Non-linear Relationships & Potential Problems

Section 3.3

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Outline

Two Important Assumptions

1. The **additive** assumptions means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.

2. The **linear** assumptions means that the change in the response Y due to a one-unit change in X_j is constants, regardless of the value of X_J.

Non-Linear Relationships

- The linear regression model assumes a linear relationship between the response and the predictors.
- The true relationship between the response and the predictors my be non-linear.
- The *polynomial regression* is a very simple way to directly extend the linear model to accommodate non-linear relationships.

Polynomial Regression

- **Polynomial regression** is a form of regression analysis in which the relationship between the predictor x and the response y is modeled as an n^{th} degree polynomial in x.
- Model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_m x_i^m + \epsilon_i$$

  for $i = 1, 2, \ldots, n$.
- We need to keep $m < n$
- In `R` we use `lm(y~poly(x,m))`.
- For more information see: `https://datascienceplus.com/fitting-polynomial-regression-r/`

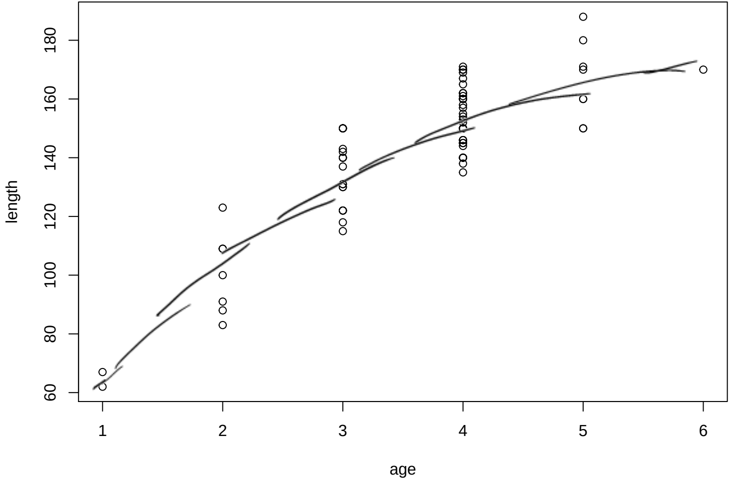
Non-Linear Relationship Example

In 1981, n = 78 bluegills were randomly sampled from Lake Mary in Minnesota. The researchers (Cook and Weisberg, 1999) measured and recorded the following data (`https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/bluegills/index.txt`):

- Response (y): length (in mm) of the fish
- Potential predictor (x1): age (in years) of the fish

The researchers were primarily interested in learning how the length of a bluegill fish is related to it age.

Scatterplot



Linear Summary

```
> summary(fish.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.649      5.755   10.89   <2e-16 ***
age           22.312      1.537   14.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.51 on 76 degrees of freedom
Multiple R-squared:  0.7349,Adjusted R-squared:  0.7314
F-statistic: 210.7 on 1 and 76 DF,  p-value: < 2.2e-16
```

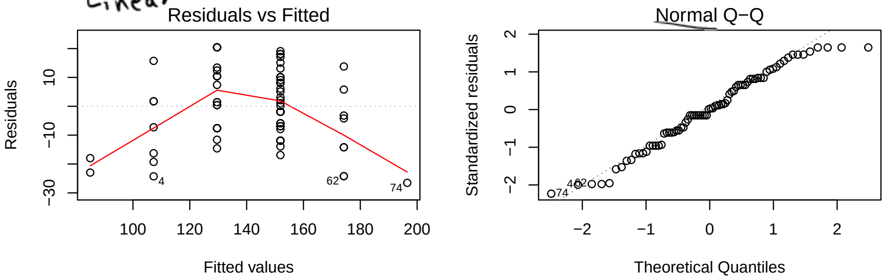
$\widehat{\text{length}} = 62.649 + 22.312 \times \text{age}$

