Chapter 6: Diagnostics

Diagnostics

- Checking error assumptions
  - Linearity
  - Normality
  - Constant variance
  - Independent predictors
- Finding unusual points (outlier, leverage)

Checking Error Assumptions

Assumption made so far: $\epsilon \sim N(0, \sigma^2 I)$

This includes

- $\mathbb{E}(\epsilon) = 0$
- $\text{Var}(\epsilon) = \sigma^2 I$
- ϵ's are independent, identically distributed, normal

Graphical and numerical diagnostic methods

Graphical methods

- Scatter Plot
- Residual vs. Fitted plot
- Normal QQ - plot
- Cook's Distance plot
- Standardized Residual vs. Fitted plot

Residual vs. Fitted Plot

Plot $\hat{\epsilon}$ against $\hat{y}$. Can show

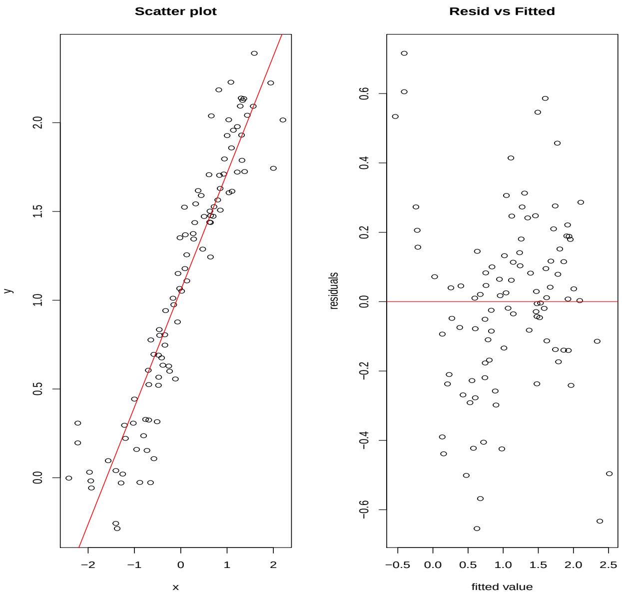
- Homoscedasticity (constant variance)
- Heteroscedasticity (non-constant variance)
- Non-linearity

Residual vs. Fitted Plot

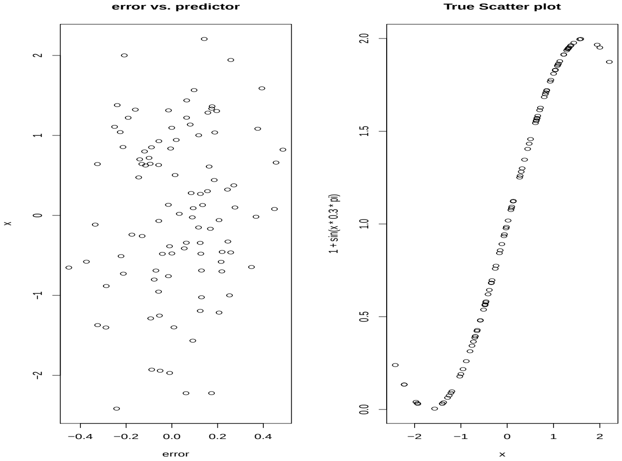
Limitations

- The Residual vs. Fitted Plot is very useful in general.
- It is sometimes NOT enough.
  - Big errors
  - Combination of assumption violations
  - No threshold

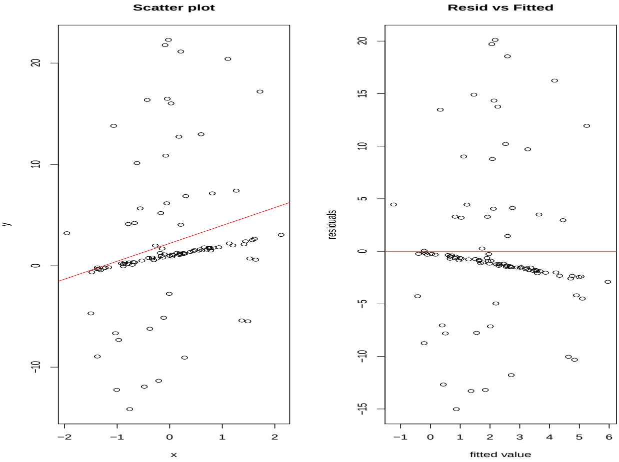
Residual vs. Fitted Plot:



Residual vs. Fitted Plot: Non-linearity

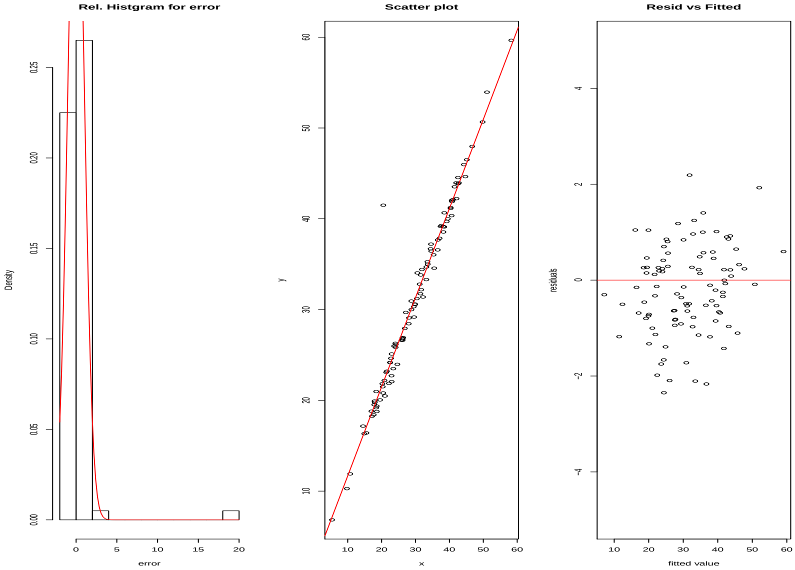


Residual vs. Fitted Plot:

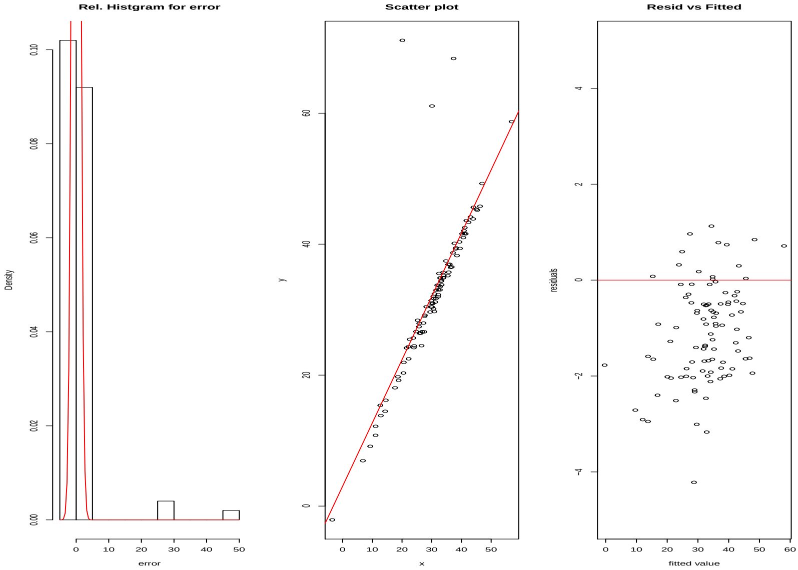


- What is the problem?

