Regression Analysis

Chapter 3. Statistical Inference

Prof. dr. Thomas Neyens

Course notes: Prof. dr. Mia Hubert & Prof. dr. Stefan Van Aelst

The use of a regression model includes:

- Testing whether the regression model is a good model for the data at hand
- Testing if specific predictors are considerably associated with the outcome
- Prediction
- Can we say something about the mean response?

Assumptions for this chapter

$$\boxed{\epsilon \sim N_n(\mathbf{0}, \sigma^2 I_n).} \tag{1}$$

Under this condition, the general linear model satisfies:

$$\begin{aligned} \mathbf{y} \sim & \quad N_n(X\beta, \sigma^2 I_n) && (2) \\ \rightarrow y_i \sim & \quad N(\mathbf{x}_i^t \beta, \sigma^2) \\ \hat{\beta}_{LS} \sim & \quad N_p(\beta, \sigma^2 (X^t X)^{-1}). && (3) \end{aligned}$$

The assumption on the errors is often violated and should be investigated carefully (residual plots).

The overall F-test

Test whether there is a regression relation between the response variable Y and the set of X-variables $X_1, \ldots, X_{p-1}$

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$
$$H_1 : \text{not all } \beta_j \text{ equal zero}$$

It can be shown that:

$$\boxed{F = \frac{\text{MSR}}{\text{MSE}} \sim_{H_0} \quad F_{p-1,n-p}} \tag{4}$$

Or, equivalently:

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}.$$

The partial F-test

Test whether a group of parameters is significant.

$$H_0 : \beta_{p-q} = \beta_{p-q+1} = \ldots = \beta_{p-1} = 0$$
$$H_1 : \text{not all } \beta_j \text{ equal zero } (j = p-q, \ldots, p-1)$$

Under H_0 we obtain the reduced model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-q-1} x_{i,p-q-1} + \epsilon_i.$$

Let SSE_{p-q} denote the error sum of squares under this reduced model and SSE_p under the full model. It can be shown that

$$\boxed{F = \frac{(\text{SSE}_{p-q} - \text{SSE}_p)/q}{\text{SSE}_p/(n-p)} \sim_{H_0} F_{q,n-p}} \tag{5}$$

The partial F-test

This test statistic can as well be described using the extra sum of squares. Since

$$\mathrm{SSE}_{p-q} - \mathrm{SSE}_p = \mathrm{SSR}(X_{p-q}, \ldots, X_{p-1} | X_1, \ldots, X_{p-q-1}),$$

(5) becomes:

$$\boxed{F = \frac{\mathrm{MSR}(X_{p-q}, \ldots, X_{p-1} | X_1, \ldots, X_{p-q-1})}{\mathrm{MSE}(X_1, \ldots, X_{p-1})}} \qquad (6)$$

Moreover,

$$\mathrm{MSR}(X_{p-q}, \ldots, X_{p-1} | X_1, \ldots, X_{p-q-1}) = \frac{1}{q}\big(\mathrm{SSR}(X_{p-q} | X_1, \ldots, X_{p-q-1}) + \mathrm{SSR}(X_{p-q+1} | X_1, \ldots, X_{p-q}) + \ldots + \mathrm{SSR}(X_{p-1} | X_1, \ldots, X_{p-2})\big)$$

The partial F-test

Example: Body Fat Data.

- amount of body fat (Y)
- triceps skinfold thickness (X_1),
- thigh circumference (X_2)
- midarm circumference (X_3).
- measurements on 20 healthy women between

Assume we want to test:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_1 : \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal zero}$$

The partial F-test

Example: Body Fat Data.

TABLE 7.4 **ANOVA Table with Decomposition of *SSR*—Body Fat Example with Three Predictor Variables.**

Source of Variation	*SS*	*df*	*MS*
Regression	396.98	3	132.33
X_1	352.27	1	352.27
$X_2 \mid X_1$	33.17	1	33.17
$X_3 \mid X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

The partial F-test

Example: Body Fat Data.

$$F = \frac{(33.17 + 11.54)/2}{98.41/16} = 3.63.$$

As $F_{2,16,0.05} = 3.63$, the p-value of our test is 5% and we are at the boundary of the decision rule.

