5 Regression Diagnostic

5.1 Introduction

The conclusions we derived so far are all based on the assumption that the model is an appropriate description of the population.
This assumption includes:

1. The observations are independent.
2. The **error** is normally distributed.
3. The error has mean 0 and variance σ^2 independent from the values of the regressors (homoscedasticity assumption).
4. Regressors and response are related linearly.

Before the results from the analysis can be trusted these assumptions have to be checked to be reasonable. Should there be sufficient evidence from the sample that they are violated the model is found inappropriate and needs adjustment. All tests and confidence intervals become invalid. Given this statement, one might want to consider to start with the model diagnostics before proceeding with the tests. In addition to checking the assumptions one should also check for outliers and influential values, which might indicate a data input error, or an experimental error.
The first requirement can only be checked from the description of the experiment. Were the measurements independently taken? If yes, we can continue, if no, then the covariance structure has to be fixed. See General Multiple Linear Regression later in the course.
The remaining requirements are checked with the help of the (different) residuals for an estimated model.

5.2 Residuals

The residual vector $\vec{e}$ was previously defined as

$$\vec{e} = \vec{Y} - \hat{Y}$$

If the model is accurate the residual vector is multivariate normal with mean $\vec{0}_n$ and Covariance matrix $\sigma^2(I_n - H)$. Since this matrix is in general not a diagonal matrix, this means that the residuals are in general not independent. If h_{ij} is the (i, j) entry of H, then $Var(e_i) = \sigma^2\ (1 - h_{ii})$, and $Cov(e_i, e_j) = -\sigma^2\ h_{ij}$.

The residuals can be viewed as realizations or measured values of the model error, and as such are helpful for checking the assumptions concerning the model error as listed above.

Different approaches exist to standardize or scale the residuals, to be able to judge for example the existence of outliers which can not be explained by the normal distribution.

Definition 5.1.

1. The standardized residuals are defined as
$$d_i = \frac{e_i}{\hat{\sigma}},\ 1 \leq i \leq n$$
2. The internally Studentized residuals are defined as
$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}},\ 1 \leq i \leq n$$
where h_{ii} is the i^{th} diagonal element of the hat matrix, $H = X'(X'X)^{-1}X$.
3. The PRESS (Prediction Sum of Squares) residuals are defined as
$$e_{(i)} = Y_i - \hat{Y}_{(i)}$$
where $\hat{Y}_{(i)}$ is the fitted value of the i^{th} response based on all observations but the i^{th}.
4. The jackknife or externally studentized residuals are defined as
$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},\ 1 \leq i \leq n$$
where $\hat{\sigma}_{(i)}$ is the estimate for σ based on all measurements but the i^{th}.
5. h_{ii} is called the leverage of a point.

Lemma 1.

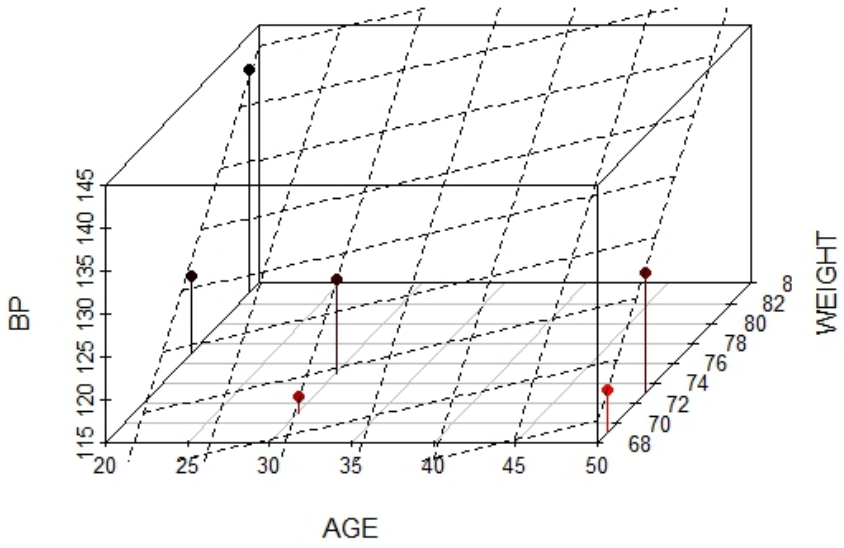
- $\sum_{i=1}^{n} d_i = 0,\ E(d_i) = 0, 1 \leq i \leq n,\ \frac{1}{n}\sum_{i=1}^{n} Var(d_i) = 1.$
- $E(r_i) = 0,\ Var(r_i) = 1, 1 \leq i \leq n.$
- (Proof see text book.)
$$e_{(i)} = \frac{e_i}{1-h_{ii}},\ 1 \leq i \leq n$$
and $E(e_{(i)}) = 0,\ Var(e_{(i)}) = \sigma^2/(1-h_{ii}).$
-
$$t_i = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}},\ 1 \leq i \leq n$$
and $E(t_i) = 0,\ Var(t_i) = 1.$
- If a measurement Y_i follows the MLRM, then t_i is t-distributed with $df = n-p-1$.
- h_{ii} measures how far the point $(x_{i1}, \ldots, x_{ik})$ falls away from the centroid of the data.

