

STAT 210
Applied Statistics and Data Analysis
Week 9 - Summary

Joaquin Ortega

King Abdullah University of Science and Technology

V30: Simple Linear Regression 4: Coefficient of Determination

The analysis of variance is based on the following decomposition for the sum of squares:

$$SST = SSE + SSR.$$

Since the sums are non-negative we have that $SSE \leq SST$.

Observe that they are equal only if there is no relation between the two variables: $SSR = 0$ means that $\hat{y}_i = \bar{y}$ for all i and for this to be true, the regression line must be horizontal, so $\beta_1 = 0$ and $y = \beta_0$.

The regression sum of squares SSR is usually interpreted as the amount of variability in Y that is explained by the regression line.

The **coefficient of determination** R^2 is defined as

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

and represents the proportion of the variation that is explained by the regression model.

Recall the summary for the regression in the first example:

```
summary(lm1)
```

```
##
## Call:
## lm(formula = FL ~ CL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86395 -0.51746 -0.02826  0.50456  1.77009
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15316    0.23477   0.652   0.515
## CL          0.48060    0.00714  67.313 <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579
## F-statistic: 4531 on 1 and 198 DF, p-value: < 2.2e-16
```

Trees example.

```
summary(lm4)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.274  -9.894  -2.894   12.068   29.852
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976  0.005835 **
## Height       1.5433     0.3839   4.021  0.000378 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784
```

Coefficient of Determination

```
summary(lm5)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48 < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

The first of these models has a low R^2 of 35.8% while the second has a much better value of 93.5%.

Relation with the Correlation Coefficient

Proposition 1 Let ρ be the correlation coefficient for the sample $(x_i, y_i), i = 1, \dots, n$. Then

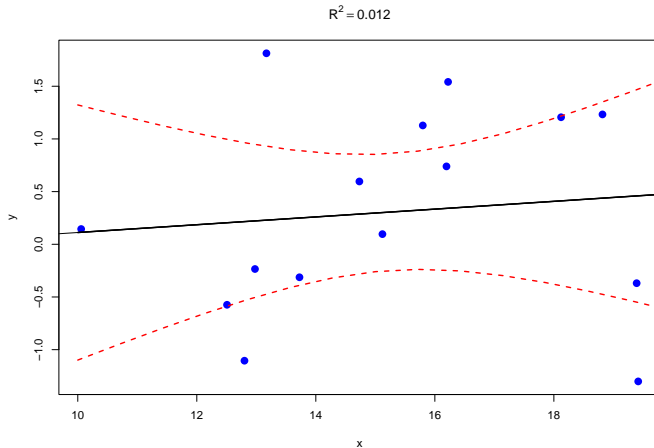
$$R^2 = \rho^2.$$

It is important to observe that this relation is only true for simple regression. It does not hold in the multivariate case.

Coefficient of Determination

As an example let us look at some simulated data. First we look at purely random values.

R squared: 0.012

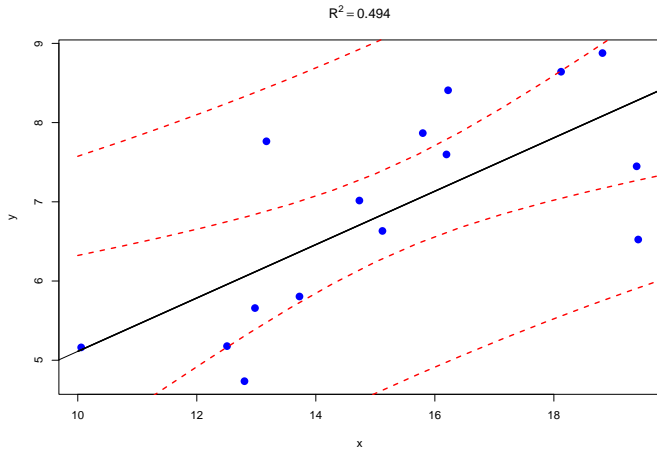


Coefficient of Determination

```
set.seed(98765)
xx <- runif(15,10,20)
zz <- rnorm(15)
r.sq <- round(summary(lm(zz~xx))$r.squared,3)
cat('R squared: ', r.sq)
plot(xx,zz,pch=16, xlab='x', ylab='y')
abline(lm(zz~xx))
title(main= bquote(R^2 == .(r.sq)))
xx.new <- data.frame(xx=seq(10,20, length.out = 15))
pc <- predict(lm(zz~xx),xx.new, int='c')
matlines(xx.new$xx, pc, lty=c(1,2,2),
         col=c('black','red','red'))
pp <- predict(lm(zz~xx),xx.new, int='p')
matlines(xx.new$xx, pp, lty=c(1,2,2),
         col=c('black','red','red'))
```

Coefficient of Determination

R squared: 0.494



Coefficient of Determination

```
yy1 <- 2 + 0.3*xx + zz
plot(xx,yy1,pch=16, xlab='x', ylab='y')
abline(lm(yy1~xx))
r.sq <-round(summary(lm(yy1~xx))$r.squared,3)
cat('R squared: ', r.sq)
title(main= bquote(R^2 == .(r.sq)))
pc <- predict(lm(yy1~xx),xx.new, int='c')
matlines(xx.new$xx, pc, lty=c(1,2,2),
         col=c('black','red','red'))
pp <- predict(lm(yy1~xx),xx.new, int='p')
matlines(xx.new$xx, pp, lty=c(1,2,2),
         col=c('black','red','red'))
```

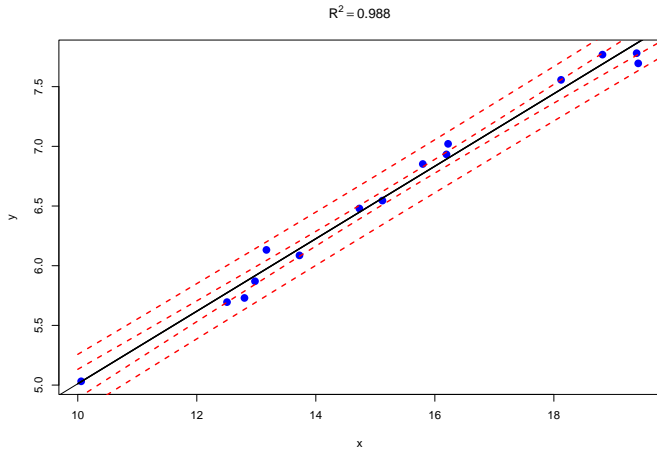
Coefficient of Determination

Now there is a linear relation between y and x but the variability due to the variance of the noise makes the explained variability to be only about 50%. As a third and final example, let us reduce noise variability by rescaling it.

