Lab Notes

2017-4-28

Influential Observations

Outliers are sometimes called influential observations. We can say that an observation is influential if omitting it from the sample creates a substantial change in one of our parameter estimates.

But how do we identify an outlier? We can make an attempt to identify it visually, but we can also do so more systematically.

An observation with an unusual x value or an unusual y value does not necessarily mean it is influential. Moreover, an observation with an unusual x and y value does not necessarily mean it is influential. We can think of such observations that, when removed, do not create a substantial change in our estimate of the slope or intercept.

We define influence as follows:

$$\textbf{Influence} = \textbf{Leverage} \times \textbf{Discrepancy}$$

Leverage

An observation has high leverage if its x is far from the mean of x. Leverage is function of this distance. We can think of leverage as the potential to influence the regression line. But, an observation with high leverage in itself does not necessarily mean it has a strong influence on the regression line.

One common way to measure this distance is with the hat matrix, which is derived as follows:

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \\
\hat{\mathbf{y}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \\
\hat{\mathbf{y}} &= \mathbf{H}\mathbf{y} \\
\mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'
\end{aligned}$$

In other words, the hat matrix is what transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. You can think of it as what is required for $\mathbf{y}$ to become $\hat{\mathbf{y}}$.

Each element of the hat matrix describes the impact that each y has on each $\hat{y}$. The diagonal elements of the hat matrix describe the impact that each y has on each $\hat{y}$ for the same observation. If it is large, then it means that the observation's observed value is important in determining its own fitted value.

This implies that the observation is "far away" from the others on x. (Why?) The diagonal elements are therefore proportional to the distance between x_i and $\bar{x}_i$. They are a simple measure of the leverage of observation i.

We can also standardize these diagonals (divide by their mean) to ease interpretation. It becomes interpreted as how many times the average leverage each observation has. A standardized hat score greater than 2 or 3 is often said to be high.

Discrepancy

Discrepancy is mainly a function of the size of the residual for an observation. But we face two problems. First, observations with high influence reduce their own residual. We can therefore standardize the residuals for the effect of leverage:

$$\hat{\epsilon}_i^{stand} = \frac{\hat{\epsilon}_i}{\sqrt{\sum_i \hat{\epsilon}_i^2/(n-k-1)}\sqrt{1-h_i}}$$

Note that the residuals are now in standard deviation units. The second problem is that the discrepancy of the observation itself reduces the residual. We can overcome this by computing studentized residuals:

$$\hat{\epsilon}_i^{stud} = \frac{\hat{\epsilon}_i}{\sqrt{\sum_{-i} \hat{\epsilon}_{-i}^2/(n-k-1)}\sqrt{1-h_i}}$$

The first part of the demoninator is the variance of the residuals from a regression without observation i. Studentized residuals greater than 2 or less than -2 are considered large: they are more than two standard deviations away from the mean.

Cook's Distance

Cook's Distance considers both the discrepancy and the leverage in one summary measure:

$$CD_i = \frac{\hat{e_i}^2}{k \times MSE}\left[\frac{h_{ii}}{(1-h_{ii}^2)}\right]$$

The first part is a measure of discrepancy; the second part is a measure of leverage. There is no statistical test of D_i, but if $D_i > \frac{4}{n-k-1}$, then it is often considered to be large.

Influential Observations for the Regression (From Assignment 2)

For this use *M2*:

1. For each observation, calculate and explain the following:

- hat value (`hatvalues`)

```
library(stats)
hatvalues(res) # Gives hat scores, where res is the result from lm()
hatvalues(res)/mean(hatvalues(res)) # Gives standardized hat scores
```

- standardized error (`rstandard`)

```
rstandard(res) #Gives standardized residuals
```

- studentized error (`rstudent`)

```
rstudent(res) #Gives studentized residuals
```

- Cook's distance (`cooksd`)

```
cooks.distance(res) #Gives Cook's Distance
```

2. Create an outlier plot and label any outliers. See the example

```
plot(hatscore,rstu, xlab="Standardized hat-values", ylab="Studentized Residuals",
main="Influence Plot")
abline(h=c(-2,2), lty=2)
abline(v=c(2,3), lty=c(2,3))
```