Continue Example. ??
Continue the example on the effect of age and weight on blood pressure and the analysis of the model:

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

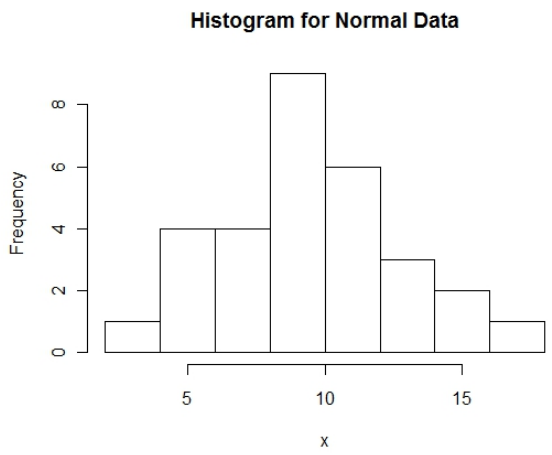
The following table displays the residuals for the model:

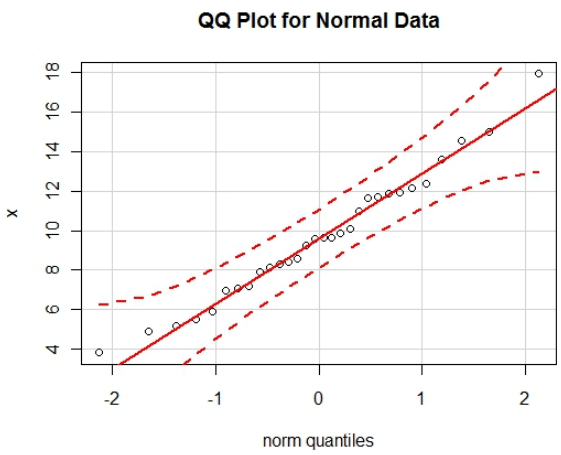
i	$e_i = y_i - \hat{y}_i$	$d_i = e_i/\hat{\sigma}$	$r_i = e_i/\hat{\sigma}\sqrt{1-h_{ii}}$	h_{ii}	$e_{(i)} = e_i/(1-h_{ii})$	t_i
1	0.005	0.004	0.005	0.488	0.010	0.004
2	0.853	0.615	1.267	0.764	3.622	1.52
3	-1.869	-1.346	-1.720	0.387	-3.051	-12.18
4	0.503	0.362	0.401	0.185	0.616	0.34
5	1.021	0.736	1.111	0.562	2.328	1.18
6	-0.513	-0.370	-0.595	0.614	-1.329	-0.52