```
yy2 <- 2 + 0.3*xx + zz/10
plot(xx,yy2,pch=16, xlab='x', ylab='y')
abline(lm(yy2~xx))
r.sq <-round(summary(lm(yy2~xx))$r.squared,3)
cat('R squared: ', r.sq)
title(main= bquote(R^2 == .(r.sq)))
pc <- predict(lm(yy2~xx),xx.new, int='c')
matlines(xx.new$xx, pc, lty=c(1,2,2),
          col=c('black','red','red'))
pp <- predict(lm(yy2~xx),xx.new, int='p')
matlines(xx.new$xx, pp, lty=c(1,2,2),
          col=c('black','red','red'))
```

Coefficient of Determination

R squared: 0.988



V 31: Simple Linear Regression 5: Model Assessment

The quality of a model depends on the veracity of the assumptions we have made, which are the basis for the estimation of the parameters.

We need to check the goodness-of-fit of the model and the possible presence of outliers or highly influential data points.

The techniques we will consider are mainly graphical. Graphs are a fundamental tool for statistical practice and in particular for model assessment.

We start by recalling Anscombe's quartet. The code that follows is from the R documentation of `anscombe`:

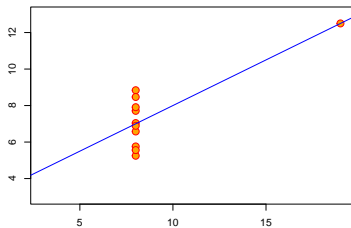
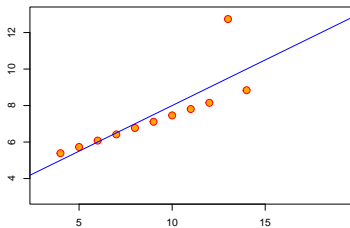
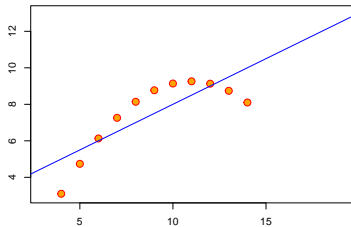
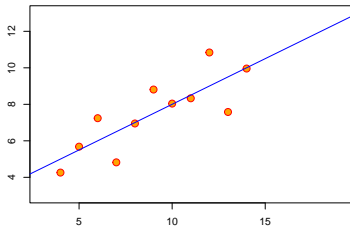
```
## Analysis of Variance Table
##
## Response: y1
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1          1  27.510   27.5100    17.99 0.00217 **
## Residuals    9   13.763    1.5292
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: y2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x2          1  27.500   27.5000    17.966 0.002179 **
## Residuals    9   13.776    1.5307
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: y3
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x3          1  27.470   27.4700    17.972 0.002176 **
## Residuals    9   13.756    1.5285
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: y4
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x4          1  27.490   27.4900    18.003 0.002165 **
## Residuals    9   13.742    1.5269
## ---
## Signif. codes:
```

```
##           lm1           lm2           lm3           lm4
## (Intercept) 3.0000909 3.000909 3.0024545 3.0017273
## x1          0.5000909 0.500000 0.4997273 0.4999091

## $lm1
##           Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 3.0000909  1.1247468  2.667348 0.025734051
## x1          0.5000909  0.1179055  4.241455 0.002169629
##
## $lm2
##           Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 3.000909  1.1253024  2.666758 0.025758941
## x2          0.500000  0.1179637  4.238590 0.002178816
##
## $lm3
##           Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 3.0024545  1.1244812  2.670080 0.025619109
## x3          0.4997273  0.1178777  4.239372 0.002176305
##
## $lm4
##           Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 3.0017273  1.1239211  2.670763 0.025590425
## x4          0.4999091  0.1178189  4.243028 0.002164602
```

Model Assessment

Anscombe's 4 regression data sets



Facts About Residuals

We have assumed that the errors $\epsilon_i, i = 1, \dots, n$ have a centered Gaussian distribution with constant variance σ^2 and are independent.

The residuals are defined as

$$\hat{\epsilon}_i = y_i - \hat{y}_i,$$

the difference between the observed and the predicted values for $x = x_j$.

Unlike the errors, **the residuals are not independent and do not have constant variance.**

They cannot be independent because we have shown that $\sum_i \hat{\epsilon}_i = 0$.

Recall that the regression parameters are given by

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

and once we have these parameters, the fitted values are given by

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

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is usually denoted by **H** and is known as the **hat** matrix, because it carries the observed vector **y** into the fitted values vector $\hat{\mathbf{y}}$.

$$\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}$$

If h_{ij} are the elements of matrix \mathbf{H} then

$$\hat{y}_i = \sum_j h_{ij} y_j.$$

The h_{ij} are the 'weights' needed to go from the observed values to the regression values, and the bigger h_{ij} is, the more influential the observed value y_j will be in the determination of \hat{y}_i .

The hat matrix gives a measure of the 'leverage' of the observations on the fitted model.

In general, the greatest impact of y_i occurs for \hat{y}_i and hence we will focus on the diagonal elements of \mathbf{H} .

The **leverage** h_{ii} is the i -th entry in the diagonal of \mathbf{H} .

Observe that

$$\begin{aligned}\text{Cov}(\hat{\epsilon}) &= \text{Cov}(\mathbf{Y} - \hat{\mathbf{Y}}) = \text{Cov}(\mathbf{Y} - \mathbf{H}\mathbf{Y}) \\ &= \text{Cov}((\mathbf{I} - \mathbf{H})\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})'.\end{aligned}$$

Now, since \mathbf{H} is symmetric, $\mathbf{I} - \mathbf{H}$ is also symmetric and it is easy to see that $(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H}$ so $\mathbf{I} - \mathbf{H}$ is also idempotent.

Therefore, we get that $\text{Var}(\hat{\epsilon}_i) = \sigma^2(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$.

Since the h_{ii} need not be equal, we see that the residuals do not have the same variance.

Also, since \mathbf{H} need not be a diagonal matrix, the $\hat{\epsilon}_i$ are usually correlated and not independent.

It is possible to prove that $0 \leq h_{ii} \leq 1$ and that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}.$$

The **standardized** residuals are defined as

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}(1 - h_{ii})^{1/2}}$$

where $\hat{\sigma}$ is the estimated error standard deviation.

Homoscedasticity and Linearity

Homoscedasticity and Linearity

The first graph that is usually drawn for model evaluation is a plot of residuals against fitted values

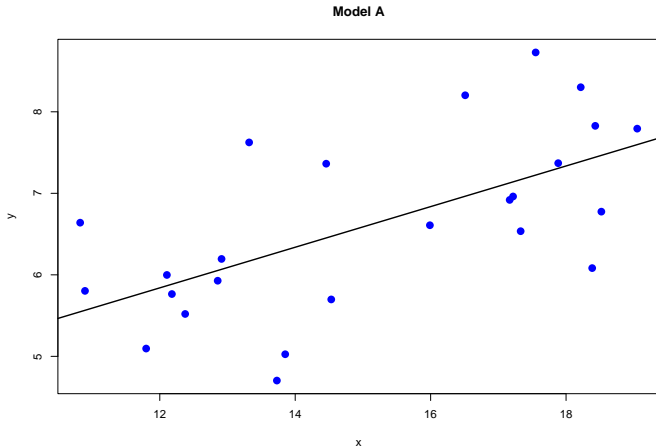
If the model assumptions are correct and we have captured as much variability as possible with the model, in this graph, we would expect to see no patterns but points scattered at random over the plotting region.

Patterns may indicate that the assumption of equal variance (homoscedasticity) does not hold or that there are still possible improvements in the model.

Let us see an example of this situation with simulated data. This will be modelA.

Homoscedasticity and Linearity

$$y = 2 + 0.3 \cdot x + \varepsilon.$$

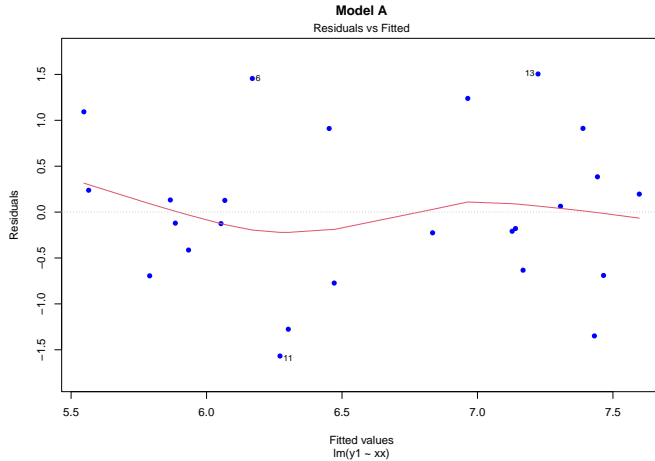


Homoscedasticity and Linearity

