Regression Diagnostics

Merlise Clyde

September 6, 2017

Outline

- Leverage

Outline

- Leverage
- Standardized Residuals

Outline

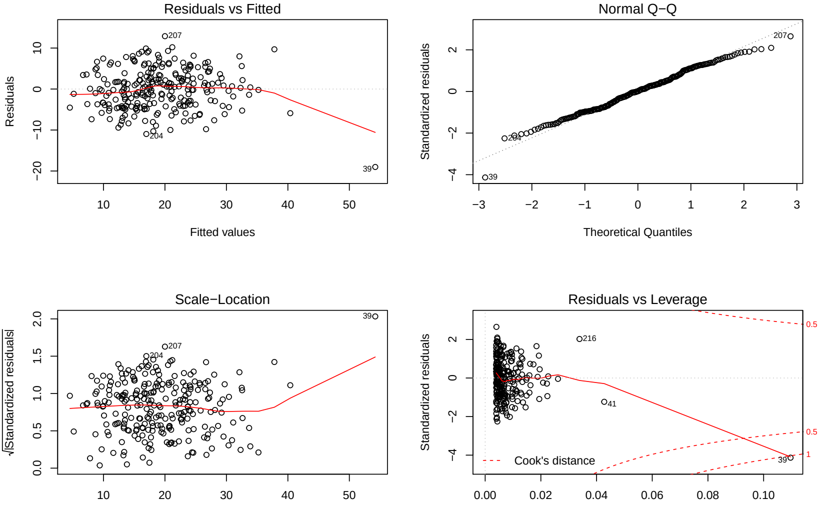
- Leverage
- Standardized Residuals
- Outlier Test

Outline

- Leverage
- Standardized Residuals
- Outlier Test
- Cook's Distance

Residual Plots

```r
bodyfat.lm = lm(Bodyfat ~ Abdomen, data=bodyfat)
par(mfrow=c(2,2))
plot(bodyfat.lm, ask=F)
```



Features of Plots

- Residuals versus fitted values

Features of Plots

- Residuals versus fitted values
- Normal Quantile: check normality of residuals or look for heavier tails than normal where observed quantiles are larger than expected under normality

Features of Plots

- Residuals versus fitted values
- Normal Quantile: check normality of residuals or look for heavier tails than normal where observed quantiles are larger than expected under normality
- Scale-Location plot:
  Detect if the spread of the residuals is constant over the range of fitted values. (Constant variance with mean)

Features of Plots

- Residuals versus fitted values
- Normal Quantile: check normality of residuals or look for heavier tails than normal where observed quantiles are larger than expected under normality
- Scale-Location plot:
  Detect if the spread of the residuals is constant over the range of fitted values. (Constant variance with mean)
- standardized residuals versus leverage with contours of Cook's distance: shows influential points where points greater than 1 or $4/n$ are considered influential

Features of Plots

- Residuals versus fitted values
- Normal Quantile: check normality of residuals or look for heavier tails than normal where observed quantiles are larger than expected under normality
- Scale-Location plot:
  Detect if the spread of the residuals is constant over the range of fitted values. (Constant variance with mean)
- standardized residuals versus leverage with contours of Cook's distance: shows influential points where points greater than 1 or $4/n$ are considered influential
- Case 39 appears to be influential and have a large standardized residual!

Hat Matrix

- predictions

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

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

Hat Matrix

- predictions
$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$
- Hat Matrix or Projection Matrix

Hat Matrix

- predictions
$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$
- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{HH} = \mathbf{H}$

Hat Matrix

- predictions

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

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

- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{H}\mathbf{H} = \mathbf{H}$
  - symmetric

Hat Matrix

- predictions

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

- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{HH} = \mathbf{H}$
  - symmetric
  - leverage values are the diagonal elements h_{ii}: $0 \leq h_{ii} \leq 1$

Hat Matrix

- predictions

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

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

- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{HH} = \mathbf{H}$
  - symmetric
  - leverage values are the diagonal elements h_{ii}: $0 \leq h_{ii} \leq 1$
- Predictions

$$\hat{Y}_i = h_{ii} Y_i + \sum_{i \neq j} h_{ij} Y_j$$

Hat Matrix

- predictions

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

- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{HH} = \mathbf{H}$
  - symmetric
  - leverage values are the diagonal elements h_{ii}: $0 \leq h_{ii} \leq 1$
- Predictions

$$\hat{Y}_i = h_{ii} Y_i + \sum_{i \neq j} h_{ij} Y_j$$

  - leverage values near 1 imply $\hat{Y}_i = Y_i$

Hat Matrix

- predictions

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

- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{HH} = \mathbf{H}$
  - symmetric
  - leverage values are the diagonal elements h_{ii}: $0 \leq h_{ii} \leq 1$
- Predictions