Inference for individual parameters

From (3), we obtain

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^t X)^{-1}_{jj})$$

and its estimated standard error

$$\mathrm{s}(\hat{\beta}_j) = s\sqrt{(X^t X)^{-1}_{jj}}.$$

Moreover it can be shown that

$$(n-p)\frac{s^2}{\sigma^2} \sim \chi^2_{n-p}$$

and that $\hat{\beta}_j$ and s^2 are independent. Consequently,

$$\boxed{\frac{\hat{\beta}_j - \beta j}{\mathrm{s}(\hat{\beta}_j)} \sim t_{n-p}}$$

Inference for individual parameters

Under the null hypothesis

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

it then holds that

$$\boxed{t = \frac{\hat{\beta}_j}{\mathrm{s}(\hat{\beta}_j)} \sim_{H_0} t_{n-p}} \tag{7}$$

Calculate t value, find p value (or compare t with critical value from t distribution) and conclude.

Inference for individual parameters

Equivalently, we can construct a $(1-\alpha)100\%$ confidence interval for β_j:

$$\mathsf{CI}(\beta_j, \alpha) = [\hat{\beta}_j - t_{n-p,\frac{\alpha}{2}}\mathsf{s}(\hat{\beta}_j), \hat{\beta}_j + t_{n-p,\frac{\alpha}{2}}\mathsf{s}(\hat{\beta}_j)]$$

and reject H_0 if 0 does not belong to $\mathsf{CI}(\beta_j, \alpha)$. Note that the quantile $t_{n-p,\alpha/2}$ satisfies

$$P(T > t_{n-p,\frac{\alpha}{2}}) = \frac{\alpha}{2} \text{ with } T \sim t_{n-p}.$$

Remember that α is the probability of a type I error, i.e.

$$\boxed{\alpha = P(H_0 \text{ is rejected } | H_0 \text{ is correct}).}$$

Inference for individual parameters

Example: Fuel data

```
> attach(fuel.frame)
> Fuelfit <- lm(Fuel~Weight+Disp.)
> Fuelsum <- summary(Fuelfit)
> Fuelsum

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4789731  0.3417877   1.401    0.167
Weight      0.0012414  0.0001720   7.220 1.37e-09 ***
Disp.       0.0008544  0.0015743   0.543    0.589

Residual standard error: 0.3901 on 57 degrees of freedom
Multiple R-squared: 0.7438,     Adjusted R-squared: 0.7348
F-statistic: 82.75 on 2 and 57 DF,  p-value: < 2.2e-16
```

Partial F test vs. t test

- We now have two tests to assess the effect of one or a group of variables.
- When testing the effect of a group of regressors, one could:
  1. use a partial F-test
  2. do several t-tests, e.g., by investigating if 0 lies in the $(1-\alpha)100\%$ CIs of each regressor
- The latter approach should be done with care, due to an unwanted inflation of Type I-error.

Type I-error inflation

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal zero}$$

$$\begin{aligned}
&P(H_0 \text{ is accepted } | H_0 \text{ is correct}) \\
&\quad = P(0 \in \mathrm{CI}(\beta_2, \alpha) \text{ and } 0 \in \mathrm{CI}(\beta_3, \alpha)) \\
&\quad = 1 - P(0 \notin \mathrm{CI}(\beta_2, \alpha) \text{ or } 0 \notin \mathrm{CI}(\beta_3, \alpha)) \\
&\quad = 1 - P(0 \notin \mathrm{CI}(\beta_2, \alpha)) - P(0 \notin \mathrm{CI}(\beta_3, \alpha)) \\
&\qquad + P(0 \notin \mathrm{CI}(\beta_2, \alpha) \text{ and } 0 \notin \mathrm{CI}(\beta_3, \alpha)) \\
&\quad \geqslant 1 - P(0 \notin \mathrm{CI}(\beta_2, \alpha)) - P(0 \notin \mathrm{CI}(\beta_3, \alpha)) \\
&\quad = 1 - \alpha - \alpha = 1 - 2\alpha.
\end{aligned}$$

