Chapter 9

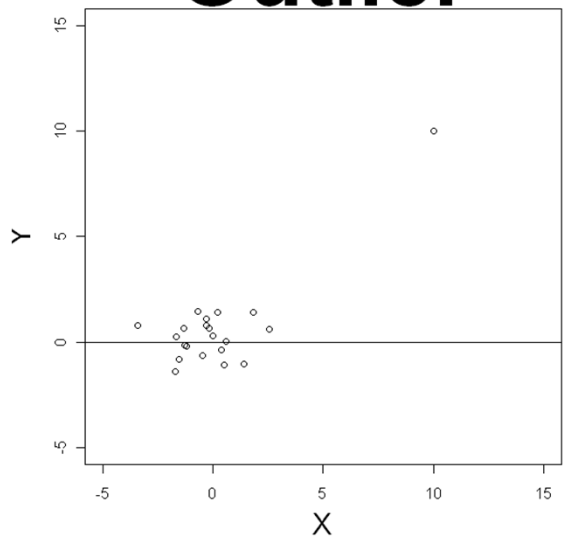
Outlier and Influence

9.1. Outlier and Influence

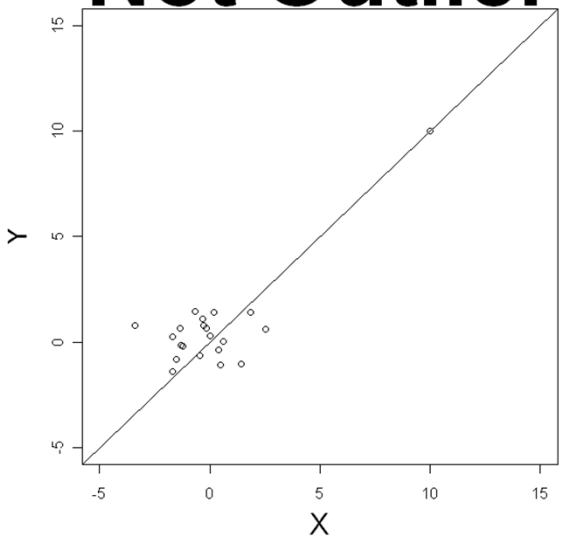
- Outlier
  - Particular case(s) that do not follow the same model as the rest of the data
  - Usually no precise definition
- In one sample distribution
  - Outlier = case of 3 s.d. away from the mean
- In regression
  - Outlier = case with large residual

9.1. Outlier and Influence

- Outlier are defined with respect to a model
  - **Outlier**



  - **Not Outlier**



  - Model: y=e
  - Model: y=a+bx+e

9.1. Outlier and Influence

- Test for outliers – Two methods
  - Propose a model for outliers, and see if the model is adequate.
    - If the model is adequate, then outlier exists
  - Compare two quantities
    - a) y_i
    - b) $\hat{y}_{i(i)}$, which is obtained by
      - Delete the i-th case, fit a regression and get $\hat{\beta}_{(i)}$
      - Predict the i-th case using the x of the i-th case: $\hat{y}_{i(i)} = x'_i \hat{\beta}_{(i)}$
    - If a) and b) differ by a lot, then outlier exists
  - Here the outlier is to be interpreted w.r.t. the regression model y=Xβ+e.

9.1. Outlier and Influence

- Method 1. Propose a model for outliers, and see if the model is adequate.
  - If we suspect that the i-th observation is an outlier,
    - Model $Y_j = X_j\beta + \delta 1_{\{j=i\}} + e_j$

      $$i.e. \begin{cases} Y_j = X_j\beta + e_j & \text{for } j \neq i \\ Y_i = X_i\beta + \delta + e_i \end{cases}$$

    - T-test for δ
      - H_0: δ=0
      - H_A: δ≠0
        - Reject H_0 → Outlier model is adequate → Outlier
    - Question: what is the d.f. of the t-distribution to use?

9.1. Outlier and Influence

- Method 2. Compare an observation to its prediction using other observations.
  - If we suspect that the i-th observation is an outlier,
    - Step 1: Delete the i-th case, fit a regression, get $\hat{\beta}_{(i)}$, $\hat{\sigma}^2_{(i)}$ and $\hat{y}_{i(i)} = x_i \hat{\beta}_{(i)}$, where

$$Y_{(i)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_{i+1} \\ \vdots \\ x_n \end{pmatrix} \beta + \begin{pmatrix} e_1 \\ \vdots \\ e_{i-1} \\ e_{i+1} \\ \vdots \\ e_n \end{pmatrix} = X'_{(i)}\beta + e_{(i)} \quad , \quad x_j = (1, x_{1j}, x_{2j}, ..., x_{pj})$$

9.1. Outlier and Influence

- Method 2. Compare an observation to its prediction using other observations.
  - Step 2: Compare y_i and $\hat{y}_{i(i)}$
    - H_0 : The i-th case is not an outlier
    - H_A : The i-th case is an outlier
      - Under H_0, $y_i - \hat{y}_{i(i)} \approx 0$
      - $Var(y_i - \hat{y}_{i(i)}) = Var(y_i) + Var(\hat{y}_{i(i)}) = \sigma^2 + Var(x_i \hat{\beta}_{(i)})$

