Chapter 5: Model Evaluation and Residual Analysis

CZ

Fall, 2014

What is a good model?

Model fitting is easy, while finding a good model is hard.

- The true relationship between y and x's is unknown.
- A regression model is a mathematical approximation to the true but unknown relationship.
- *George Box: "All models are wrong, but some are useful."*
- A good model should be
  - complex enough to provide good fit to the observed data
  - simple enough for interpretation and further usage
- Principle of Parsimony: A simpler model is always preferred if it works!

Assumptions for Linear Regression Model

Goal: to assess the regression assumptions and model fitting. Recall

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{1}$$

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$
- ϵ_i and ϵ_j are uncorrelated (independent)
- For statistical inference, $\epsilon_i \sim N(0, \sigma^2)$.

Residuals

Diagnostics for a given model are usually carried out indirectly through an examination of the residuals.

- Residual: the difference between the observed y and the fitted y:

$$r_i = y_i - \hat{y}_i$$

- Random Error: the difference between the observed y and the expected y:

$$\epsilon_i = y_i - E(y_i)$$

If the model is appropriate for the data at hand, r_i should then reflect the properties assumed for the ϵ_i.

Properties of Residuals

We have: $\mathbf{r} = (\mathbf{I} - \mathbf{H})\epsilon$

- $E(\mathbf{r}) = 0$
- $\text{Var}(\mathbf{r}) = \sigma^2(\mathbf{I} - \mathbf{H})$
- Under normality assumption, $\mathbf{r} \sim MVN(0, \sigma^2(\mathbf{I} - \mathbf{H}))$

Standardized Residuals (Studentized Residuals)

$$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, i = 1, 2, \ldots, n$$

where h_{ii} is the ith diagonal element of $\mathbf{H}$. Under the assumptions of random erros, $d_1, \ldots, d_n$ are approximately N(0,1). (Are they independent?)

Residual plots for checking $E(\epsilon_i) = 0$

Potentially, the most important assumption for linear regression models is $E(\epsilon_i) = 0$. The likely causes for violation of this assumption are:

- Effect of x's on y is not in fact linear
- Omission of some important predictors

We shall consider four types of plot for checking this assumption:

- Residuals versus x_j, $j = 1, \ldots, p$
- Partial residuals versus x_j, $j = 1, \ldots, p$
- Residuals versus $\hat{y}$
- Added-variable plots

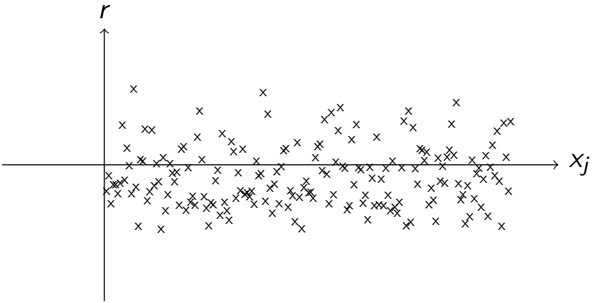
Residuals versus x_j

Suppose we fit a multiple regression model and calculate

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}).$$

- If x_j does have a linear effect on y, the regression model will remove these linear effects from y's, and the residuals are expected to be randomly scattered around 0.
- When we plot raw residuals $r_1, \ldots, r_n$ against the n values $x_{1j}, \ldots, x_{nj}$, we expect to see a random scatter for $j = 1, \ldots, p$.

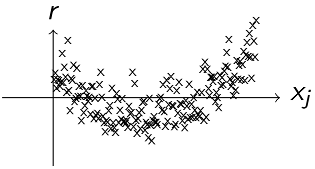
Residuals versus x_j



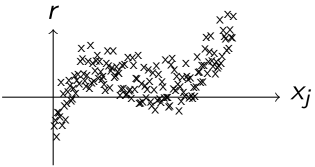
Points fall into a horizontal band around 0, and there is NO special trends.

Residuals versus x_j

On the other hand, if we see any obvious non-random pattern, it suggests non-linearity and we could adapt the way x_j is modelled. For example,



or



may require higher order terms (e.g. x_j^2, x_j^3) in the model $\Longrightarrow$ Polynomial Regression.

Partial residuals versus x_j

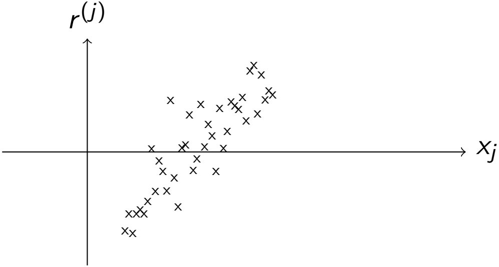
- Plots of the raw residuals are some times difficult to interpret because we have to decide whether the scatter looks random or not.
- Partial residuals: judge whether the plot looks linear - this is often easier.
- For each x_j, the partial residuals $r_i^{(j)}$ is defined as

$$\begin{aligned} r_i^{(j)} &= r_i + \hat{\beta}_j x_{ij} \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}) + \hat{\beta}_j x_{ij} \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i,j-1} + \hat{\beta}_{j+1} x_{i,j+1} + \cdots + \hat{\beta}_p x_{ip}) \end{aligned}$$

$i = 1, \ldots, n.$ The estimated linear effect of x_j is added back into the residuals.