$RSE = 12.51$

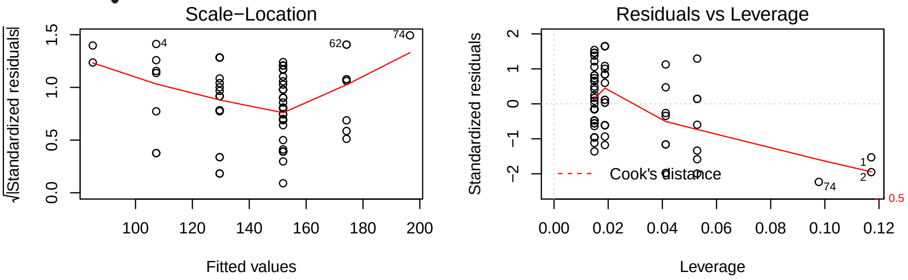
$R^2 = 73.14\%$

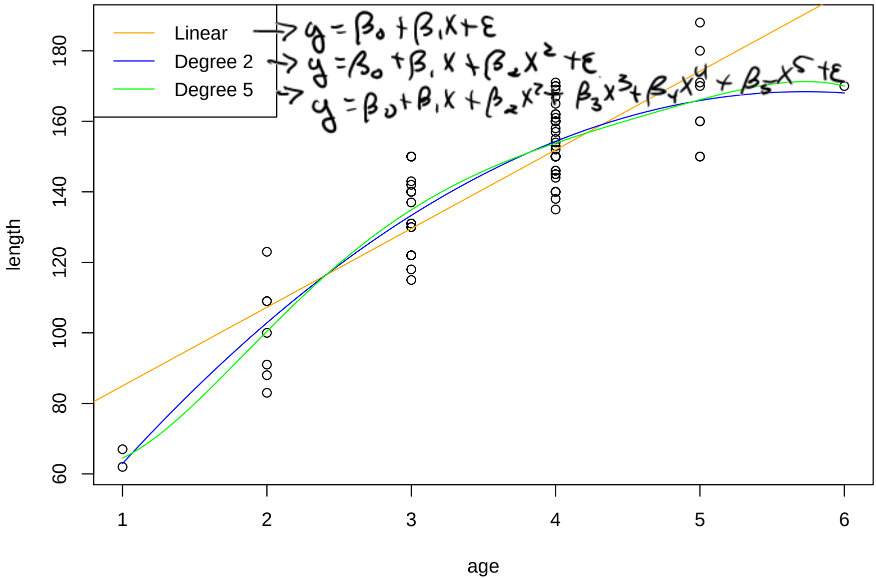
Assumptions Linear, Indep, Normal, Equal variance

Linear



Equal Variance




Linear
Degree 2
Degree 5
$y = \beta_0 + \beta_1 X + \varepsilon$
$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \varepsilon$
length
60
80
100
120
140
160
180
1
2
3
4
5
6
age

Polynomial Regression R

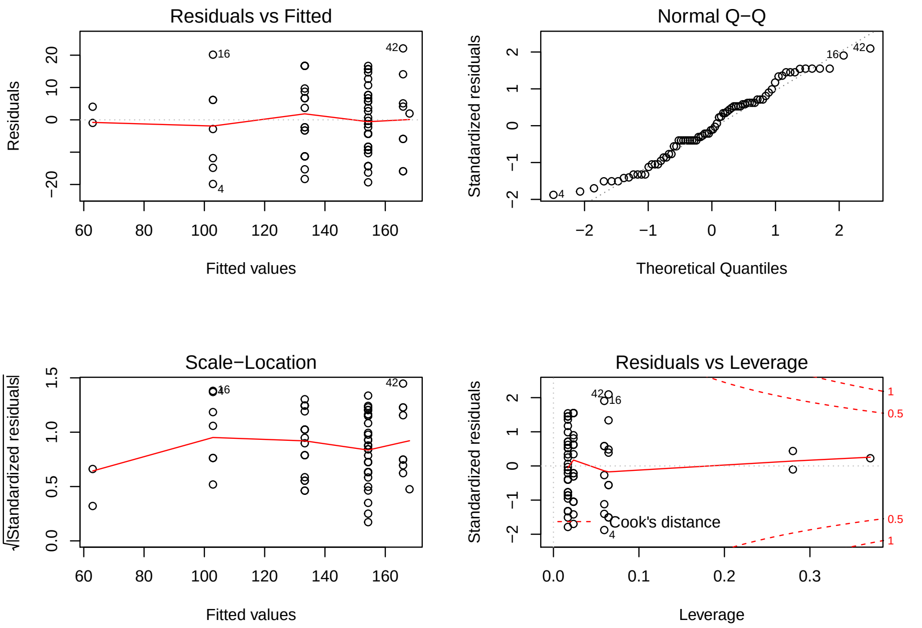
degree

```
> fish.lm2 = lm(length~poly(age,2),data = index)
> summary(fish.lm2)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      143.603      1.235 116.290  < 2e-16 ***
poly(age, 2)1    181.565     10.906  16.648  < 2e-16 ***
poly(age, 2)2    -54.517     10.906  -4.999 3.67e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.91 on 75 degrees of freedom
Multiple R-squared:  0.8011,Adjusted R-squared:  0.7958
F-statistic: 151.1 on 2 and 75 DF,  p-value: < 2.2e-16
```