$$\hat{Y}_i = h_{ii}Y_i + \sum_{i \neq j} h_{ij}Y_j$$

  - leverage values near 1 imply $\hat{Y}_i = Y_i$
  - potentially influential

Hat Matrix

- predictions

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

- Hat Matrix or Projection Matrix
  - idempotent $\mathbf{HH} = \mathbf{H}$
  - symmetric
  - leverage values are the diagonal elements h_{ii}: $0 \leq h_{ii} \leq 1$
- Predictions

$$\hat{Y}_i = h_{ii}Y_i + \sum_{i \neq j} h_{ij}Y_j$$

  - leverage values near 1 imply $\hat{Y}_i = Y_i$
  - potentially influential
- Leverage: measure of how far x_i is from center of data

$$h_{ii} = 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^T \left((\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T)^T(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T)\right)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

Residual Analysis

- residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathrm{var}(e_i) = \hat{\sigma}^2(1 - h_{ii})$$

Residual Analysis

- residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathrm{var}(e_i) = \hat{\sigma}^2(1 - h_{ii})$$

- Standardized residuals:

$$r_i = e_i / \sqrt{\mathrm{var}(e_i)}$$

Residual Analysis

- residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathrm{var}(e_i) = \hat{\sigma}^2(1 - h_{ii})$$

- Standardized residuals:

$$r_i = e_i / \sqrt{\mathrm{var}(e_i)}$$

- if leverage is near 1 then residual is near 0 and variance is near 0 and r_i is approximately 0 (may not be helpful)

Predicted Residual

- Estimates without Case (i):

$$\hat{\beta}_{(i)} = (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^T \mathbf{Y}_{(i)}$$
$$= \hat{\beta} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i e_i}{1 - h_{ii}}$$

*Standardized predicted residual is

$$\frac{e_{(i)}}{\sqrt{\mathrm{var}(e_{(i)})}} = \frac{e_i/(1-h_{ii})}{\hat{\sigma}/\sqrt{1-h_{ii}}} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

these are the same as standardized residual!

Predicted Residual

- Estimates without Case (i):

$$\begin{aligned}\hat{\beta}_{(i)} &= (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} \\ &= \hat{\beta} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\end{aligned}$$

- Predicted residual

$$e_{(i)} = y_i - \mathbf{x}_i^T\hat{\beta}_{(i)} = \frac{e_i}{1-h_{ii}}$$

*Standardized predicted residual is

$$\frac{e_{(i)}}{\sqrt{\mathrm{var}(e_{(i)})}} = \frac{e_i/(1-h_{ii})}{\hat{\sigma}/\sqrt{1-h_{ii}}} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

these are the same as standardized residual!

Predicted Residual

- Estimates without Case (i):

$$\hat{\beta}_{(i)} = (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^T \mathbf{Y}_{(i)}$$
$$= \hat{\beta} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i e_i}{1 - h_{ii}}$$

- Predicted residual

$$e_{(i)} = y_i - \mathbf{x}_i^T \hat{\beta}_{(i)} = \frac{e_i}{1 - h_{ii}}$$

- variance

$$\mathrm{var}(e_{(i)}) = \frac{\sigma^2}{1 - h_{ii}}$$

*Standardized predicted residual is

$$\frac{e_{(i)}}{\sqrt{\mathrm{var}(e_{(i)})}} = \frac{e_i/(1 - h_{ii})}{\hat{\sigma}/\sqrt{1 - h_{ii}}} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

these are the same as standardized residual!

Standardized Residuals with External Estimate of σ

- Both the standardized residual and standardized predicted residual use all of the data in estimating σ

Standardized Residuals with External Estimate of σ

- Both the standardized residual and standardized predicted residual use all of the data in estimating σ
- if case i is an outlier, should also exclude it from estimating σ^2

Standardized Residuals with External Estimate of σ

- Both the standardized residual and standardized predicted residual use all of the data in estimating σ
- if case i is an outlier, should also exclude it from estimating σ^2
- Estimate $\hat{\sigma}^2_{(i)}$ using data with case i deleted

$$\mathrm{SSE}_{(i)} = \mathrm{SSE} - \frac{e_i^2}{1 - h_{ii}}$$

$$\hat{\sigma}^2_{(i)} = \mathrm{MSE}_{(i)} = \frac{\mathrm{SSE}_{(i)}}{n - p - 1}$$