Partial residuals versus x_j

For each x_j, when we plot $(r_1^{(j)}, \ldots, r_n^{(j)})$ versus $(x_{1j}, \ldots, x_{nj})$, we expect to see a linear trend if the model with a linear term in x_j is adequate. Hence the typical pattern of partial residual plots when the assumption is not violated is



for $j = 1, \ldots, p$.

Partial residuals versus x_j

Comment: In simple linear regression, we can see relationship between y and x simply by plotting the two variables. In a multiple regression situation, plotting y versus each x_j does not show the marginal effect of x_j. The y values are also affected by the remaining predictors. In the partial residuals plot, we remove the estimated effect of remaining predictors hence attempting to uncover the marginal effect of x_j on y. We can then judge whether this marginal effect is linear or not.

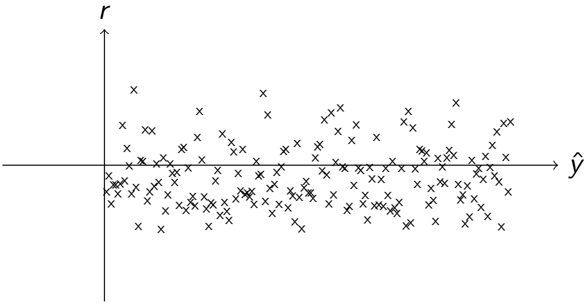
Partial residuals versus x_j

In R, a function `crPlots( )` in the `car` package has been made available to you to produce partial residual plots. See the example later.

Residuals versus $\hat{y}$

If the model is adequate ($E(\epsilon) = 0$), we have $\text{cov}(\mathbf{r}, \hat{\mathbf{Y}}) = \mathbf{0}$ (we have shown this!)

- If we plot r_i against $\hat{y}_i$, the residuals should lie within a horizontal band around zero and should not exhibit any special pattern.

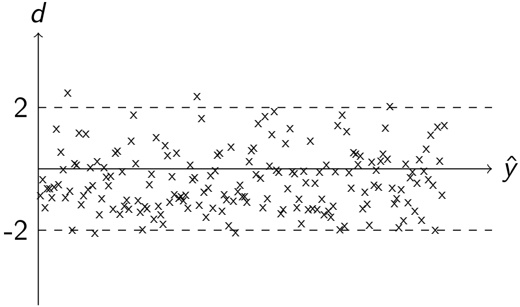


Studentized residuals versus $\hat{y}$

Sometimes, instead of plotting raw residuals against $\hat{y}_i$, we plot studentized residuals d_i against $\hat{y}_i$. If the model is adequate,

- the studentized residuals should lie within a horizontal band around zero and should not exhibit any special pattern.
- approximately 95% of d_i should be within (-2,2), and almost all of them should be within (-3,3).

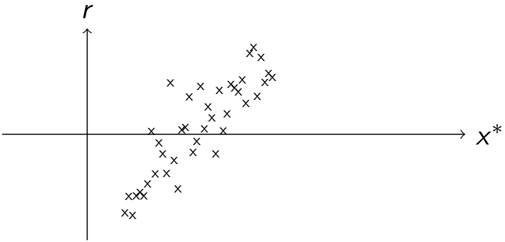
Studentized residuals versus $\hat{y}$



Points fall into a horizontal band between (-2, 2), and there is NO special trends.

Added-Variable Plots

Sometimes, we might suspect that an important predictor has not been included in the model. Consider the plot of residuals versus a new explanatory variable x^* (that currently is not in the model),



it suggests that the addition of x^* may improve the model.

Added-Variable Plots

When deciding whether a new predictor should be included, an added variable plot turns out to be a more powerful graph. To produce the added-variable plot for a new explanatory variable x^*, we

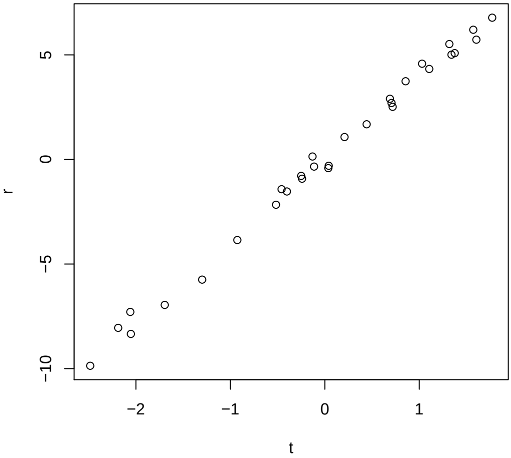
- regress y on all the current predictors $x_1, \ldots, x_p$ and denote the residual vector $\mathbf{r}$
- regress x^* on all of $x_1, \ldots, x_p$ and denote the residual vector $\mathbf{t}$.
- Plot residual vector $\mathbf{r}$ versus $\mathbf{t}$. If there is a pattern in this plot, it indicates that the predictor x^* may be a useful addition to the model.

