STAT302 Methods of Data Analysis 1
Module 8: Problematic observations

Austin Brown

University of Toronto, Department of Statistical Sciences

November 6, 2024

Lecture 1

Writing lecture

Dorey will discuss some nontechnical strategies and tools for writing your final project.

Lecture 2

Review

Brief Review

Learning objectives

- Define problematic observations
- Understanding and addressing outliers
- Understanding and addressing high leverage points

A new step in our regression analysis

- Fit a model with acceptable diagnostics (transformations, polynomial terms, interactions possibly handled to get here)
- Variable selection (Overall and Partial F-tests)
- **Generally at this stage, investigate extreme observations and possible effects on the regression fit**
- Inference with final model (Confidence intervals, prediction, etc.)

Some causes of problematic observations

- Observation is rare for the population (i.e. follows the population but is a rare event)
- Observation does not follow the population model
- Incorrect measurement such as bad instrumentation (i.e. physical instruments, lab measurements).
- Data was input into the database incorrectly (i.e. human error)
- etc.

Lecture 2: Extreme response values

Outliers

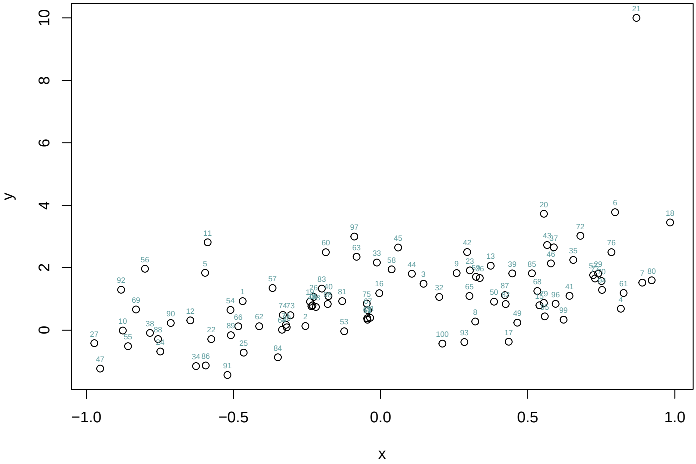
An **outlier** is an observation $y_i, \boldsymbol{x}_i$ with an extreme response value. An extreme y_i value does not align or fit the pattern of the rest of the observations. An extreme response value will have a large standardized residual $|r_i|$.

Examples:

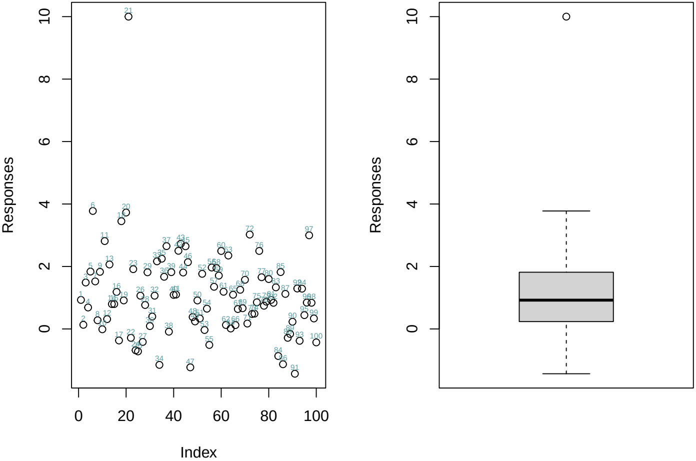
- The production time for one observation is very large
- The restaurant price is very expensive

Lecture 3: detecting outliers

Detect the outlier



Detect the outlier



Detecting outliers

We expect an outlier to have a large (observed) residual:

$$\hat{e}_i = Y_i - \hat{Y}_i.$$

So we anticipate an outlier to have a large standardized residual

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$$

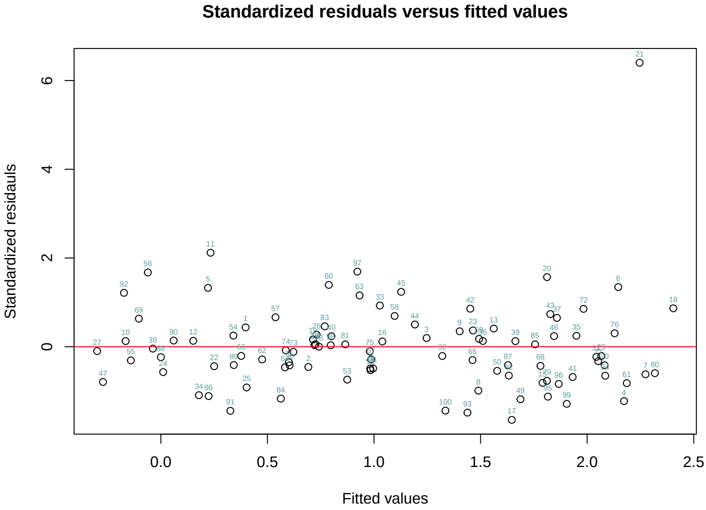
to be much larger. Recall the variance of the residual is $\mathsf{Var}(\hat{e}_i | \boldsymbol{X}) = \sigma^2(1 - h_{i,i})$.