```
summary(modelA)
```

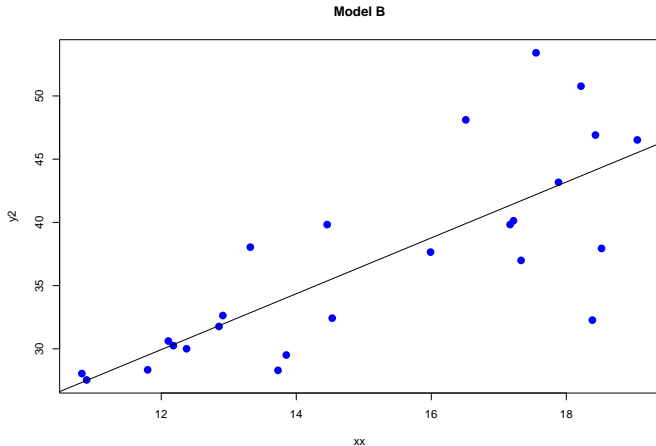
```
##
## Call:
## lm(formula = y1 ~ xx)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5680 -0.6330 -0.1202  0.3848  1.5047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.84963    0.99632   2.860 0.008853 **
## xx             0.24921    0.06488   3.841 0.000834 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8659 on 23 degrees of freedom
## Multiple R-squared:  0.3908, Adjusted R-squared:  0.3643
## F-statistic: 14.76 on 1 and 23 DF,  p-value: 0.0008336
```

Homoscedasticity and Linearity



Homoscedasticity and Linearity

Model B: $y = 2 + 2.3 \cdot x + (x - 10) \cdot \varepsilon$.

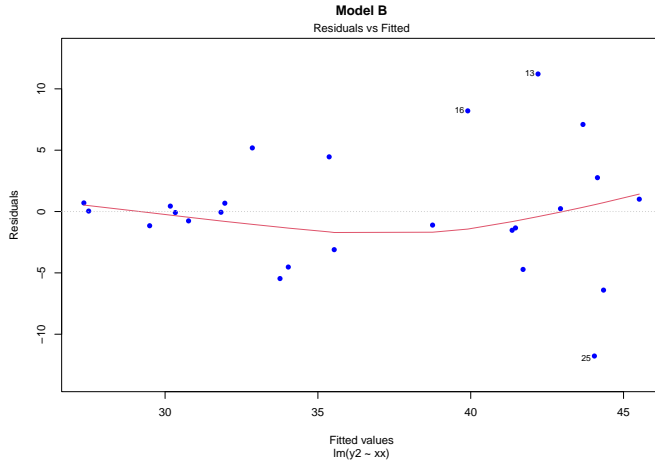


Homoscedasticity and Linearity

```
summary(modelB)
```

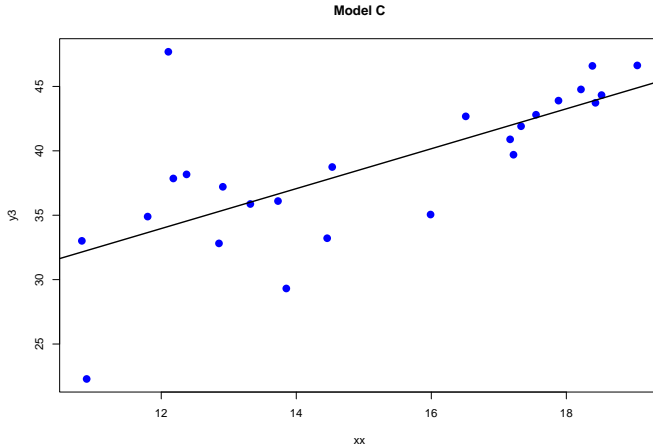
```
##  
## Call:  
## lm(formula = y2 ~ xx)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -11.7769  -1.5265  -0.0589   1.0082  11.2117   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   3.4322     5.7244   0.600    0.555      
## xx            2.2090     0.3727   5.926 4.84e-06 ***  
## ---  
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.975 on 23 degrees of freedom  
## Multiple R-squared:  0.6043, Adjusted R-squared:  0.5871   
## F-statistic: 35.12 on 1 and 23 DF,  p-value: 4.843e-06
```

Homoscedasticity and Linearity



Homoscedasticity and Linearity

Model C: $y = 2 + 2.3 \cdot x + (x - 20) \cdot \varepsilon$.

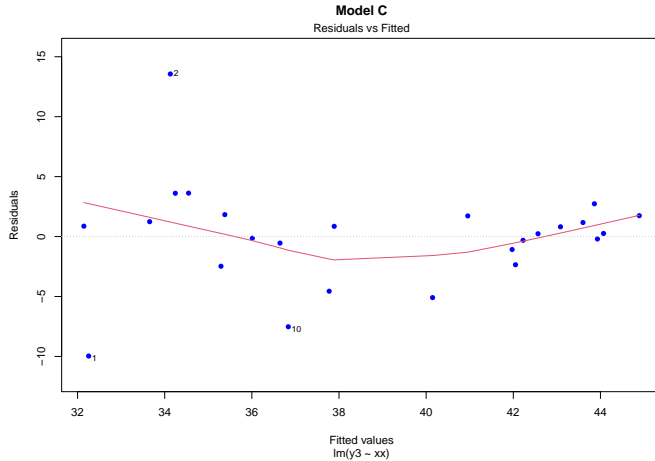


Homoscedasticity and Linearity

```
summary(modelC)
```

```
##  
## Call:  
## lm(formula = y3 ~ xx)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -9.9667 -1.0808  0.2575  1.7215 13.5612   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  15.3724     5.0786   3.027 0.005999 **   
## xx           1.5495     0.3307   4.686 0.000102 ***  
## ---  
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.414 on 23 degrees of freedom  
## Multiple R-squared:  0.4884, Adjusted R-squared:  0.4661   
## F-statistic: 21.95 on 1 and 23 DF,  p-value: 0.0001019
```

Homoscedasticity and Linearity



Homoscedasticity and Linearity

In both cases we have a 'funnel' shape, although with different orientations. This is an indication that the variance is not constant.

A possible way to deal with this problem is to transform the data. Useful transformations in this case are the Box-Cox transformations. We won't go into any detail about this but for positive data the transformations are given by

$$T_{\lambda}(y) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

The command `boxcox` in R calculates the optimal transformation for a given data set.

The graphs of residuals against fitted values are also useful to detect cases in which the model does not explain all the structure present in the data.

Example Q

As an example let us consider a quadratic relation between two variables that we try to model as a linear relation.

```
set.seed(4567)
xx <- runif(25,10,20)
zz <- rnorm(25,sd=4)
y4 <- 2 + 1.3*xx + 3*(xx-10)^2+zz
modelD <- lm(y4~xx)
```

Homoscedasticity and Linearity

```
summary(modelD)
```

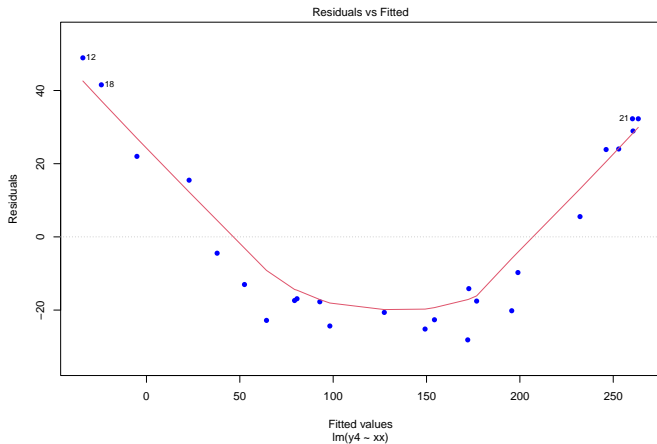
```
##
## Call:
## lm(formula = y4 ~ xx)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.14  -20.18  -13.02   23.87   48.93
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -350.746     26.814  -13.08 3.88e-12 ***
## xx             31.569       1.718   18.37 3.06e-15 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.13 on 23 degrees of freedom
## Multiple R-squared:  0.9362, Adjusted R-squared:  0.9334
## F-statistic: 337.5 on 1 and 23 DF,  p-value: 3.057e-15
```


In the summary for the regression, we see that slope and intercept are significant with a very low p -value and that the coefficient of determination R^2 has a (high) value of 0.936.

However, if we look at the summary data for the residuals, we see that the values do not correspond to a symmetric distribution, as one would expect if they followed a (centered) normal distribution.

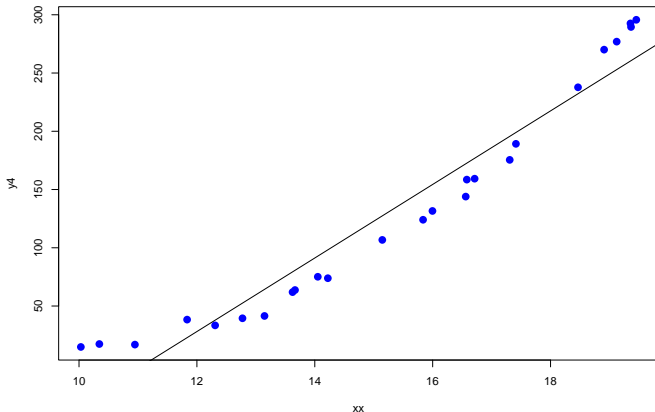
Homoscedasticity and Linearity

```
plot(modelD, which = 1, col = 'blue', pch = 16)
```



Homoscedasticity and Linearity

Indeed, if we had looked at the data in the first place -something one should always do- we would have seen that a linear relation is not adequate for this data.



Homoscedasticity and Linearity

We can add a quadratic term to the regression to include this structure into account. We will look at multiple regression in detail later on, but for completeness, let's fit a quadratic model.