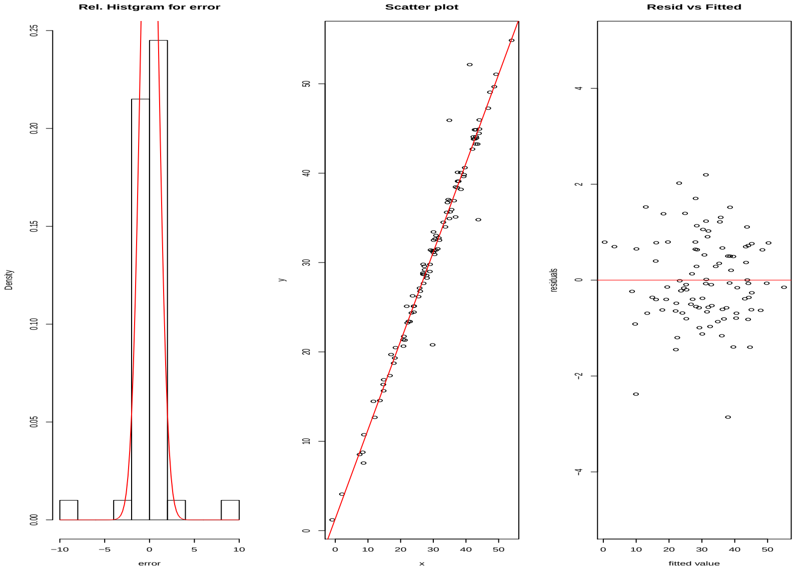
Residual vs. Fitted Plot: Unusual point, Outlier



Residual vs. Fitted Plot



Residual vs. Fitted Plot



Remaining Question

- How to determine if there are outliers
  - Studentized (Standardized) Residual
  - Cook's Distance
- How to fix the assumptions

What to Do

- Non-constant variance
  - Nothing
  - Weighted least squares (Ch 8.2)
  - Transformation of the response (Ch 9.1, 9.2)
- Nonlinearity: change the model (e.g., polynomial model (Ch 9.4))
- Unusual point: either remove it or do nothing

Checking Constant Variance: Savings Example

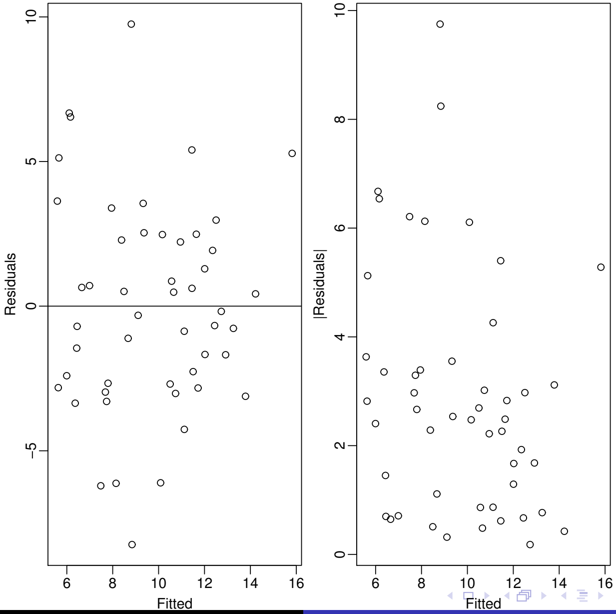
- 50 different countries, 1960 – 1970
- Response: aggregate personal saving divided by disposable income (*sr*)
- Predictors: per capital disposable income (*dpi*), percentage rate of change in per capita disposable income (*ddpi*), percentage of population under 15 (*pop15*), percentage of population over 75 (*pop75*)

```
> data(savings)
> result <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
   savings)
```

Savings Example Ctd

```
> plot(result, which = 1)
> plot(result$fitted.values, abs(result$residuals),
xlab = "Fitted", ylab = "|residuals|")
```

Savings Example Ctd



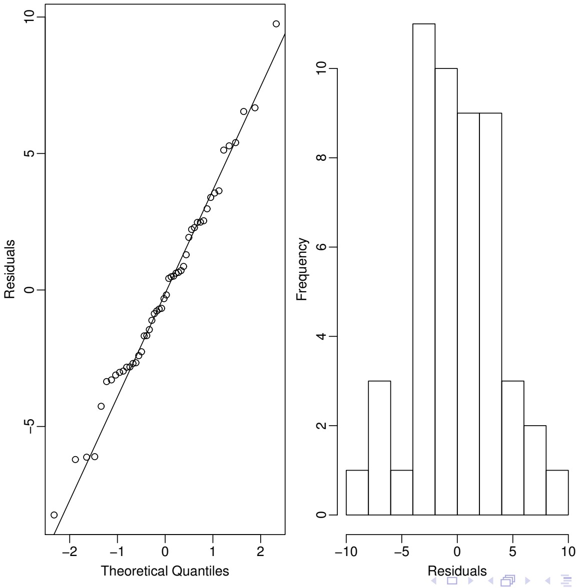
Checking Normality

QQ-plot

1. Sort the residuals $\hat{\epsilon}_{[1]} \leq \hat{\epsilon}_{[2]} \cdots \leq \hat{\epsilon}_{[n]}$
2. Compute $u_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$
3. Plot $\hat{\epsilon}_{[i]}$ against u_i.

```
## QQ-plot
> qqnorm(result$residual, ylab="Residuals")
> qqline(result$residual)
## Histogram
> hist(result$residual, xlab="Residuals")
```

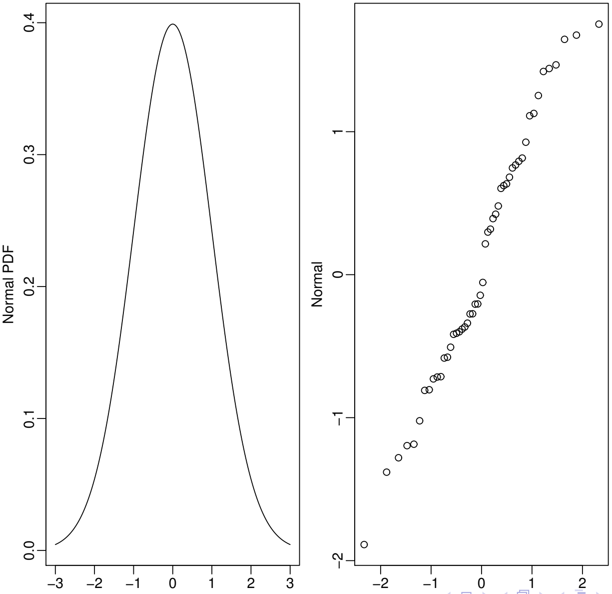
QQ-plot Example



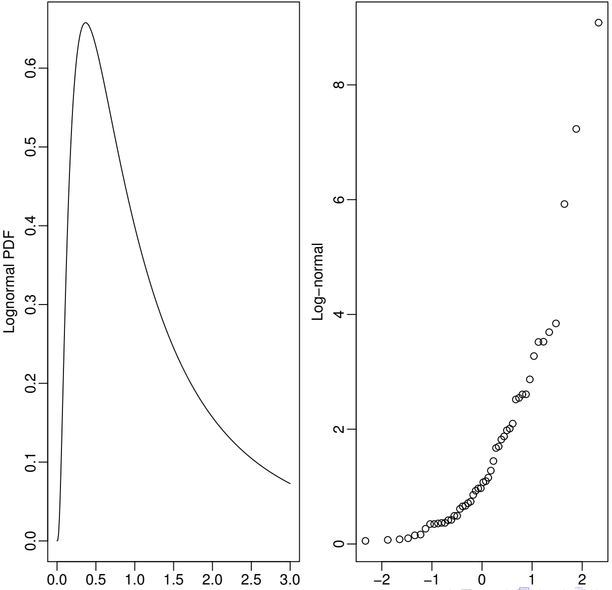
Non-Normality

- Skewed distribution (e.g., log-normal)
- Long-tailed distribution (e.g., Cauchy)
- Short-tailed distribution (e.g., uniform)