$\widehat{\text{length}} = 143.603 + 181.565\,\text{age} - 54.517\,\text{age}^2$

$RSE = 10.91$

adj $R^2 = 79.6\%$


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Cubic Regression

response predictor

```
> #Cubic Polynomial
> fish.lm3 = lm(length~poly(age,3),data = index)
> summary(fish.lm3)
```

dreege

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     143.603      1.243 115.544  < 2e-16 ***
poly(age, 3)1   181.565     10.976  16.541  < 2e-16 ***
poly(age, 3)2   -54.517     10.976  -4.967 4.25e-06 ***
poly(age, 3)3     2.234     10.976   0.203    0.839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.98 on 74 degrees of freedom
Multiple R-squared:  0.8012,Adjusted R-squared:  0.7932
F-statistic: 99.44 on 3 and 74 DF,  p-value: < 2.2e-16
```

$$\widehat{length} = 143.603 + 181.565\,age - 54.517\,age^2 + 2.234\,age^3$$

Lab Questions

We will use the Boston data for these questions. Make sure you load the `MASS` library.

```
library(MASS)
```

1. Perform a linear regression on medv (response variable) onto lstat (predictor). What is the adjusted R^2 for this model?

   a) 0.5441
   b) 0.5432
   c) 0.0002
   d) 0.95

2. Draw a scatterplot between medv (y) and lstat (x). Does it appear linear?

   a) Yes
   b) No

In R type in