Added-Variable Plots

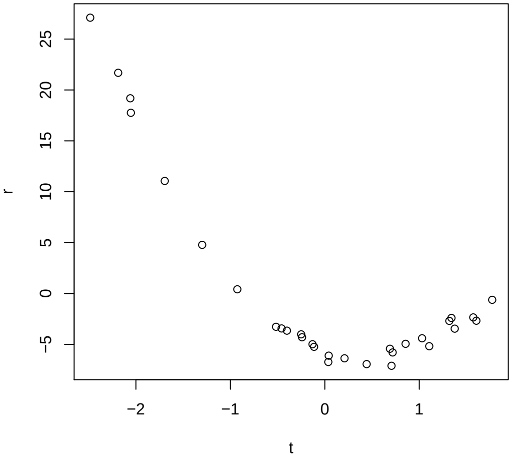
In `R`, the `car` package provides a function `avPlots(model, variable, ...)` to produce added-variable plots, where

- `model` is the model object produced by `lm()`.
- `variable` is the name of the new explanatory variable that is not included in the `lm()` fitting. The observations of this new variable shall be included in your data set though.

Example of Added-Variable Plots



Example of Added-Variable Plots

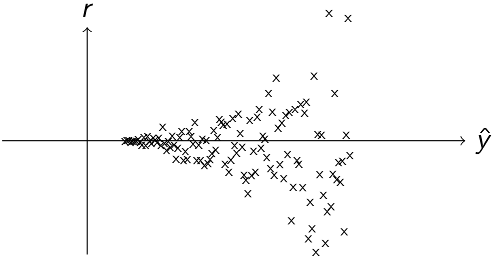


Residual plots for checking constant variance $V(\epsilon_i) = \sigma^2$

Once we are satisfied about the assumption that $E(\epsilon_i) = 0$, which is equivalently to saying that we are modelling $E(y)$ adequately as a linear function of the predictors, we can move on to the assumptions that $V(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$. This assumption states that the error variance, or equivalently $V(y_i)$ is a constant.

Residual plots for checking constant variance $V(\epsilon_i) = \sigma^2$

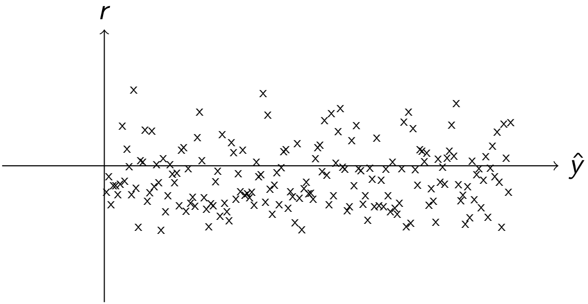
To detect non-constant variance (heteroscedasticity), a standard diagnostic is to plot the residuals against the fitted values: r_i versus $\hat{y}_i$. We examine this plot to see if the residuals appear to be have constant variability with respect to the fitted values. A pattern like



suggests that the constant variance assumption is violated.

Residual plots for checking constant variance $V(\epsilon_i) = \sigma^2$

A random scatter (below), on the other hand, supports the assumption.



Residual plots for checking Normality of ϵ_i's

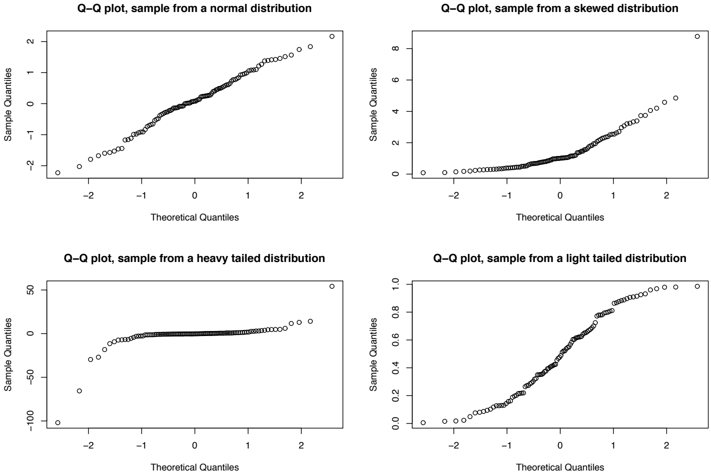
- We probably do not want to worry about the normality assumption until the other, more serious assumptions have been checked and fixed.
- If all assumptions are valid, including the normality assumptions, and we have sufficient degrees of freedom for the residuals, then the residuals should look approximately like a sample from a normal distribution.