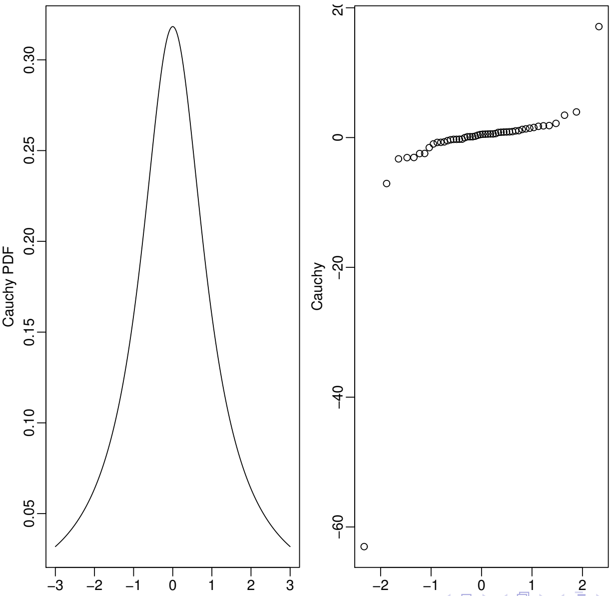
QQ-plot of Normal



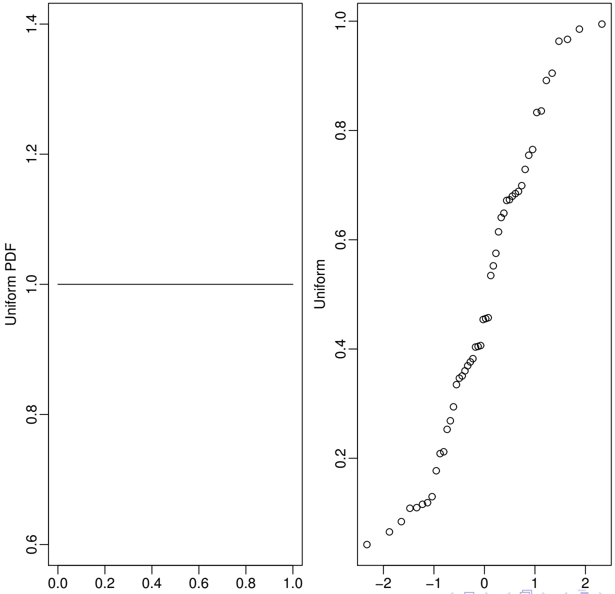
QQ-plot of Log-normal



QQ-plot of Cauchy



QQ-plot of Uniform



Shapiro-Wilk test for normality

```
> shapiro.test(result$residual)
        Shapiro-Wilk normality test
data:  result$residual
W = 0.987, p-value = 0.8524
```

H_0 : samples are normally distributed $\quad H_A : H_0$ is not true.

Shapiro-Wilk test for normality

Not very helpful (QQ plots are better).

- Small n – little power
- Large n – non-normality is less important

What to do about non-normal errors

- Transformation of the response (Ch 9)
- Robust methods (long-tailed distribution (difficult))
- Nothing

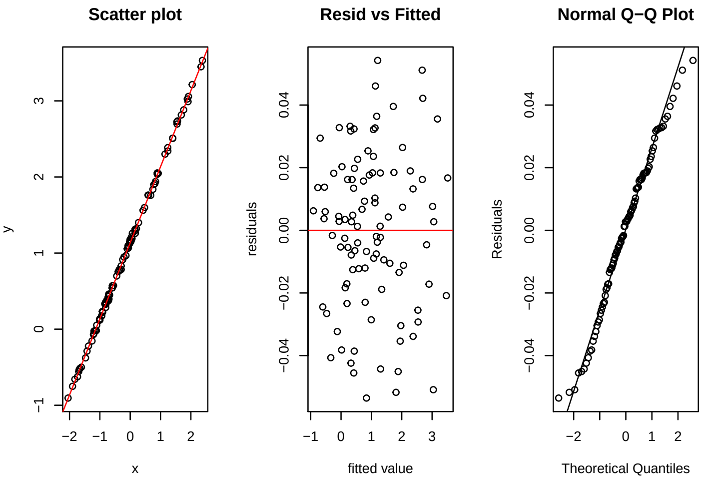
Correlated Errors

Temporally related data

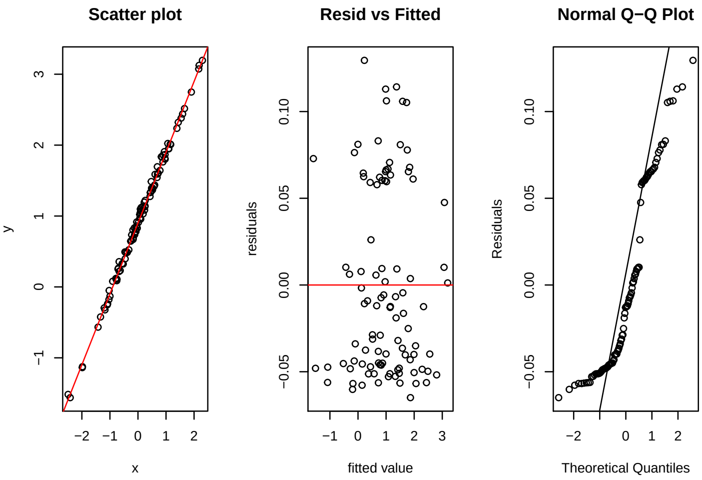
- Plot $\hat{\epsilon}$ against time
- Plot $\hat{\epsilon}_i$ against $\hat{\epsilon}_{i-1}$
- Time series analysis may be more appropriate

No temporal relationship or other ordering in the variables $\Rightarrow$ checking independence is very hard.

Correlated Errors (1)



Correlated Errors (2)



Correlated Errors: Globwarm Example

- 145 samples , 1856 – 2000
- Response: temperature (nhtemp)
- Predictors: wusa, jasper, westgreen, chesapeake, tornetrask, urals, mongolia, tasman,
- 'Year' is not considered as a predictor

```
> data(globwarm)
> head(globwarm)
     nhtemp  wusa jasper westgreen chesapeake tornetrask urals mongolia tasman year
1000     NA -0.66  -0.03      0.03      -0.66       0.33 -1.49     0.83  -0.12 1000
1001     NA -0.63  -0.07      0.09      -0.67       0.21 -1.44     0.96  -0.17 1001
1002     NA -0.60  -0.11      0.18      -0.67       0.13 -1.39     0.99  -0.22 1002
1003     NA -0.55  -0.14      0.30      -0.68       0.08 -1.34     0.95  -0.26 1003
1004     NA -0.51  -0.15      0.41      -0.68       0.06 -1.30     0.87  -0.31 1004
1005     NA -0.47  -0.15      0.52      -0.68       0.07 -1.25     0.77  -0.37 1005
```

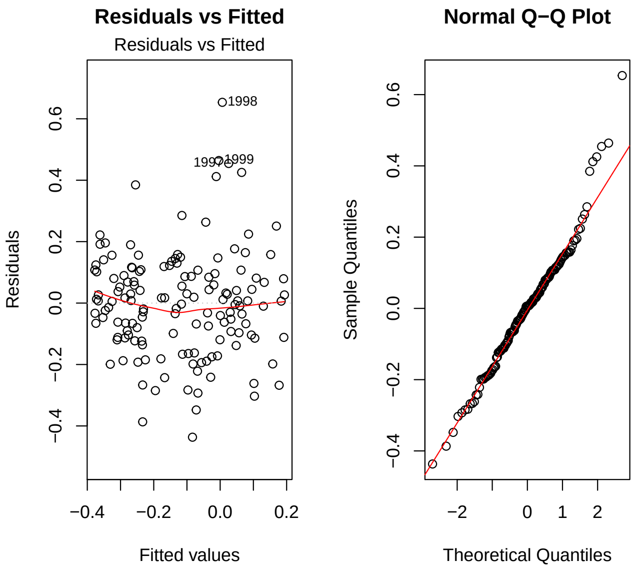
Correlated Errors: Globwarm Example

```
## fitting model
> result = lm( nhtemp ~ wusa+  jasper+ westgreen+ chesapeake
    + tornetrask+ urals+ mongolia+ tasman, data = globwarm)

## fitted model
> summary(result)
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.242555   0.027011  -8.980 1.97e-15 ***
wusa         0.077384   0.042927   1.803 0.073647 .
jasper      -0.228795   0.078107  -2.929 0.003986 **
westgreen    0.009584   0.041840   0.229 0.819168
chesapeake  -0.032112   0.034052  -0.943 0.347346
tornetrask   0.092668   0.045053   2.057 0.041611 *
urals        0.185369   0.091428   2.027 0.044567 *
mongolia     0.041973   0.045794   0.917 0.360996
tasman       0.115453   0.030111   3.834 0.000192 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Correlated Errors: Globwarm Example

