Statistical Sleuth*.