Residual Plots for Checking Normality of ϵ_i's

We use Q-Q (quantile-quantile) plot to check normality. The *ordered* residuals (ordered r_i's) are plotted against the expected standard normal order statistics ($z_i' s$). If the normality assumption is correct, a Q-Q plot should show an approximately straight line.

Q-Q Plot

Figure 5.4.1 Examples of Q–Q Plots



Q–Q plot, sample from a normal distribution

Q–Q plot, sample from a skewed distribution

Q–Q plot, sample from a heavy tailed distribution

Q–Q plot, sample from a light tailed distribution

Q-Q Plot

In R, the function `qqnorm( )` can be used to produce the Q-Q plot of residuals.

```
qqnorm(r)
```

Non-constant Variance

- Transformations of the response are used to deal with non-constant variance in the errors.
- Consider the regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where $\mu_i = E(y_i) = \beta_0 + \beta x_{i1} + \cdots + \beta_p x_{ip}$.
Suppose that y_i has non-constant variance:

$$V(y_i) = \mu_i^\alpha \sigma^2$$

- Now we want to find a transformation $g(y_i)$ such that $g(y_i)$ has a constant variance.

Variance-Stabilizing Transformations

We approximate $g(y_i)$ by a first-order Taylor series expansion around μ_i:

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)[\frac{dg(y_i)}{dy_i}]_{y_i=\mu_i}.$$

Denote $[\frac{dg(y_i)}{dy_i}]_{y_i=\mu_i}$ as $g'(\mu_i)$. Then, the variance of $g(y_i)$ can be approximated as

$$V(g(y_i)) \approx g'(\mu_i)^2 V(y_i - \mu_i) = g'(\mu_i)^2 \mu_i^\alpha \sigma^2$$

To stabilize the variance, we need to choose transformation g(.) such that

$$[g'(\mu_i)]^2 = \frac{1}{\mu_i^\alpha}$$

Box-Cox Transformation

Box-Cox Transformation:

$$g(y_i) = \begin{cases} \frac{y_i^{1-\alpha/2}}{1-\alpha/2} & \alpha \neq 2 \\ log(y_i) & \alpha = 2 \end{cases}$$

Consequently, we have:

$$g'(y_i) = \begin{cases} y_i^{-\alpha/2} & \alpha \neq 2 \\ \frac{1}{y_i} & \alpha = 2 \end{cases}$$

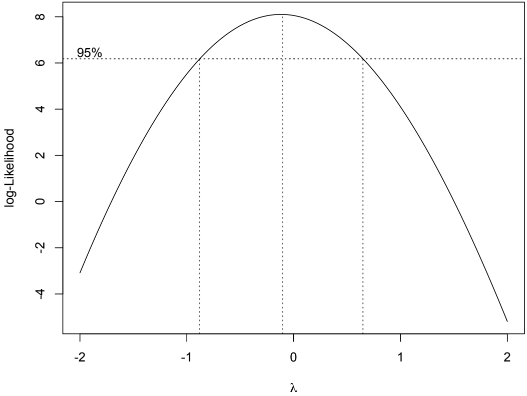
We have $V(g(y_i)) = \sigma^2$.

Box-Cox Transformation

- If the variance increases with the mean, then $\alpha > 0$; y should be either raised to a power $1 - \alpha/2$, such as square root, reciprocal or logged.
- If the variance decreases with the mean (which is not very common), then $\alpha < 0$; we want a function where $g(y)$ increases faster with y, like y^2.
- "Try and Error".

Box-Cox Transformation

- In R, the function boxcox(y $\sim$ x) in *library(MASS)* can be used to find the best transformation.



Based on the plot, log-transformation is appropriate.

log Transformation

Suppose we fit the model

$$\log y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

- Estimation:

  $$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{Y}.$$

  where $\mathbf{Y} = (\log y_1, \log y_2, \ldots, \log y_n)^T$.

- Interpretation of the model:

  $$y = e^{\beta_0} e^{\beta_1 x_1} \times \cdots \times e^{\beta_p x_p} \times e^{\epsilon}$$

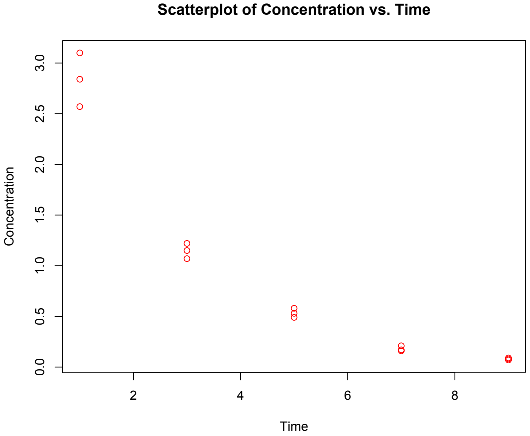
  Predictors have multiplicative effects instead of additive effects on y. Such models might be useful for exponential growth, e.g., of bacteria, of internet use, etc.

log Transformation

Interpretation of β_j:
$100(e^{\beta_j} - 1)\%$ is interpreted as the percentage change in the average value of response per unit increase in x_j, when other predictors are held constant.

