Diagnostics (Part I)

Lecture 8

Alexandra Chronopoulou

I ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES
Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will:

- discuss how we detect unusual observations:
  - high leverage points
  - outliers
  - highly influential points

Regression Model Assumptions

Recall, that we can write the MLR model as:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$$

- Error: assumed to be iid, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Model: assumed to be linear in the parameters, i.e., $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$.

We might have unusual observations.

We will use both, graphical and numerical tools for diagnosis.

Finding Unusual Observations

Why we discuss unusual observations first?

- Least squares regression is very sensitive to individual data points.
- It is possible the inference, p-values, parameter estimation, CI's are all driven by a single data point.
- Sometimes, the estimated parameters and other related statistics (such as R^2) depend heavily on one observation, in the sense that if that observation was removed, the result of the analysis would change.

Types of Unusual Observations

- High leverage points: We will define a measure called "leverage" which quantifies how far a data point is from the center of the whole sample (remember the Mahalanobis distance?). Points with a large value of leverage are flagged as the *high leverage points*. High leverage points could be "good" or "bad".

- Outliers: data points that do not fit the model as the other data points. We will introduce a formal testing procedure to identify outliers.

- Highly influential points: How does each individual observation affect the estimation of the model? We will define some measure, "*Cook's distance*", to quantify the aforementioned change for each data point, and data points with a large value of Cook's distance are called high influential points.

Leverage

- The diagonal elements of $\mathbf{H}$,

$$h_i = H_{ii}$$

  are called leverages and are very useful diagnostics.

- h_i gives a measure (invariant under any affine transformation of $\mathbf{X}$) of how far the i-th observation is from the center of the data (in the X-space).

Leverage Points

- For simple linear regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

- In general:

$$\begin{aligned} h_i &= \mathbf{x_i}^\mathbf{T} (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1} \mathbf{x_i} && (1) \\ &= \frac{1}{n} + \frac{1}{n-1} (\mathbf{z}_i - \bar{\mathbf{z}})^T \hat{\Sigma}^{-1} (\mathbf{z}_i - \bar{\mathbf{z}}) && (2) \end{aligned}$$

where

$$\hat{\Sigma}^{-1}_{(p-1)\times(p-1)} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{z} - \bar{\mathbf{z}})(\mathbf{z} - \bar{\mathbf{z}})^T$$

is the sample covariance of the $(p-1)$ predictor variables. The second term in the right hand side of (2) is the so-called Mahalanobis distance from $\mathbf{z}_i$ to the data center $\bar{\mathbf{z}}$

Properties of the Leverage

- Recall that the hat matrix is idempotent $\mathbf{H} = \mathbf{H}\mathbf{H}^\top$ and has $tr(\mathbf{H}) = p$.
- These imply that

$$\sum_i h_i = p \text{ and } \sum_j H_{ij}^2 = h_i.$$

- For a given i we can decompose the last sum as follows:

$$\sum_j H_{ij}^2 = H_{ii}^2 + \sum_{i \neq j} H_{ij}^2 = h_i$$

$$\Rightarrow \sum_{i \neq j} H_{ij}^2 = h_i(1 - h_i) \Rightarrow h_i(1 - h_i) > 0$$

- From this we can conclude the following properties of h_i:

$$0 < h_i < 1, \; \sum_i h_i = p$$

Fitted Values and Leverage

- Recall the equation $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.
- In matrix form:

$$\begin{pmatrix} \hat{y}_1 \\ \dots \\ \hat{y}_i \\ \dots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} H_{11} & \dots & H_{1n} \\ \dots & \dots & \dots \\ H_{i1} & \dots & H_{in} \\ \dots & \dots & \dots \\ H_{n1} & \dots & H_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix}$$

$$\begin{aligned} \hat{y}_i &= H_{i1} y_1 + \ldots + H_{ii} y_i + \cdots + H_{in} y_n \\ &= H_{i1} y_1 + \ldots + h_i y_i + \cdots + H_{in} y_n \end{aligned}$$

Fitted Values and Leverage

- Note that the LS fit, $\hat{y}_i$, is a linear combination of the n data points:

$$\hat{y}_i = h_i y_i + \sum_{i \neq j} H_{ij} y_j$$

  This means that $h_i = \frac{d\hat{y}_i}{dy_i}$

- When h_i is large (close to 1), $\hat{y}_i$ relies heavily on y_i (instead of using the information from other data points), therefore $\hat{y}_i$ will be "forced" to be close to the observed y_i.

