2.7 Further Model Checking

2.7.1 Outliers and influential observations

An outlier, in the context of regression is an observation whose standardized residual is large (in absolute value) compared with the rest of the data. Recall the definition of the standardized residuals:

$$r_i = \frac{e_i}{S\sqrt{1-h_{ii}}}, \qquad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

An outlier will usually be apparent from any of the residual plots.

Minitab prints a warning about observations with a standardized residual greater than (in absolute value) 2. However, with a large number of observations there is more chance that a strange observation will occur in a data set. So, we need to be cautious when deciding about such values.

If we find an outlier we should check whether the observation was misrecorded or miscopied, if so correct it. If it seems correctly recorded we should rerun the analysis excluding the outlier. If the conclusions from the second analysis differ substantially from the first one we should report both.

As well as outliers in the y values, we sometimes have values of x which are different to the rest. To detect an observation with an unusual x value we use the *leverage*. This is defined as the h_{ii} value (as in the definition of the standardized residual).

Note that

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

so on average an observation will have a leverage of $2/n$. We shall regard an observation with $h_{ii} > 4/n$ as having a large leverage and with $h_{ii} > 6/n$ as a very large leverage.

An observation with a large leverage is not a wrong observation (although if the leverage is very large it is probably worth checking wether the x value has been recorded correctly). Rather, it is a *potentially influential observation*, i.e., one whose omission would cause a big change in the parameter estimates.

We can use a statistic called Cook's distance to measure the influence of an observation.

For a simple linear regression model consider omitting the ith observation (x_i, y_i) and refitting the model. Denote the new fitted values by $\hat{y}^{(i)}$. We define Cook's statistic for case i to be

$$D_i = \frac{1}{2s^2}\sum_{j=1}^{n}(\hat{y}_j^{(i)} - \hat{y}_j)^2.$$

It can be shown that

$$D_i = \frac{e_i^2}{2s^2}\frac{h_{ii}}{(1-h_{ii})^2}.$$

This shows that D_i depends on both the size of the residual e_i and the leverage h_{ii}. So a large value of D_i can occur due to large e_i or large h_{ii}.

A common technique to determine if D_i is unusually large is to determine whether D_i is bigger than the 50th percentile of an $\mathcal{F}_{p,n-p}$ distribution, where p is the number of parameters in the model. If so it has a major influence on the fitted value. Even if the largest D_i is not bigger than this value the corresponding observation could still be considered influential if it is a lot larger than the second largest.

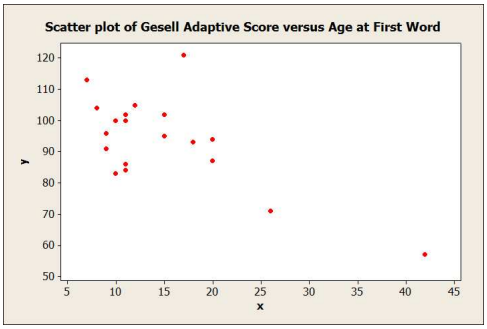
It is not recommended that influential observations be removed, but they indicate that some doubt should be expressed about the conclusions since without the influential observations the conclusions might be rather different.

MINITAB
To store values of h_{ii} and of D_i

Stat → Regression → Regression...
Storage
✓ Hi (leverages)
✓ Cook's distance

Example 2.6. Gesell's Score
The following data give Age at First Word (X) and Gesell Adaptive Score (Y) for 21 individuals from an investigation into cyanotic heart disease.

Obs.	x	y	Obs.	x	y
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100



The data represent the Gesell's adaptive scores (y) versus age of infants (x, in months) at first word. The scatter plot indicates two unusual observations: one is a large value of y compared to other values at a similar x and one is a large value of x, which is far from all the other x values. See the details of this example on the separate sheet given on-line on the course website.

2.7.2 Transformation of the response

Example 2.7. Plasma level of polyamine.
The plasma level of polyamine (Y) was observed in 25 children of age 0 (new-

$x = 0$	20.12	16.10	10.21	11.24	13.35
$x = 1$	8.75	9.45	13.22	12.11	10.38
$x = 2$	9.25	6.87	7.21	8.44	7.55
$x = 3$	6.45	4.35	5.58	7.12	8.10
$x = 4$	5.15	6.12	5.70	4.25	7.98

Table 2.1: Plasma levels data

born) to 4 years old (X). The results are given in Table 2.1. We are interested whether the level of polyamine decreases linearly while the age of children increases up to four years. See the details of this example on the separate sheet given on-line on the course website. □

If the model checking suggests that the variance is not constant, or that the data are not from a normal distribution (these often happen together) then it might be possible to obtain a better model by transforming the observations y_i. Commonly used transformations are

- $\ln y$; this is particularly good if $\mathrm{Var}(Y_i) \propto [E(Y_i)]^2$.
- $\sqrt{y}$; this is particularly good if $\mathrm{Var}(Y_i) \propto E(Y_i)$.
- $1/y$.

They are special cases of a large family of transformations, the Box-Cox transformation,

$$\begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{when } \lambda \neq 0; \\ \ln y, & \text{when } \lambda = 0. \end{cases}$$

MINITAB uses a simpler version of this transformation, that is y^λ when $\lambda \neq 0$ and also $\ln y$ when $\lambda = 0$. The Box-Cox transformation estimates lambda that minimizes the standard deviation of a standardized transformed variable. Trigonometric functions are also used in some cases, in particular the arc-sine or arc-tangent. In practice the log transformation is often the most useful and is generally the first transformation we try, but note all values of y_i need to be positive.

