STAT 480 Statistical Computing Applications

Unit 3. Linear Regression

Lecture 3. Diagnostics

Department of Statistics, Iowa State University

Spring 2019

Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.

Introduction

- One related issue is the importance of each case on estimation and other aspects of the analysis.
- In some datasets, the fitted regression line may change in important ways if just one case is deleted from the data.
- We will develop methods for detecting and identifying such influential cases. This will involve two types of diagnostic statistics, distance measures and leverages.
- We will use some graphical approaches to detect those influential cases.

Leverage During Regression

- During a regression, some data points have more leverage than others
  - Leverage points = data with an extreme value of the predictor variable (x)
- Like outliers, high leverage data can have outsized influence on the regression results
  - Well define “influence” more exactly later
- Leverage can be quantified by looking at the distances between x's (accounting for correlation).
- For two or more predictor variables (multiple regression), we use matrices to calculate the leverage values.

Residuals and Leverage

- Recall that predicted values $\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{X}\hat{\beta}$, where

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

  is called the "hat matrix".

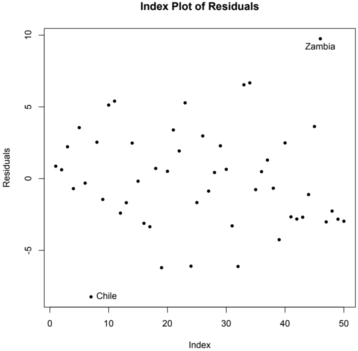
- Residual: $\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.
- $\mathrm{var}(\hat{\boldsymbol{\varepsilon}}) = (\mathbf{I} - \mathbf{H})\sigma^2$ assuming $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.
- We see that although the errors may have equal variance and be uncorrelated, the residuals do not.

Leverage

- $h_i = \mathbf{H}_{ii}$ are called leverages and are useful diagnostics.
- We see that $\mathrm{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$, so that a large leverage for h_i will make $\mathrm{var}(\hat{\varepsilon}_i)$ small, i.e., the fit will be "forced" be to close to Y_i.
- h_i depends only on $\mathbf{X}$, not on Y.
- $\sum_{i=1}^n h_i = p$.
- $h_i \geq \frac{1}{n}$ for all i.
- An average value for h_i and a "rule of thumb" is that leverages of more than $\frac{2p}{n}$ should be looked at more closely.

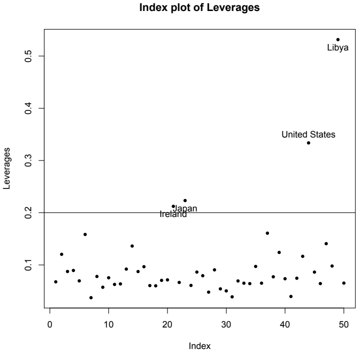
Plot of Residuals

```
> m=lm(sr~pop15+pop75+dpi+ddpi,data=LifeCycleSavings)
> plot(m$res,ylab="Residuals",main="Index Plot of Residuals",
+  pch=20)
> countries=row.names(LifeCycleSavings)
> identify(1:n,m$res,countries)
```



Plot of Leverages

```
> x = model.matrix(m)
> lev = hat(x)
> plot(lev,ylab="Leverages", main = "Index plot of Leverages",
+ pch=20)
> abline(h=2*5/50)
> identify(1:50,lev,countries)
```



Studentized Residuals

- As we have seen $\mathrm{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$, this suggests the use of

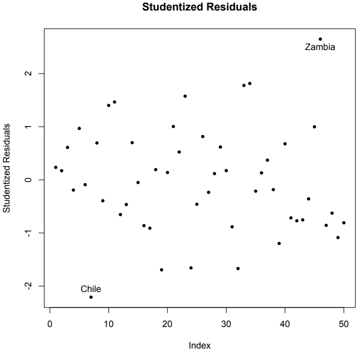
  $$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

  which are called (internally) studentized residuals.

  - If the model assumptions are correct, $\mathrm{var}(r_i) = 1$ and $\mathrm{cor}(r_i, r_j)$ tends to be small.
  - Studentized residuals are sometimes preferred in residual plots as they have been standardized to have equal variance.
  - Studentized residuals offer an alternative criterion for identifying outliers.