Detecting outliers

Get the standardized residuals in R:

```
rstandard(fit)
```

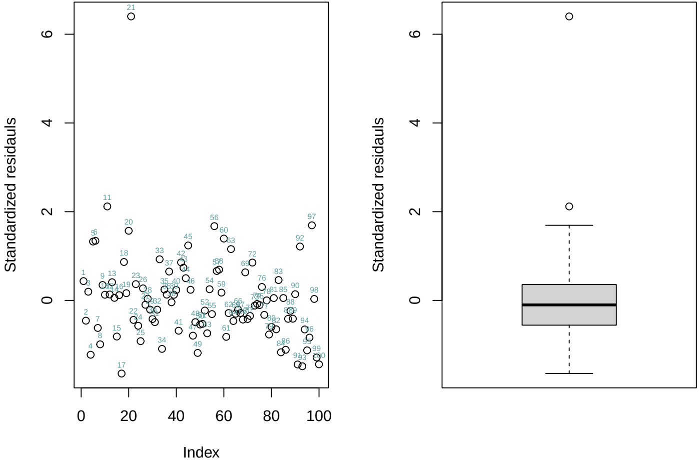
Detect the outlier



Could also use the square root standardized residuals that R creates.

Detect the outlier

Other plots can accomplish the same goal:



Detecting outliers

A **cutoff rule** can be used to determine an outlier for standardized residuals larger than

$$|r_i| \geq 2$$

for medium sized data and

$$|r_i| \geq 4$$

for large data (See [Sheather, 2009] page 155). This is only a guideline and not always a precise cutoff value.

Lecture 2: Extreme predictor values

The hat matrix

Minimizing the RSS in regression geometrically is precisely projecting $\boldsymbol{Y}$ onto the linear subspace generated by the columns of $\boldsymbol{X}$. The hat matrix is a projection matrix defined

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

This can be seen here by

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

Notice that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ and $\boldsymbol{H}^T = \boldsymbol{H}$.

The hat matrix

$$E(\hat{\boldsymbol{e}}|\boldsymbol{X}) = 0$$

The hat matrix

$$\mathsf{Cov}(\hat{\boldsymbol{e}}|\boldsymbol{X}) = \sigma^2(I - \boldsymbol{H})$$

Leverage

The **leverage** values are the diagonal elements of the hat matrix $h_{i,i}$. A direct matrix calculation (work this on your own!) yields

$$h_{i,i} = \boldsymbol{x_i}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x_i}$$

and large leverage values give a measure of extreme predictor values even in multiple regression. We can write

$$\hat{\boldsymbol{Y}}_i = \sum_{j=1}^{n} h_{i,j} \boldsymbol{Y}_j$$

to see how high leverage points can change the fitted values.

Leverage

An observation with high **leverage** is an observation that has a high leverage value $h_{i,i}$. This is representing an extreme $\boldsymbol{x}_i$ value. The value does not fit the pattern of the rest of the observations.

Examples:

- The production size for one order is much larger than the rest
- The restaurant decor, foot, and service rating for one restaurant are much lower than the rest

Lecture 3: detecting high leverage

Detecting outliers

Get the leverage values in R:

```
hatvalues(fit)
```

Hat matrix continued

In SLR,

$$h_{i,i} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_{XX}}$$

Hat matrix continued

Detect the high leverage point

Since it is a projection matrix, the average

$$\mathsf{Avg}(\mathsf{diag}(H)) = \frac{1}{n}\sum_{i=1}^{n} h_{i,i} = (p+1)/n.$$

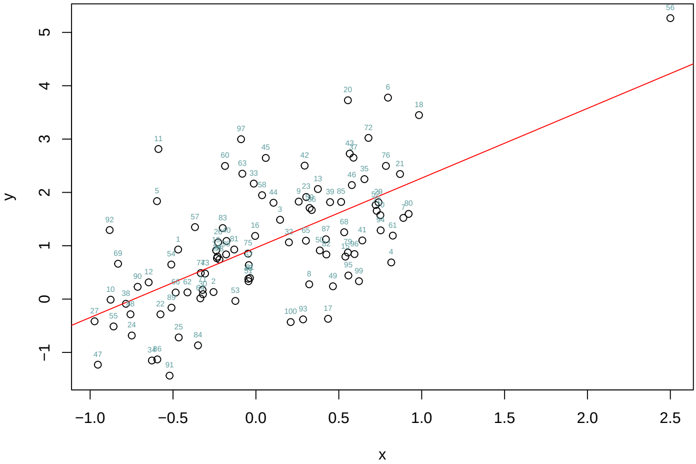
This leads to a cutoff value as one guideline