```
## Diagnotic Plots
> par(mfrow = c(1,2))
> plot(result, which = 1, main = "Residuals vs Fitted")
> qqnorm(residuals(result))
> qqline(residuals(result), col ="red")
```

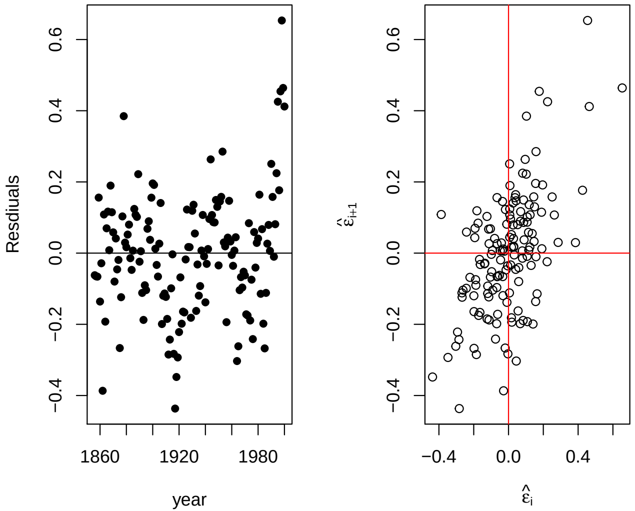
Correlated Errors: Globwarm Example



Correlated Errors: Globwarm Example

```
## Scatter Plots: Residual vs Year and i th Residual vs i+1 th R
> n = length(residuals(result))
> plot(residuals(result) ~year, na.omit(globwarm), ylab ="Resdiu
> abline(h= 0)
> plot(tail(residuals(result), n-1) ~head(residuals(result), n-1
> abline(h=0, v=0, col = "red")
```

Correlated Errors: Globwarm Example



Durbin-Watson test for Correlated Errors

H_0 : the errors are uncorrelated *vs.* $H_A : H_0$ is not true.

```
## load library
> require(lmtest)

## Durbin-Watson test
> dwtest( nhtemp ~ wusa+  jasper+ westgreen+ chesapeake
   + tornetrask+ urals+ mongolia+ tasman, data = globwarm)

Durbin-Watson test

DW = 0.81661, p-value = 1.402e-15
alternative hypothesis: true autocorrelation is greater than 0
```

Correlated Errors: Quadratic Relationship Example

Suppose that there is quadratic relationship between a predictor and the response.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon.$$

However we consider the simple linear model (missing predictor).

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

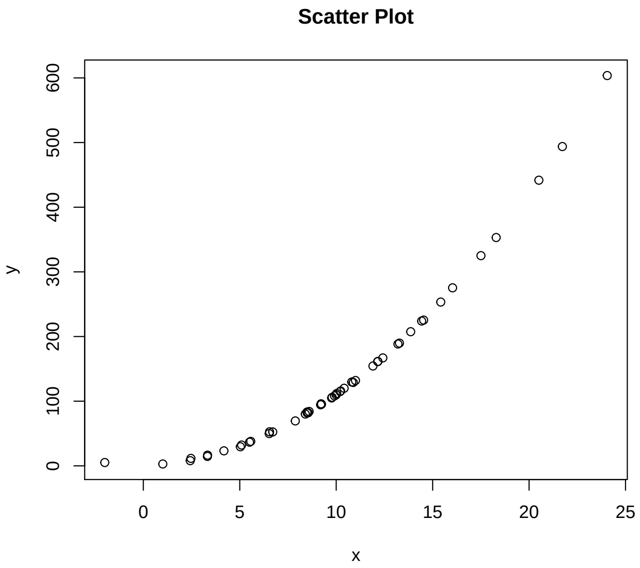
Correlated Errors: Quadratic Relationship Example

```
## Sample Size
n = 50

## Predictor
x = rnorm(n, 10, 5)
z = x^2

## Reponse Variable
y = 1 + x + z + rnorm(n, 0, 1)

## Fitting Model (Missing z variable)
result= lm(y ~ x)
summary(result)
```

Correlated Errors: Quadratic Relationship Example

```
> summary(result)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -91.633     12.742  -7.192 3.74e-09 ***
x             22.936      1.135  20.209  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 41.49 on 48 degrees of freedom
Multiple R-squared:  0.8948,Adjusted R-squared:  0.8926
F-statistic: 408.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

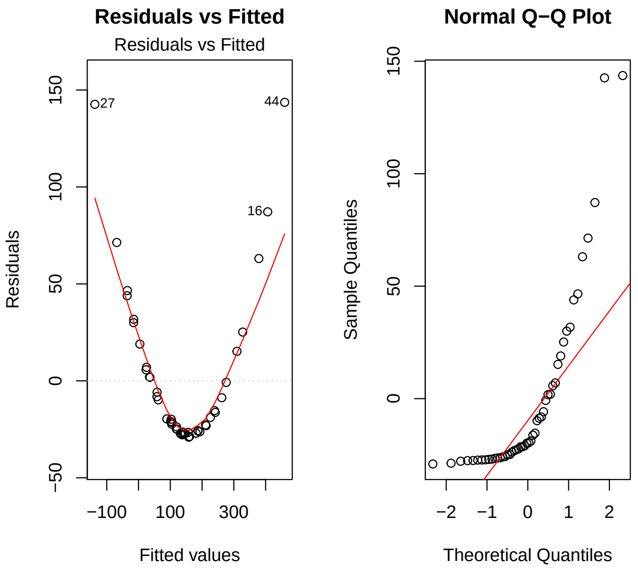
Correlated Errors: Quadratic Relationship Example



Correlated Errors: Quadratic Relationship Example

```
> par(mfrow = c(1,2))
> plot(result, which = 1, main = "Residuals vs Fitted")
> qqnorm(residuals(result))
> qqline(residuals(result), col ="red")
```

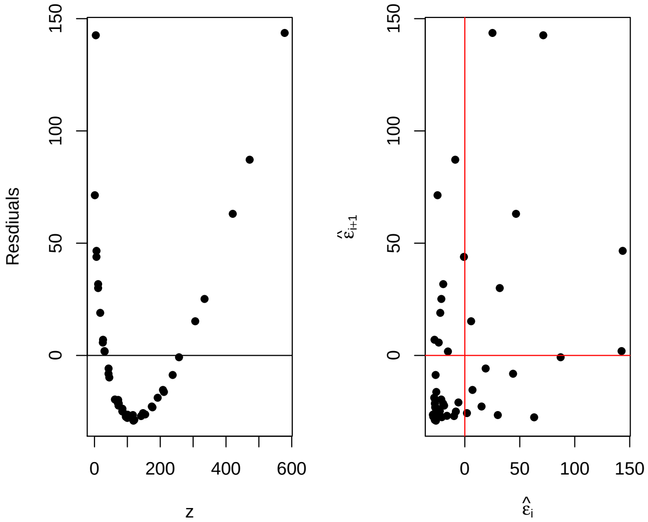
Correlated Errors: Quadratic Relationship Example



Correlated Errors: Quadratic Relationship Example

```
> par(mfrow = c(1,2))
> n = length(residuals(result))
> plot(residuals(result) ~ z, ylab ="Resdiuals", pch = 16 )
> abline(h= 0)
> plot(tail(residuals(result), n-1) ~head(residuals(result), n-1
    , xlab = expression(hat(epsilon)[i])
    , ylab = expression(hat(epsilon)[i+1]), pch = 16 )
> abline(h=0, v=0, col = "red")
```

Correlated Errors: Quadratic Relationship Example



Durbin-Watson test for Correlated Errors

H_0 : the errors are uncorrelated *vs.* $H_A : H_0$ is not true.

```
## Durbin-Watson test
> dwtest( y ~ x)


Durbin-Watson test


data:  y ~ x
DW = 1.2681, p-value = 0.003285
alternative hypothesis: true autocorrelation is greater than 0
```

Studentized Residuals

Since $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$ where $h_i = H_{ii}$, let