```
modelE <- lm(y4~xx+I(xx^2))  
summary(modelE)
```

```
##  
## Call:  
## lm(formula = y4 ~ xx + I(xx^2))  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -9.2420 -1.8221  0.0683  2.8383  9.8580   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  309.9679    26.9643   11.50 9.03e-11 ***  
## xx          -59.5669     3.6675  -16.24 9.81e-14 ***  
## I(xx^2)       3.0235     0.1212   24.95 < 2e-16 ***  
## ---  
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.747 on 22 degrees of freedom  
## Multiple R-squared:  0.9978, Adjusted R-squared:  0.9976   
## F-statistic: 5039 on 2 and 22 DF, p-value: < 2.2e-16
```

Homoscedasticity and Linearity

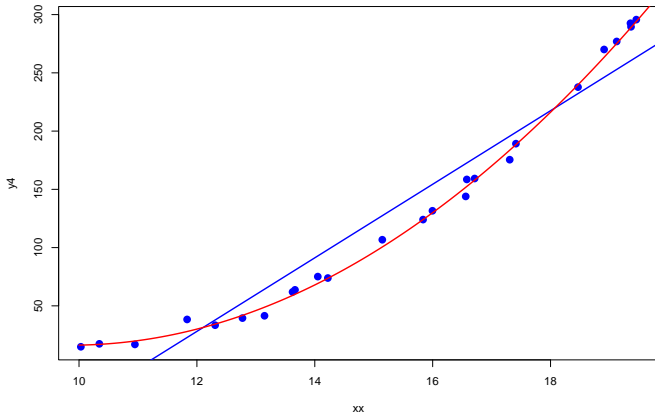
From the summary data for the regression we see that linear and quadratic terms are significant and that the summary data for the residuals is consistent with a symmetric distribution.

Also, the R^2 has increased to 0.998.

Next we plot the data, the regression line (from the first regression model) and the quadratic curve we have just fitted.

Homoscedasticity and Linearity

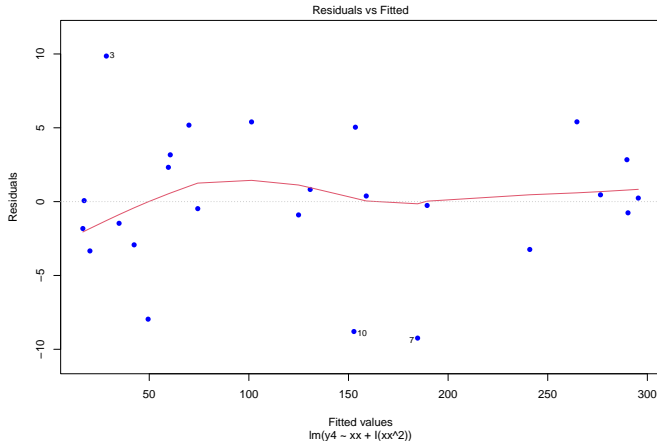
```
plot(xx,y4,pch=16, col = 'blue', cex = 1.5)
abline(modelD, col='blue', lwd=2)
curve(309.99-59.57*x+3.02*x^2,10,20, add=T,
      col='red', lwd=2)
```



Homoscedasticity and Linearity

Finally, the graph to evaluate the new model looks much better than the those for the previous model.