- Consequently, the variance for the residual r_i will be small, and the variance for the fit $\hat{y}_i$ will be large (since the fit from another data set would be quite different).

$$\mathrm{Var}\left(\hat{\mathrm{y}}_{\mathrm{i}}\right) = \sigma^2 \mathrm{h}_{\mathrm{i}}, \quad \mathrm{Var}\left(\mathrm{r}_{\mathrm{i}}\right) = \sigma^2(1 - \mathrm{h}_{\mathrm{i}})$$

High-leverage points

Since $\sum_i h_i = p$, a rule-of-thumb is that observations with leverages more than $2p/n$ (twice the mean leverage) should be flagged as high-leverage points and should be examined closely.

- Good high-leverage points: its y point follows the pattern of the rest of the data, but with an x_i value that is far away from the sample mean.
- Bad high-leverage points: its y value does not follow the pattern suggested by the rest of the data; the LS fitting might change a lot if we remove this point.

Example: Leverages in Bike Shares data set

Use the function influence to extract the leverages, and the function halfnorm to plot the leverages in increasing order.

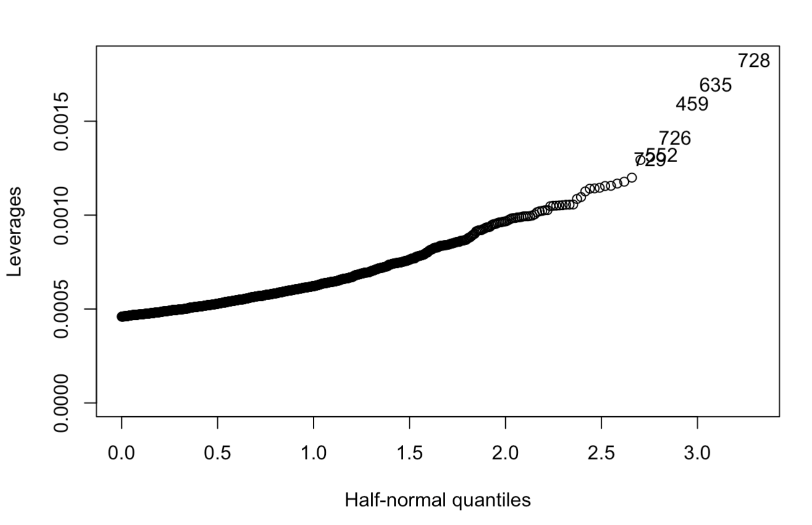
```
n=dim(bikeshares.reg)[1]; # sample size
p=4; # 3 predictors we have in the model plus the intercept

bikeshare.mlr = lm(cnt ~ t1 +  hum + wind_speed, data=bikeshares.reg )

# Compute Leverages
lev=influence(bikeshare.mlr)$hat

# Determine which exceed the 2p/n threshold
newlev = lev[lev>2*p/n]

# Prepare a half-normal plot
halfnorm(newlev, 6, labs=as.character(1:length(newlev)), ylab="Leverages")
```

Half-Normal Plots

- Designed to identify unusually large values and assess positive data.
- Plot the data against the positive normal quantiles. Specifically,

  1. Sort the data:
  $$x_{[1]} \leq \ldots \leq x_{[n]}.$$
  2. Compute the quantiles:
  $$u_i = \Phi^{-1}\left(\frac{n+i}{2n+1}\right)$$
  3. Plot $x_{[i]}$ against u_i.

Example: Leverages in Bike Shares data set



Residuals

The residuals $r_i = y_i - \hat{y}_i$ do not have a constant variance (WHY?). So they need to be standardized. There are two versions of the residuals:

– Standardized Residuals r_i^*: They are internally standardized. Under the model assumptions they follow approximately a Normal distribution.

– Studentized residuals t_i: They are externally standardized. They follow a T distribution and will be used in our outlier test.