        $= \sigma^2 + x_i Var(\hat{\beta}_{(i)}) x'_i = \sigma^2 + \sigma^2 x_i (X'_{(i)} X_{(i)})^{-1} x'_i$

      - t-stat = $\dfrac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + x_i (X'_{(i)} X_{(i)})^{-1} x'_i}}$

9.1. Example

- y=c(1,2,0,2,15,-1,1); x1=c(0,3,2,1,4,2,1); x2=c(8,7,8,9,10,8,9)
- Method 1. Propose a model for outliers, and see if the model is adequate.
  - x3=c(0,0,0,0,1,0,0)
  - fit1=lm(y~x1+x2+x3)
  - summary(fit1)
- Method 2. Propose a model for outliers, and see if the model is adequate.
  - y.d=y[-5]; x1.d=x1[-5]; x2.d=x2[-5]; X.d=cbind(1,x1.d,x2.d); x.i=c(1,x1[5],x2[5])
  - fit2=lm(y.d~x1.d+x2.d)
  - y.hat.d=x.i%*%fit2$coef
  - s=summary(fit2)$sigma
  - t.stat=(y[5]-y.hat.d)/s/sqrt(1+x.i%*%solve(t(X.d)%*%X.d)%*%x.i)
  - P.value=2*pt(-abs(t.stat),3)
- What can you observe?

9. Outlier and Influence

- Optional. Equivalence b/w two tests for outliers
  - Method 1: $Y_j = X_j\beta + \delta 1_{\{j=i\}} + e_j$

$Y_j = X_j\beta + \delta 1_{\{j=i\}} + e_j$, i.e. $Y = X'\beta + z\delta + e, \quad z = (0,0,...0,1,0...0)'$

Fit $z = X'\alpha + e$ gives $\tilde{\alpha} = (X'X)^{-1}X'z$

$$\hat{\delta} = \frac{(z - X\tilde{\alpha})'Y}{(z - X\tilde{\alpha})'(z - X\tilde{\alpha})} = \frac{z'(I-H)Y}{z'(I-H)'(I-H)z} = \frac{z'\hat{e}}{z'(I-H)z} = \frac{\hat{e}_i}{1-h_{ii}}$$

$$Var(\hat{\delta}) = \frac{\sigma^2}{(z - X\tilde{\alpha})'(z - X\tilde{\alpha})} = \frac{\sigma^2}{1-h_{ii}}$$

$$t_1 = \frac{\hat{\delta}}{\sqrt{Var(\hat{\delta})}} = \frac{\hat{e}_i}{\hat{\sigma}_f\sqrt{1-h_{ii}}}$$

$$(n-p-1)\hat{\sigma}^2 = (Y - X\tilde{\beta})^{\otimes 2} = (Y - X\hat{\beta} - z\hat{\delta})^{\otimes 2} + \hat{\delta}'(z - X\tilde{\alpha})'(z - X\tilde{\alpha})\hat{\delta}$$

$$= (n-p-2)\hat{\sigma}_f^2 + \frac{\hat{e}_i^2}{1-h_{ii}}$$

9. Outlier and Influence

- Optional. Equivalence b/w two tests for outliers
  - Method 2: $\dfrac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i (X'_{(i)} X_{(i)})^{-1} x'_i}}$

First verify that

$$(X'_{(i)} X_{(i)})^{-1} = (X'X - x_i x'_i)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} x_i x'_i (X'X)^{-1}}{1 - x'_i (X'X)^{-1} x_i}$$

$$X'_{(i)} Y_{(i)} = X'Y - x_i y_i$$

Then using the above we have

$$\hat{\beta}_{(i)} = (X'_{(i)} X_{(i)})^{-1} X'_{(i)} Y_{(i)}$$

$$= \hat{\beta} - (X'X)^{-1} x_i y_i + \frac{(X'X)^{-1} x_i x'_i \hat{\beta}}{1 - h_{ii}} - \frac{(X'X)^{-1} x_i x'_i (X'X)^{-1} x_i y_i}{1 - h_{ii}}$$

$$= \hat{\beta} - (X'X)^{-1} x_i y_i + \frac{(X'X)^{-1} x_i \hat{y}_i}{1 - h_{ii}} - \frac{(X'X)^{-1} x_i h_{ii} y_i}{1 - h_{ii}} = \hat{\beta} - \frac{(X'X)^{-1} x_i \hat{e}_i}{1 - h_{ii}}$$

9. Outlier and Influence

- Optional. Equivalence b/w two tests for outliers
  - Method 2: $\dfrac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i (X'_{(i)} X_{(i)})^{-1} x'_i}}$

From $\hat{\beta}_{(i)} = \hat{\beta} - \dfrac{(X'X)^{-1} x_i \hat{e}_i}{1 - h_{ii}}$, multiply x'_j to get

$$\begin{cases} \hat{y}_{j(i)} = \hat{y}_j - \dfrac{h_{ji}\hat{e}_i}{1 - h_{ii}} \\ \hat{e}_{j(i)} = \hat{e}_j + \dfrac{h_{ji}\hat{e}_i}{1 - h_{ii}} \end{cases}$$