```
par(mfrow = c(2,2))
plot(fit.lm)
```

3. Do we see a pattern in the residuals?
   a) Yes
   b) No

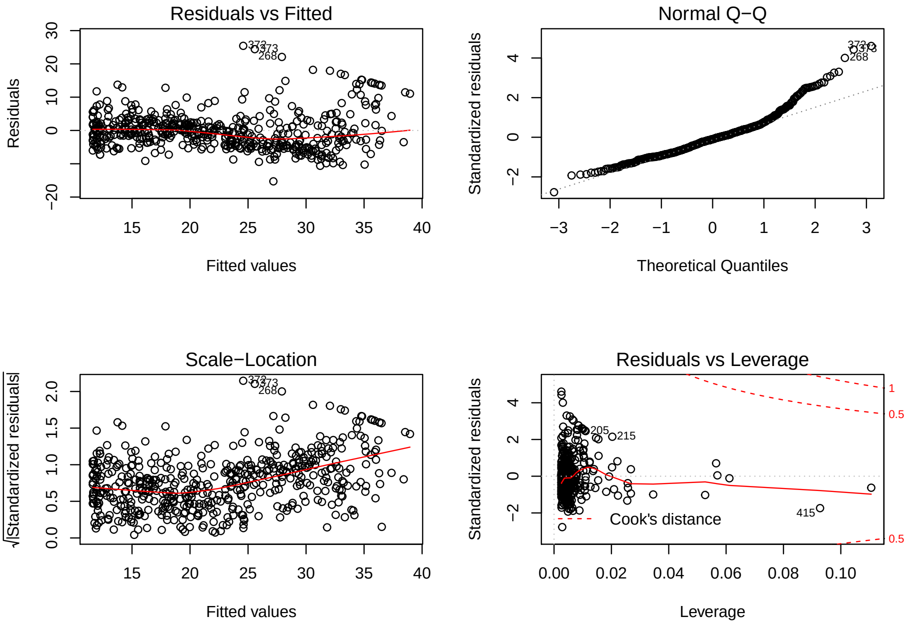
   If there is a pattern then the linear model may not be the best model.

4. Type in R and run the summary:
   `fit.lm2 = lm(medv ~ poly(lstat,2),data = Boston)`.
   What is the adjusted R^2 for this model?

   a) 0.0002
   b) 0.055
   c) 0.6407
   d) 0.6393

5. Run a cubic model. What is the adjusted R^2?

   a) 0.6558
   b) 0.6578
   c) 0.0002
   d) 0.054


Residuals vs Fitted
Residuals
Fitted values
372
373
268
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
205
215
415
Cook's distance
1
0.5
0.5

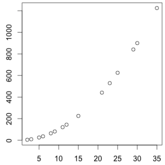
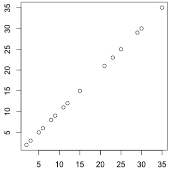
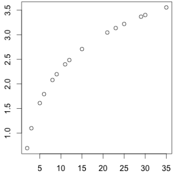
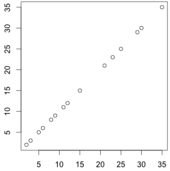
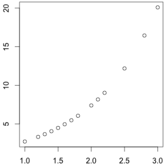
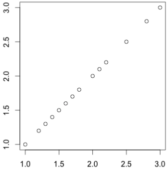
Some Warnings About Polynomial Models

- The following list is from `https://online.stat.psu.edu/stat462/node/158/`.
- The fitted model is more reliable when it is built on a larger sample size.
- Consider how large the size of the predictors(s) will be when incorporating higher degree terms as this may cause numerical overflow for the statistical software being used.
- Do not go strictly by low *p*-values to incorporate a higher degree term, but rather just uses these to support your model only if the resulting residual plots looks reasonable.
- As a standard practice if you have an n^{th} degree polynomial, always include the each X^j such that $j < n$.

Potential Problems in Linear Regression

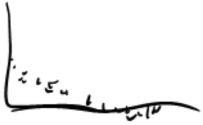