```
plot(modelE, which = 1, col = 'blue', pch = 16)
```

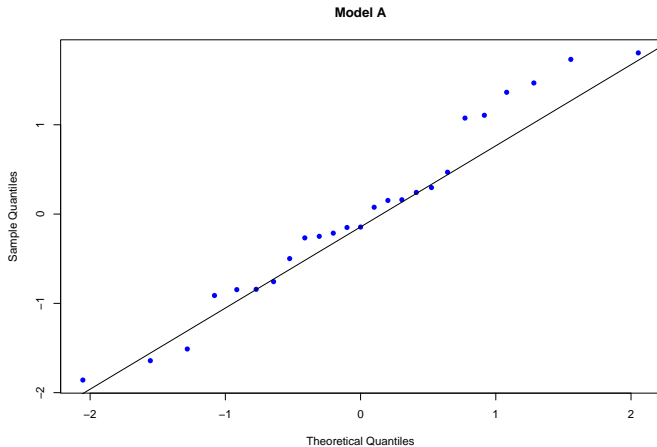


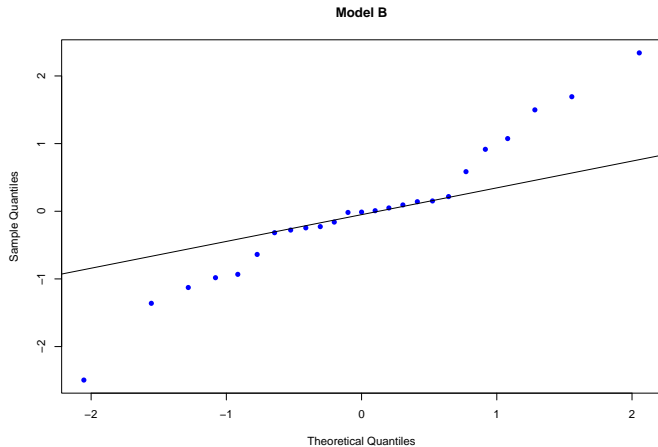
Gaussianity

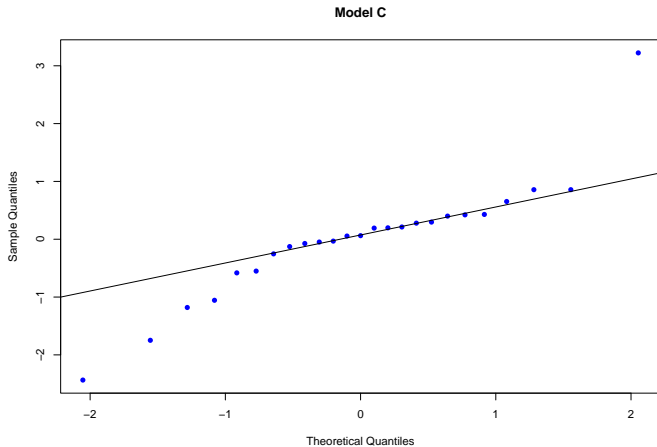
Another important assumption we have made is that the errors have a normal distribution.

To check this assumption it is usual to draw a quantile plot for the residuals. However, since we have seen that the residuals do not have constant variance, it is usual to plot the standardized residuals.

In R, standardized residuals are obtained with the `rstandard` command acting on an `lm` object

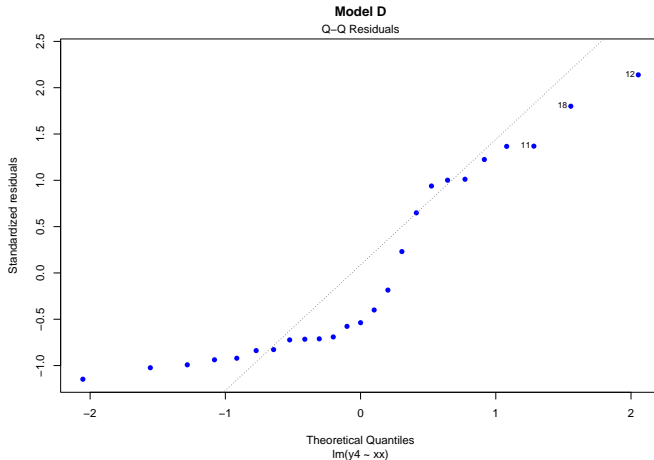




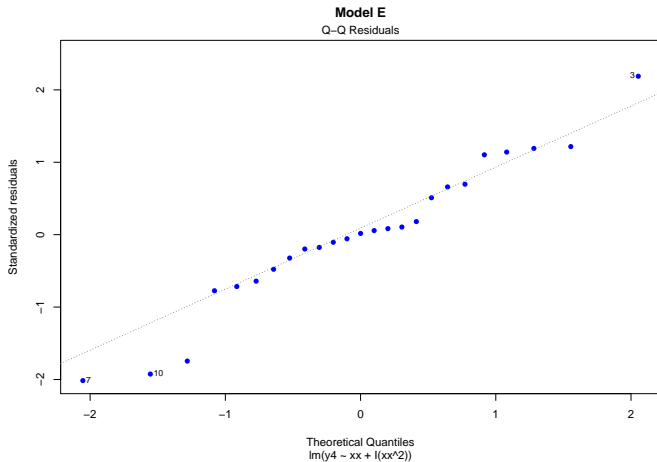


Gaussianity: Example Q

Let's go back to example Q and graph the quantiles plots before and after adding the quadratic term.



Gaussianity: Example Q



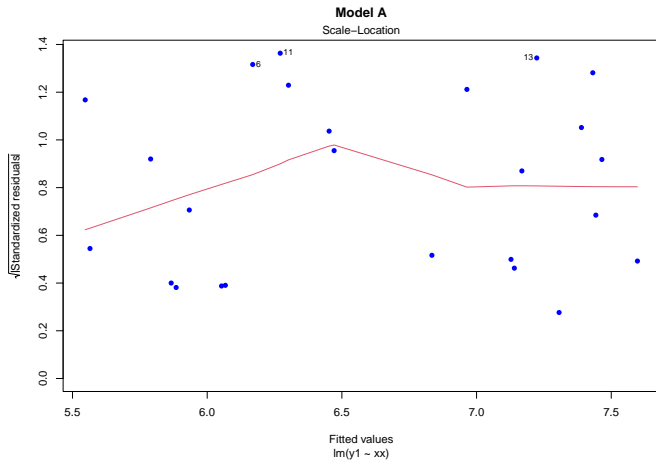
A third graph that is also useful for detecting departures from the assumptions, is similar to the first one on a different scale.

Instead of the residuals, the square root of the absolute value of the standardized residuals is plotted against fitted values, so all values in the y axis are positive.

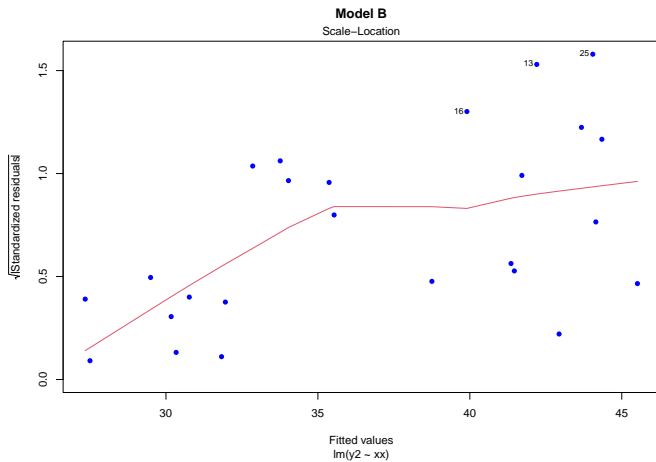
Again, we expect to see no structure or patterns, but random points scattered on the graph.

Additionally, since the residuals have been standardized, large values indicate possible atypical points.

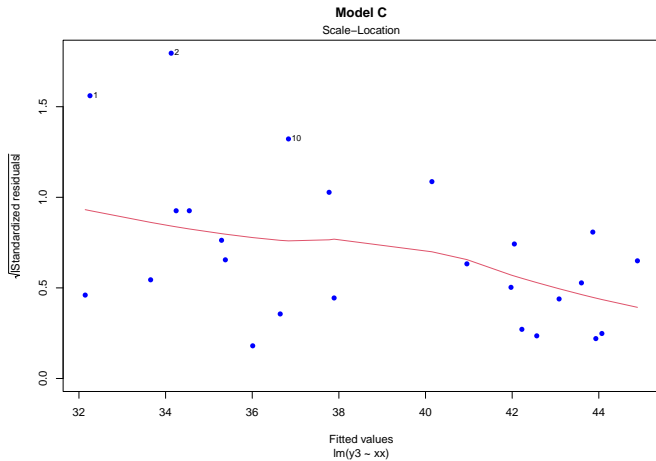
Diagnostic Plots



Diagnostic Plots



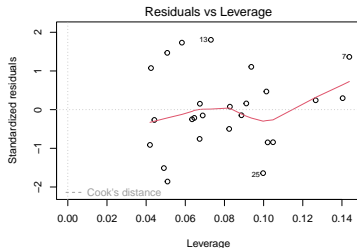
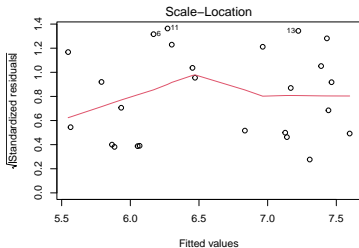
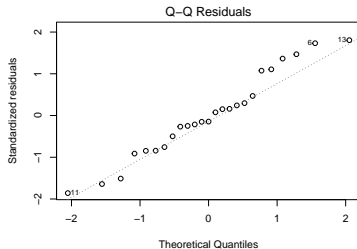
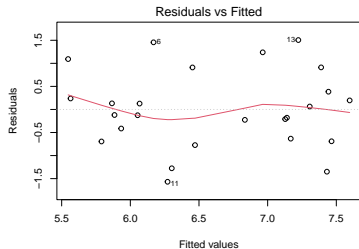
Diagnostic Plots



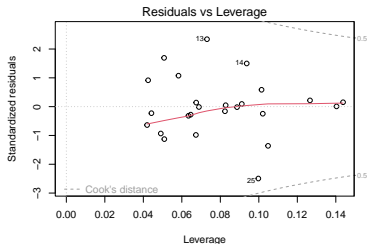
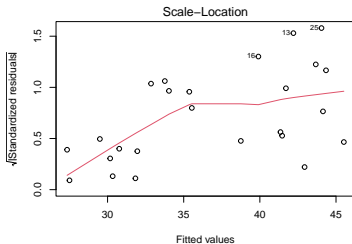
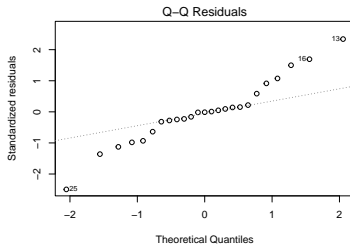
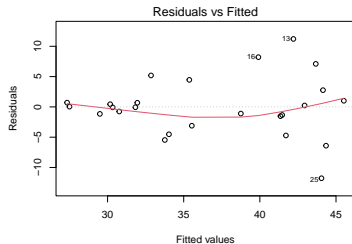
Finally, a graph of standardized residuals against leverage is usually drawn. This plot highlights the values that have highest influence on the parameter estimates.

As we have seen before, these four graphs can be obtained from an `lm` object using the `plot` function if the screen has been previously partitioned into four, as the following instructions illustrate for the first two models we fitted previously.

Diagnostic Plots: Model A



Diagnostic Plots: Model B



V32: Simple Linear Regression 6: Influential Points and Transformations

Influential and Atypical Points

Atypical points are data with large residuals in relation to the residuals of the rest of the observations.

Influential points are points that have a strong influence on the model. By this, we mean that if the model is fitted, excluding these points, the model changes substantially.

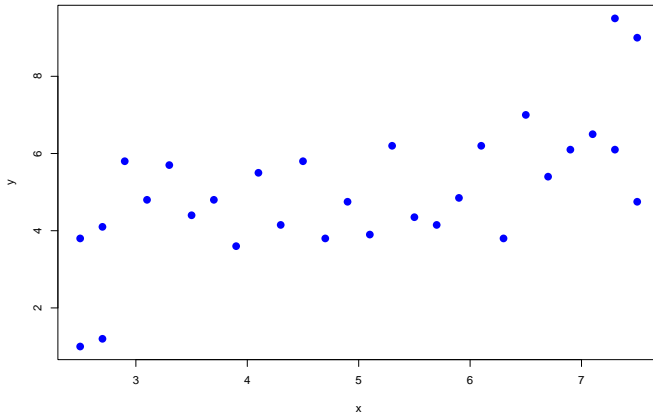
This example of the effect of influential points on regression is taken from Chatterjee, Hadi, & Price's book¹.

The data came from a study that measured the effect of the previous program's rating on the audience of a newscast.

¹Chatterjee, Hadi & Price *Regression Analysis by Example, 3rd Edition*, Wiley, 1999

Influential and Atypical Points