In matrix notations, we write (with $z = (0,0,...,1,..0)'$, i^{th} entry $= 1$)

$$\hat{e}_{(i)} = \hat{e} + \frac{Hz\hat{e}_i}{1 - h_{ii}} = (I - H)Y + \frac{\hat{e}_i}{1 - h_{ii}} Hz$$

Therefore

$$(n - p - 2)\hat{\sigma}^2_{(i)} = \sum_{j \neq i} \hat{e}^2_{j(i)} = \hat{e}_{(i)}' \hat{e}_{(i)} - \hat{e}^2_{i(i)} = Y'(I - H)Y + \left(\frac{\hat{e}_i}{1 - h_{ii}}\right)^2 z' Hz - \left(\frac{\hat{e}_i}{1 - h_{ii}}\right)^2$$

$$= (n - p - 1)\hat{\sigma}^2 - \frac{\hat{e}^2_i}{1 - h_{ii}}$$

9. Outlier and Influence

- Optional. Equivalence b/w two tests for outliers
  - Comparing method 1 and 2, we have
    - $(n-p-2)\hat{\sigma}^2_{(i)} = (n-p-1)\hat{\sigma}^2 - \frac{\hat{e}_i^2}{1-h_{ii}}$ and $(n-p-1)\hat{\sigma}^2 = (n-p-2)\hat{\sigma}^2_f + \frac{\hat{e}_i^2}{1-h_{ii}}$

      i.e. $\hat{\sigma}^2_{(i)} = \hat{\sigma}^2_f$
    - $y_i - \hat{y}_{i(i)} = \hat{e}_{i(i)} = \hat{e}_i + \frac{h_{ii}\hat{e}_i}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}}$
    - From $(X'_{(i)}X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}}{1-x'_i(X'X)^{-1}x_i}$, we have

      $x_i(X'_{(i)}X_{(i)})^{-1}x'_i = x_i(X'X)^{-1}x'_i + \frac{x_i(X'X)^{-1}x_i x'_i (X'X)^{-1}x'_i}{1-x'_i(X'X)^{-1}x_i} = h_{ii} + \frac{h_{ii}^2}{1-h_{ii}}$
  - Therefore

$$t_2 = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1+x_i(X'_{(i)}X_{(i)})^{-1}x'_i}} = \frac{\hat{e}_i/(1-h_{ii})}{\hat{\sigma}_f\sqrt{1+h_{ii}+\frac{h_{ii}^2}{1-h_{ii}}}} = \frac{\hat{e}_i}{\hat{\sigma}_f\sqrt{1-h_{ii}}} = t_1$$

9. Outlier and Influence

- Easy formula for the outlier test
  - $$t = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{\frac{1}{2}}, \qquad \text{where} \quad r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$
  - Advantage: Only fit regression once!
  - Proof (optional)

$$t = \frac{\hat{e}_i}{\hat{\sigma}_f \sqrt{1-h_{ii}}} = \frac{\hat{e}_i}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2 - \frac{\hat{e}_i^2}{1-h_{ii}}}{(n-p-2)}}\sqrt{1-h_{ii}}}$$

$$= \frac{\hat{e}_i}{\hat{\sigma}\sqrt{\frac{(n-p-1)-r_i^2}{(n-p-2)}}\sqrt{1-h_{ii}}} = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{\frac{1}{2}}$$

9. Outlier and Influence

- Significance levels for outlier test
  - Two situations
    1. Before fitting the regression, you suspect in advance that the i-th case is an outlier. (rare)
    2. You first examine the residual plots and then suspect that the case with large residual is an outlier. (often)
  - What's the difference?
    - Let t_i be the t-stat for case-i, cv is the critical value.
    - Case 1:
      - $P(|t_i|>cv)=0.05$
    - Case 2:
      - $P(\max(|t_i|)>cv)>>>0.05$

9. Outlier and Influence

- Significance levels for outlier test
  - Usually we first examine the residual plots and then suspect that the case with large residual is an outlier.
    - Need to approximate P(max(|t_i|)>c)
  - Bonferroni adjustment (conservative)

$$P\left(\max_{i=1,..n} |t_i| > c_{\alpha/2}\right) = P\left(\bigcup_{i=1}^{n} \{|t_i| > c_{\alpha/2}\}\right) \le \sum_{i=1}^{n} P\left(|t_i| > c_{\alpha/2}\right) = n\alpha$$

    - Thus we compare max|t_i| with the α/2n quantile of t-dist
    - $$P\left(\max_{i=1,..n} |t_i| > c_{\frac{\alpha}{2n}}\right) \le n\frac{\alpha}{n} = \alpha$$
    - Equivalently, we can multiply the p-value by n.