Residuals are very useful in regression diagnostics. Some authors recommend using the standardized version of the residuals instead of the raw residuals in *all* diagnostic plots.

Difference between ε and r

ε: true residuals (our theoretical quantities)

r: estimated residuals

- Both residuals are normally distributed, but:

$$\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad \mathbf{r} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 (\mathbf{I_n} - \mathbf{H}))$$

  where **H** is the projection/hat matrix.

- The errors ε_i's have equal variance and are independent, while the residuals r_i's have unequal variance and are correlated.

- $\mathbb{E}(\varepsilon) = \mathbb{E}(\mathbf{r}) = \mathbf{0}$. But

$$\sum_i \varepsilon_i \neq 0, \quad \sum_i r_i = 0$$

  (by default we assume an intercept is included in the model)

Standardized Residuals

Since $r_i \sim \mathcal{N}(0, \sigma^2(1 - h_i))$, it is reasonable to consider a standardization of the residuals in this form:

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_i}}, \quad i = 1, \ldots, n$$

- $\sum_i r_i^*$ is no longer zero.
- Since the r_i is not independent of $\hat{\sigma}$, each r_i^* is *not distributed as a T distribution*.
- As an approximation, we can view the r_i^*'s as *iid* $\mathcal{N}(0, 1)$ random variables, although they are *not* Normally distributed and they are slightly correlated.

Studentized Residuals

- The studentized residuals are based on the idea of leave-one-out (also know as *jackknife* residuals).
- Here is the leave-one-out idea:
  1. Run a regression model on the $(n-1)$ samples with the i-th sample (x_i, y_i) removed.
  2. Denote the leave-one-out estimates of the regression coefficient and error variance by $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$, where the notation (i) means "excluding the i-th observation."
  3. Then, check the discrepancy between observations y_i and the fitted value $\hat{y}_{(i)} = \mathbf{x}^T \hat{\beta}_{(i)}$

- Define the Studentized Residuals as:

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \left(1 + x_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} x_i\right)^{1/2}} = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 - h_i}}$$

  which follows a T_{n-p-1} distribution if $y_i \sim \mathcal{N}(\mathbf{x}_i^T \beta, \sigma^2)$.

- One can also show that r_i^* and t_i are a monotone transformation of each other.

- We do not need to run the model n times to get the estimates $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$ since it can be shown that:

$$t_i = r_i^* \left(\frac{n - p - 1}{n - p - {r_i^*}^2} \right)^{1/2}$$

An Outlier test

- Outliers are observations that do not fit the model, but Outliers are not necessarily observations with large residuals.
- An outlier test is a useful tool to distinguish observations that have large residuals from outliers. We need to used the studentized residuals for the outlier test.
- Under the Null hypothesis H_0,

$$t_i \sim T_{n-p-1}$$

 distribution. So we can use a t-test to test whether the *i*-th observation is an outlier or not.

- Generally, we would want to perform this outlier test for all n observations, doing the tests one at a time.
- If we perform the test on the largest observed residuals this would be an example of data snooping, unless somehow these cases were identified before data collection.
- In order to be certain that the overall type I error rate is no greater than α, the Bonferroni correction may be used. When doing so, each case would be tested at level α/n.

Bonferroni Correction

- Suppose we are testing m hypothesis sinultaneously.
- For each test, we use a significant level α. That is, the chance to make a Type I error is α.
- Suppose we want to control the overall type I error rate (for all m tests) to be 95%.
- We should set the individual significance levels to be $\alpha = 5\%/m$

What we should do with outliers?

- Delete them? When?
- Points should not be routinely deleted simply because they do not fit the model. No data snooping!
- Outliers, as well as other unusual observations discussed here, often flag *potential problems* of the current model. Instead of dropping them, maybe, try a new alternative model.