2.7.3 Lack of Fit Test

We have seen that the residuals for the plasma data are not likely to be a sample from a normal distribution with a constant variance. One of the reasons can be that the straight line is not a good choice of the model. This fact can be easily seen here, but we can also test lack of fit. The test function is also based on the model assumptions so we should not see clear evidence against the assumptions for the test to be valid.

The test is possible when we have replications, that is more than one observation for some values of the explanatory variable. In Example 2.7 we have five observations for each age x_i.

Notation:
Denote by Y_{ij} the j-th response at x_i, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, that is the number of all observations is $n = \sum_{i=1}^{m} n_i$. The average response at x_i is

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

We denote the fitted response at x_i by $\widehat{Y}_i$, which is the same for all observations at x_i. □

The residuals e_{ij} are

$$e_{ij} = Y_{ij} - \widehat{Y}_i.$$

These differences arise for two reasons. Firstly the j-th observation of a given x_i is an outcome of a random variable. Observations obtained for the same value of X may produce different values of Y. Secondly the model we fit may not be a good one.

How could we distinguish between the random variation and the lack of fit? We need more than one observation at x_i to be able to do it.

The difference

$$Y_{ij} - \bar{Y}_i$$

indicates the random variation at x_i; it is called *pure error*. The difference between the mean and the fitted response, i.e.,

$$\bar{Y}_i - \widehat{Y}_i,$$

indicates *lack of fit* at x_i.

Using the double index notation we may write the sum of squares for residuals as

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{Y}_i)^2.$$

We can also define the *pure error sum of squares* as

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

and the *lack of fit sum of squares* as a measure of lack of fit:

$$\begin{aligned} SS_{LoF} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{Y}_i - \widehat{Y}_i)^2 \\ &= \sum_{i=1}^{m} n_i(\bar{Y}_i - \widehat{Y}_i)^2. \end{aligned}$$

Theorem 2.6. *In the simple linear regression model we have*

$$SS_E = SS_{LoF} + SS_{PE}.$$

Proof.

$$\begin{aligned} SS_E &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{Y}_i)^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\{(Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \widehat{Y}_i)\}^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{Y}_i - \widehat{Y}_i)^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \widehat{Y}_i) \\ &= SS_{PE} + SS_{LoF} + 2\sum_{i=1}^{m}(\bar{Y}_i - \widehat{Y}_i)\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i) \\ &= SS_{PE} + SS_{LoF} \end{aligned}$$

since $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i) = 0$. □

This theorem shows how the residual sum of squares is split into two parts, one due to the pure error and one due to the model lack of fit. To work out the split of the degrees of freedom, note that to calculate SS_{PE} we must calculate m sample means $\bar{Y}_i, i = 1, \ldots, m$. Each sample mean takes up one degree of freedom. Thus the degrees of freedom for pure error are $n - m$. By subtraction, the degrees of freedom for lack of fit are

$$\nu_{LoF} = \nu_E - \nu_{PE} = (n-2) - (n-m) = m - 2.$$

This can be included in the Analysis of variance table as follows:

ANOVA table

Source of variation	d.f.	SS	MS	VR
Regression	1	SS_R	MS_R	$\frac{MS_R}{MS_E}$
Residual	$n-2$	SS_E	$MS_E = \frac{SS_E}{n-2}$	
Lack of fit	$m-2$	SS_{LoF}	$MS_{LoF}=\frac{SS_{LoF}}{m-2}$	$\frac{MS_{LoF}}{MS_{PE}}$
Pure Error	$n-m$	SS_{PE}	$MS_{PE}=\frac{SS_{PE}}{n-m}$	
Total	$n-1$	SS_T		

We will see later that

$$\mathrm{E}[SS_{PE}] = (n-m)\sigma^2$$

whether the simple linear regression model is true or not.

It can also be shown that if the simple linear regression model is true then

$$\mathrm{E}[SS_{LoF}] = (m-2)\sigma^2.$$

Hence, both MS_{PE} and MS_{LoF} give us unbiased estimators of σ^2, but the latter one only if the model is true.

Let
H_0 : simple linear regression model is "true"
H_1 : $\neg H_0$

Then, under H_0,

$$\frac{(m-2)MS_{LoF}}{\sigma^2} \underset{H_0}{\sim} \chi^2_{m-2}.$$

Also

$$\frac{(n-m)MS_{PE}}{\sigma^2} \sim \chi^2_{n-m}$$

whatever the model.

Hence, under H_0, the ratio of these two independent statistics divided by the respective degrees of freedom is distributed as $\mathcal{F}_{m-2,n-m}$, namely

$$F = \frac{MS_{LoF}}{MS_{PE}} \underset{H_0}{\sim} \mathcal{F}_{m-2,n-m}.$$

Note that we can only do this lack of fit test if we have replications. These have to be true replications, not just repeated measurements on the same sampling unit.

Example 2.8. Plasma level continued.
To illustrate these ideas we return to the plasma example. We have seen that the residual plots show some evidence that a transformation is necessary. The analysis of variance table for the plasma data after the log transformation of the response variable is following.

MINITAB
we can get the decomposition of the residual sum of squares into pure error sum of squares and lack of fit sum of squares by clicking on pure error within Options under Stat → Regression → Regression.

Source	DF	SS	MS	F	P
Regression	1	2.6554	2.6554	60.63	0.000
Residual Error	23	1.0073	0.0438		
Lack of Fit	3	0.0885	0.0295	0.64	0.597
Pure Error	20	0.9188	0.0459		
Total	24	3.6627			

The p-value is 0.597 so the numerical output shows no reason to doubt the fit of this model.
□