$$h_{i,i} \geq 2(p+1)/n$$

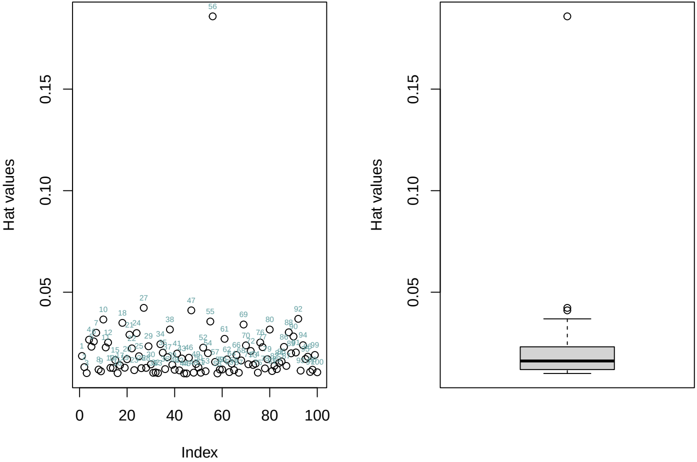
is high leverage if it is larger than 2 times the average leverage values (See [Sheather, 2009] page 154).

Detect the high leverage point

For SLR, we can just plot it, but for MLR this is not possible generally.



Detect the high leverage point



Lecture 3

Learning objectives

- Understand influential points

Lecture 3: influential points

Influential point

A data point is **influential** if it influences the regression analysis (i.e. the predictions, estimated coefficients change drastically). Outliers and high leverage data points are not necessarily influential and we have to investigate to determine whether or not they are influential.

Examples:

- Removing the observation of a restaurant with an extreme decor rating changes the estimated coefficient dramatically.
- Removing the observation of a restaurant with extreme ratings changes the RSS and predictions dramatically.

Important differences

- An outlier observation is not necessarily a high leverage point
- An outlier observation is not necessarily an influential point
- A high leverage observation is not necessarily an outlier
- A high leverage observation is not necessarily influential.

Addressing problematic observations

The most important steps for this course:

- Identify and classify the observation as outlier, high leverage, or influential.
- Comment on the limitations and understand the effect of the observation on the regression model fit

Things to consider and depend on the specific problem and dataset:

- Remove an observation if it is justifiable (incorrect data input, etc.).
- Do not remove an observation.
- Removing an influential point without proper justification may lead to incorrect/dishonest/unethical conclusions.

Ethical concerns on removing observations

Consider a prescription drug medical trial where one person in the trial experienced severe side effects and was removed and left out of the final reporting. What can be problematic about this?

What to do?

Diagnostics can identify an high leverage points, outliers, and influential point. Identification and understanding how the observation influences the analysis may be enough. (See [Sheather, 2009] page 57) What to do is problem dependent:

- Identify and do nothing
- Identify and remove the observation
- Identify and fit a different regression model. Add polynomial terms, etc.

Cook's distance

Identify influential observations with Cook's distance. Gives a measure of the influence of the observation i on all of the fitted values

$$D_i = \frac{1}{(p+1)\hat{\sigma}^2}[\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)}]^T[\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)}]$$

where $\hat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)}$ where $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares solution with observation i removed.

Cook's distance

This is the same as:

$$D_i = \frac{1}{(p+1)\hat{\sigma}^2}[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}]^T \boldsymbol{X}^T \boldsymbol{X} [\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}]$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares solution with observation i removed.

The intuition here is that this is of a similar form as the F statistic with $p+1$ and $n-p-1$ degrees of freedom.

Cook's distance

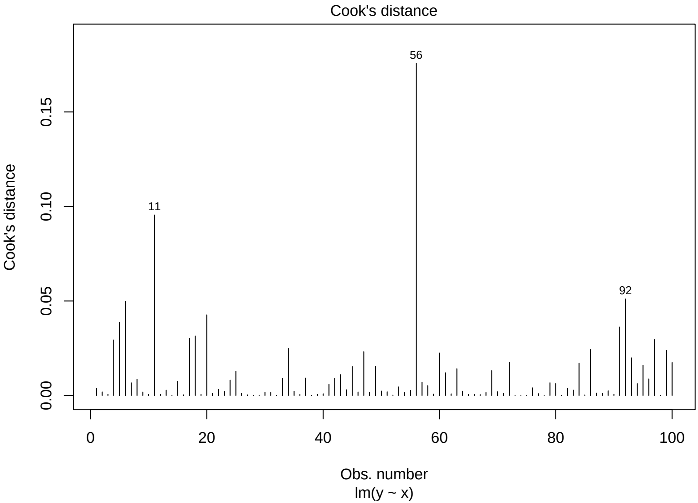
Get Cook's distances D_i in R.

```r
cooks.distance(fit)

plot(fit, which = 4)

# All diagonistic plots
par(mfrow = c(2, 2))
plot(fit, which = c(1, 2, 3, 4))
```