1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

We have talked already about how to overcome some of these problems.

Original Scatter-plot	Type of plot and Transformation	Modified Scatter-plot
	← looks like $y = x^2$ transform by changing (x, y) into $(x, \sqrt{y})$ →	
	← looks like $y = \log x$ transform by changing (x, y) into (x, e^y) →	
	← looks like $y = e^x$ transform by changing (x, y) into $(x, \log y)$ →	

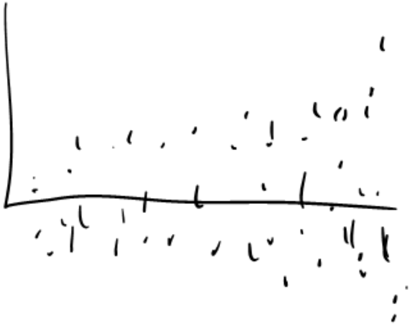
Correlation of Error Terms

- Assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated.
- For example if ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- If there is correlation among the error terms then the estimated standard errors will tend to underestimate the true standard errors. This results in narrower confidence and prediction intervals.
- Time series data is an example of correlation among the error terms.
- Residual plot is best way to tell if there is correlation.



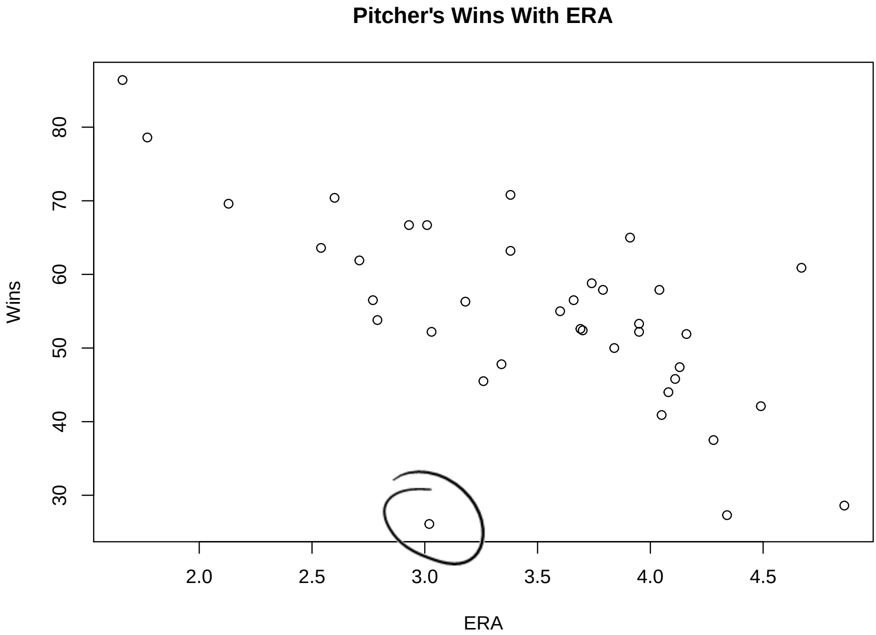
Non-constant Variance of Error Terms

- **Heterscedasticity** is where the variances of the error terms increase with the value of the response. This will appear as a *funnel shape* in the residual plot.
- Possible solution is to transform the Y using $log(Y)$ or $\sqrt{Y}$.

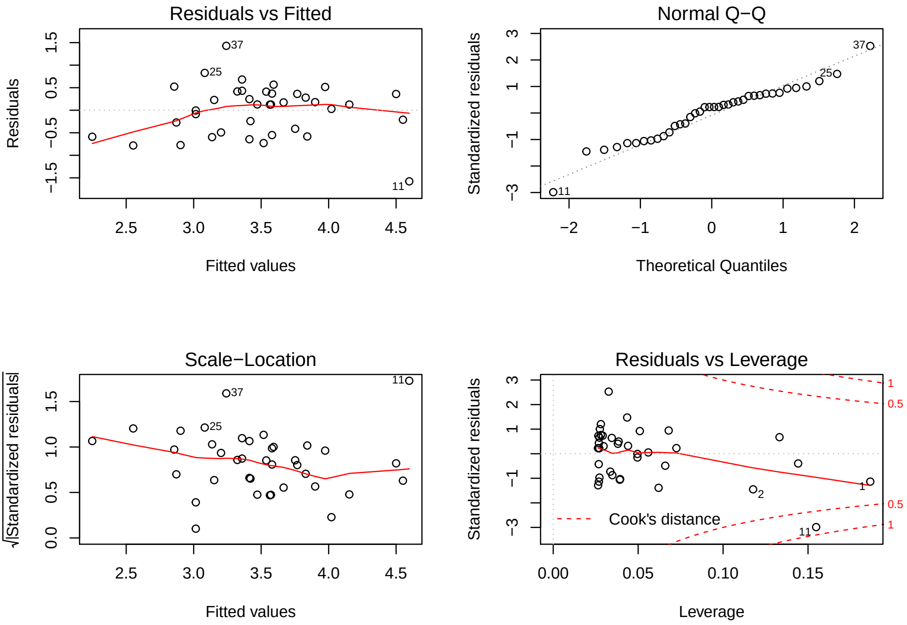


Outliers

- An **outlier** is a point for which y_i is far from the value predicted by the model.
- Outliers can have an effect on the estimated regression parameters, RSE and R^2.
- Scatterplot and residual plot would be the best to detect outliers.

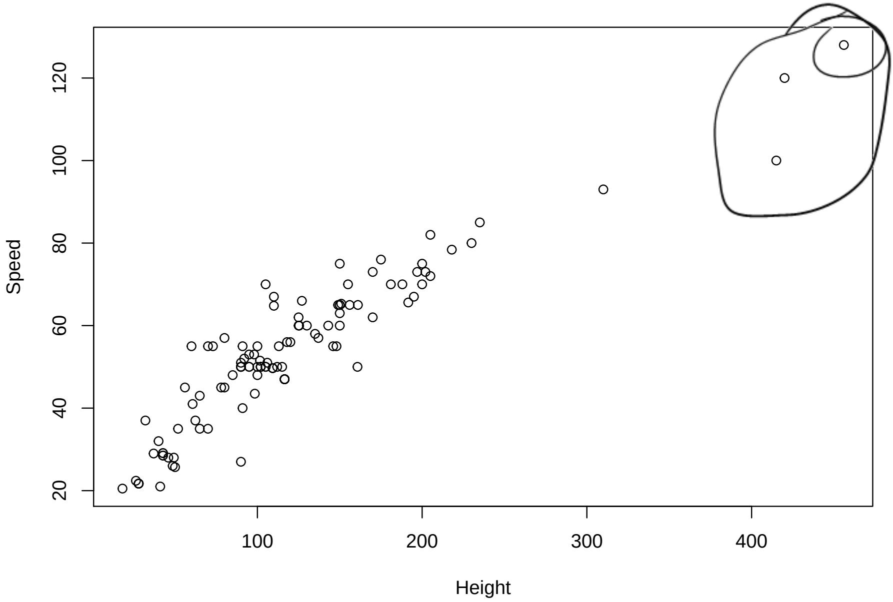

Pitcher's Wins With ERA
Wins
ERA
30
40
50
60
70
80
2.0
2.5
3.0
3.5
4.0
4.5

