Influential Observations

Fundamental Techniques in Data Science with R

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Influential Observations

Influential observations contaminate analyses in two ways:

1. Exert too much influence on the fitted regression model
2. Invalidate estimates/inferences by violating assumptions

There are two distinct types of influential observations:

1. Outliers
    - Observations with extreme outcome values, relative to the other data.
    - Observations with outcome values that fit the model very badly.

2. High-leverage observations
    - Observation with extreme predictor values, relative to other data.

OUTLIERS

Outliers

Outliers can be identified by scrutinizing the residuals.

- Observations with residuals of large magnitude may be outliers.
- The difficulty arises in quantifying what constitutes a ”large” residual.

If the residuals do not have constant variance, then we cannot directly compare them.

- We need to standardize the residuals in some way.

Detecting Outliers

We are specifically interested in *externally studentized residuals*.

- We can't simply standardize the ordinary residuals.
  - *Internally studentized residuals*
  - Outliers can pull the regression line towards themselves.
  - The internally studentized residuals for outliers will be too small.

Begin by defining the concept of a *deleted residual*:

$$\hat{\varepsilon}_{(n)} = Y_n - \hat{Y}_{(n)}$$

- $\hat{\varepsilon}_{(n)}$ quantifies the distance of Y_n from the regression line estimated after excluding the nth observation.

Studentized Residuals

If we standardize the deleted residual, $\hat{\varepsilon}_{(n)}$, we get the externally studentized residual:

$$t_{(n)} = \frac{\hat{\varepsilon}_{(n)}}{SE_{\hat{\varepsilon}_{(n)}}}$$

The externally studentized residuals have two very useful properties:

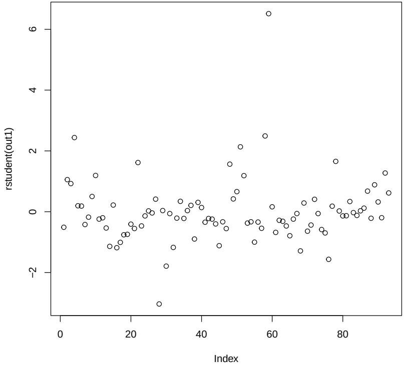
1. Each $t_{(n)}$ is scaled equivalently.
   - We can directly compare different $t_{(n)}$.
2. The $t_{(n)}$ are *Student's t* distributed.
   - We can quantify outliers in terms of quantiles of the t distribution.
   - $|t_{(n)}| > 3.0$ is a common rule of thumb for flagging outliers.

Studentized Residual Plots

```r
out1 <- lm(Price ~ Horsepower, data = Cars93)
rstudent(out1) |> plot()
```

Index plots of the externally studentized residuals can help spotlight potential outliers.

- Look for observations that clearly "stand out from the crowd."



HIGH-LEVERAGE CASES

High-Leverage Points

We identify high-leverage observations through their *leverage* values.

- An observation's leverage, h_n, quantifies the extent to which its predictors affect the fitted regression model.

- Observations with X values very far from the mean, $\bar{X}$, affect the fitted model disproportionately.

High-Leverage Points

We identify high-leverage observations through their *leverage* values.

- An observation's leverage, h_n, quantifies the extent to which its predictors affect the fitted regression model.
- Observations with X values very far from the mean, $\bar{X}$, affect the fitted model disproportionately.

In simple linear regression, the nth leverage is given by:

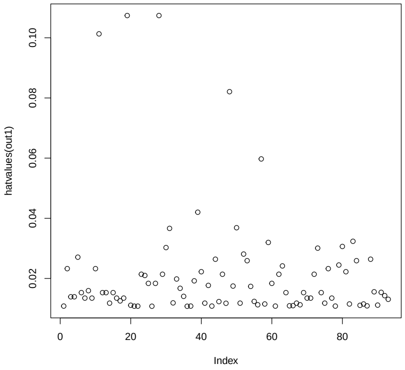
$$h_n = \frac{1}{N} + \frac{\left(X_n - \bar{X}\right)^2}{\sum_{m=1}^{N} \left(X_m - \bar{X}\right)^2}$$

Leverage Plots

```r
hatvalues(out1) |> plot()
```

Index plots of the leverage values can help spotlight high-leverage points.

- Again, look for observations that clearly "stand out from the crowd".



Influential Observations

Outliers & Leverages → Influential Points

Observations with high leverage or large (externally) studentized residuals are not necessarily influential.

- High-leverage observations tend to be more influential than outliers.
- The worst problems arise from observations that are both outliers and have high leverage.

Measures of influence simultaneously consider extremity in both X and Y dimensions.

- Observations with high measures of influence are very likely to cause problems.

Measures of Influence

All measures of influence use the same logic as the deleted residual.

- Compare models estimated from the whole sample to models estimated from samples excluding individual observations.

One of the most common measures of influence is *Cook's Distance*.

$$D_n = \frac{\sum_{n=1}^{N}\left(\hat{Y}_n - \hat{Y}_{(n)}\right)^2}{(P+1)\,\hat{\sigma}^2}$$

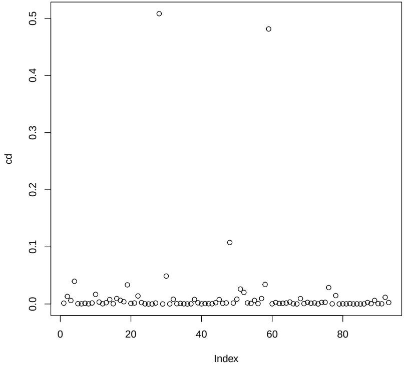
$$= (P+1)^{-1} t_n^2 \frac{h_n}{1-h_n}$$

Plots of Cook's Distance

```r
cd <- cooks.distance(out1)
plot(cd)
```

Index plots of Cook's distances can help spotlight the influential points.

