Model checking

- **Univariate Diagnostics and Graphs:** The first step in a regression analysis is generally to examine all of the variables in the model.
- One dimensional graphs such as histograms or boxplots are also very useful to see if there are any outliers, a point which deviates from the model, in the covariates or response.
- Outliers in a covariate can often indicate a point that will have very high influence on the fitted model. Outliers in the response are often points for which the model will not fill well. In some cases it may be wise to omit them from the analysis.

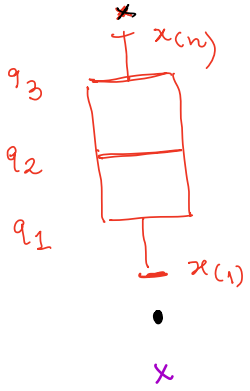
Box plot – variable x : sample $\underset{\sim}{x} = (x_1, \dots, x_n)$

sort sample : $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$

1st quartile = q_1 ; median = $q_2 = me$; 3rd quartile

2nd quartile

IQR = $q_3 - q_1$



if $x > q_3 + 1.5\,IQR$ or $x < q_1 - 1.5\,IQR$ $\Rightarrow$ mild outlier (•)

if $x > q_3 + 3.0\,IQR$ or $x < q_1 - 3.0\,IQR$ $\Rightarrow$ strong outlier (x)

- Model checking

- Residual Analysis

the residuals from the M.R. model $e_i = y_i - \hat{y}_i$ play an important role in judging model adequacy. Analysis of the residuals is important to check the assumption: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

e_i Plots:

- histogram: e_i Normality
- QQ-plot (in R qqPlot command)

} check e_i Normality

Obs: $e_i = (y_i - \hat{y}_i)$

$E(e_i) = 0$ but since this is not a new observation as we did in Prediction, y_i and $\hat{y}_i$ are not independent. So

$$Var(y_i - \hat{y}_i) = Var(y_i) + Var(\hat{y}_i) - 2\,Cov(y_i, \hat{y}_i)$$

It can be show that $Cov(y_i, \hat{y}_i) = \sigma^2 \underset{\sim}{x}_i^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_i = Var(\hat{y}_i)$

so $Var(y_i - \hat{y}_i) = \sigma^2 + \sigma^2 \underset{\sim}{x}_i^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_i - 2\sigma^2 \underset{\sim}{x}_i^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_i =$

$= \sigma^2 - \sigma^2 x_i^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} x_i = \sigma^2 (1 - \underset{\sim}{x}_i^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_i)$

Notation: $\underset{\sim}{H} = \underset{\sim}{X} (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} X^T$ matrix Hat

$\hat{\underset{\sim}{Y}} = \underset{\sim}{X} \hat{\underset{\sim}{\beta}} = X (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{Y} = \underset{\sim}{H} \underset{\sim}{Y}$ the matrix $\underset{\sim}{H}$ transforms the observed values of $\underset{\sim}{Y}$ into a vector of fitted values $\hat{\underset{\sim}{Y}}$

the i-th diagonal element of the hat matrix is:

$$h_{ii} = \underset{\sim}{x}_i^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_i$$

So; $Var(e_i) = \sigma^2 (1 - h_{ii})$

$e_i \overset{iid}{\sim} N(0, \sigma^2 (1 - h_{ii}))$

Properties of $\underset{\sim}{H}$ matrix:

Idempotent: $\underset{\sim}{H} \cdot \underset{\sim}{H} = \underset{\sim}{H}$

$\underset{\sim}{H}^T = \underset{\sim}{H}$

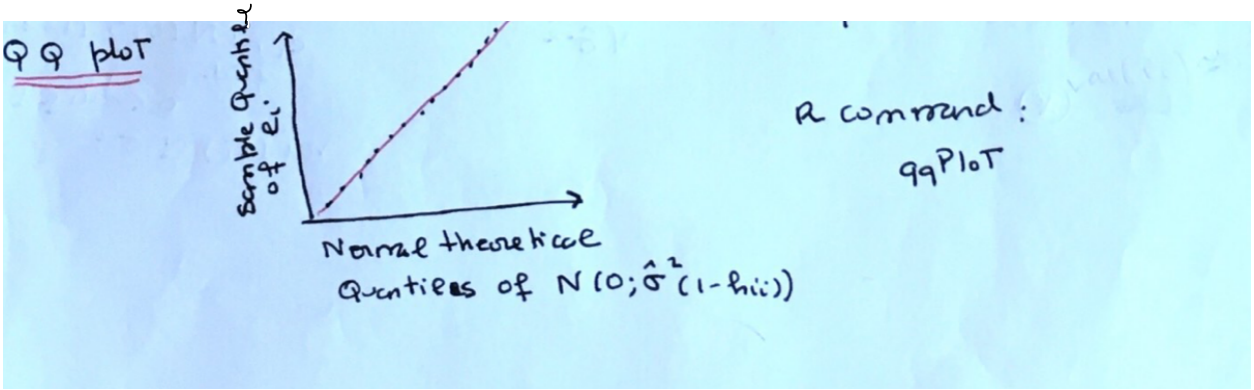
and $\sum_{i=1}^{n} h_{ii} = \text{trace}(\underset{\sim}{H}) = p$

and

$0 < h_{ii} \leq 1$

Model checking-QQ-Plot

- QQ-plot of e_i to check the normality assumption and possible outliers. In statistics, a QQ-plot ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). If the two distributions being compared are similar, the points in the QQ-plot will approximately lie on the line $y = x$.