```
tv <- read.csv('TV.csv'); plot(tv, col = 'blue', pch = 16, cex =
```



The graph shows a possible linear relationship between the variables with a positive slope.

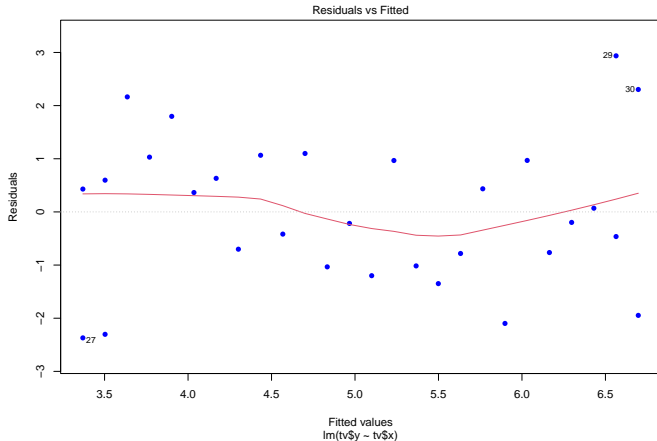
Influential and Atypical Points

```
fittv <- lm(tv$y ~ tv$x)
summary(fittv)
```

```
##
## Call:
## lm(formula = tv$y ~ tv$x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36994 -0.95755 -0.06405  0.96824  2.93634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7065     0.8172   2.088 0.045977 *
## tv$x          0.6654     0.1552   4.287 0.000194 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 28 degrees of freedom
## Multiple R-squared:  0.3963, Adjusted R-squared:  0.3747
## F-statistic: 18.38 on 1 and 28 DF,  p-value: 0.0001939
```

Influential and Atypical Points

```
plot(fittv, which=1, col = 'blue', pch = 16)
```



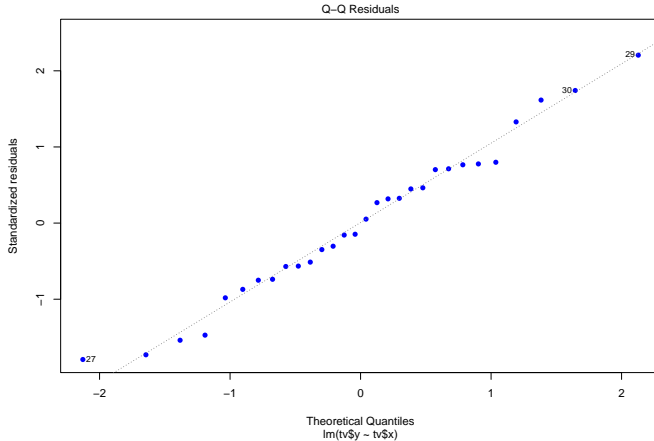
Influential and Atypical Points

In this plot, we see that there are four points, two on the upper right-hand corner with numbers 29 and 30, and two on the lower left-hand corner, one of which is numbered 27, which seem to 'tilt' the graph because the rest of the points seem to follow a decreasing pattern with a small negative slope.

Next, we plot the other three diagnostic graphs we have considered before.

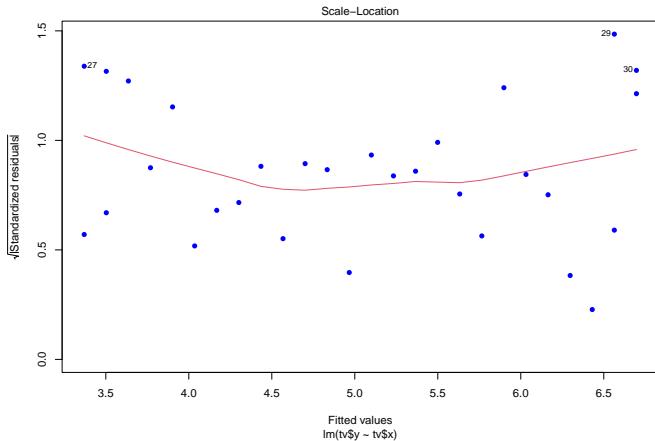
Influential and Atypical Points

```
plot(fittv, which=2, col = 'blue', pch = 16)
```



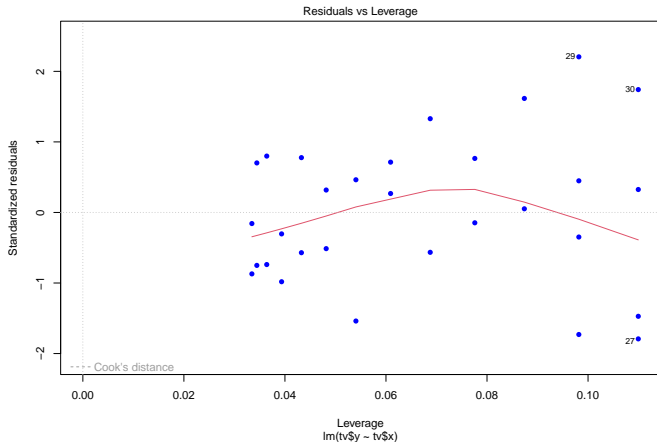
Influential and Atypical Points

```
plot(fittv, which=3, col = 'blue', pch = 16)
```



Influential and Atypical Points

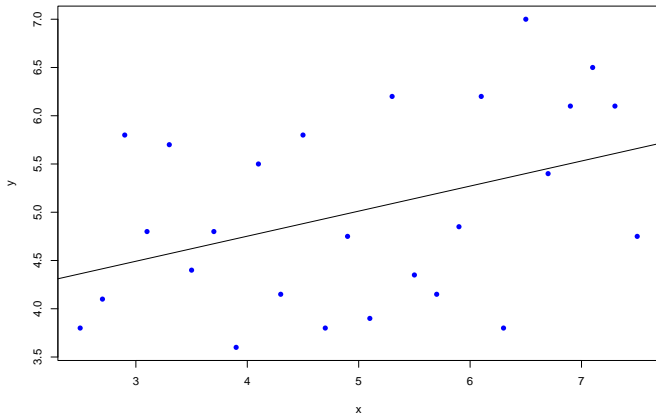
```
plot(fittv, which=5, col = 'blue', pch = 16)
```



Influential and Atypical Points

As an exercise we fit the model excluding these four points.