9. Outlier and Influence

- Bonferroni adjustment (conservative)

$$P\left(\max_{i=1,..n} |t_i| > c\right) = P\left(\bigcup_{i=1}^{n} \{|t_i| > c\}\right) \le \sum_{i=1}^{n} P(|t_i| > c) = n\alpha$$

  - Thus we compare max|t_i| with the α/2n quantile of t-dist
  - Equivalently, we can multiply the p-value by n.
- Example

$$t = r_i \left(\frac{n-p-2}{n-p-1-r_i^2}\right)^{\frac{1}{2}}, \qquad \text{where} \quad r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

  - From the data with n=17, df=14, max|t_i|=12.4
    - M1) critical value = qt(1-0.05/2/17,14) = 3.593 → Reject
    - M2) p.value = 17*2*pt(-12.4,14) =1.04e-07 <0.05 →Reject
    - Therefore the case corresponds to t_i=12.4 is an outlier!

9.2 Influence analysis

- Cook's distance
- Aim: study the influence of observations
  - Key idea: for each of i, comparing
    - $\hat{\beta} = (X'X)^{-1}X'Y$ Least Square estimate
    - $\hat{\beta}_{(i)} = (X_{(i)}^{'}X_{(i)})^{-1}X_{(i)}^{'}Y_{(i)}$ LSE with the i-th observation deleted
  - Cook's distance

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

or

$$D_i = \frac{(\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}})'(\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}})}{(p+1)\hat{\sigma}^2} \qquad \begin{pmatrix} \hat{\mathrm{Y}} = \mathrm{X}\hat{\beta} \\ \hat{\mathrm{Y}}_{(i)} = \mathrm{X}\hat{\beta}_{(i)} \end{pmatrix}$$

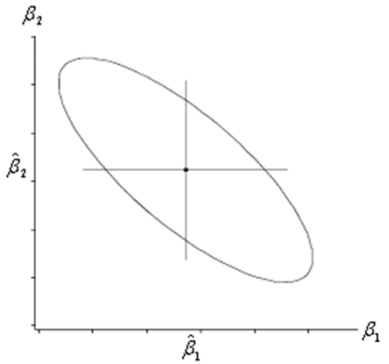
  - Interpretations
    - Normalized Distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$
    - Distance between $\hat{Y}$ and $\hat{Y}_{(i)}$
- Large Cook's distance → i-th case is influential

9.2 Interpretation of Cook's distance

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

- How large is a "large" D_i?
- Recall: (1-α) Confidence Region for $\hat{\beta}$

$$\left\{ \beta : \frac{(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})}{(p+1)\hat{\sigma}^2} \le F(\alpha, p+1, n-p-1) \right\}$$



- If $D_i = F(\alpha, p+1, n-p-1)$, it means
  - deleting the i-th case will pull the estimate of β to the edge of the (1-α)C.R. from the complete data
- In practice: compare D_i to 1
  - For most F-distribution, the 50 percentile is 1
    - qf(0.5,5,10)=0.93, qf(0.5,10,10)=1, (Try qf(0.5,m,n) for some m,n!)
  - Deleting the i-th case pull the estimate to 0.5CR's edge

9.2 Cook's distance

- Computational formula for D_i

$$D_i = \frac{(\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}})'(\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}})}{(p+1)\hat{\sigma}^2} = \left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2 \frac{h_{ii}}{(p+1)\hat{\sigma}^2}$$

- Advantage: Only fit regression once!
  - Proof (optional)
    - From p.11 we obtained

$$\hat{e}_{(i)} = \hat{e} + \frac{Hz\hat{e}_i}{1-h_{ii}}, \quad \mathrm{z} = (0,0...1,..0)' \quad [\mathrm{i}\text{ - th entry} = 1]$$

    - So

$$\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}} = \hat{e}_{(i)} - \hat{e} = \frac{Hz\hat{e}_i}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}}(h_{1i},...,h_{ii},...h_{ni})'$$

$$(\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}})'(\hat{\mathrm{Y}}_{(i)} - \hat{\mathrm{Y}}) = \left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2 \sum_i h_{1i}h_{1i} = \left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2 h_{ii} \quad (\mathrm{coz\ HH} = \mathrm{H})$$

9.2 Example

- Rat data (rat.txt), n=19
  - Study the factor affecting the amount of drug in rats' liver
- Variables
  - Y -- % of drug in the liver
  - Dose – dose of drug, range: 0-1
  - BodyWt – Body weight
  - LiverWt – Liver weight
- Regression: $Y = \beta_0 + \beta_1 Dose + \beta_2 BodyWt + \beta_3 LiverWt$
  - Outlier test:

$$t = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{\frac{1}{2}}, \qquad \text{where} \quad r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

  - Leverage:

$$h_{ii} = x'_i (X'X)^{-1} x_i, \qquad x_i = (1, x_{i1}, x_{i2}.., x_{ip})$$

  - Cook's Distance:

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{(p+1)\hat{\sigma}^2} = \left(\frac{\hat{e}_i}{1-h_{ii}} \right)^2 \frac{h_{ii}}{(p+1)\hat{\sigma}^2}$$