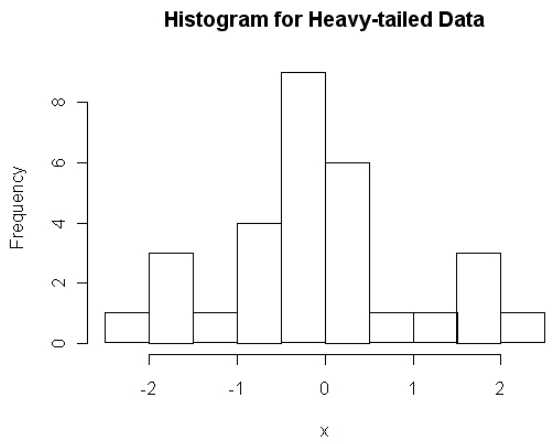


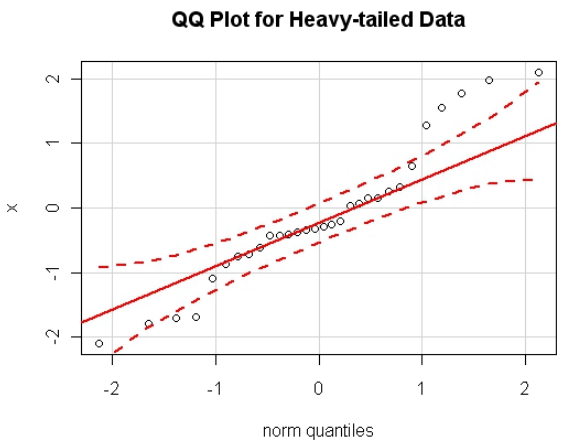
5.3 Check for Normality

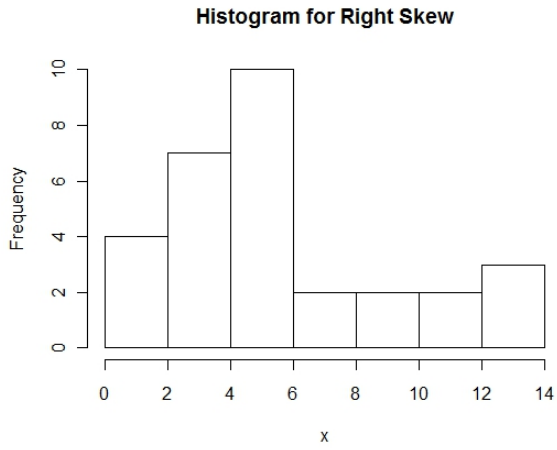
To check the normality assumption a normal probability plot (QQ-plot) and/or a histogram for the residuals should be produced.
The normal probability gives a visual comparison of the percentiles from the residuals with percentiles from a normal distribution. If these match the graph will show a straight line. See graphs below for explanations for deviations from this pattern

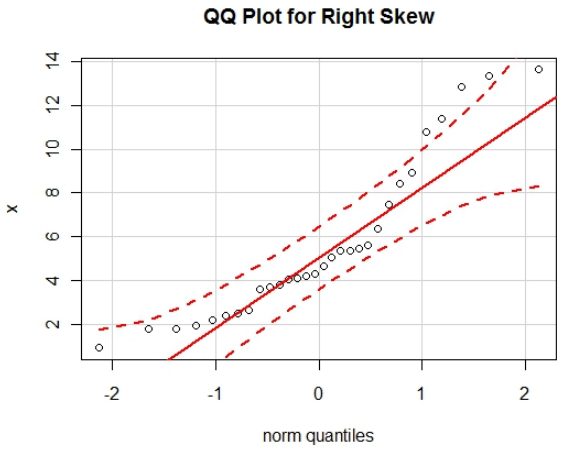








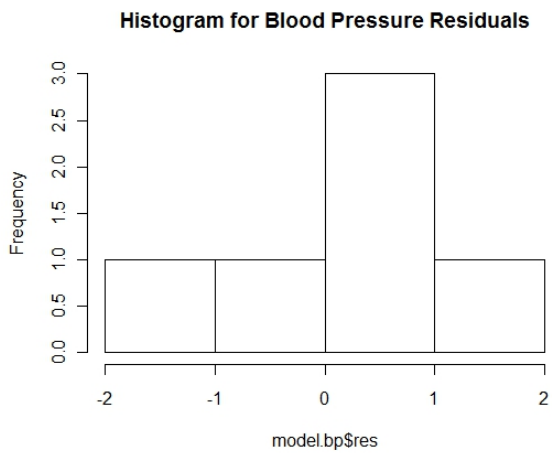


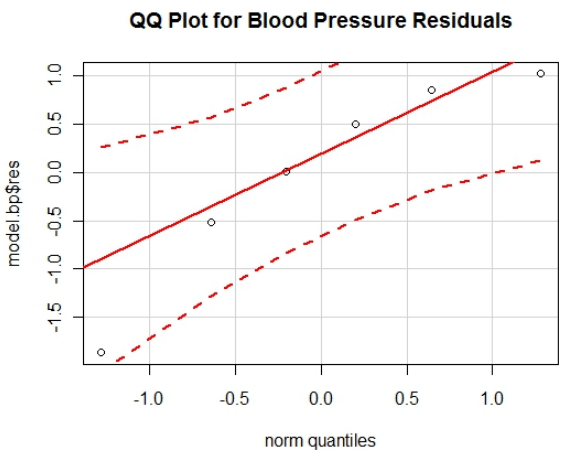


Continue Example. ??