```
tv2 <- tv[1:26,]; plot(tv2, col = 'blue', pch = 16)
fittv2 <- lm(tv2$y ~ tv2$x)
abline(fittv2)
```



Influential and Atypical Points

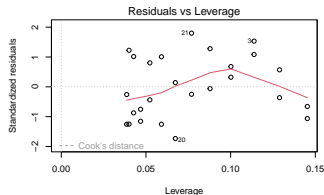
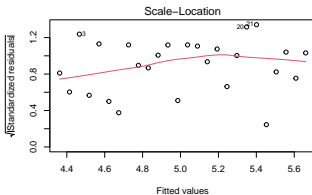
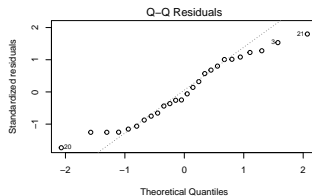
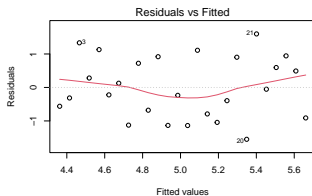
```
summary(fittv2)
```

```
##  
## Call:  
## lm(formula = tv2$y ~ tv2$x)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1.5491 -0.7635 -0.1375  0.8577  1.5990   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   3.7132     0.6314   5.881 4.56e-06 ***  
## tv2$x         0.2597     0.1209   2.147  0.0421 *    
## ---  
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.9251 on 24 degrees of freedom  
## Multiple R-squared:  0.1611, Adjusted R-squared:  0.1262   
## F-statistic: 4.609 on 1 and 24 DF,  p-value: 0.04211
```


Influential and Atypical Points

The results change radically. The slope, which is reduced to 0.26, is moderately significant, and the R^2 goes down to 0.1611.

```
par(mfrow = c(2,2))  
plot(fittv2)
```



Transformed Data

The data in the Bacteria file represent the number (in hundreds) of marine bacteria that survived 200 kilovolt X-ray exposure for periods ranging from 1 to 15 6-minute intervals.

The experiment was conducted to test the hypothesis that bacterial deaths occur when their 'vital center' is struck by a ray. This type of bacteria does not form groups or chains, so the number of bacteria can be estimated by plate counts.

If the theory is correct, the logarithm of the number of survivors must have a linear relationship to the length of the exposure.

If n_t represents the number of surviving bacteria at time t

$$n_t = n_0 e^{\beta t}, \quad t > 0,$$

where n_0 and β are the model parameters and have simple interpretations: n_0 is the number of bacteria at the beginning of the experiment, and β is the rate of destruction or death of the bacteria. Taking logarithms

$$\log n_t = \log n_0 + \beta t = \alpha + \beta t$$

where $\alpha = \log n_0$, and we see that this is a linear function of t . If we add a random error to the model, we get

$$\log n_t = \alpha + \beta t + \epsilon_t$$

and we can now fit a regression model.

If we go back to the original (exponential) model the error appears multiplicatively:

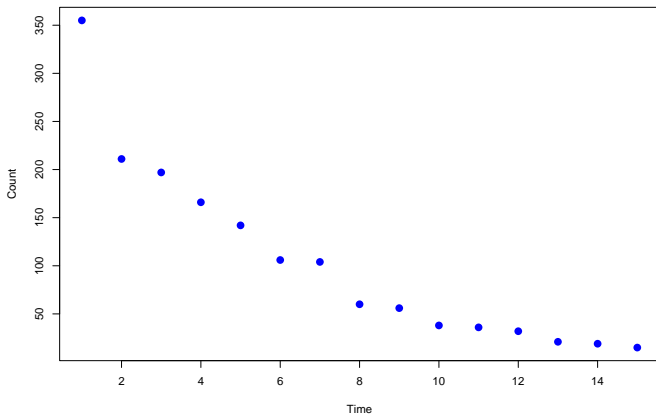
$$n_t = n_0 e^{\beta t} u_t,$$

where $u_t = e^{\epsilon_t}$ is a multiplicative error.

The model assumes that $\epsilon_t = \log u_t$ has a normal distribution and therefore u_t must have lognormal distribution.

We start graphing the data

```
Bacteria <- read.csv('Bacteria.csv')  
plot(Bacteria, col = 'blue', pch = 16, cex = 1.5)
```



The graph suggests a non-linear relationship between the two variables.

However, we proceed to fit a linear model to study the consequences. The model is

$$n_t = \alpha + \beta t + \epsilon_t,$$

```
attach(Bacteria)
fitbac <- lm(Count ~ Time)
summary(fitbac)
```

```
##
## Call:
## lm(formula = Count ~ Time)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
##	-43.867	-23.599	-9.652	10.223	114.883

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	259.58	22.73	11.420	3.78e-08 ***
## Time	-19.46	2.50	-7.786	3.01e-06 ***

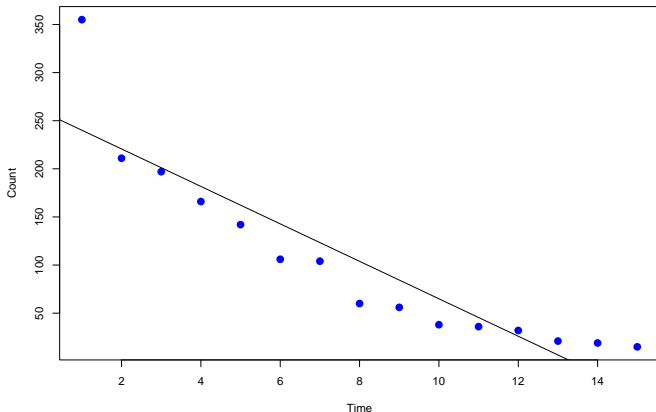
```
## ---
## Signif. codes:
```

	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1
##											

```
## Residual standard error: 41.83 on 13 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8098
## F-statistic: 60.62 on 1 and 13 DF,  p-value: 3.006e-06
```

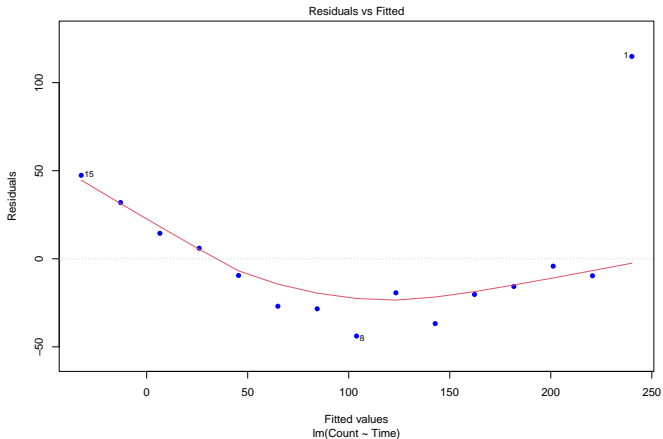

Although the regression coefficient is significant and we have a high value for R^2 , the linear model is not appropriate.

```
plot(Bacteria, col = 'blue', pch = 16, cex = 1.5)  
abline(fitbac)
```



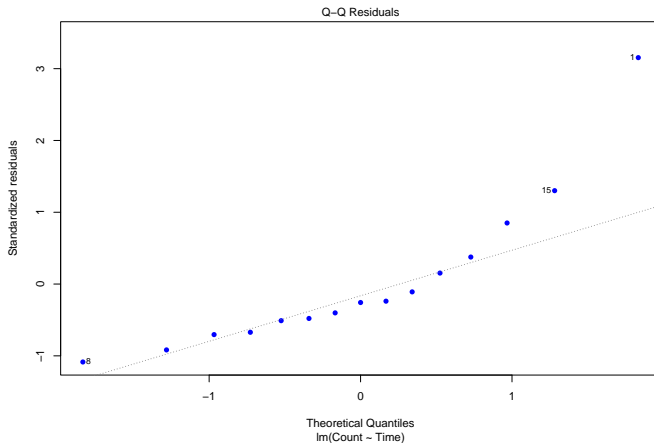
We find more evidence of this in the residuals graphs:

```
plot(fitbac, which=1, col = 'blue', pch = 16)
```