9.2 Example

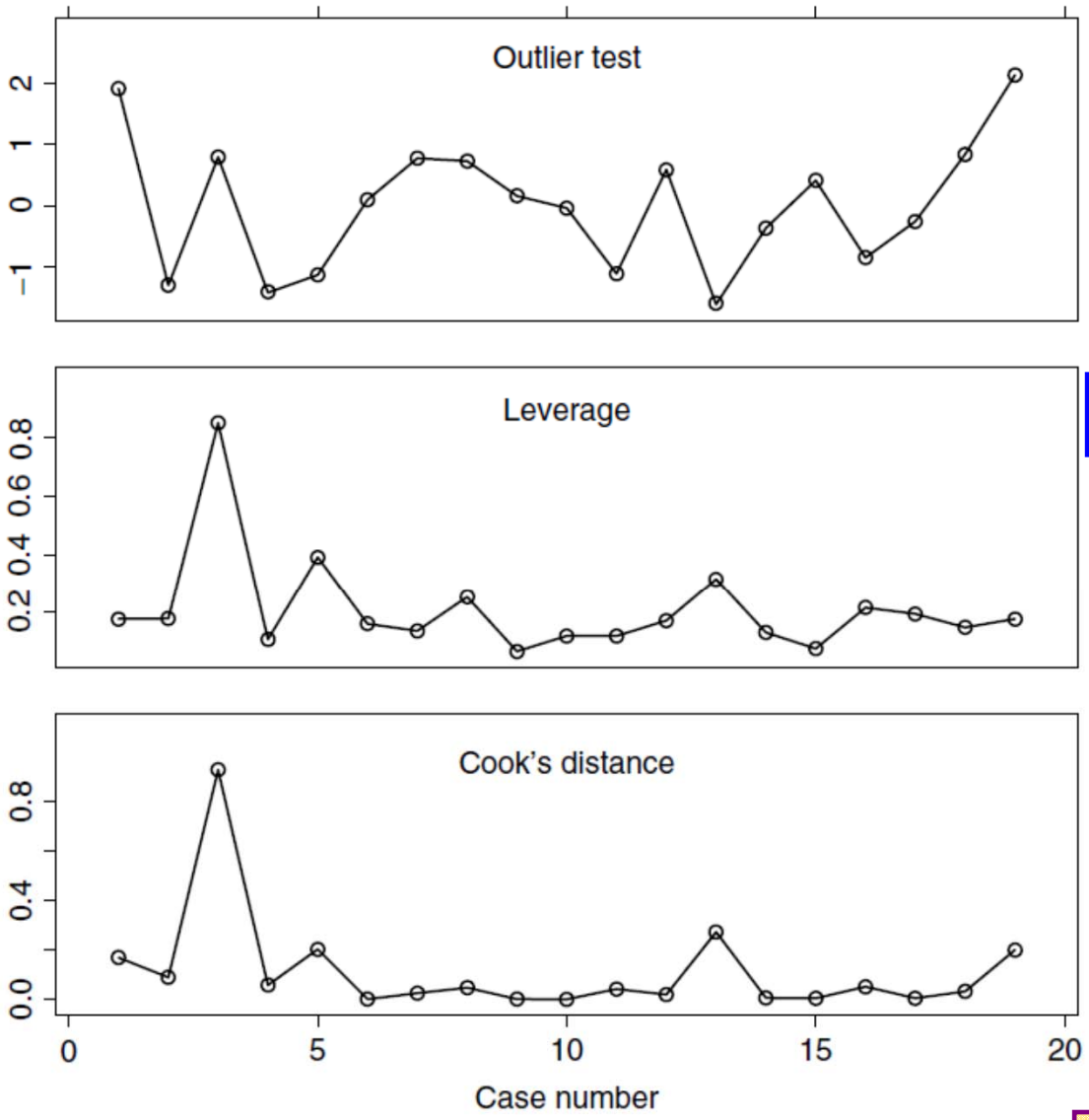


FIG. 9.3 Diagnostic statistics for the rat data.

$$t = r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{\frac{1}{2}}, \qquad \text{where} \quad r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Compare to:

qt(1-0.025/19,14) = 3.65

$$h_{ii} = x'_i (X'X)^{-1} x_i, \qquad x_i = (1, x_{i1}, x_{i2}.., x_{ip})$$

Compare to: 1

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{(p+1)\hat{\sigma}^2} = \left(\frac{\hat{e}_i}{1-h_{ii}} \right)^2 \frac{h_{ii}}{(p+1)\hat{\sigma}^2}$$

Compare to: 1

Which case is Outlier/Influential ?