QQ-plot for the residuals from the blood pressure data in the model

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

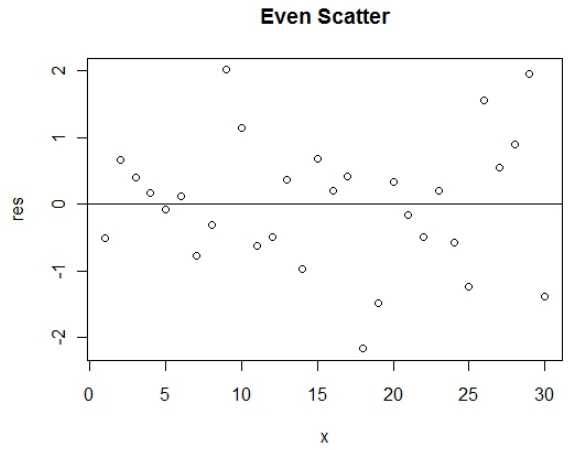


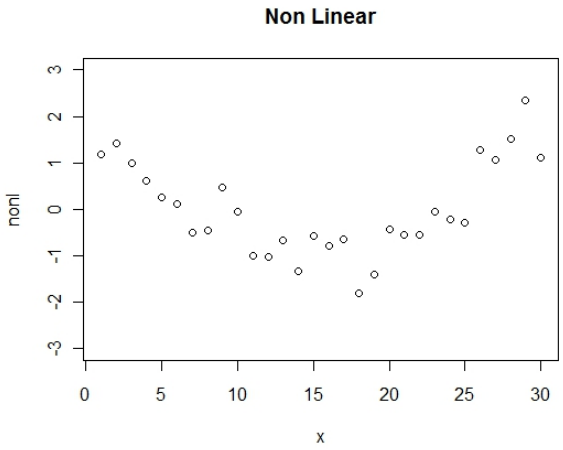


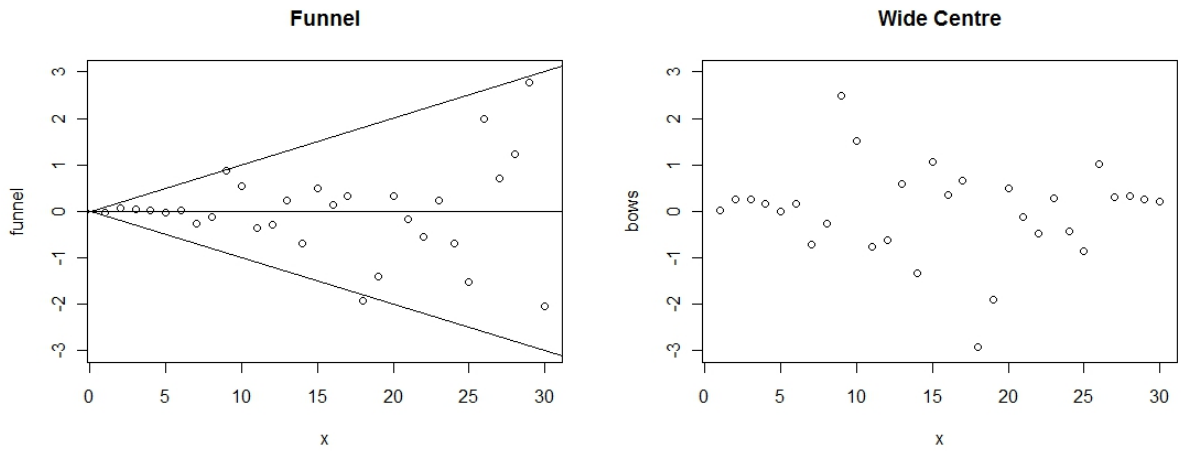
Some word of caution about the Shapiro Wilk Test:
The Shapiro Wilk Test tests the null-hypothesis that a sample originates from a normal population. For the model to be appropriate the test should not find evidence against the null hypothesis.
But as we know, a non-significant test result does not mean that there is evidence for H_0 to be true, only lack of evidence against it. In consequence my recommendation is to assess the visual tools (histogram and QQ-plot) for outliers and symmetry to check for normality and skip the Shapiro Wilk test, or at least to not put too much emphasis on its result.