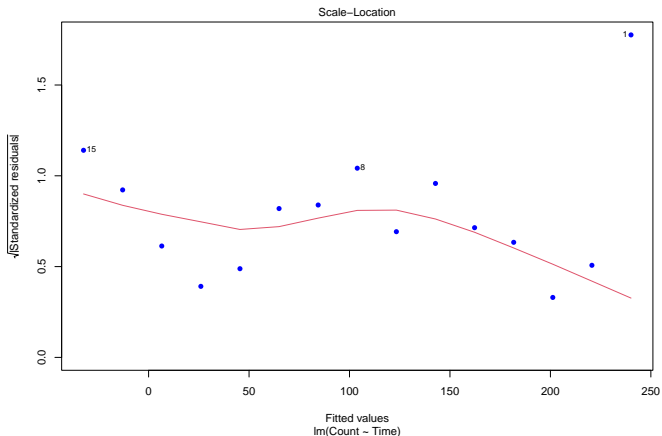
We find more evidence of this in the residuals graphs:

```
plot(fitbac, which=2, col = 'blue', pch = 16)
```



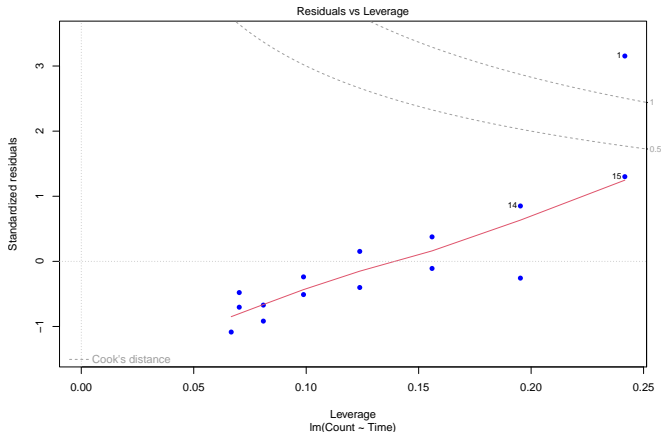
We find more evidence of this in the residuals graphs:

```
plot(fitbac, which=3, col = 'blue', pch = 16)
```



We find more evidence of this in the residuals graphs:

```
plot(fitbac, which=5, col = 'blue', pch = 16)
```



The first graph, residuals vs. fitted values, shows that this model does not explain all the relation existing between these two variables.

The quantile plot shows disagreement on the right tail of the distribution.

The third graph shows again that there is a structure in the residuals that has not been included in the model.

Finally, the last graph shows that there are some highly influential points in the regression.

Let's look now at the model using a logarithmic transformation

```
fitlogbac <- lm(log(Count) ~ Time)
summary(fitlogbac)
```

```
##
## Call:
## lm(formula = log(Count) ~ Time)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
##	-0.18445	-0.06189	0.01253	0.05201	0.20021

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	5.973160	0.059778	99.92	< 2e-16 ***
## Time	-0.218425	0.006575	-33.22	5.86e-14 ***

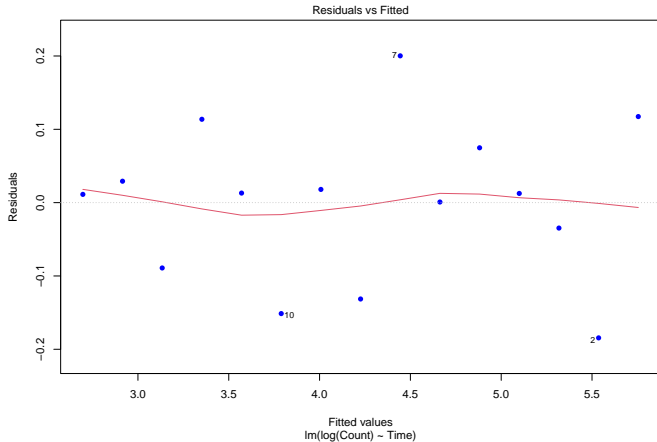
```
## ---
## Signif. codes:
```

##	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1
----	---	-------	-------	------	------	-----	------	-----	-----	-----	---

```
##
## Residual standard error: 0.11 on 13 degrees of freedom
## Multiple R-squared: 0.9884, Adjusted R-squared: 0.9875
## F-statistic: 1104 on 1 and 13 DF, p-value: 5.86e-14
```

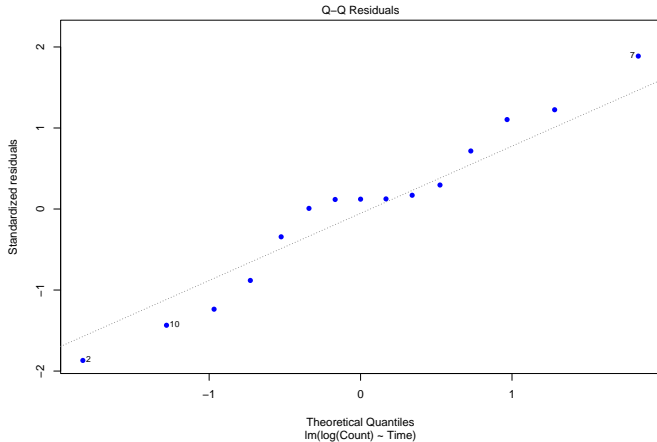
Transformed Data

```
plot(fitlogbac, which=1, col = 'blue', pch = 16)
```



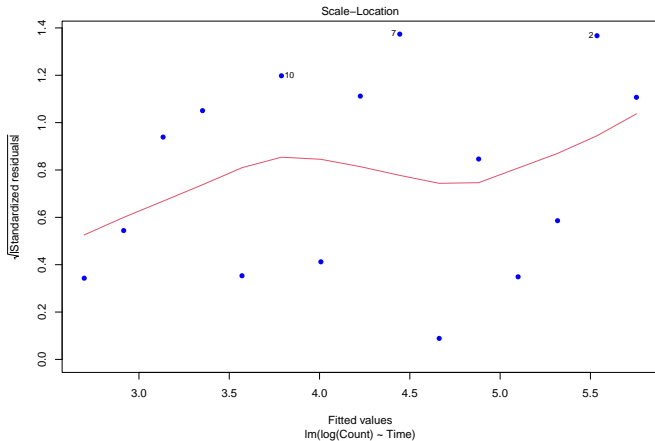
Transformed Data

```
plot(fitlogbac, which=2, col = 'blue', pch = 16)
```



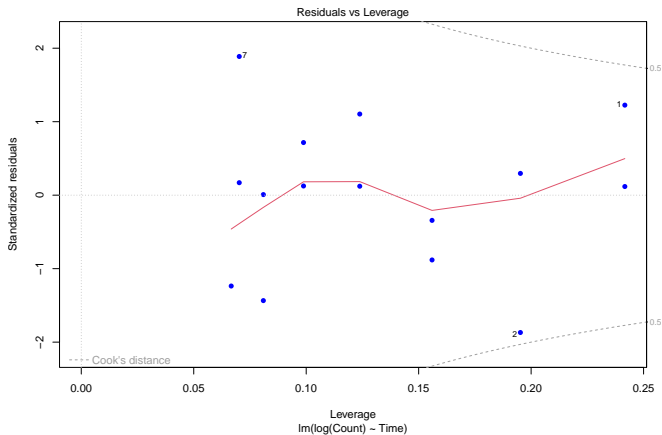
Transformed Data

```
plot(fitlogbac, which=3, col = 'blue', pch = 16)
```



Transformed Data

```
plot(fitlogbac, which=5, col = 'blue', pch = 16)
```



We now see that the regression coefficients are significant, standard errors are reasonable, and the model explains about 98% of the variation in the data.

The residual graphs also show a considerably improved fit.

The residuals appear to be randomly distributed, the fit of the experimental data and model predictions is good, and the dispersion of the residuals has been considerably reduced.

The only graph that is not entirely satisfactory is the normal qq-plot, but this may be because we have little data.

The linear model for $\log(\text{Count})$ is

$$\log(\text{Count}) = 5.97 - 0.219 \cdot t$$

where t is time, and the exponential model is

$$\text{Count} = e^{5.97 - 0.219 \cdot t} = 392.75e^{-0.219 \cdot t}$$

Problem List 8