9.2 Example

- Regression: Y =β0+ β1Dose + β2BodyWt + β3LiverWt

Regression:

Complete data

TABLE 9.1 Regression Summary for the Rat Data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
BodyWt      -0.021246   0.007974  -2.664   0.0177
LiverWt      0.014298   0.017217   0.830   0.4193
Dose         4.178111   1.522625   2.744   0.0151

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-Squared: 0.3639
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

Regression:

Case 3 removed

TABLE 9.2 Regression Summary for the Rat Data with Case 3 Deleted

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.311427   0.205094   1.518    0.151
BodyWt      -0.007783   0.018717  -0.416    0.684
LiverWt      0.008989   0.018659   0.482    0.637
Dose         1.484877   3.713064   0.400    0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-Squared: 0.02106
F-statistic: 0.1004 on 3 and 14 DF,  p-value: 0.9585
```

9.2 Example

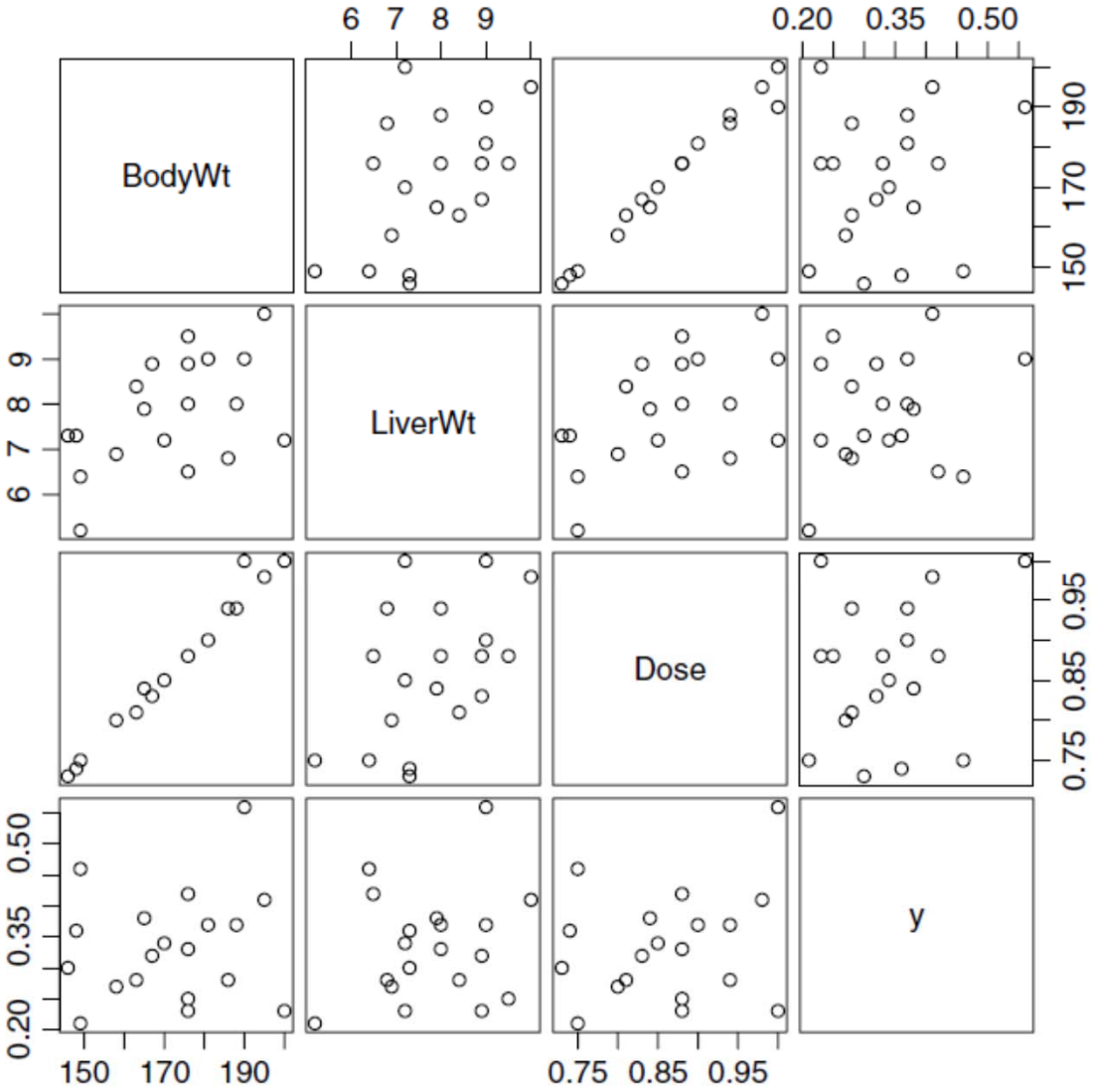


FIG. 9.2 Scatterplot matrix for the rat data.

Scatterplot matrix:

Not easy to figure out outlier

9.2 Example

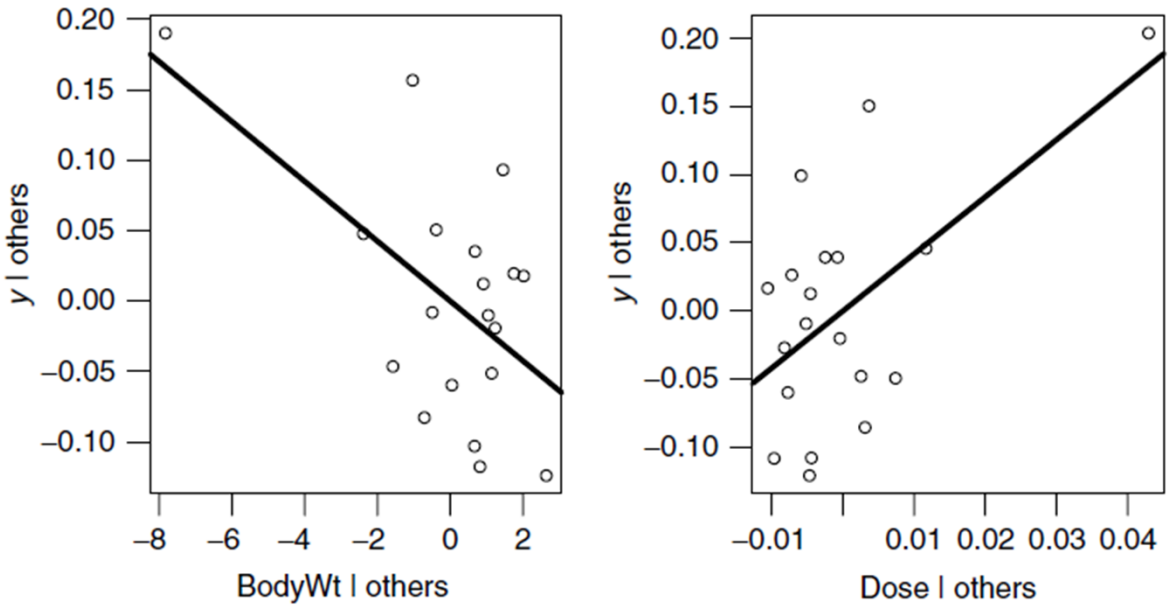


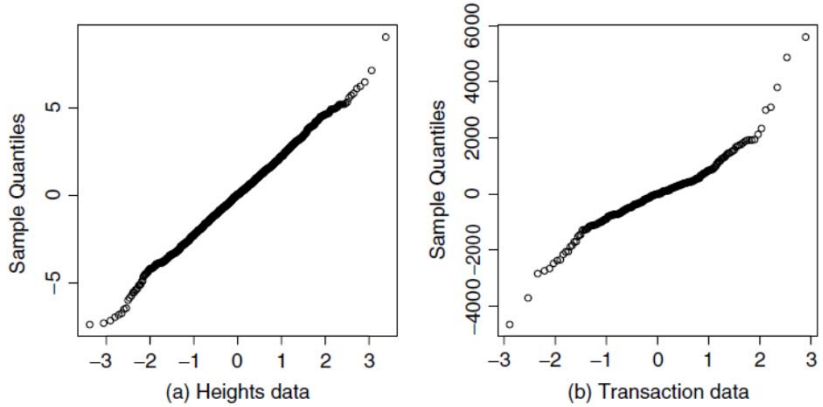
FIG. 9.4 Added-variable plots for *BodyWt* and *Dose*.

Added-variable plot:

Easier to figure out outlier

– The outlier is also influential, creating false significance in the complete data set

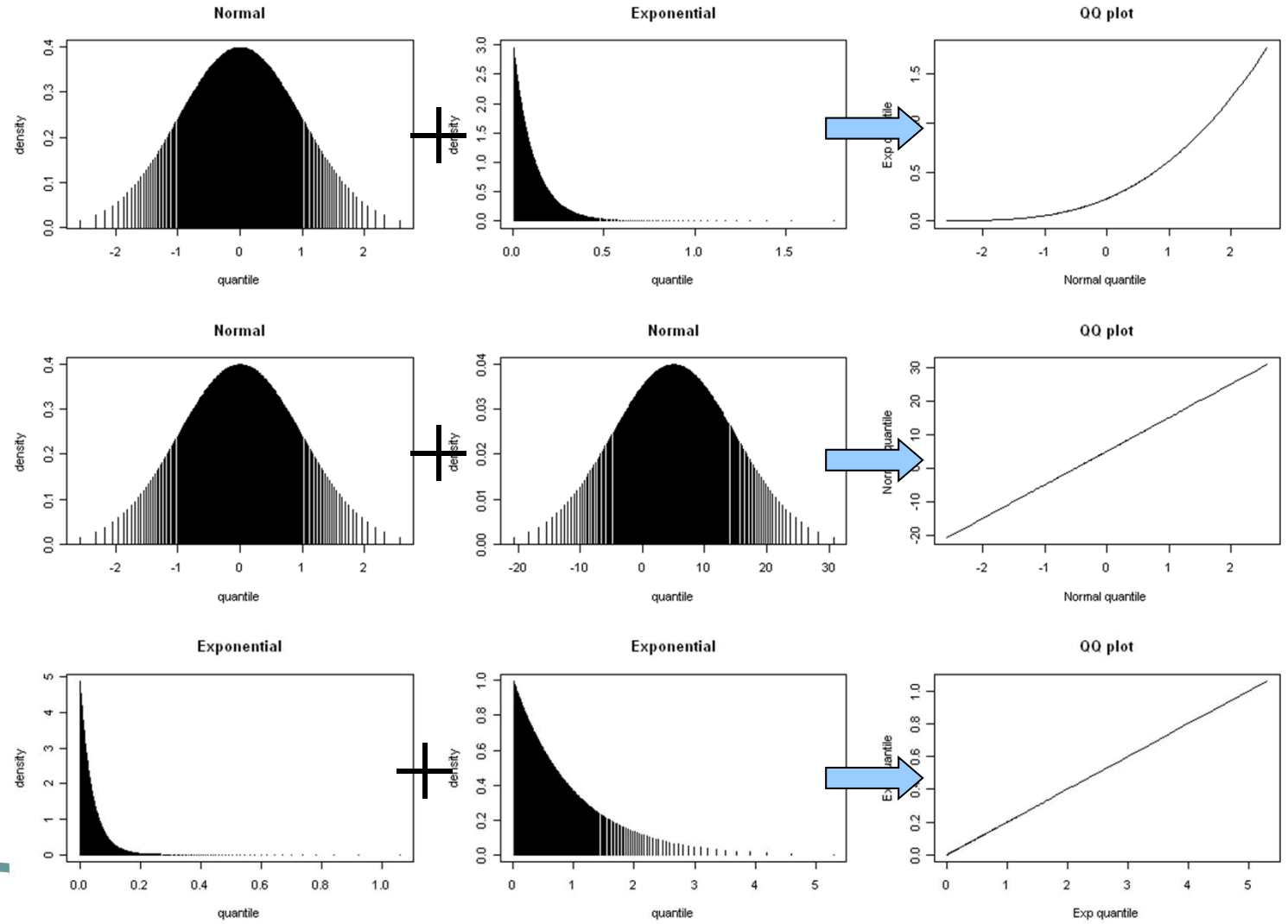
9.3 Normal Probability Plot (QQ plot)



- Aim:
  - Check the normality of residuals
  - In general, check whether the distribution of a sample agrees with an hypothetical distribution