5.4 Checking Homoscedasticity

To check if the assumption of equal variance throughout the range is justified, plot the residuals against the different regressors and the fitted values.
The following patterns are common results



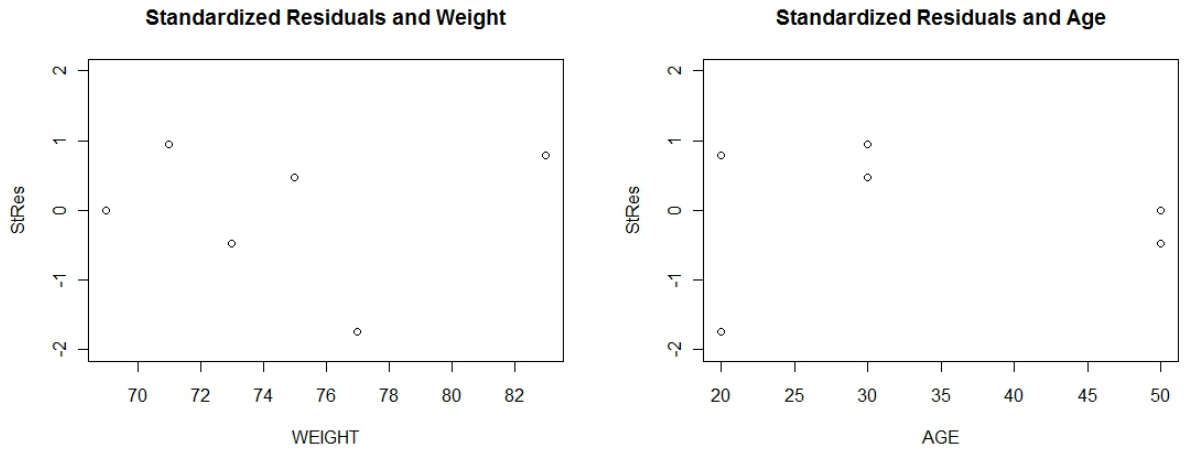


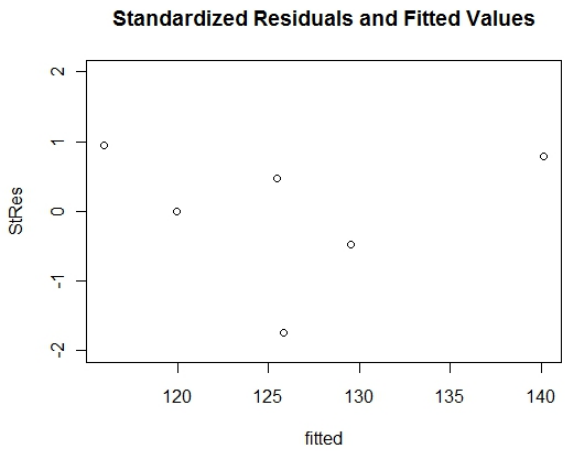


If the residual plots resemble even scatter there is no reason for concern about a violation of the homoscedasticity assumption. If a pattern similar to the nonlinear graph is displayed this indicates, that the relation ship between this regressor and the response is no well modeled. A transformation of the regressor or the response might be required to fix this problem. It could also indicate that an important variable is missing in the model.
The two graphs at the bottom show possible patterns indicating a violation of homoscedasticity. It might be necessary to weight the data point differently (see a later chapter).

Continue Example. ??
Residual Plots from the blood pressure data in the model

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$





Conclusion?

The third plot is called the *Residual Plot* (residuals against the fitted values). This plot is not directly related to the assumptions but is still sometimes helpful to detect outliers or model violations. They are interpreted in the same way as the residuals and predictor scattergrams.

5.5 Checking Model Structure

As explained in the subsection above, those plots might indicate a problem with the model structure, but more effectively for detect a misspecification of the model are Partial Residual Plots.