Hence,

$$P(H_0 \text{ is rejected } | H_0 \text{ is correct}) \leqslant 1 - (1 - 2\alpha) = 2\alpha.$$

The Bonferroni correction for Type I-error inflation

If we want to be sure that

$$P(H_0 \text{ is rejected } | H_0 \text{ is correct}) \leqslant \alpha,$$

we can apply the Bonferroni correction, where you construct simultaneous confidence intervals for testing $g \leqslant p$ parameters with a confidence of at least $1 - \alpha$ that are given by

$$[\hat{\beta}_j - t_{n-p,\frac{\alpha}{2g}}\ \mathrm{s}(\hat{\beta}_j), \hat{\beta}_j + t_{n-p,\frac{\alpha}{2g}}\ \mathrm{s}(\hat{\beta}_j)].$$

The Bonferroni correction for Type I-error inflation

Disadvantages:

- Because the simultaneous confidence intervals are wider than the individual confidence intervals, they yield a larger type II error, especially when there are many parameters to be tested.

$$\boxed{P(\text{type II error}) = P(H_0 \text{ is accepted} | H_1 \text{ is correct})}$$

- The separate tests do not take into account the correlation between the parameter estimates

Test for all parameters

A $(1-\alpha)100\%$ joint confidence region for the unknown $\boldsymbol{\beta} \in \mathbb{R}^p$ is given by an ellipsoid with center $\hat{\boldsymbol{\beta}}_{LS}$:

$$E_\alpha = \{\mathbf{x} \in \mathbb{R}^p | \frac{(\mathbf{x} - \hat{\boldsymbol{\beta}}_{LS})^t (X^t X)(\mathbf{x} - \hat{\boldsymbol{\beta}}_{LS})}{ps^2} \leqslant F_{p,n-p,\alpha}\}.$$

This ellipsoid can be used to test hypotheses of the form:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$
$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$

for some fixed vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. If $\boldsymbol{\beta}_0$ does not belong to E_α, we reject the H_0 hypothesis at the α significance level.

Test for all parameters

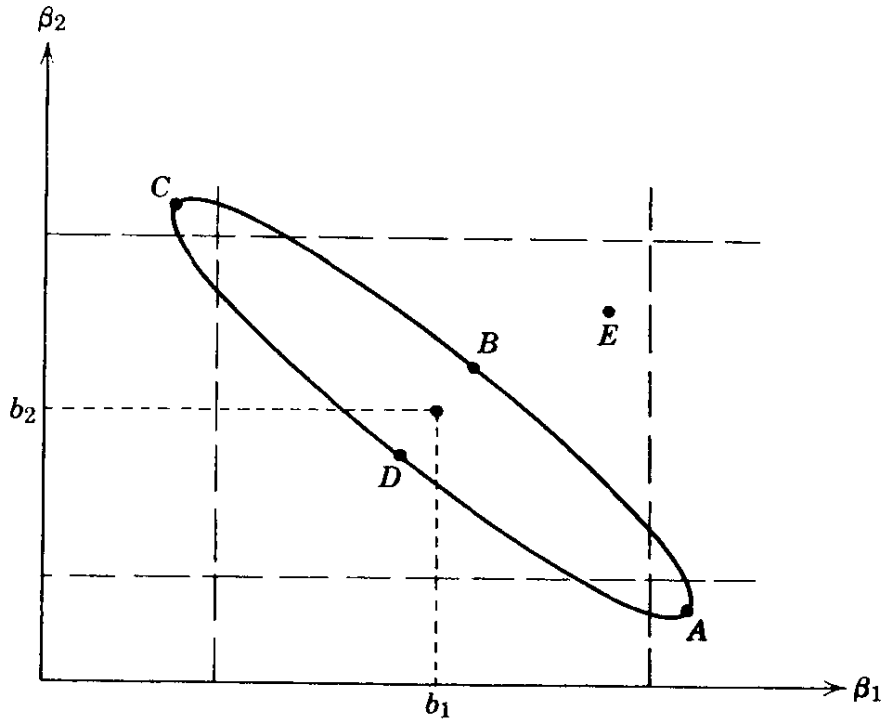


Figure 5.1. Joint and individual confidence statements. The point (b_1, b_2) defined by the least squares estimates is at the center of both ellipse and rectangle.