- Idea
  - Compare the quantile of the sample distribution to the quantile of the normal distribution.

9.3 Normal Probability Plot

- Compare the quantile of two distributions

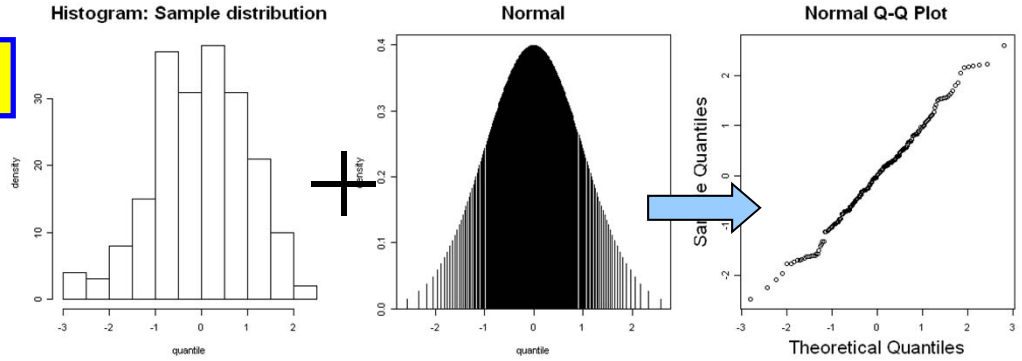


9.3 Normal Probability Plot

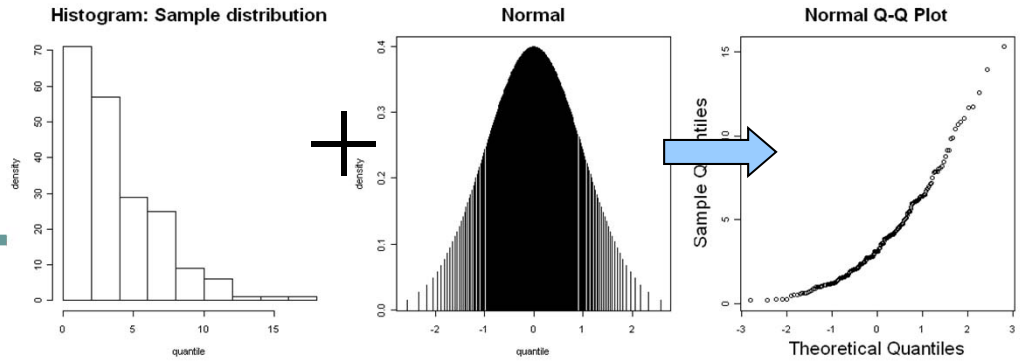
- Compare the quantile of the sample distribution and the quantile of the normal distribution

- **Sample: $x=(x_1,x_2 \ldots x_n)$**
- **Ordered sample: $x_{(1)},x_{(2)} \ldots x_{(n)}$ = quantile of sample dist: $q_{1/(n+1)}$, $q_{2/(n+1)},\ldots,q_{n/(n+1)}$**
- **Normal quantile : qnorm(1/(n+1)), qnorm(2/(n+1)), …, qnorm(n/(n+1))**
- **QQ plot = Scatterplot of the pairs ($x_{(i)}$, qnorm(i/(n+1))), all i**

R: qqnorm(x)



Straight line → Normal



Not Straight → nonNormal