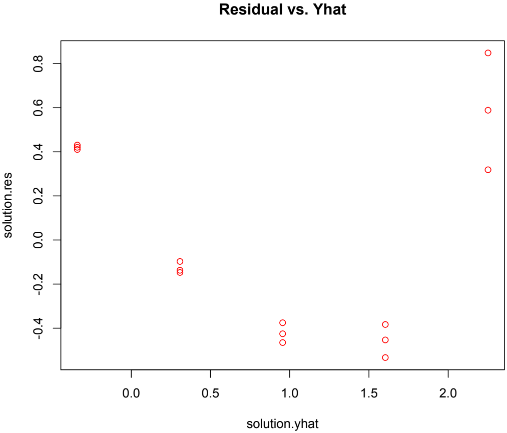
Example

Consider the relationship between $y =$ the concentration of a chemical solution, $x =$ time since preparation of the solution. A graph of this data is shown below:



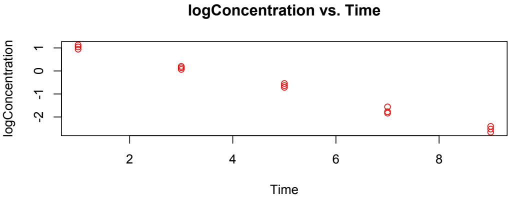
Example

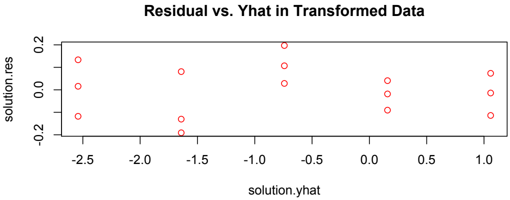
Fit y as a linear function of x and plot residuals vs. $\hat{y}$



Example

The residual plot shows non-constant variance. We need to transform y. Try log transformation and refit the data.





Example

- The fitted regression line is:
  Estimated log(Concentration)$=1.508-0.45\times$ Time
- We find the 95% confidence interval for $E(\log y)$ at $x=2$ is: (0.515, 0.702).
  Therefore, an approximate 95% confidence interval for $E(y)$ at $x=2$ is:
  $$(e^{0.515}, e^{0.702})=(1.674, 2.018).$$
- Note: When transforming y and computing CI or PI at a given x, reverse-transforming the intervals does not give exactly correct CI or PI for $E(Y)$.

Deal with Non-Linearity

Fit $y = \beta_0 + \beta_1 x + \epsilon$ and then plot r_i vs x_i, for $i = 1, \ldots, n$

- Non-linear but constant variance ⇒ include higher order terms, such as x^2, x^3 or transform x.
- Non-linear and non-constant variance ⇒
  - Transform y first, since usually by correcting one direction of departure from the assumptions the other departure may also be modified.
  - If transforming y only stablizes the variance but not totally fix the problem of non-linearity, then we can include higher order terms x^2, x^3 or transform x.

Deal with Non-Linearity

Include higher order terms: Polynomial Regression

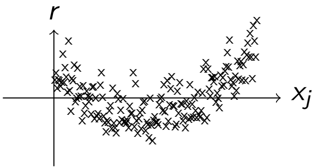
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

- Rule 1: If x^2 is in the model, x should be in the model. In general, if a higher order term is in, all lower order terms should also be in the model.
- Rule 2: We include a higher order term only if the new model is significantly better.

Deal with Non-Linearity

What about multiple regression?

- Plot r v.s. a particular x_j to check the need of x_j^2 and x_j^3, etc.



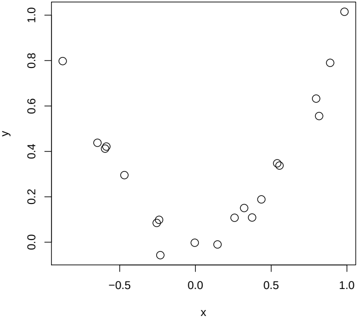
- need back-and-forth trials to reach a better model.

Polynomial Regression

For a simple linear regression model, we assume:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

which implies, $E(y_i) = \beta_0 + \beta_1 x_i$. However, the following graph shows the general trend between x and y is not linear.



Polynomial Regression

Instead, we may assume a polynomial model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

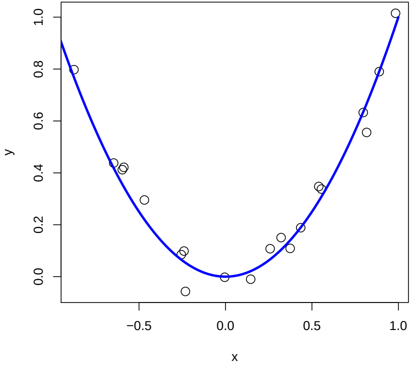
where

$$\mathbf{X} = \left(\begin{array}{ccc} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & .. & .. \\ 1 & x_n & x_n^2 \end{array} \right), \boldsymbol{\beta} = \left(\begin{array}{c} \beta_0 \\ \beta_1 \\ \beta_2 \end{array} \right)$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}.$$

Polynomial Regression

The blue line represents the fitted regression line using a polynomial model of order 2 (quadratic).