If the correlation between the parameter estimates is large, the ellipsoid will be more elongated

Test for all parameters

- disadvantage: if the H_0 hypothesis is rejected, we can not directly deduce statements about the individual parameters.

Therefore, individual or simultaneous confidence intervals for β_{0j} are still useful.

A general linear hypothesis

All previous hypotheses can be generalised as

$$H_0 : C\boldsymbol{\beta} = \mathbf{0} \qquad (8)$$
$$H_1 : C\boldsymbol{\beta} \neq \mathbf{0}$$

with C a $(q \times p)$ matrix with $\text{rank}(C) = q \leqslant p$.

- Example 1:

$$H_0 : \beta_1 = \beta_2, \beta_3 = 0$$

is equivalent to (8) with

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$$

A general linear hypothesis

- Example 2:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

is equivalent to (8) with $C = (\mathbf{0} \;\; I_{p-1})$.

Inference about the mean response

- At a fixed point $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0,p-1})^t$, the (unknown) *mean response* is denoted as $E[Y_0|\mathbf{x}_0] = \mathbf{x}_0^t\beta$.
- An unbiased estimator of the mean response is given by $\hat{y}_0 = \mathbf{x}_0^t\hat{\beta}$ with $\text{Var}(\hat{y}_0) = \mathbf{x}_0^t\Sigma\hat{\beta})\mathbf{x}_0$ (estimated by $s^2\mathbf{x}_0^t(X^tX)^{-1}\mathbf{x}_0$).

A $(1-\alpha)100\%$ confidence interval for the mean response $E[Y_0|\mathbf{x}_0]$ is then given by

$$\hat{y}_0 \pm t_{n-p,\frac{\alpha}{2}} s\sqrt{\mathbf{x}_0^t(X^tX)^{-1}\mathbf{x}_0}.$$

Inference about the mean response

- A new point $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0,p-1})^t$
- A confidence interval for the unknown response

$$y_0 = \mathbf{x}_0^t \boldsymbol{\beta} + \epsilon_0$$

is constructed as follows.

Consider the random variable $\hat{y}_0 - y_0 = \mathbf{x}_0^t \hat{\boldsymbol{\beta}} - \mathbf{x}_0^t \boldsymbol{\beta} - \epsilon_0$. It holds that