Standardized Residuals with External Estimate of σ

- Both the standardized residual and standardized predicted residual use all of the data in estimating σ
- if case i is an outlier, should also exclude it from estimating σ^2
- Estimate $\hat{\sigma}^2_{(i)}$ using data with case i deleted

$$\mathrm{SSE}_{(i)} = \mathrm{SSE} - \frac{e_i^2}{1 - h_{ii}}$$

$$\hat{\sigma}^2_{(i)} = \mathrm{MSE}_{(i)} = \frac{\mathrm{SSE}_{(i)}}{n - p - 1}$$

- Externally Standardized residuals

$$t_i = \frac{e_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}} = \frac{y_i - \mathbf{x}_i^T \hat{\beta}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}} = r_i \left(\frac{n - p - 1}{n - p - r_i^2} \right)^{1/2}$$

Distribution of Externally Standardized Residuals

$$t_i = \frac{e_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1-h_{ii})}} = \frac{y_i - \mathbf{x}_i^T\hat{\beta}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1-h_{ii})}} \sim \mathrm{St}(n-p-1)$$

Outlier Test

- Regression $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \beta$

Outlier Test

- Regression $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \beta$
- Hypotheses:

Outlier Test

- Regression $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- Hypotheses:
  - H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus

Outlier Test

- Regression $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- Hypotheses:
  - H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus
  - H_a: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_i$ (different mean)

Outlier Test

- Regression $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- Hypotheses:
  - H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus
  - H_a: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_i$ (different mean)
- Show that t-test for testing H_0: $\alpha_i = 0$ is equal to t_i

Outlier Test

- Regression $\text{E}[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- Hypotheses:
  - H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus
  - H_a: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_i$ (different mean)
- Show that t-test for testing H_0: $\alpha_i = 0$ is equal to t_i
- if p-value is small declare the ith case to be an outlier: $\text{E}[Y_i]$ not given by $\mathbf{X}\beta$ but $\mathbf{X}\beta + \delta_i \alpha_i$

Outlier Test

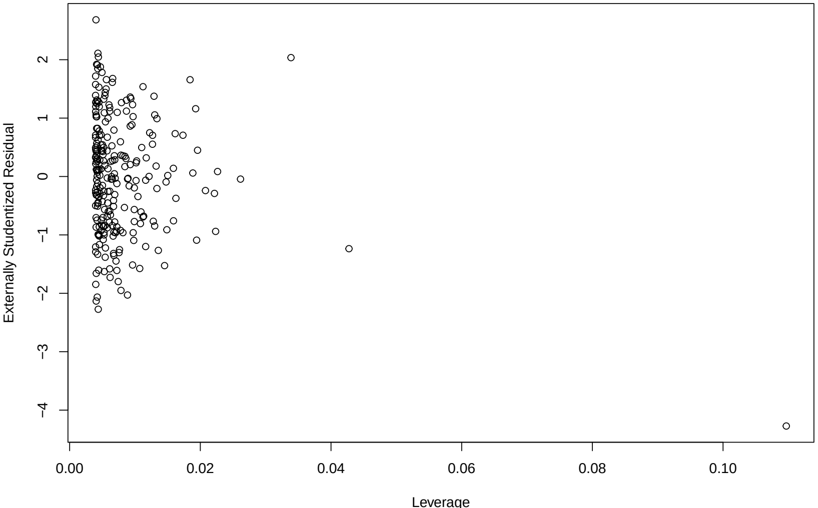
- Regression $\mathsf{E}[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- Hypotheses:
  - H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus
  - H_a: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_i$ (different mean)
- Show that t-test for testing H_0: $\alpha_i = 0$ is equal to t_i
- if p-value is small declare the ith case to be an outlier: $\mathsf{E}[Y_i]$ not given by $\mathbf{X}\beta$ but $\mathbf{X}\beta + \delta_i \alpha_i$
- Can extend to include multiple δ_i and δ_j to test that case i and j are both outliers

Outlier Test

- Regression $\mathrm{E}[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- Hypotheses:
  - H_0: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ versus
  - H_a: $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \alpha_i$ (different mean)
- Show that t-test for testing H_0: $\alpha_i = 0$ is equal to t_i
- if p-value is small declare the ith case to be an outlier: $\mathrm{E}[Y_i]$ not given by $\mathbf{X}\beta$ but $\mathbf{X}\beta + \delta_i \alpha_i$
- Can extend to include multiple δ_i and δ_j to test that case i and j are both outliers
- Extreme case $\boldsymbol{\mu} = \mathbf{X}\beta + \mathbf{I}_n \boldsymbol{\alpha}$ all points have their own mean!