Adequacy of a Regression Model

In what ways could a model fail to be adequate?

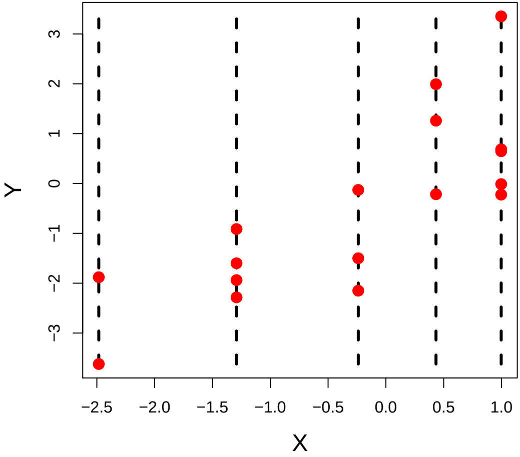
- Functional form could be inadequate, i.e. choice of predictors, linear/non-linear form, interactions not included
- The error term might be inadequate. i.e. non-constant variance
- Unusual observations my have large influence on fit

Testing for lack of fit

- Previously, we use residuals to estimate σ^2. If the model is incorrect, this estimate is not appropriate.
- If we have repeated observations from the same predictors, we can do a better job of checking model adequacy.
- We can estimate σ^2 without assuming any model.
- Suppose we have data of the form (one predictor):

$$\begin{array}{llllll} x_1: & y_{11} & y_{12} & \cdots & y_{1n_1} \\ x_2: & y_{21} & y_{22} & \cdots & y_{2n_2} \\ & & \vdots & & \\ x_k: & y_{k1} & y_{k2} & \cdots & y_{k2n_k} \end{array}$$

Example of Repeated Observations From the Same Predictors



Testing for lack of fit

- Think of model as

$$y_{ij} = \mu_i + \epsilon_{ij}$$

- Write as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_k)^T$ and $\mathbf{Z}$ is matrix of zero and one indicating group membership

- LSE is

$$\widehat{\boldsymbol{\beta}} = (\bar{y}_1, \ldots, \bar{y}_k)^T$$

  where $\bar{y}_j$ is the sample mean for the jth group

- SSE is

$$SSE(\widehat{\beta}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Testing for lack of fit

- This SSE is sometimes called the pure error sum of squares (PESS) since it makes no functional assumptions about the relationship between μ_i and x_i.
- Now consider a regression model:

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$$

and compute $SSE(\hat{\boldsymbol{\beta}}_A) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i)$ for the assumed model

- We can show

$$SSE(\widehat{\boldsymbol{\beta}}_A) - SSE(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{k} n_i (\bar{y}_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

Testing for lack of fit

- $SSE(\widehat{\beta}_A) - SSE(\widehat{\beta})$ measures the lack of fit of the linear model and is called lack-of-fit sum of squares (LFSS).
- We can use this to form a lack of fit test based on the F-distribution
- The additional sum of squares $SSE(\widehat{\beta}_A) - SSE(\widehat{\beta})$ has $(n-2)-(n-k) = k-2$ degrees of freedom
- Use the F statistic

$$F = \frac{(SSE(\widehat{\beta}_A) - SSE(\widehat{\beta}))/(k-2)}{SSE(\widehat{\beta})/(n-k)}$$

- If $F > F_\alpha(k-2, n-k)$, there is lack of fit in the linear model.

Example

An experiment was conducted to study the relationship between the yield from a chemical reaction (y) an the reaction temperature (x).

Temperature, x	Yield (gs), y	$\bar{y}_i$
60	51	51.0
70	82, 78	80.0
80	91, 96	93.5
90	98, 89, 99	95.3
100	82, 83	82.5
110	54, 52	53.0

Example

The Pure Error Sum of Squares is given by:

$$\begin{aligned} SSE(\hat{\beta}) =& (51-51)^2 + (82-80)^2 + (78-80)^2 \\ & + \cdots + (54-53)^2 + (52-53)^2 = 83.67 \end{aligned}$$

Example

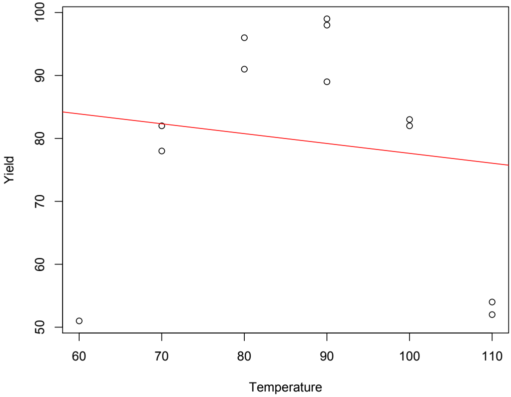
Fit the following model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- $SSE(\hat{\beta}_A) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 3393.6$
- $SSE(\hat{\beta}_A) - SSE(\hat{\beta}) = 3393.6 - 83.67 = 3309.93$
- $F = \frac{3309.93/4}{83.67/6} = 59.34$
- Since $59.34 > F_{0.05}(4, 6) = 4.53$, we reject H_0 and conclude there is lack of fit in the linear model.