Let $X_{(1)}$ be the design matrix, excluding the column $\vec{x}_1$.

1. Find $\vec{e}[\vec{Y}|X_{(1)}]$, the residuals from the model $\vec{Y} = X_{(1)}\vec{\beta}_{(1)} + \vec{\varepsilon}_{(1)}$
2. Find $\vec{e}[\vec{x}_1|X_{(1)}]$, the residuals from the model $\vec{x}_1 = X_{(1)}\vec{\beta}_{x,(1)} + \vec{\varepsilon}^{\#}$,
3. Plot $\vec{e}[\vec{Y}|X_{(1)}]$ against $\vec{e}[\vec{x}_1|X_{(1)}]$.

The least squares slope for the model

$$\vec{\varepsilon}[\vec{Y}|X_{(1)}] = \beta_1 \vec{e}[\vec{x}_1|X_{(1)}] + \varepsilon^*$$

is equal to $\hat{\beta}_1$.

The rationale for the partial residual plot is:

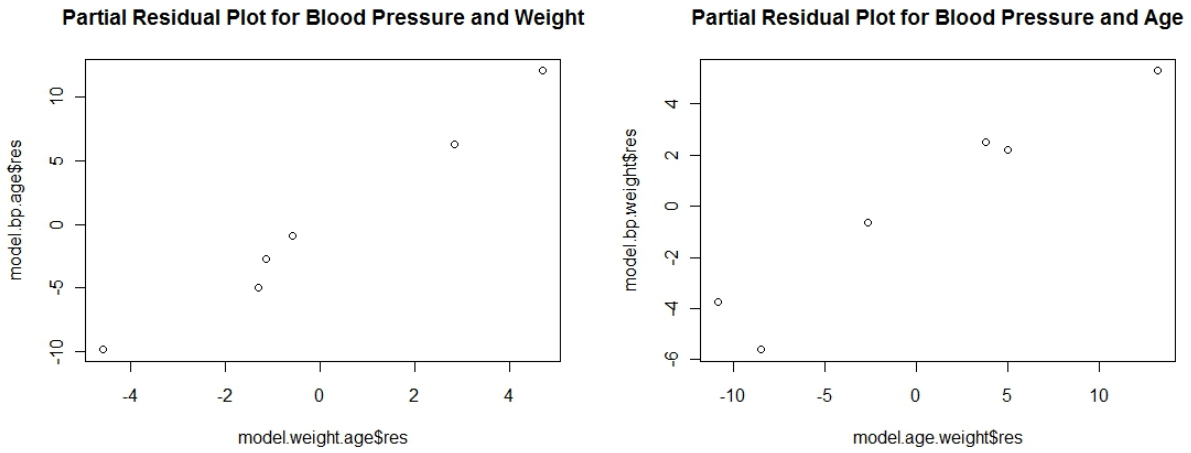
1. First take out of $\vec{Y}$, what can be explained by all regressors but x_1.
2. Second find what is "unique" in x_1, by eliminating from x_1, what can be explained by the remaining regressors.
3. Now investigate, if the "unique" information from x_1 has additional information about what in the response can not be explained by the other regressors. The plot will display the nature of the relationship between the unique parts in x_1 and Y. If the model is appropriate then the graphs should show a linear pattern.

Continue Example. Investigate the true nature of the relationship between weight and blood pressure.

1. Fit the model
$$bp = \beta_0 + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$
the residuals are $\vec{e}[\vec{bp}|X_{(1)}] = (2.7, -12.1, 4.9, 0.9, 9.9, -6.3)$.

2. Fit the model
$$weight = \beta_0^{\#} + \beta_2^{\#}(age) + \varepsilon^{\#}, \quad \varepsilon^{\#} \sim \mathcal{N}(0, \sigma^2)$$
the residuals are $\vec{e}[weight|X_{(1)}] = (-1.1, 4.7, -1.3, -0.6, -4.6, 2.9)$.