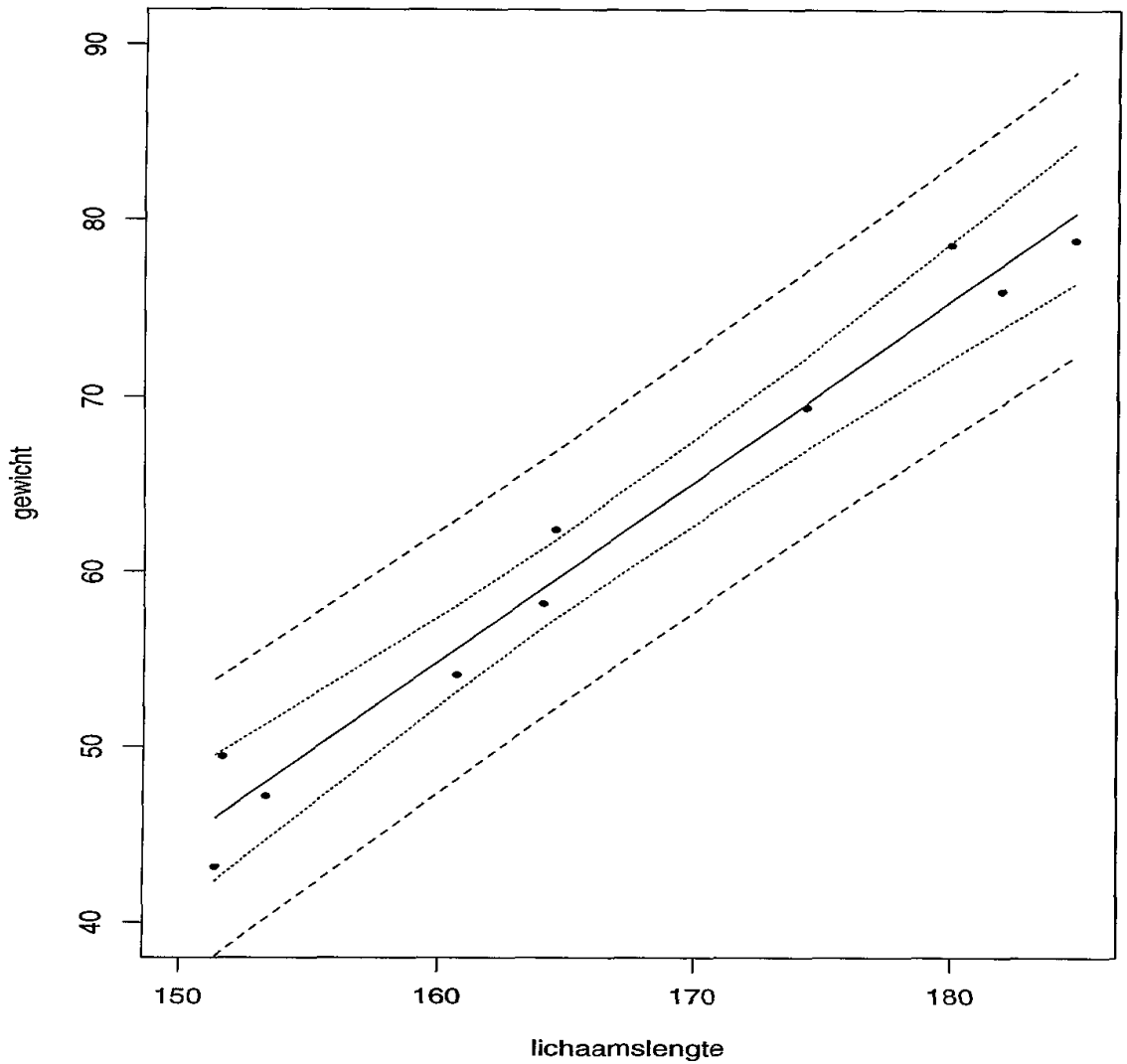
$$E[\hat{y}_0 - y_0] = \mathbf{x}_0^t E[\hat{\boldsymbol{\beta}}] - \mathbf{x}_0^t \boldsymbol{\beta} = 0$$
$$\mathrm{Var}[\hat{y}_0 - y_0] = \sigma^2 \mathbf{x}_0^t (X^t X)^{-1} \mathbf{x}_0 + \sigma^2$$

because $\hat{\boldsymbol{\beta}}$ and ϵ_0 are independent and $\epsilon_0 \sim N(0, \sigma^2)$.
A $(1-\alpha)100\%$ prediction interval for the unknown response y_0 is then given by

$$\hat{y}_0 \pm t_{n-p, \frac{\alpha}{2}} s \sqrt{\mathbf{x}_0^t (X^t X)^{-1} \mathbf{x}_0 + 1}.$$

This interval is larger than the confidence interval for the mean response because it also includes the uncertainty given by ϵ_0.

CI for mean response vs. PI



Residual plots

Helpful to check the validity of our model assumptions!

1. non-normality
2. time effects (correlation)
3. nonconstant variance (and transformations of Y)
4. curvature (and transformations of the regressors)
5. outlier detection ...

Residual plots

Normal quantile plot We usually assume condition (1) which says that the errors are normally distributed with zero mean.