Example

The red curve is the fitted line from the linear model.



Lack of fit test for more than one predictors

- Data

$$\begin{array}{lllllll}
\text{At} & x_{11}, x_{12}, \ldots, x_{1p}: & y_{11} & y_{12} & \cdots & y_{1n_1} \\
\text{At} & x_{21}, x_{22}, \ldots, x_{2p}: & y_{21} & y_{22} & \cdots & y_{2n_2} \\
 & & & \vdots & & \\
\text{At} & x_{k1}, x_{k2}, \ldots, x_{kp}: & y_{k1} & y_{k2} & \cdots & y_{k2n_k}
\end{array}$$

- The lack of fit statistic is:

$$F = \frac{(SSE(\widehat{\beta}_A) - SSE(\widehat{\beta}))/(k-p-1)}{SSE(\widehat{\beta})/(n-k)}$$

Need a condition: $k > p+1$

- $SSE(\widehat{\beta}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$
- $SSE(\widehat{\beta}_A) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$

Example

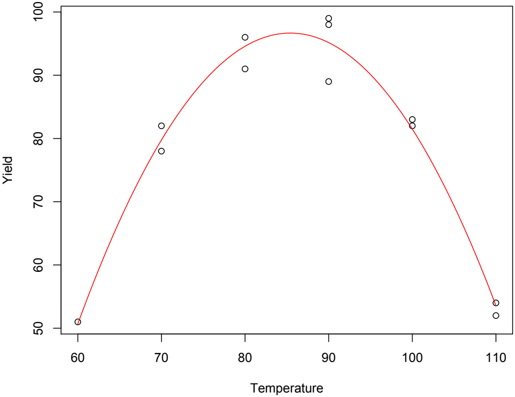
Fit the following model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- $SSE(\hat{\beta}_A) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2 = 89.01$
- $SSE(\hat{\beta}_A) - SSE(\hat{\beta}) = 89.01 - 83.67 = 5.34$
- $F = \frac{5.34/3}{83.67/6} = 0.13$
- Since $0.13 < F_{0.05}(3, 6) = 4.76$, we don't have enough evidence to reject H_0 and conclude the quadratic model is adequate.

Example

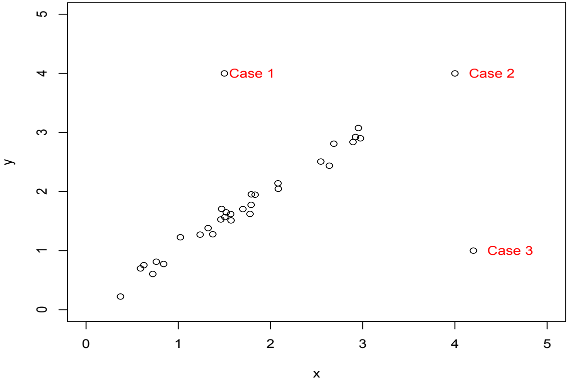
The red curve is the fitted curve from the quadratic model.



Extreme observations in the data set

- Frequently in regression analysis, the data set at hand may contain some observations that are extreme;
- An outlier is a particular case with unusual value in y or in x;
- These outlying observations often have significant effects on the fitted regression equation. It is therefore important for us to determine whether they should be retained or eliminated.

Outliers in the data set



A case may be outlying with respect to its y value (e.g. Case 1), with respect to its x value(s) (e.g. Case 2), or both (e.g. Case 3).

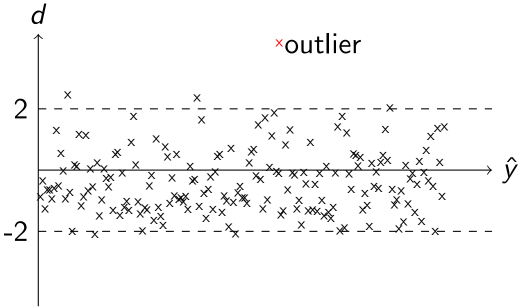
Detecting outliers in response

- Simple diagnostic tool: studentized residuals

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

where h_{ii} is the *ith* diagonal element of the hat matrix $\mathbf{H}$.

- Large values of d_i (e.g. $|d_i| > 2.5$) indicate outliers in y.