3. Using the plot and rules of thumb identify outliers and influential observations

Influential Observations for a Coefficient

1. Run *M2*, deleting each observation and saving the coefficient for `outratiodirff`. This is a method called the jackknife. You can use a for loop to do this, or you can use the function `jackknife` in the package resamplr.

```
library(bootstrap)
library(car)
data <- Duncan #Dataset
model <- formula(income ~ education + prestige) #Model formula

theta <- function(x, xdata, coefficient){ #Function to extract coefficients
              coef(lm(model, data=xdata[x,]))[coefficient] }

jackknife.apply <- function(x, xdata, coefs) #Function to repeat extraction
{
sapply(coefs,
       function(coefficient) jackknife(x, theta, xdata=xdata,
                                       coefficient=coefficient),
        simplify=F)
}

results <- jackknife.apply(1:nrow(data), data, c("(Intercept)", "education",
                                       "prestige"))

jackknifeCOEF <- matrix(NA, ncol=3, nrow=45)
colnames(jackknifeCOEF) <- c("Intercept", "education", "prestige")
for (i in 1:nrow(data)){
  jackknifeCOEF[i,] <- coef((lm(income ~ education + prestige, data=data[-i,])))
}
```

- For which observations is there the largest change in the coefficient on `outratiodiff`?
  1. Which observations have the largest effect on the estimate of `outratiodiff`?
  2. How do these observations compare with those that had the largest effect on the overall regression as measured with Cook's distance?
  3. Compare the results of the jackknife to the `dfbeta` statistic for `outratiodiff`

The DFBETA is defined as follows:

$$DFBETA_{-i} = \frac{\hat{\beta}_{\mathbf{k}} - \hat{\beta}_{k(-i)}}{SE(\hat{\beta}_{(-i)})}$$

It is the ratio between the difference in parameter estimates after removing observation i and the standard error of the estimates after removing observation i

```
dfbeta(res) #Gives the dfbetas (a value greater than 2/sqrt(n) is considered large)
dfbetas(res) #Gives the standardized dfbetas
```

2. Aronow and Samii note that the influence of observations in a regression coefficient is different than the the influence of regression observations in the entire regression. Calculate the observation weights for `outratiodiff`.
   1. Regress `outratiodiff` on the control variables

2. The weights of the observations are those with the highest squared errors from this regression. Which observations have the highest coefficient values?
3. How do the observations with the highest regression weights compare with those with the highest changes in the regression coefficient from the jackknife?

Omitted Variable Bias

An informal way to assess the potential impact of omitted variables on the coeficient of the variable of interest is to coefficient variation when covariates are added as a measure of the potential for omitted variable bias (Oster 2016). Nunn and Wantchekon (2011) (Table 4) calculate a simple statistic for omitted variable bias in OLS. This statistic "provide[s] a measure to gauge the strength of the likely bias arising from unobservables: how much stronger selection on unobservables, relative to selection on observables, must be to explain away the full estimated effect."

1. Run a regression without any controls. Denote the coefficient on the variable of interest as $\hat{\beta}_R$.
2. Run a regression with the full set of controls. Denote the coefficient on the variable of interest in this regression as $\hat{\beta}_F$.
3. The ratio is $\hat{\beta}_F/(\hat{\beta}_R - \hat{\beta}_F)$. Calculate this statistic for *M2* and interpret it.

Heteroskedasticity

1. Run *M2* and *M3* with a heteroskedasticity consistent (HAC), also called robust, standard error. How does this affect the standard errors on `outratio` coefficients? Use the **sandwich** package to add HAC standard errors (Zeileis 2004).

```
library(sandwich)
vcovHAC(res) #Gives the estimated HAC covariance matrix
sqrt(diag(vcovHAC(res))) #Gives the HAC standard errors
```

2. Model *M3* is almost equivalent to running separate regressions on each combination of `Type` and `Year`.
    1. Run a regression on each subset of combination of `Type` and `Year`.
    2. How do the coefficients, standard errors, and regression standard errors (σ) differ from those of *M3*.
    3. Compare the robust standard errors in *M3* to those in the subset regressions. What is the relationship between heteroskedasticity and difference between the single regression with interactions (*M3*) and the multiple regressions.

Weighted Regression

1. Run *M2* and *M3* as weighted regressions, weighted by the population (`Popn`) and interpret the coefficients on `outratiodiff` and interactions. Informally assess the extent to which the coefficients are different. Which one does it seem to affect more?
2. What are some rationales for weighting by population? See the discussion in Solon, Haider, and Wooldridge (2013) and Angrist and Pischke (2014).