We may also standardize the residuals $d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}$ $i = 1, \ldots, n$

If the errors ε_i are normal distributed approximately 95%. of the standardize residuals should fall in the interval $(-2; +2)$. Residuals that are far outside this interval may indicate the presence of an outlier

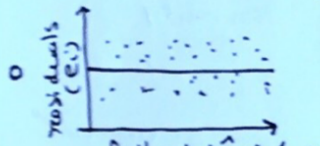
- Standardized residuals (d_i): residuals: $e_i = y_i - \hat{y}_i$ with $E(e_i) = 0$ and sample variance $\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^n e_i^2}{n-p} = MSE = \hat{\sigma}^2$
- $\boxed{d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}}$, $i = 1, \ldots, n$ with $E(d_i) = 0$ and $\text{VAR}(d_i) \approx 1$

Note: $e_i \sim N(0, \sigma^2(1 - h_{ii}))$ but since $0 < h_{ii} \leq 1 \Rightarrow$ $\text{var}(e_i) \leq \sigma^2$ (i.i.d.)

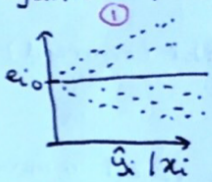
- If the fitted regression model is adequate, we expected the standardized residuals to look like independent draws from an $N(0, 1)$.

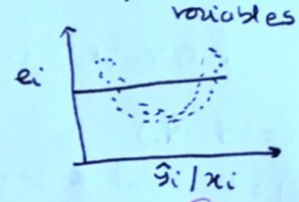
- Residual plots: The residuals e_i or standardized residuals $d_i = e_i/\sqrt{\hat{\sigma}^2}$ are used to obtain various residual plots for model checking:
  1. Plot e_i against the fitted model $\hat{y}_i = \mathbf{x}_i\hat{\beta}$, where $\mathbf{x}_i$ is the i-th row of the design matrix $\mathbf{X}$. This plot can be used to check the constant variance of ϵ_i. If the model is appropriate for the data the plot should show an even scatter. Any discernible pattern in the plot means that the regression equation does not describe the data correctly, since pattern forms when the residuals are unevenly distributed about the regression line. Outliers may also get identified in such a plot.

Plot residuals against fitted values ($\hat{y}_i$) / explanatory variables (x_i)



Ideal situation





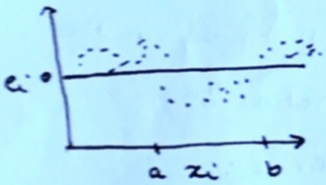
$i = 1, \dots, p-1$

Anomalies

① variance increasing in time or magnitude of x_i or y_i

transformation of response: $\sqrt{y}$; $\ln y$ or $1/y$

② Model inadequacy: transformation on x_i or y variable or both or other regressor should be considered



$x_i < a$ and $x_i > b$ } $e_i > 0 \Rightarrow$ model underpredicts the response

$a < x_i < b$; $e_i < 0 \Rightarrow$ model overpredicts the response

other type of scaled residuals:

- Studentized residuals: $R_i = \dfrac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ $\quad i = 1, \dots, n$

Better statistic to examinate potential outliers.

- Studentized residuals(r_i): Since $\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H})\mathbf{y}$, $\text{COV}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ and $\text{VAR}(e_i) = \sigma^2(1 - h_{ii}), i = 1, \ldots, n$.
- $$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, \quad h_{ii} = \mathbf{x}_i^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{x}_i$$

Model checking

- However, in simple linear regression, an observation that is unconditionally unusual in either its y or x value is called a univariate outlier, but it is not necessarily a regression outlier.
- While it is relatively easy to find outliers in univariate datasets and in simple regression, it is harder in multiple regression. An outlier in a regression setting may not be an outlier in any of the individual variables.
- Generally, regression outliers will have a large standardized residual in absolute value. Typically we will examine a point with $|d_i| > 2$ as a possible outlier.

High leverage points and influential observations

- It is helpful to distinguish leverage points and influential cases.
- **Leverage:** refers to the influence of an observation because it is outlying in the x-direction. Leverage statistics are based on the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, we have that:

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j.$$

  In addition, it can be shown that $0 < h_{ii} < 1$ for all i and $tr(\mathbf{H}) = p$.
- Thus, if h_{ii} is large relatively to other h_{ij} (in magnitude), then y_i will be a major contributor to the fitted value $\hat{y}_i$. Large leverage points have been defined as those bigger than $2p/n$.