- Look for observations that clearly "stand out from the crowd".



Treating Influential Observations

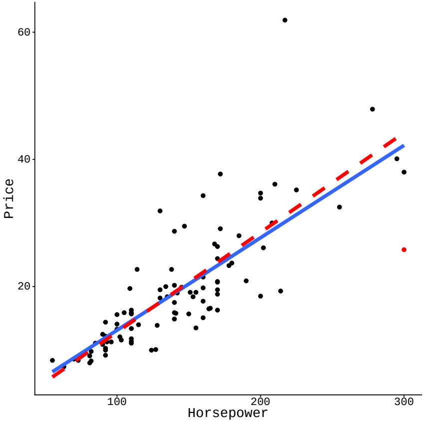
Removing Influential Observations

```
(maxD <- which.max(cd))

28
28
```

Observation number 28 was the most influential according to Cook's Distance.

- Removing that observation has a small impact on the fitted regression line.
- Influential observations don't only affect the regression line, though.



Removing Influential Observations

```
## Exclude the influential case:
Cars93.2 <- Cars93[-maxD, ]

## Fit model with reduced sample:
out2 <- lm(Price ~ Horsepower, data = Cars93.2)

round(summary(out1)$coefficients, 6)

             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.398769   1.820016 -0.768548 0.444152
Horsepower   0.145371   0.011898 12.218325 0.000000

round(summary(out2)$coefficients, 6)

             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -2.837646   1.806418 -1.570868 0.119722
Horsepower   0.156750   0.011996 13.066942 0.000000
```

Removing Influential Observations

```
partSummary(out1, 2)

Residuals:
    Min      1Q  Median      3Q     Max
-16.413  -2.792  -0.821   1.803  31.753

partSummary(out2, 2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4069  -3.0349  -0.5912   1.8530  30.7229
```

Removing Influential Observations

```r
summary(out1)[c("sigma", "r.squared", "fstatistic")] |>
    unlist() |>
    head(3)

         sigma         r.squared fstatistic.value
      5.976953          0.621287       149.287468

summary(out2)[c("sigma", "r.squared", "fstatistic")] |>
    unlist() |>
    head(3)

         sigma         r.squared fstatistic.value
     5.7243112         0.6548351      170.7449721
```

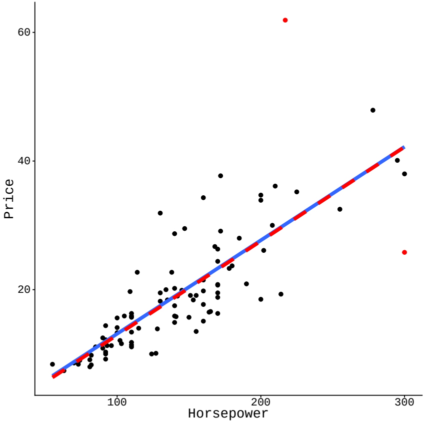
Removing Influential Observations

```r
(maxDs <- sort(cd) |> names() |> tail(2) |> as.numeric())
```

```
[1] 59 28
```

If we remove the two most influential observations, 59 and 28, the fitted regression line barely changes at all.

- The influences of these two observations were counteracting one another.
- We're probably still better off, though.



Removing Influential Observations

```r
## Exclude influential cases:
Cars93.2 <- Cars93[-maxDs, ]

## Fit model with reduced sample:
out2.2 <- lm(Price ~ Horsepower, data = Cars93.2)

round(summary(out1)$coefficients, 6)

              Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.398769   1.820016 -0.768548 0.444152
Horsepower   0.145371   0.011898 12.218325 0.000000

round(summary(out2.2)$coefficients, 6)

              Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.695315   1.494767 -1.134166  0.25977
Horsepower   0.146277   0.009986 14.648807  0.00000
```

Removing Influential Observations

```
partSummary(out1, 2)

Residuals:
    Min      1Q  Median      3Q     Max
-16.413  -2.792  -0.821   1.803  31.753

partSummary(out2.2, 2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3079  -2.5786  -0.6084   1.9775  14.5793
```

Removing Influential Observations

```r
summary(out1)[c("sigma", "r.squared", "fstatistic")] |>
    unlist() |>
    head(3)

          sigma         r.squared fstatistic.value
       5.976953          0.621287       149.287468

summary(out2.2)[c("sigma", "r.squared", "fstatistic")] |>
    unlist() |>
    head(3)

          sigma         r.squared fstatistic.value
      4.7053314         0.7068391      214.5875491
```

Treating Influential Points

The most common way to address influential observations is simply to delete them and refit the model.

- This approach is often effective—and always simple—but it is not fool-proof.
- Although an observation is influential, we may not be able to justify excluding it from the analysis.

Robust regression procedures can estimate the model directly in the presence of influential observations.

- Observations in the tails of the distribution are weighted less in the estimation process, so outliers and high-leverage points cannot exert substantial influence on the fit.
- We can do robust regression with the `rlm()` function from the **MASS** package.