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

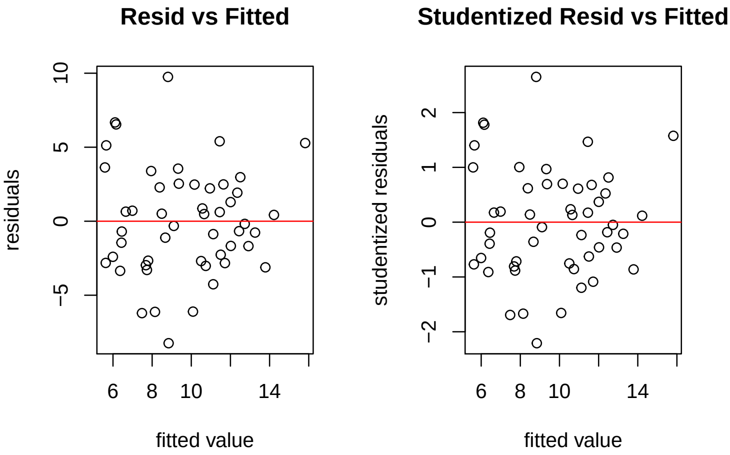
These are called (internally) studentized residuals

- It is better to use studentized residuals for diagnostic plots (QQ-plot and testing constant variance)
- In practice, usually little difference

Savings Example

```
## Compute studentized residuals
> result.s <- summary(result)
> sigma.s <- result.s$sig
> hat.s <- lm.influence(result)$hat
> stud.res <- result$residuals/(sigma.s * sqrt(1-hat.s))
> par(mfrow = c(1,2))
> plot(result$fitted.values, result$residuals, xlab = "fitted va
ylab = "residuals", main = "Resid vs Fitted")
> abline(h = 0, col = "red")
> plot(result$fitted.values, stud.res, xlab = "fitted value",
ylab = "studentized residuals", main = "Studentized Resid vs Fit
>abline(h = 0, col = "red")
```

Studentized Residual vs. Fitted Plot



Studentized Residuals

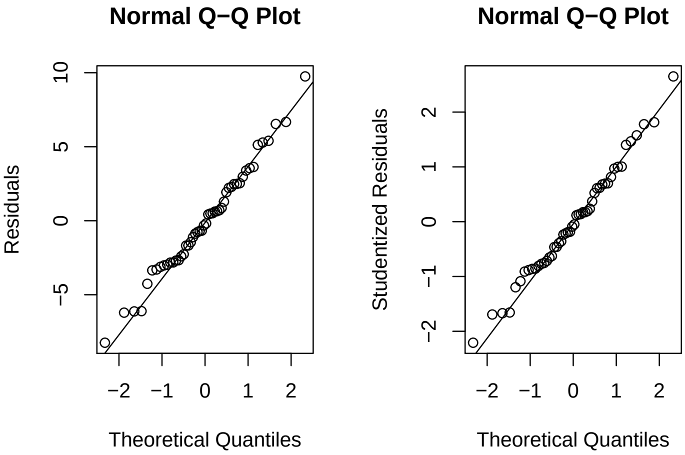
If absolute values of studentized residuals are greater than 2.5, they are more likely to be unusal points.

```
> which.max( abs(stud.res) )
Zambia
46
> stud.res[46]
Zambia
2.650915
```

Savings Example

```
> qqnorm(result$residual, ylab="Residuals")
> qqline(result$residual)
> qqnorm(stud.res, ylab="Studentized Residuals")
> qqline(stud.res)
```

Studentized Residual QQ Plot



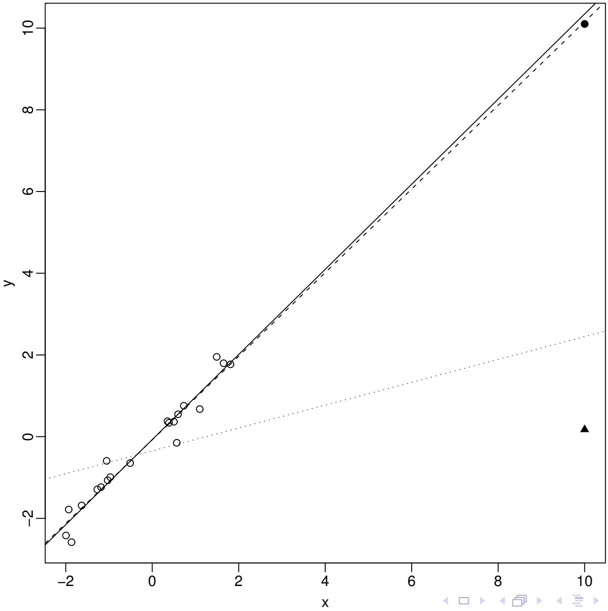
Finding Unusual Points

1. Outliers – do not fit the model well
2. Leverage – extreme in the predictor space, but not necessarily influence the fit

A point can be none, one, or both of these.

Influential points – affect the fit of the model substantially

Which Point is an Outlier?



Leverage

Recall the hat matrix $H = X(X^TX)^{-1}X^T$.

Leverage of point i: $h_i = H_{ii}$.

- h_i depends only on X
- $var(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$
- $\sum_i h_i = p + 1$
- Average of h_i is $\frac{p+1}{n}$

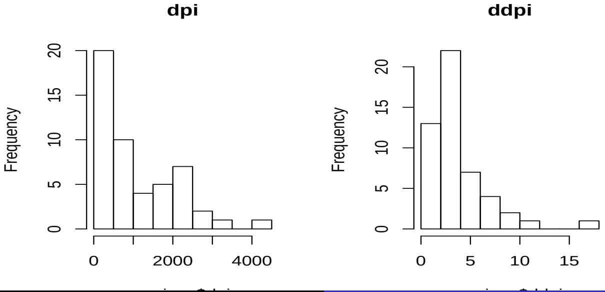
Rule of thumb: Leverages greater than $2 \times \frac{p+1}{n}$ are considered high.

Savings Example

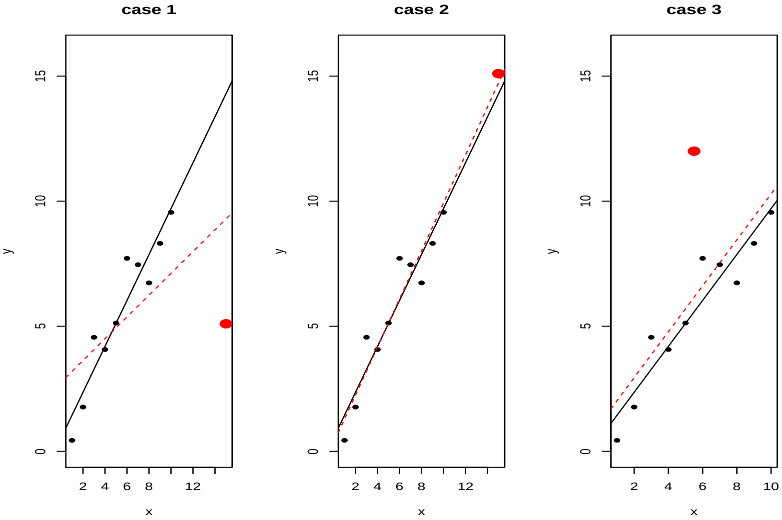
```
> hat.s <- lm.influence(result)$hat
> 2 * (4 + 1)/50
[1] 0.2
> which(hat.s > 0.2)
Ireland          Japan United States          Libya
21            23            44            49
```

Savings Example

```
> savings$dpi[c(44)]
[1] 4001.89
> savings$ddpi[c(23, 49)]
[1]  8.21 16.71
> par(mfrow= c(1,2))
> hist(savings$dpi, main ="dpi")
> hist(savings$ddpi, main ="ddpi")
```



Examples: Outliers



Outliers

How do we distinguish between truly unusual points and large residuals?

Compare two models:

Model 1: $Y = X\beta + \epsilon$, vs. Model 2: $Y_{-i} = X_{-i}\beta + \epsilon'$,

$\iff$

Compare residuals for i^{th} observation.

Outliers

How do we distinguish between truly unusual points and large residuals?

- Exclude point i, recompute $\hat{\beta}_{(i)}$ and $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$.
- If $|y_i - \hat{y}_{(i)}|$ is large, then observation i is an outlier; but how large is large?