Detecting outliers in predictors

Recall $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T} = (h_{ij})_{n \times n}$ and $\hat{\mathbf{Y}} = \mathbf{HY}$. Therefore, the fitted value

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

- h_{ii} indicates how heavily y_i contributes to $\hat{y}_i$
- Fact: $h_{ii}(1 - h_{ii}) = \sum_{j \neq i} h_{ij}^2$, $0 \leq h_{ii} \leq 1$
- If h_{ii} is large (compared to other h_{ij}'s) , $h_{ii} y_i$ dominates $\hat{y}_i$
- If $h_{ii} \to 1$, $\hat{y}_i \to y_i$
- This implies that when h_{ii} is large, the fitted line will be forced to go very close to the ith observation. We say that the case i has high leverage on the fitted line.

Detecting outliers in predictors

- The **leverage** of the *ith* observed predictor(s) is just h_{ii}.
- It reflects the distance between $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and the others.
- The leverage h_{ii} is small for cases with $(x_{i1}, \ldots, x_{ip})$ near the centroid $(\bar{x}_1, \ldots, \bar{x}_p)$, that is determined by all cases. The leverage h_{ii} will be large if $(x_{i1}, \ldots, x_{ip})$ is far from the centroid.
- Recall $0 \le h_{ii} \le 1$; $i = 1, ..., n$ and $\sum_{i=1}^{n} h_{ii} = p + 1$ (Why?).

  $h_{ii} > 2(p+1)/n \Leftrightarrow$ significant outlier x_i

Leverage for Simple Linear Regression

For a simple linear regression,

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

The leverage is the smallest when $x_i = \bar{x}$ and it is large if x_i is far from $\bar{x}$.

Outliers vs. Influential Points

- The real issue is not whether an observation is an outlier or not
- It is whether an observation has a major influence on the fitted regression equation (Influential points)

Detecting Influential Points

- The *i*th observation is influential if its exclusion may cause major changes in the fitted regression coefficients. The change can be denoted as:

  $$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}$$

  where $\hat{\boldsymbol{\beta}}_{(-i)}$ is the estimated regression coefficients when the *i*th observation is not used to estimate the regression equation.

- To identify influential points, we use Cook's Distance. For *i*th observation,

$$\begin{aligned} D_i &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^T (\hat{\sigma}^2 (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}})^{-\mathbf{1}} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-\mathbf{i})})}{p+1} \\ &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^T \mathbf{X}^\mathbf{T}\mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-\mathbf{i})})}{\hat{\sigma}^2 \cdot (p+1)} \end{aligned}$$

Cook's Distance

We can prove

$$D_i = \frac{h_{ii} d_i^2}{(1 - h_{ii})(p + 1)}$$

where $d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ is the studentized residual.

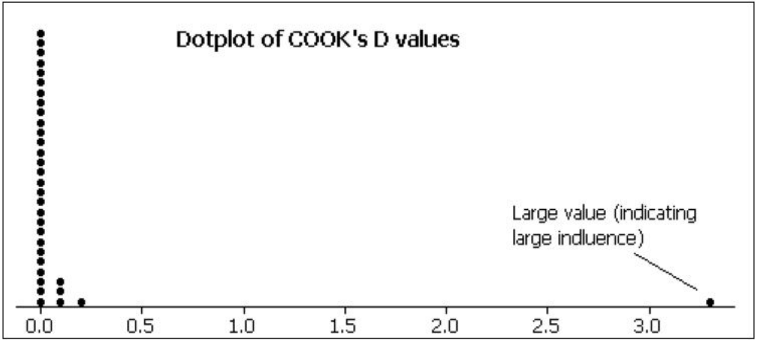
- If h_{ii} is large, but d_i is small, the influence will be small.
- If d_i is large, but h_{ii} is small, the influence will also be small.
- Cook's D is an overall measure.

Cook's Distance: check influence

- A large value indicates that the observation has a large influence on the results.
- Cook suggested that a Cook's Distance is significantly large when it is greater than $F_{0.5}(p+1, n-p-1)$
- Many data analysts just look to see if a Cook's D_i value is > 1 or obviously greater than others.

Cook's Distance

Below is a dotplot of the Cook's D values for a real data. The one large value is for the point that is influential. The interpretation is that the inclusion (or deletion) of this point will have a large influence on the overall result.



Summary

(1) Outliers: visual detection through residual plots of standardized residual versus $\hat{y}$.

(2) Influential cases:

- Judged by leverage h_{ii}: case i is potentially influential if

$$h_{ii} > 2(p+1)/n$$

  Note: an observation with high leverage value is NOT necessarily influential.

- Measured by Cook's D:

$$D_i = \frac{h_{ii} d_i^2}{(1-h_{ii})(p+1)}$$

If $D_i > F_{0.5}(p+1, n-p-1)$, case i is an influential point.

Conclusion: remove or not remove?

For outliers and influential points,

- Correct/remove for obvious errors due to data entry mistakes.
- A careful decision on whether keep or remove them because the target population may be changed due to inclusion/exclusion of certain cases.