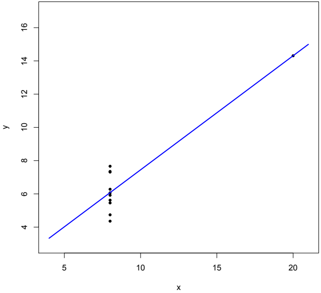
Plot of Studentized Residuals

```
> # Plot of Studentized Residuals
> ms = summary(m)
> stud = m$res/(ms$sig*sqrt(1-lev))
> plot(stud,ylab="Studentized Residuals",
+  main="Studentized Residuals",pch=20)
> identify(1:50,stud,countries)
```



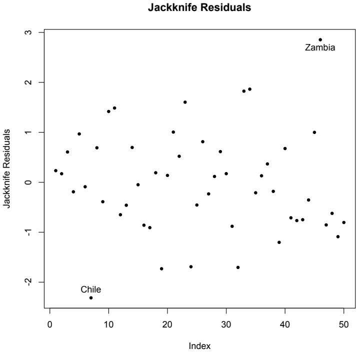
An Outlier Test

- An outlier is a point does not fit the current model. We need to be aware of such exceptions.
- An outlier test is useful because it enables us to distinguish between truly unusual points and residuals which are large but not exceptional.
    - Outliers may affect the fit.
    - Outliers can conceal themselves.



Plot of Jackknife Residuals

```
> # Plot of Jackknife Residuals
> jack = rstudent(m)
> plot(jack,ylab="Jackknife Residuals",
+  main="Jackknife Residuals",pch=20)
> identify(1:50,jack,countries)
```



What Should Be Done About Outliers?

- Check for a data entry error.
- Examine the physical context – why did it happen?

  One example of the importance of outliers is in the statistical analysis of credit card transactions. Outliers in this case may represent fraudulent use.
- Exclude the point from the analysis but try re-including it later if the model is changed. One should be careful before removing outliers:
  - The error distribution may not be normal and so larger residuals maybe expected.

Influential Observations

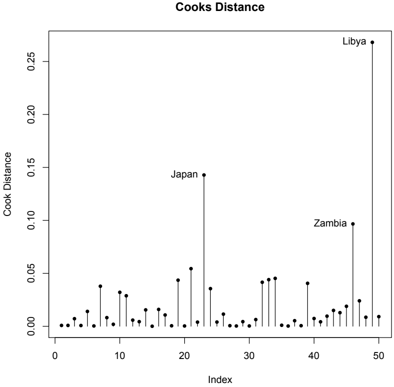
- An influential point is one whose removal from the dataset would cause a large change in the fit.
- An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least on of those two propoerties.
- The most popular measure of influence is Cook's distance:

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

The first term, r_i^2, is the residual effect and the second is the leverage. The combination of the two leads to influence.

Plot of Cooks Distance

```
> cook=cooks.distance(m)
> plot(cook,ylab="Cook Distance",main="Cooks Distance",pch=20)
> segments(1:50,0,1:50,cook)
> identify(1:50,cook,countries)
```

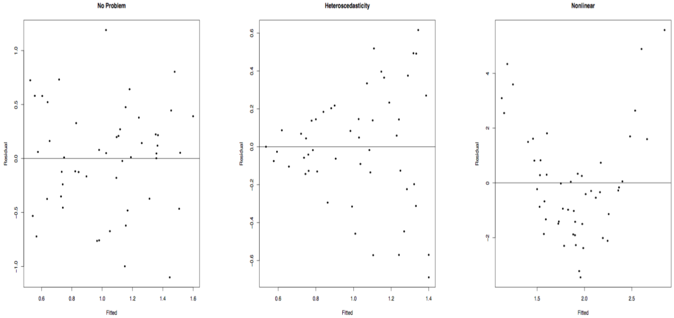


Residual Plots

Outliers and influential points indicate cases that are in some way individually unusual but we also need to check the model assumptions. For that reason, we need to plot the residuals.