R Code

```r
plot(rstudent(bodyfat.lm) ~ hatvalues(bodyfat.lm),
     ylab="Externally Studentized Residual",
     xlab="Leverage")
```



P-Value

- P-value for test that observation with largest studentized residual is an outlier

```
abs.ti = abs(rstudent(bodyfat.lm))
pval= 2*(1- pt(max(abs.ti), bodyfat.lm$df - 1))
```

P-Value

- P-value for test that observation with largest studentized residual is an outlier

```
abs.ti = abs(rstudent(bodyfat.lm))
pval= 2*(1- pt(max(abs.ti), bodyfat.lm$df - 1))
```

- Issues with multiple comparisons if we compare each p-value to $\alpha = 0.05$

P-Value

- P-value for test that observation with largest studentized residual is an outlier

```
abs.ti = abs(rstudent(bodyfat.lm))
pval= 2*(1- pt(max(abs.ti), bodyfat.lm$df - 1))
```

- Issues with multiple comparisons if we compare each p-value to $\alpha = 0.05$
- Bonferroni compares p-values to α / n

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true
- The **familywise error rate** (FWER) is the probability of rejecting at least one true H_i (making at least one type I error).

$$\begin{aligned}\text{FWER} &= P\left\{\bigcup_{i=1}^{n_0}\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq \sum_{i=1}^{n_0}\left\{P\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq n_0\frac{\alpha}{n} \leq n\frac{\alpha}{n} \\ &= \alpha\end{aligned}$$

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true
- The **familywise error rate** (FWER) is the probability of rejecting at least one true H_i (making at least one type I error).

$$\begin{aligned} \text{FWER} &= P\left\{\bigcup_{i=1}^{n_0}\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq \sum_{i=1}^{n_0}\left\{P\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq n_0\frac{\alpha}{n} \leq n\frac{\alpha}{n} \\ &= \alpha \end{aligned}$$

- This does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true.

Bonferroni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true
- The **familywise error rate** (FWER) is the probability of rejecting at least one true H_i (making at least one type I error).

$$
\begin{aligned}
\text{FWER} &= P\left\{\bigcup_{i=1}^{n_0}\left(p_i \le \frac{\alpha}{n}\right)\right\} \le \sum_{i=1}^{n_0}\left\{P\left(p_i \le \frac{\alpha}{n}\right)\right\} \le n_0\frac{\alpha}{n} \le n\frac{\alpha}{n} \\
&= \alpha
\end{aligned}
$$

- This does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true.
- Link https://en.wikipedia.org/wiki/Bonferroni_correction

Bonferroni Correction

```
abs.ti = abs(rstudent(bodyfat.lm))
pval= 2*(1- pt(abs.ti, bodyfat.lm$df - 1))
min(pval) < .05/nrow(bodyfat)
```

```
## [1] TRUE
```

```
sum(pval < .05/nrow(bodyfat))
```

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

- Bonferroni multiplicity adjustment compare each p-value to α/n and reject null (point is not an outlier) if the p-value is less than α/n

Bonferroni Correction

```
abs.ti = abs(rstudent(bodyfat.lm))
pval= 2*(1- pt(abs.ti, bodyfat.lm$df - 1))
min(pval) < .05/nrow(bodyfat)
```

```
## [1] TRUE
```

```
sum(pval < .05/nrow(bodyfat))
```

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

- Bonferroni multiplicity adjustment compare each p-value to α/n and reject null (point is not an outlier) if the p-value is less than α/n
- Start with max absolute value of t_i (or min p-value)

Bonferroni Correction

```
abs.ti = abs(rstudent(bodyfat.lm))
pval= 2*(1- pt(abs.ti, bodyfat.lm$df - 1))
min(pval) < .05/nrow(bodyfat)
```

```
## [1] TRUE
```

```
sum(pval < .05/nrow(bodyfat))
```

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

- Bonferroni multiplicity adjustment compare each p-value to α/n and reject null (point is not an outlier) if the p-value is less than α/n
- Start with max absolute value of t_i (or min p-value)
- Case 39 would be considered an outlier based on Bonferroni or other multiplicity adjustments. no other outliers

Cook's Distance

- Measure of influence of case i on predictions

$$D_i = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{(i)}\|^2}{\hat{\sigma}^2\, p}$$

after removing the ith case

Cook's Distance

- Measure of influence of case i on predictions

$$D_i = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{(i)}\|^2}{\hat{\sigma}^2 \, p}$$

  after removing the ith case

- Easier way to calculate

$$D_i = \frac{e_i^2}{\hat{\sigma}^2 p} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right],$$

$$D_i = \frac{r_{ii}}{p} \frac{h_{ii}}{1 - h_{ii}}$$

Model Assessment

- Always look at residual plots!