High leverage points and influential observations

- The leverage h_{ii} has another interpretation. It measures the distance of $\mathbf{x}_i$ to the center of the explanatory variables, where $\mathbf{x}_i$ is the ith data point of the design matrix $\mathbf{X}$. For instance, consider the simple linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$. It can be shown that:

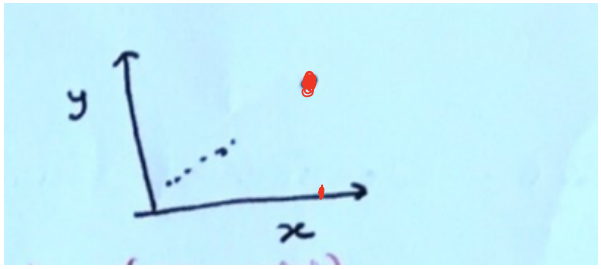
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Consequently, if the i-th data point of the explanatory variables is far away from the center, then it has a high leverage and pulls the model fit toward itself. It is, therefore, useful in linear regression analysis to check the high leverage points.

- Leverage points (outlying in x-direction)

  h_{ii} : Large : $h_{ii} > \frac{2p}{n}$

  R: hatvalues command



High leverage points and influential observations

- **Influential observations** of a linear regression model are defined as those points that significantly affect the inferences drawn from the data. Methods for assessing the influence are often derived from the change in the $\hat{\beta}$ if the observations are removed from the data.

- Influential point (Influential to estimation of $\hat{y}_i$)

  Cook Distance : Is a measure of the squared distance between the usual Least square estimate of β based in all observations and the estimate obtained when the i-th point is removed ($\beta_{(i)}$)

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2} = \frac{e_i^2 h_{ii}}{p\hat{\sigma}^2(1-h_{ii})^2} = \frac{r_i^2(\hat{\sigma}^2(1-h_{ii}))h_{ii}}{p\hat{\sigma}^2(1-h_{ii})^2}$$

$$= \frac{r_i^2 h_{ii}}{p(1-h_{ii})}, \quad i = 1, \ldots, n$$

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

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

Large value of D_i implies that the i-th point is influential

rule: $D_i > \frac{4}{n-p}$ Montegomery $D_i > 1$

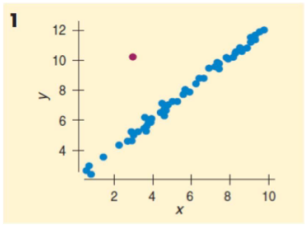
R: cooks.distance(reg.model) command

High leverage points and influential observations

- Cook's distance can thus be understood to depend on three quantities:
  1. The number of variables, p;
  2. A component reflecting how well the model fits y, r_i^2;
  3. A measure of how much an observation is discrepant from the rest of the data in the independent variables, h_{ii}.

Outliers, Leverage and Influence

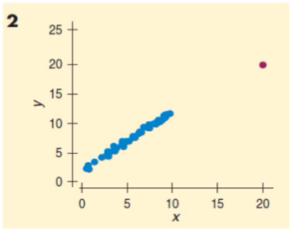
Tell whether the point is a high-leverage point, if it has a large residual and if it is influential.



- Not high-leverage
- Large residual
- Not very influential (maybe?)

Outliers, Leverage and Influence

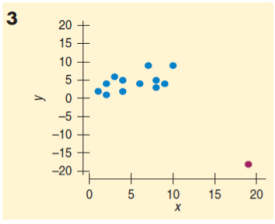
Tell whether the point is a high-leverage point, if it has large residual and if it is influential.



- High-leverage
- Small residual
- Not very influential

Outliers, Leverage and Influence

Tell whether the point is a high-leverage point, if it has a large residual and if it is influential.



- High-leverage
- Medium/Large residual
- Very influential (omitting the red point will change the slope dramatically!)

High leverage points, influential observations and outliers

To summarize:

- An observation might be an outlier in either the x or the y direction;
- Outliers in x are called leverage points. These are diagnosed with the leverage statistic, h_{ii};
- Outliers in the y direction can be diagnosed with studentized residuals; / standardize residuals
- An outliers may not be influential in the regression results,
- High leverage points are necessary but not sufficient conditions for influential observations;
- Observations are influential when the regression results change a lot as a consequence of leaving an observation out of the analysis. These are diagnosed with the Cook's distance statistics, D_i.

Multicollinearity

- If several explanatory variables are highly correlated (0.90 and above), some of the diagonal components of the inverse matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{C}$ will be very large. Consequently, the confidence intervals for some of the coefficients (say the β_j) tend to be very wide. This problem is called multicollinearity.

- For example, let suppose that we have just two explanatory variables x_1 and x_2. Now, suppose that the model fits well to the the data set. Then the overall test should reject $H_0 : \beta_1 = \beta_2 = 0$.

Multicollinearity

- Nevertheless, if x_1 and x_2 are highly correlated, the individual tests may lead to accept $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. This paradoxical situation can occur because both variables convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot.

- Multicollinearity occurs because the variables contain redundant information. If one of the variables doesn't seem logically essential to your model, removing it may reduce or eliminate multicollinearity.