3. Plot the two residual vectors.

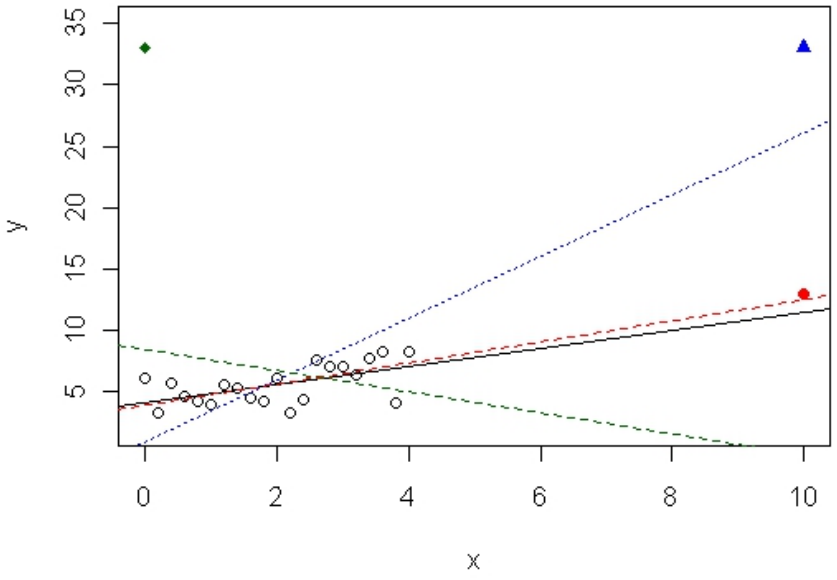


In a similar way the partial residual plot for age was found. Both plots show a linear pattern and support the model.

The usefulness of partial residual plots is limited when several variables have been misspecified. They also do not indicate the presence of interaction effect between regressors. They can be misleading in the presence of multicollinearity.

5.6 Identifying Outliers and Influential Values

An outlier is an extreme observation. Outliers can be outliers in regard to the response variable (Y-space) or in regard to the values of the regressor variables (X-space). High influence measurements are those where small changes in X or Y have a large influence on the estimated regression function. All measurements which have "large" residuals and/or fall apart from the general pattern in the residual plots are potential (Y-space) outliers and should be investigated. If they are high influence measurements, then they might have to be removed from the data-set and separately reported.



- The solid circle represents an outlier in x-space and is a high leverage point (a point falling away from the other measurements), but is not highly influential, the least squares line is not very much influenced the presence of the point, compare the solid line with the dashed line.
- The solid triangle is an outlier in x space and high leverage point as well, and is also highly influential. The intercept and slope change significantly when adding this observation, compare the solid line with the dotted line.
- The solid diamond is an outlier in $y - space$, and is highly influential on the least squares estimators of slope and intercept. Compare the line with the negative slope with the solid line.

It is important to identify all outliers and double check data entry, highly influential observations should be removed and reported separately.

Identification of outliers start with the inspection of the different residuals. For example large standardized residuals or studentized residuals (larger than 3) point to potential outliers. The comparison of jackknife residuals with the studentized residuals is especially helpful to identify outliers and highly influential measurements.
In the example the third measurement has a very high jackknife residual (absolute and in comparison with the studentized residual), and therefore a test for it being an outlier seems to be appropriate. Since the jackknife residuals are t-distributed if the MLRM model describes the observation, these can be used to conduct an outlier test.
Since it is appropriate to test each measurement it is important to apply a Bonferroni correction when conducting this test, since conducting n tests can result in a large overall error if no correction is applied.

Continue Example.

1. H_0 : third measurement fits the model (based on all measurements but the third) versus H_a : H_0 is not true

   $\alpha = 0.05/n = 0.05/6 = 0.0083$

2. See above

3. $t_0 = -12.18$, $df = 6 - 3 - 1 = 2$

4. P-value= $2P(t > 12.18) = 0.00667$ (using R: 2*(1-pt(12.18,2)))

5. Since P-value is smaller than α, Reject H_0.

6. According to the test measurement 3 should be considered an outlier.

What is causing this being an outlier? Error when entering data? Wrong measurement? Another explanatory variable? Or just an uncommon value?

Comment:

1. Two or more outliers close to each other can hide each other, and are not detected by using jackknife residuals.

2. Single outliers are not so much a problem in large data sets.

What to do with outliers?

1. Check for data entry mistake.

2. Go back to the experiment and try to answer what caused the measurement.

3. Exclude the measurement from the analysis and report separately, for statistical honest reporting.

Reoccurring outliers can be systematic:

Example 5.1.
(Reported in Faraway) A NASA satellite was launched to record and monitor atmospheric information, but missed for many years (until 1985) the decline in ozone above the Antarctic, because the data analysis tool automatically removed outliers from the analysis. The low ozone values therefore were treated as outliers (and not reported) until a British survey reported the decline in ozone above the Antarctic in 1985.