Outliers

Note that

- $$\widehat{\text{var}}(\hat{y}_{(\text{pred})}) = \hat{\sigma}\sqrt{1 + x_{new}^T \left(X^T X\right)^{-1} x_{new}}$$

- $$\widehat{\text{var}}(y_i - \hat{y}_{(i)}) = \hat{\sigma}_{(i)}\sqrt{1 + x_i^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_i}$$

- $$\frac{\hat{\cdot} - \cdot}{\hat{se}(\cdot)} \sim t_{n-(p+1)}.$$

Externally Studentized Residuals

It turns out

$$
\begin{aligned}
t_i &= \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + x_i^T \left(X_{(i)}^T X_{(i)} \right)^{-1} x_i}} \\
&\sim t_{(n-1)-(p+1)}
\end{aligned}
$$

The book also calls these jackknife or cross-validated residuals.

Multiple Hypothesis Tests

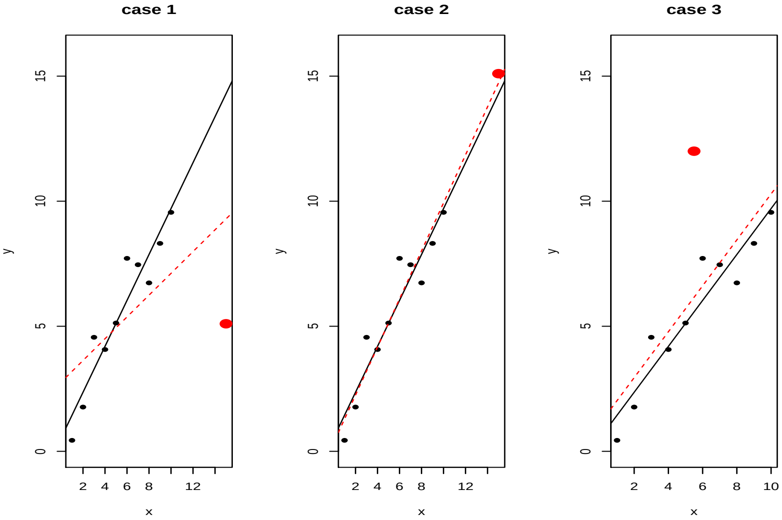
- If $|t_i|$ is too large, reject and conclude observation i is an outlier (because p-value is $P(|t_{(n-1)-(p+1)}| > |t_i|)$.
- For each observation i, compare $|t_i|$ with $t^{\alpha/2}_{n-(p+1)-1}$.
- Will fail to reject too many points. Why?

Bonferroni Correction

$$\begin{aligned} \text{Type I Error} &= Pr_{H_0}(\text{reject at least one test}) \\ &\leq \sum_i Pr_{H_0}(\text{reject test } i) \\ &= n\alpha \end{aligned}$$

Bonferroni correction: test each hypothesis at level α/n

Examples: Outliers



Case 1

```
## Externally Studentized Residuals
> rstudent(res1)
> round( rstudent(res1), 3)
     1      2      3      4      5      6      7      8      9     10     11
-1.520 -0.914  0.224 -0.193  0.084  1.097  0.747  0.217  0.756  1.214 -6.356

## P-values
> round( 2 * ( pt( abs(rstudent(res1)), 11 - 1 - 2, lower.tail = F)), 3)
    1     2     3     4     5     6     7     8     9    10    11
0.167 0.387 0.828 0.852 0.935 0.305 0.476 0.834 0.471 0.259 0.000
```

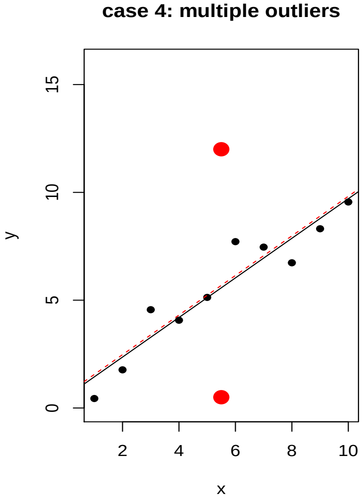
Case 2

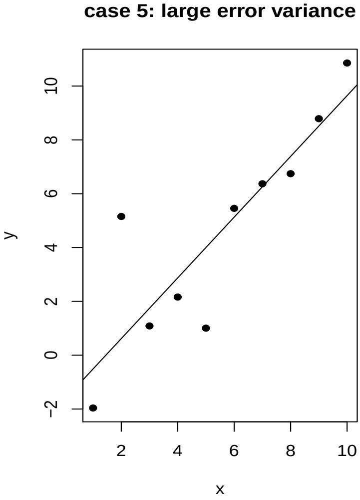
```
## Externally Studentized Residuals
> round( rstudent(res1), 3)
     1      2      3      4      5      6      7      8      9     10     11
-1.066 -0.552  1.732 -0.108 -0.002  2.150  0.437 -1.547 -0.735 -0.422  0.556

## P-values
> round( 2 * ( pt( abs(rstudent(res1)), 11 - 1 - 2, lower.tail = F)), 3)
    1     2     3     4     5     6     7     8     9    10    11
0.318 0.596 0.121 0.917 0.998 0.064 0.674 0.160 0.484 0.684 0.594
```

Case 3

```
## Externally Studentized Residuals
> round( rstudent(res1), 3)
     1      2      3      4      5      6      7      8      9     10     11
-0.857 -0.579  0.323 -0.321 -0.253  0.492 -0.034 -0.826 -0.519 -0.383  6.272

## P-values
> round( 2 * ( pt( abs(rstudent(res1)), 11 - 1 - 2, lower.tail = F)), 3)
    1     2     3     4     5     6     7     8     9    10    11
0.416 0.578 0.755 0.757 0.806 0.636 0.974 0.433 0.618 0.711 0.000
```

Examples: Outliers

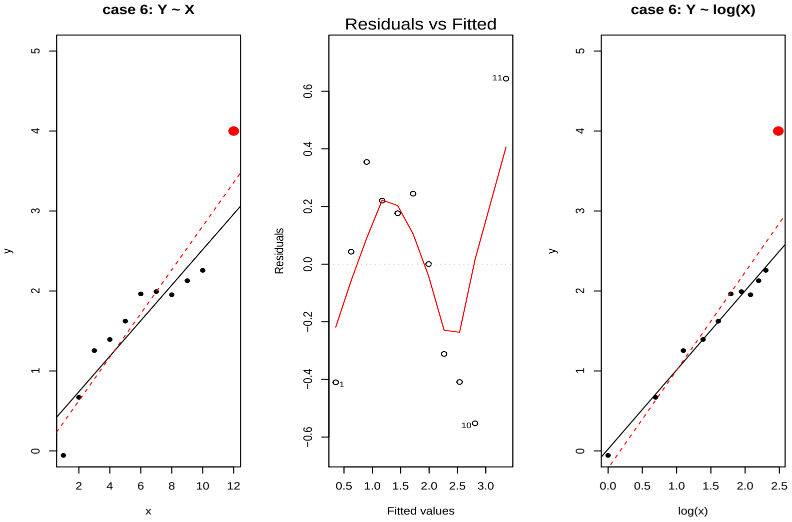




Case 4

```
## Externally Studentized Residuals
> round( rstudent(res1), 3)
     1      2      3      4      5      6      7      8      9     10     11     12
-0.480 -0.280  0.447 -0.088 -0.036  0.581  0.146 -0.479 -0.234 -0.112  3.550 -2.418

## P-values
> round( 2 * ( pt( abs(rstudent(res1)), 11 - 1 - 2, lower.tail = F)), 3)
    1     2     3     4     5     6     7     8     9    10    11    12
0.644 0.786 0.667 0.932 0.972 0.577 0.887 0.645 0.821 0.914 0.008 0.042
```