Cook's distance



Detect influential observations

Some guidelines:

- If D_i sticks out from the other values, it is almost certainly influential.
- A cutoff rule is $D_i > m$ where m is the median of the $F(p+1, n-p-1)$. We can look at the form of Cook's distance and if $D_i \sim F(p+1, n-p-1)$, then if $D_i > m$ where m is the median of the $F(p+1, n-p-1)$, then this is outside or at the edge of a 50% confidence region for the F distribution. There are differing opinions on a precise cutoff rule.

Real-data example

Example Rat data from Weisberg.

Computing Cook's distance

Use the inversion of a matrix formula:

$$(A - vv^T)^{-1} = A^{-1} + \frac{A^{-1}vv^TA^{-1}}{1 - v^TA^{-1}v}$$

to get

$$\begin{aligned}(X_{(i)}^T X_{(i)})^{-1} &= (X^TX - x_ix_i^T)^{-1} \\ &= (X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1 - h_{i,i}}.\end{aligned}$$

Computing Cook's distance

Now use $X_{(i)}^T Y_{(i)} = X^T Y - y_i x_i$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \\
&= \left[(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{i,i}} \right] \left[X^T Y - y_i x_i \right] \\
&= \hat{\boldsymbol{\beta}} + \frac{(X^T X)^{-1} x_i}{1 - h_{i,i}} \left[x_i^T \hat{\boldsymbol{\beta}} - (1 - h_{i,i}) y_i - h_{i,i} y_i \right] \\
&= \hat{\boldsymbol{\beta}} - \frac{(X^T X)^{-1} x_i}{1 - h_{i,i}} \hat{e}_i.
\end{aligned}$$

Computing Cook's distance

Now plug this into Cook's distance:

$$\begin{aligned} D_i &= \frac{1}{(p+1)\hat{\sigma}^2}[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}]^T \boldsymbol{X}^T \boldsymbol{X} [\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}] \\ &= \frac{1}{(p+1)\hat{\sigma}^2} \frac{\hat{e}_i^2}{(1-h_{i,i})^2} x_i^T (X^T X)^{-1} x_i \\ &= \frac{r_i^2}{(p+1)} \frac{h_{i,i}}{(1-h_{i,i})}. \end{aligned}$$

Very quick to compute.

Some other measures of influence

Difference in coefficient estimation

How does removing an observation change the coefficient estimates? Difference in individual betas:

$$\mathsf{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)}\sqrt{(X^T X)^{-1}_{j+1,j+1}}}$$

with cutoff guideline

$$|\mathsf{DFBETAS}_{j(i)}| \geq \frac{2}{\sqrt{n}}.$$

Difference in individual fits

How does removing an observation change the estimation for those specific predictors $\boldsymbol{x_i}$? Difference in individual fits:

$$\mathsf{DFFITS}_{(i)} = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{i,i}}}$$

with cutoff guideline

$$|\mathsf{DFFITS}_{(i)}| \geq 2\sqrt{\frac{p+1}{n}}.$$

Computing influence measures in R

Compute influence measures in R:

```
dfbetas(fit)

dffits(fit)
```

Lecture 3: Example

Example by hand

A data set contains 100 observations. $y_{10} = 8$ and $x_{10} = -0.88$. $s_{XX} = 36.265$ and $\overline{x} = 0.0728$. R is used to get a regression is fit to get $\hat{\beta}_0 = 1.07$ and $\hat{\beta}_1 = 0.9$ and residual standard error $\hat{\sigma} = 1.232$. R is also used to get $qf(.5, 2, 98) = 0.698073$. Use the cutoffs to determine:

- Compute $h_{10,10}$. Is this a high leverage value?
- Compute r_{10}. Is this an outlier?
- Compute D_{10}. Is this an influential point? r

Example using R

Now use

`module8_activity.Rmd`

to plot each diagnostic and perform the analysis.

Module outcomes

1. What is the difference between each type of problematic observations?
2. How do we quantify the various problematic observations?
3. How do we then identify the presence of each type of problematic observation?
4. What do we do if problematic observations are present?

References I

Simon Sheather. *A modern approach to regression with R*. Springer Science & Business Media, 2009.