Model Assessment

- Always look at residual plots!
- Check constant variance, outliers, influence, normality assumption

Model Assessment

- Always look at residual plots!
- Check constant variance, outliers, influence, normality assumption
- Treat e_i as "new data" - look at structure, other predictors `avplots`

Model Assessment

- Always look at residual plots!
- Check constant variance, outliers, influence, normality assumption
- Treat e_i as "new data" - look at structure, other predictors `avplots`
- Case 39 looks an influential outlier!

Model Assessment

- Always look at residual plots!
- Check constant variance, outliers, influence, normality assumption
- Treat e_i as “new data” - look at structure, other predictors `avplots`
- Case 39 looks an influential outlier!
- Impact on predictions?

Predictions with Case 39

```r
predict(bodyfat.lm, newdata=bodyfat[39,],
        se=T,  interval="prediction")
```

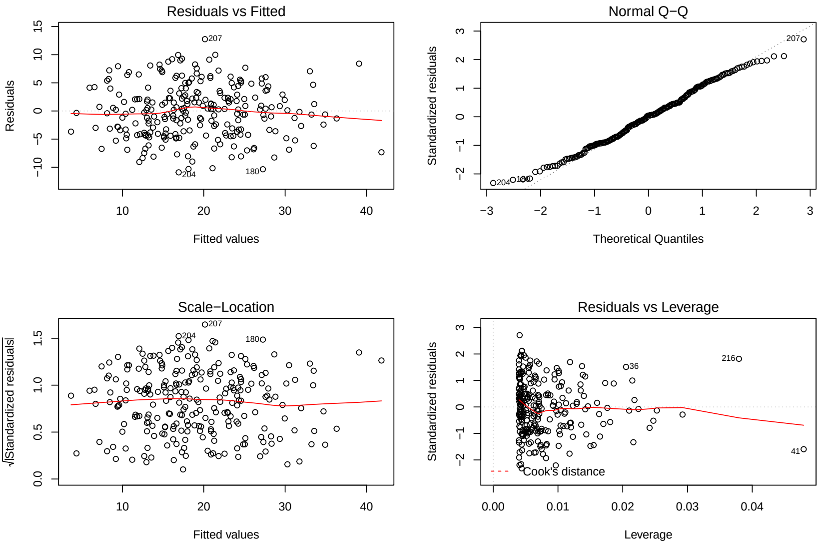
```
## $fit
##         fit     lwr      upr
## 39 54.21599 44.0967 64.33528
##
## $se.fit
## [1] 1.615311
##
## $df
## [1] 250
##
## $residual.scale
## [1] 4.877484
```

Predictions without Case 39

```
bodyfatsub.lm = lm(Bodyfat ~ Abdomen, data=bodyfat,
                   subset=c(-39))
predict(bodyfatsub.lm, newdata=bodyfat[39,],
        se=T, interval="prediction")
```

```
## $fit
##         fit      lwr      upr
## 39 56.55856 46.71172 66.40541
##
## $se.fit
## [1] 1.655744
##
## $df
## [1] 249
##
## $residual.scale
## [1] 4.717441
```

Residual Checks without Case 39



How should we proceed?

- Reproducible Research - Document removing a case

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!
- Remove statistically significant outliers if you cannot conform other data entry errors, etc

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!
- Remove statistically significant outliers if you cannot conform other data entry errors, etc
- Influential points (not outliers): report analysis with & without

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!
- Remove statistically significant outliers if you cannot conform other data entry errors, etc
- Influential points (not outliers): report analysis with & without
- If we remove Case 39, are there other outliers or influential points?

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!
- Remove statistically significant outliers if you cannot conform other data entry errors, etc
- Influential points (not outliers): report analysis with & without
- If we remove Case 39, are there other outliers or influential points?
- Model Uncertainty (more later)

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!
- Remove statistically significant outliers if you cannot conform other data entry errors, etc
- Influential points (not outliers): report analysis with & without
- If we remove Case 39, are there other outliers or influential points?
- Model Uncertainty (more later)
- Robust Models (more later)

How should we proceed?

- Reproducible Research - Document removing a case
- Adjust for multiple testing!
- Remove statistically significant outliers if you cannot conform other data entry errors, etc
- Influential points (not outliers): report analysis with & without
- If we remove Case 39, are there other outliers or influential points?
- Model Uncertainty (more later)
- Robust Models (more later)
- Next: Transformations