Examples: Outliers



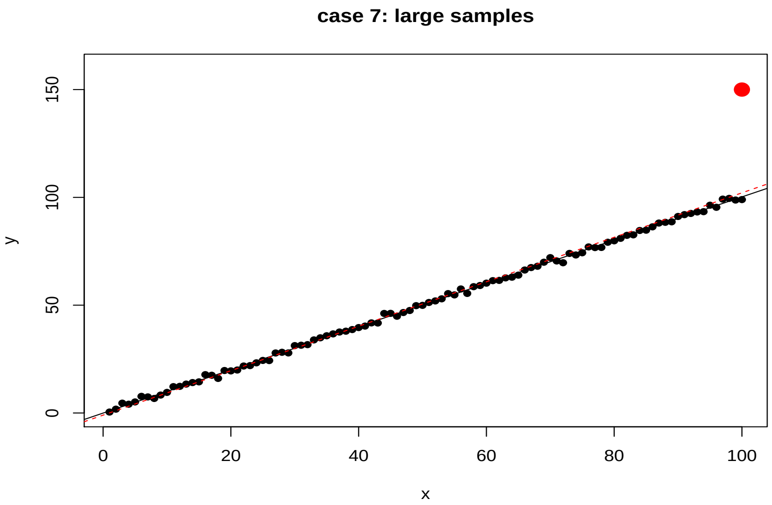
Case 6: $Y \sim X$

```
## Externally Studentized Residuals
> round( rstudent(res1), 3)
     1      2      3      4      5      6      7      8      9     10     11
-1.283  0.117  0.977  0.572  0.448  0.624  0.001 -0.822 -1.143 -1.741  2.667

## P-values
> round( 2 * ( pt( abs(rstudent(res1)), 11 - 1 - 2, lower.tail = F)), 3)
    1     2     3     4     5     6     7     8     9    10    11
0.235 0.910 0.357 0.583 0.666 0.550 0.999 0.435 0.286 0.120 0.028
```

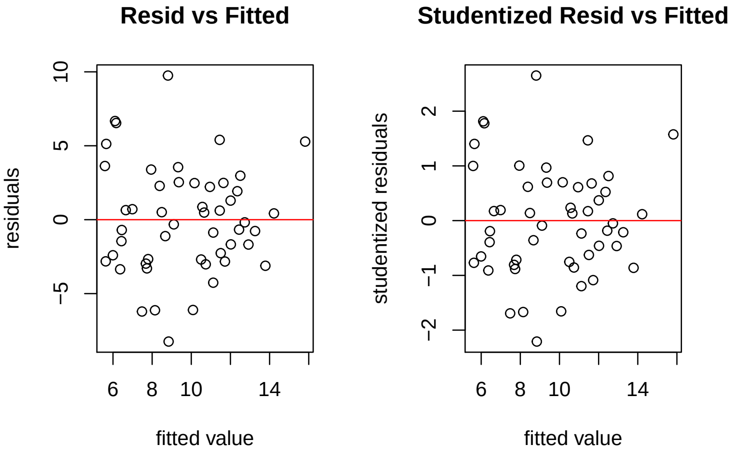
Case 6: $Y \sim \log(X)$

```
## Externally Studentized Residuals
> round( rstudent(res1), 3)
1      2      3      4      5      6      7      8      9      10     11
 0.501  0.089  0.279 -0.196 -0.293 -0.036 -0.393 -0.886 -0.819 -0.831 13.206

## P-values
> round( 2 * ( pt( abs(rstudent(res1)), 11 - 1 - 2, lower.tail = F)), 3)
1     2     3     4     5     6     7     8     9     10    11
0.630 0.931 0.787 0.850 0.777 0.972 0.704 0.401 0.437 0.430 0.000
```

Examples: Outliers



Savings Example: Studentized Residual vs. Fitted Plot



Savings Example

```
## Compute (externally) studentized residuals
> ti <- rstudent(result)
> max(abs(ti))
[1] 2.853558
> which(ti == max(abs(ti)))
Zambia
    46
## Compute p-value
> 2*(1-pt(max(abs(ti)), df=50-1-5))
[1] 0.006566663
## compare to alpha/n
> 0.05/50
[1] 0.001
```

Remarks on Outliers

- Two or more outliers can hide each other.
- Examine the context – what could it mean?
  - Occasionally data entry errors occur
  - Hidden variables may be part of the explanation
  - Something going wrong: e.g., fraudulent use of credit cards
  - A new unknown effect (you may get a Nobel prize if you can explain it!)
  - Some patterns just have exceptions...

Influential Points

An influential point is one whose removal from the dataset would cause a large change in the fit. At least one of the following:

- Outlier
- High leverage

How to measure the influence?

- Change in the coefficients $\hat{\beta} - \hat{\beta}_{(i)}$
- Change in the fit $X^T(\hat{\beta} - \hat{\beta}_{(i)}) = \hat{y} - \hat{y}_{(i)}$

Cook's Distance

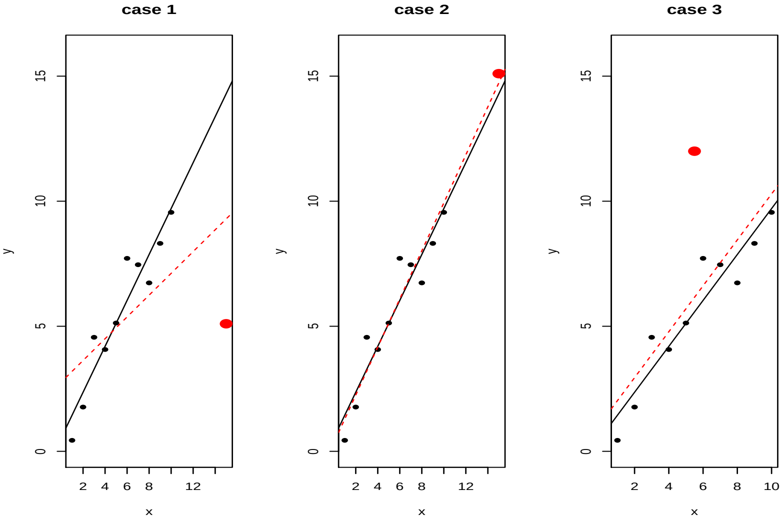
Cook statistic:

$$
\begin{aligned}
D_i &= \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{(p+1)\hat{\sigma^2}} \\
&= \frac{1}{p+1} r_i^2 \frac{h_i}{1-h_i}
\end{aligned}
$$

Combination of residual effect and leverage effect

Rule of thumb: Cook's Distance $D_i > \frac{4}{n-p-1}$ is considered large.

Examples: Influential Points



Case 1

```
## Cook's Distance
> round( cooks.distance(res1), 3)
    1     2     3     4     5     6     7     8     9    10    11
0.365 0.111 0.005 0.003 0.000 0.059 0.030 0.003 0.046 0.145 4.485

## Threshold
> 4 / (11 - 1 - 1)
0.444
```

Case 2

```
## Cook's Distance
> round( cooks.distance(res1), 3)
    1     2     3     4     5     6     7     8     9    10    11
0.202 0.043 0.233 0.001 0.000 0.166 0.011 0.124 0.044 0.020 0.200

## Threshold
> 4 / (11 - 1 - 1)
0.444
```

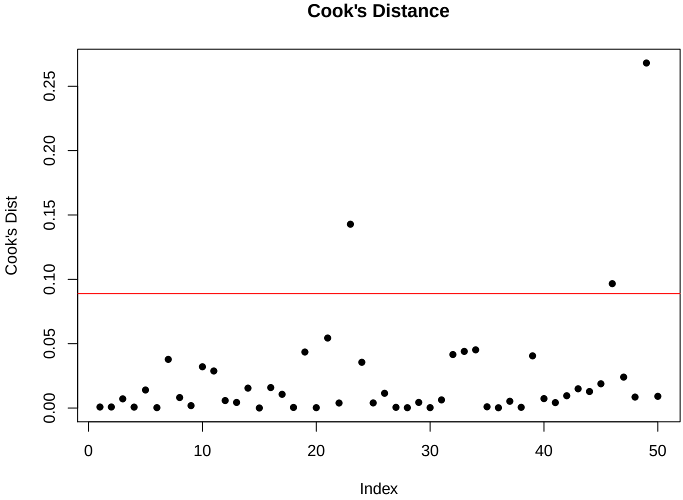
Case 3

```
## Cook's Distance
> round( cooks.distance(res1), 3)
1     2     3     4     5     6     7     8     9     10    11
0.192 0.057 0.012 0.008 0.004 0.014 0.000 0.071 0.046 0.041 0.374

## Threshold
> 4 / (11 - 1 - 1)
0.444
```

Savings Example

```
## Compute Cook's distance
> cook <- cooks.distance(result)
> plot( cook, pch = 16 , ylab ="Cook's Dist", main = "Cook's Distance"
> abline(h = 4/ (50 - 4 -1), col="red")
> which(cooks.distance(result) >4/ (50 - 4 -1) )
Japan Zambia  Libya
23     46     49
```

Savings Example Continued



Savings Example

Recall that Ireland, Japan, United States, and Libya may be leverage points. In addition, there is no outliers from t-tests.