```
> era.lm = lm(ERA ~ Wins,data = Era)
> summary(era.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.614231   0.405423  13.848 5.43e-16 ***
Wins        -0.038957   0.007229  -5.389 4.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5744 on 36 degrees of freedom
Multiple R-squared:  0.4465,Adjusted R-squared:  0.4311
F-statistic: 29.04 on 1 and 36 DF,  p-value: 4.557e-06
```

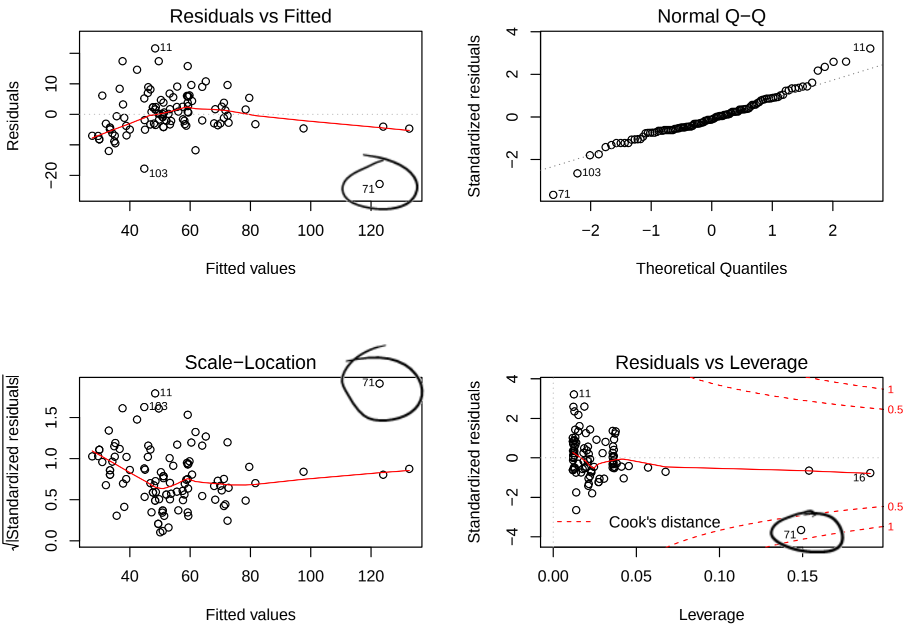

Residuals vs Fitted
Residuals
Fitted values
37
25
11
Normal Q−Q
Standardized residuals
Theoretical Quantiles
37
25
11
Scale−Location
√|Standardized residuals|
Fitted values
11
37
25
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1
0.5
0.5
1
2
11
1

High Leverage Points

- **High leverage points** have an unusual value for X.
- High leverage observations tend to have a sizable impact on the estimated regression line.
- Can be determined by scatterplots for a simple linear regression.
- In order to quantify an observation's leverage we compute the **leverage statistic**.
- In `R` we can use the Residuals vs Leverage to see if there are high leverage observations.



Speed
120
100
80
60
40
20
100
200
300
400
Height


Residuals vs Fitted
Residuals
Fitted values
11
103
71
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
16

Collinearity

- **Collinearity** refers to the situation in which two or more predictor variable are closely related to one another.
- In regression this will cause difficulty to separate out the individual effects of collinear variables on the response.
- This reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow.
- The **power** of the hypothesis test for $H_0 : \beta_i = 0$ - probability of correctly detecting a nonzero coefficient - is reduced by collinearity.

Detecting Multiconllinearity

- Check the correlation matrix. In R: $cor()$.
- The variance inflation factor (VIF). The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.
  - The smallest possible value for VIF is 1.
  - If VIF is more than 10, multicolinearity is strongly suggested.
  - The VIF for each variable can be computed using the formula:

    $$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

    Where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.
  - In R studio install the "fmsb" package. The function is VIF().

```
> library(fmsb)
> VIF(stock2.lm)
[1] 9.768829
```