Example: Outliers in Bike Shares data set

Use the function rstudent to get the studentized residuals, and the function sort to sort the residuals in decreasing order.

```
# Compute Studentized Residuals
jack=rstudent(bikeshare.mlr);

# The critical value WITH Bonferroni correction is
qt(.05/(2*n), n-p-1)
```

```
## [1] -4.681361
```

```
# The critical value WITHOUT Bonferroni correction is
qt(.05/2, n-p-1)
```

```
## [1] -1.9601
```

```
# Sort the residuals indescending order to find outliers (if any)
sort(abs(jack), decreasing=TRUE)[1:10]
```

```
##     4462     5130     5139     4471    15888     5140    15217    15385
## 6.408782 5.665958 5.499140 5.317999 4.807279 4.787554 4.746059 4.738005
##    16727    14905
## 4.661289 4.522918
```

There are 8 outliers in this data set since the first 8 values are larger than 4.681361 in absolute value.

Influential observations

- Observations whose removal greatly affects the regression analysis are called influential observations.

- An influential observation may be (or may not) an outlier or a high-leverage observation; or may be both: an outlier and a high-leverage observation.

- We will use the Cook's distance to detect influential observations.

$$D_i = \frac{||\mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_{(i)}||^2}{p\hat{\sigma}^2} = \frac{||\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}||^2}{p\hat{\sigma}^2} = \frac{{r_i^*}^2}{p}(\frac{h_i}{1-h_i})$$

  which indicates that highly influential points are either outliers (large $|r_i^*|$) or high-leverage points (large h_i) or both.

- A rule-of-thumb: observations with $D_i \geq 1$ are highly influential.

ILLINOIS

Example: Influential observations in Bike Shares data set

Use the function cooks.distance to calculate the Cook's distance for each observation and the function halfnorm to plot the CD's in increasing order.

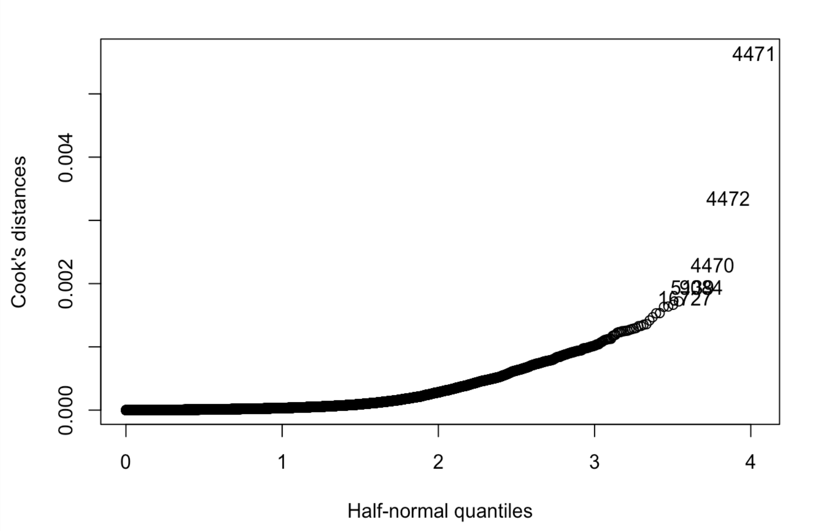
```
# Compute Cook's Distance
cook = cooks.distance(bikeshare.mlr)
# Extract max Cook's Distance
max(cook)
```

```
## [1] 0.005641587
```

```
# Prepare a Half Normal Plot of Cook's distances
halfnorm(cook, 6, labs=as.character(1:length(cook)), ylab="Cook's distances")
```

According to the rule-of-thumb ($CD \geq 1$), there are not influential observations. However, there is one observation that is too far from the rest.

Example: Influential observations in Bike Shares data set



Summary about Unusual Observations

- High-leverage points:

$$h_i = H_{ii} > 2p/n$$

High-leverage points are far away from the center of the data (in terms of the Mahalanobis distance). Keep in mind that:

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 h_i, \quad \mathrm{Var}(r_i) = \sigma^2(1 - h_i)$$

- Outliers: Perform a t-test on the studentized residuals using the Bonferroni correction.
  - This is equivalent to removing the i-th point, run LS on the remaining $(n-1)$ data points, and then form a PI at $\mathbf{x}_i$; if PI covers y_i, then the i-th point is NOT an outlier.
- Highly influential points; Use Cook's distance and check whether $D_i \geq 1$:

$$D_i = \frac{{r_i^*}^2}{p}\left(\frac{h_i}{1 - h_i}\right)$$

which indicates that high influential points are either outliers or high-leverage points or both.