- It can be shown that $\hat{\varepsilon}$ and $\hat{Y}$ are independent.
- Plot $\hat{\varepsilon}$ against $\hat{Y}$. This is the most important diagnostic plot that you can make. It should appear as a parallel band around 0. Otherwise, it would suggest model violation. Things to look for are
  - heteroscedasticity (non-constant variance): If spread of $\hat{\varepsilon}_i$ increases as $\hat{Y}$ increases, error variance of Y increases with mean of Y. Then we need a transformation of Y.
  - nonlinearity (which indicates some change in the model is necessary).

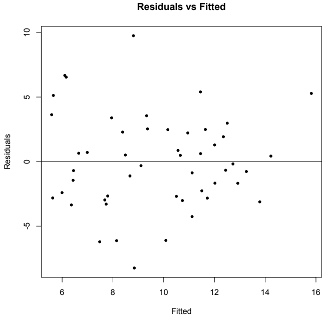
Residual Plots (Cont.)



Residuals versus fitted plots – the first suggests no change to the current model while the second shows non-constant variance and the third indicates some nonlinearity which should prompt some change in the structural form of the model.

Residual Plot for Saving Data

```
> # Residual Plot
> plot(m$fit,m$res,xlab="Fitted",ylab="Residuals",
+  pch=20,main="Residuals vs Fitted")
> abline(h=0)
```



We see a constant variance in the vertical ($\hat{\varepsilon}$) direction and the scatter is more or less symmetric about 0. No problem !

Transformation

- Transformation is needed when the errors have non-constant variance or they are not normal.
- Non-Constant Variance:

  There are two approaches to dealing with non-constant variances.

  - Weighted least squares is appropriate when the form of the non-constant variance is either known exactly or there is some known parametric form.
  - Alternatively, one can transform Y so that the variance of Y is constant.

Non-Linearity

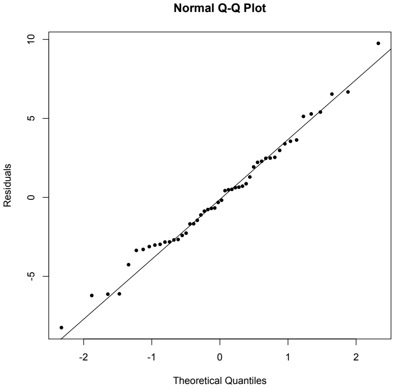
- How to check if the systematic part $(\mathrm{E}(Y) = \mathbf{X}\beta)$ is correct?
  - Regress Y and all $\mathbf{X}$ except X_i, get residuals $\hat{\delta}$. This represents Y with the other X-effect taken out.
  - Regress X_i on all $\mathbf{X}$ except X_i, get residuals $\hat{\gamma}$. This represnets X_i with the other X-effect taken out.
- There are two diagnostics plots:
  - Partial Regression or Added Variable Plot: Plot $\hat{\delta}$ against $\hat{\gamma}$.
  - Partial Residual Plot: Plot $\varepsilon + \hat{\beta}_i X_i$ against X_i.
- Non-linearity in these plots indicate that Y is not linear in X_i.

Assessing Normality

- The test and confidence intervals we use are based on the assumption of normal errors. The residuals can be assessed for normality using a *Q-Q plot*. The steps are:
    1. Sort the residuals: $\hat{\varepsilon}_{(1)} \leq \hat{\varepsilon}_{(2)} \leq \cdots \leq \hat{\varepsilon}_{(n)}$.
    2. Compute $u_i = \Phi^{-1}(\frac{i}{n+1})$.
    3. Plot $\hat{\varepsilon}_{(i)}$ against $u_{(i)}$.
- If the residuals are normally distributed an approximately straight line relationship will be observed.
- Histogram and Boxplot of the residuals are also used for checking normality, but they are not as sensitive.
- If you suspect non-normality, transformation of the response might help.

Quantile–Quantile Plot for Savings Data

```
> # QQ Plot
> qqnorm(m$res,ylab="Residuals",pch=20)
> qqline(m$res)
```


Normal Q-Q Plot
Residuals
Theoretical Quantiles