According to the choice of a test or a method, the result may be different.

Checking the Structure of the Model: Linearity

Plot $\hat{\epsilon}$ against $\hat{y}$ and x_j, but other predictors impact the relationship. Consider

- Partial regression plots
- Partial residual plots

Isolate the effect of x_j on y

Partial Regression Plots

1. Regress y on all x except x_j, get residuals $\hat{\delta}$
2. Regress x_j on all x except x_j, get residuals $\hat{\gamma}$
3. Plot $\hat{\delta}$ against $\hat{\gamma}$

Partial Regression Plots: Intuition

In the summary table below, we have an evidence that X_2 and Y have a significant relationship after removing the effects of other variables.

```
## fitting model
> result = lm( Y ~ X1+ X2+ X3 )


## fitted model
> summary(result)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.242555   0.027011  -8.980 1.97e-15 ***
X1            0.077384   0.042927   1.803 0.073647 .
X2           -0.228795   0.078107  -2.929 0.003986 **
X3            0.009584   0.041840   0.229 0.819168
```

Partial Regression Plots: Intuition

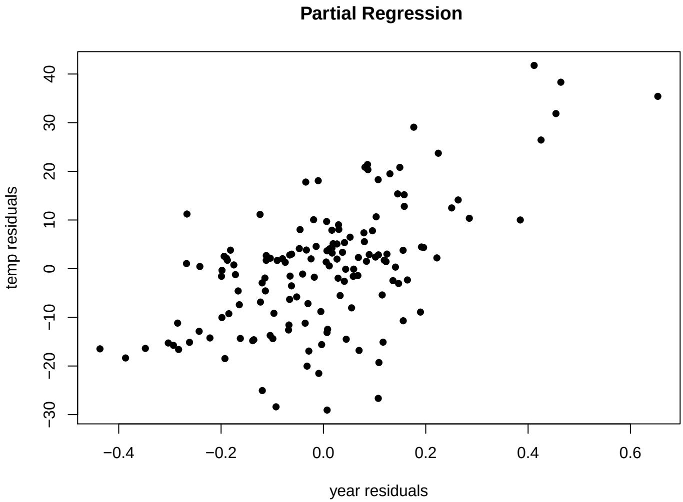
Residual is part of Y after removing the effects of predictors.

Therefore, partial regression plots show the relationship between a predictor and the response variable after removing the effects of predictors.

Global Warming Example

```
## Load Data
> data(globwarm)

# Remove Missing Values
> id =  which( is.na( globwarm$nhtemp ) == FALSE )
> globwarm = globwarm[id,]

## Fitting models
> result.a = lm( nhtemp ~ wusa+  jasper+ westgreen+
   chesapeake+ tornetrask+ urals+ mongolia+ tasman, data = globwarm)
> result.b = lm( year ~ wusa+  jasper+ westgreen+
   chesapeake+ tornetrask+ urals+ mongolia+ tasman, data = globwarm)

## Partial Regression Plot
> plot( residuals(result.a), residuals(result.b), xlab = "year residuals",
   ylab ="temp residuals", main = " Partial Regression", pch = 16 )
```

Global Warming Example



Partial Residual Plots

- Plot $\hat{\epsilon} + \hat{\beta}_j x_j$ against x_j

Where does this come from?

$$\begin{aligned} y - \sum_{j' \neq j} x_{j'} \hat{\beta}_{j'} &= \ldots \\ &= x_j \hat{\beta}_j + \hat{\epsilon} \end{aligned}$$

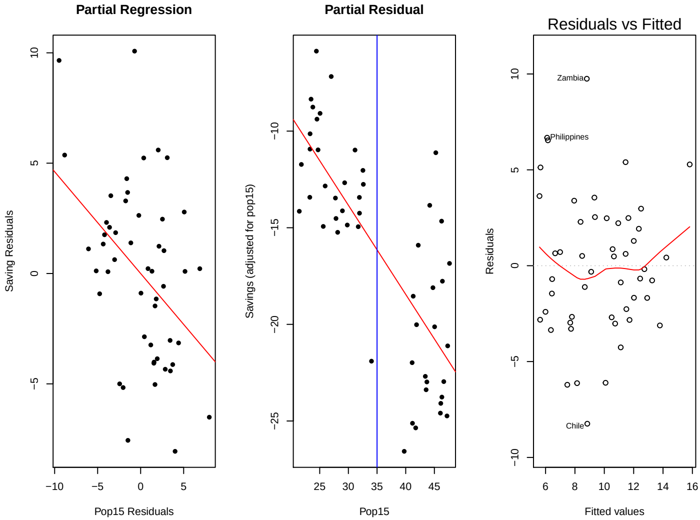
The slope is $\hat{\beta}_j$. Look for non-linearity and outliers and influential points.

Savings Example

```
## Load Data
> data(savings)

## Partial regression plot
> delta <- residuals(lm(sr ~ pop75 + dpi + ddpi, data=savings))
> gamma <- residuals(lm(pop15 ~ pop75 + dpi + ddpi, data=savings))
> plot(gamma,delta, xlab="Pop15 Residuals",  ylab="Saving Residuals",
   main = "Partial Regression", pch = 16)
> temp <- lm(delta ~ gamma)
> abline(reg=temp, col = "red")
```

```
## Partial Residual Plot
> result = lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> plot(savings$pop15, result$residuals + coef(result)[2] * savings$pop15,
   xlab="Pop15", ylab="Savings (adjusted for pop15)", main = "Partial Residual"
> abline(a=0, b=coef(result)['pop15'], col = "red")
> abline(v = 35, col = "blue")

## Residual Fitted Plot
> plot(result, which = 1)
```

Savings Example Continued



Savings Example Continued

```
## Two separate regressions on two groups
> temp1 <- lm(sr ~ pop15 + pop75 + dpi
    + ddpi, data=savings, subset=(pop15 > 35))
> temp2 <- lm(sr ~ pop15 + pop75 + dpi
    + ddpi, data=savings, subset=(pop15 < 35))
```

Summary of Assumptions

- Linearity
- Normality
- Constant Variance
- Independent Errors
- Unusual Points (influential points)
  - Leverage Points
  - Outliers

Summary of Assumptions

- Linearity

- Scatter Plot
- Residual vs Fitted Plot
- Partial Regression Plot
- Partial Residual Plot

Prediction (x), Inference (Test, CI) (x)

Summary of Assumptions

- Normality
- Normal QQ Plot
- Shapiro-Wilk's Test

Prediction (o), Inference (Test, CI) (x)

Summary of Assumptions

- Constant Variance

\- Residual vs Fitted Plot

\# Prediction (o), Inference (Test, CI) (may be...)

Summary of Assumptions

- Independent Errors
- Residual vs Fitted Plot
- ϵ_i vs ϵ_{i+1} Plot
- ϵ vs Time Plot
- Durbin-Watson test

\# Prediction (o), Inference (Test, CI) (x)

Summary of Assumptions

- Influential Points
- Histogram, Scatter Plot
- Residual vs Fitted Plot
- Leverage
- Internally and Externally Studentized Residuals
- Cook's Distance

Prediction (may be), Inference (Test, CI) (may be)

Summary of Assumptions

- Leverage Points
- Histogram, Scatter Plot
- Leverage
- Cook's Distance

Prediction (o; however be careful), Inference (Test, CI) (o)

Summary of Assumptions

- Outliers
- Histogram, Scatter Plot
- Residual vs Fitted Plot
- Externally Studentized Residuals
- Cook's Distance

\# Prediction (x), Inference (Test, CI) (x)

Summary of Diagnostics

- Just fitting a model is not enough
- Graphical diagnostics are more informative but also more subjective
- Diagnostics often suggest a change in the model and then the whole process is repeated
- Time-consuming... but worth it