Leverage:
The hat matrix, $H = X'(X'X)^{-1}X$, is proportional to the covariance matrix of $\hat{Y}$, therefore the diagonal entries, h_{ii}, measure the variance in $\hat{Y}$, or one could look at $1 - h_{ii}$ being proportional to the variance of the i^{th} residual. Therefore the (i, j) entry measures the impact or *leverage* of the i^{th} measurement on the j^{th} fitted value, $\hat{Y}_j$. The i^{th} diagonal entry indicates the distance of

the i^{th} measurement from the center of the x-space. Observation falling far away from the center potentially have a high impact on the least squares estimates of $\vec{\beta}$. This is not always true, remember the solid dot in the diagram above being an outlier in x-space but not being influential. We consider measurements with $h_{ii} > 2p/n$ of being potentially high leverage points.

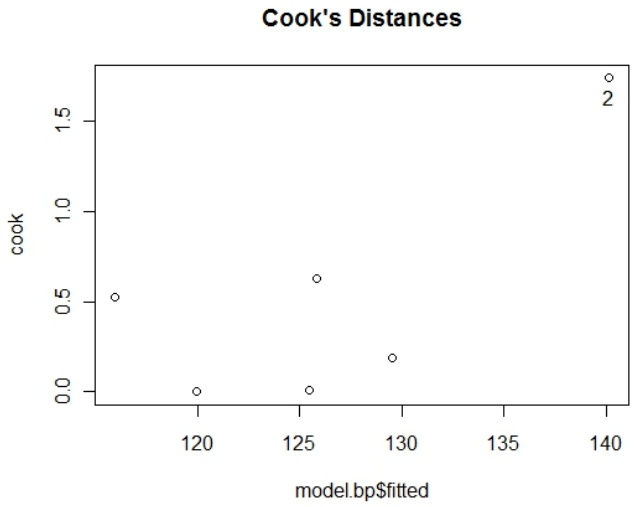
Cook's Distance:
Cook's distance measures the difference in the least squares estimator based on all measurements versus all but one measurement and in that way measures how sensitive the estimates are to the presence of this one measurement.

Definition 5.2.
Let $\hat{\beta}_{(i)}$ notify the least squares estimate of the parameter vector $\vec{\beta}$ based on all measurements but the i^{th}. The Cook's distance for measurement i is defined as

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})'(X'X)(\hat{\beta} - \hat{\beta}_{(i)})}{p\,\hat{\sigma}^2} = \frac{(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})}{p\,\hat{\sigma}^2},\ 1 \leq i \leq n.$$

Observations with large Cook's D identify highly influential measurements.
In order to judge if D_i is large it is compared to the α percentile of the F distribution with $df_1 = p$ and $df_2 = n - p$: with the interpretation:
If $D_i = F_{0.5,p,n-p}$ then just removing point i from the data set would move $\hat{\beta}_{(i)}$ to the boundary of a 50% confidence region for $\vec{\beta}$ based on the complete data set. Quite a strong effect.
Since the $F_{0.5,p,n-p} \approx 1$, a Cook's distance> 1 is suspicious. This is just a rule of thumb, since Cook's distance is not truly F-distributed, but this rule works pretty well in practice.

Continue Example.
The Cook's distances for the six measurements are:
8.8×10^{-6}, 1.73, 0.624, 0.0121, 0.526, 0.187
Plotting them against the fitted values gives (identifying the most extreme):



Cook's distance for measurement 2 is the largest and grater than 1. We now should investigate the effect of removing measurement 2 from the sample on the least squares estimates.

Regressor	$\hat{\beta}$	$\hat{\beta}_{(2)}$
(Intercept)	-66.13	-42.27
AGE	0.44	2.07
WEIGHT	2.40	0.41

Conclusion?

After identifying highly influential observations, the question arises what should be done about them. Their treatment is similar to the outlier treatment, first check if they are caused by an error in the data entry, or a wrong measurement in the experiment. If it is a legitimate observation, then it is hard to argue for it to be removed.
If the number of influential observations, which can not be discarded, is substantial, one should use a robust estimation technique.