- Least squares residuals always have zero mean, so we do not have to check for it!
- Normality can be verified using a *normal quantile plot*
- Note that residuals have different standard errors, hence different distributions
- Plot a normal quantile plot of the *standardized* residuals:

$$\boxed{e_i^{(s)} = \frac{e_i}{s\sqrt{1 - h_{ii}}}} \tag{9}$$

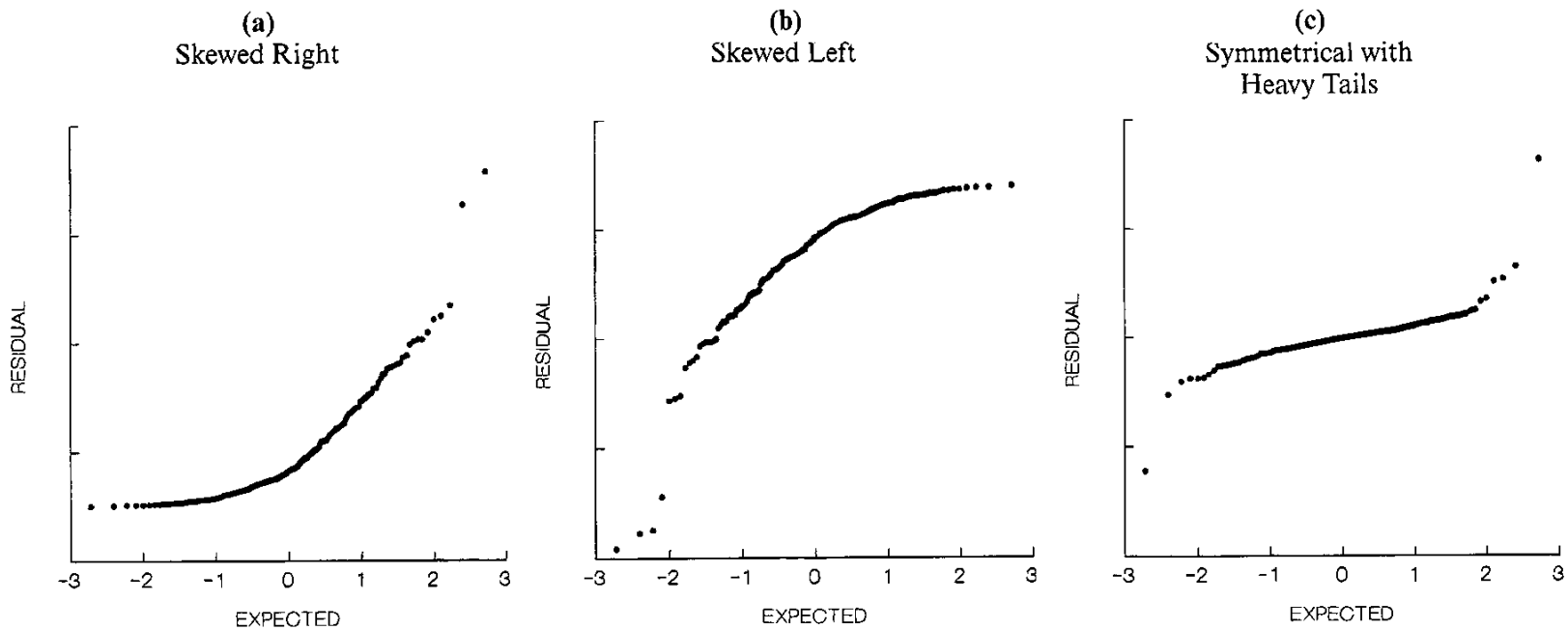
with h_{ii} the ith diagonal element of the hat matrix H.

- Formal tests possible through the Shapiro-Wilk statistic or the Kolmogorov-Smirnov test

Residual plots

Normal quantile plot

FIGURE 3.9 **Normal Probability Plots when Error Term Distribution Is Not Normal.**

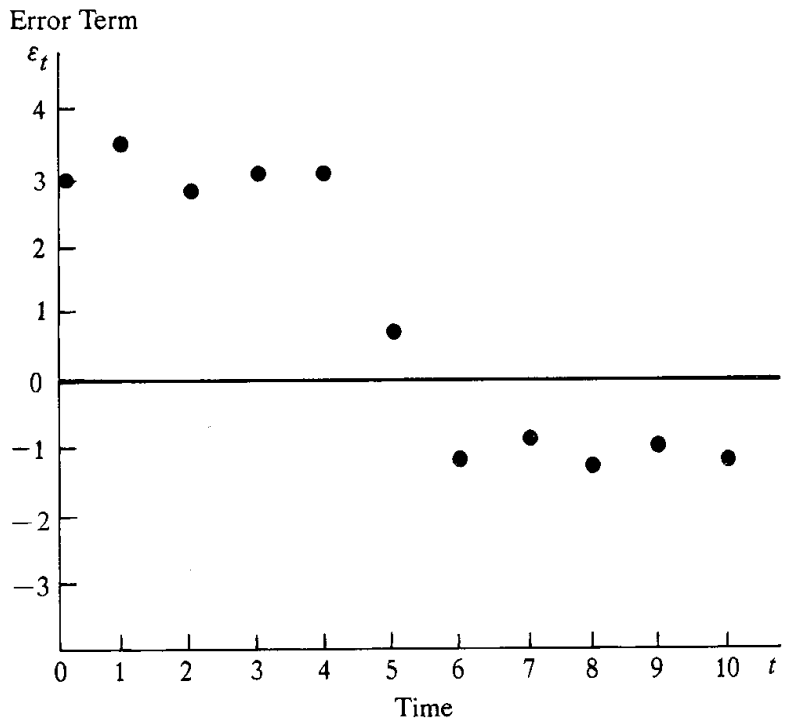


Residual plots

Plot of the residuals versus their index

- When we make a plot of the e_i versus i, we do not expect to see any pattern as the errors should be uncorrelated.
- This plot is useful if the index i has a physical interpretation, such as 'time'.

FIGURE 12.1 **Example of Positively Autocorrelated Error Terms.**



Residual plots

Plot of the residuals versus their index

- Check whether the variance is constant over time (known as homoscedasticity).

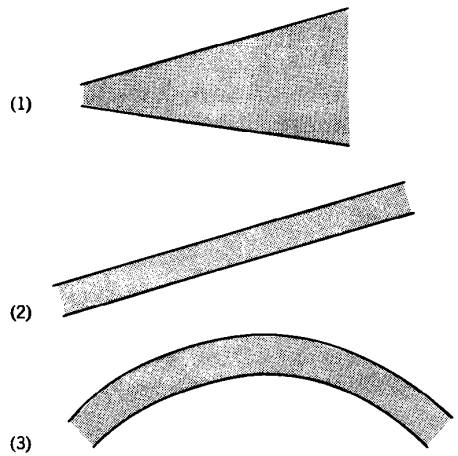


Figure 2.6. Examples of characteristics shown by unsatisfactory residuals behavior.

- Patterns (2) and (3) show that the linear model should be refined by adding first-order or second order terms.

Residual plots

Plot of the residuals versus fitted values

- If the linear model is correct, e_i are uncorrelated with the $\hat{y}_i$.
- A non-horizontal or curved band in a $(\hat{y}_i, e_i)$ plot shows that the linear model is not appropriate.
- heteroscedasticity in case of a funnel shape

Plot of the residuals versus independent variables

- Each independent variable X_j should be uncorrelated with the residuals e_i
- Can indicate that the regression fit is defective.

Residual plots

TABLE 2.5. Possible Remedies for Unsatisfactory Residuals Plots

Unsatisfactory Plot: See Figure 2.6	Plot of e_i Versus		
	Time Order	Fitted $\hat{Y}_i$	X_{ji} Values
Funnel indicating nonconstant variance	Use weighted[a] least squares	Use weighted[a] least squares or transform[b] the Y_i	Used weighted[a] least squares or transform[b] the Y_i
Ascending or descending band	Consider adding first-order term in time	Error in analysis or wrongful omission of β_0	Error in the calculations; first-order effect of X_j not removed
Curved band	Consider adding first- and second-order terms in time	Consider adding extra terms to the model or transform[b] the Y_i	Consider adding extra terms to the model or transform[b] the Y_i

Residual plots

Plot of the standardized residuals

- If they are a good approximation of the true errors, they should be approximately gaussian distributed.
- observations whose absolute standardized residual is larger than, say 2.5, can be pinpointed as outliers.

More on Outliers in Chapter 9!

Residual plots - Fuelfit example