Chapter 2

Diagnostics and Remedial Measures

2.1 Basic assumptions, departures and remedial measures

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n, \]

\( \beta_0 \) and \( \beta_1 \) are parameters
\( X_1, \ldots, X_n \) are known constants
\( \epsilon_i, \ldots, \epsilon_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \).

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Departure</th>
<th>Remedial measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear model</td>
<td>Nonlinearity of regression function</td>
<td></td>
</tr>
<tr>
<td>Constancy of error variance</td>
<td>Nonconstancy of error variance</td>
<td></td>
</tr>
<tr>
<td>Independence of error terms</td>
<td>Nonindependence of error terms</td>
<td></td>
</tr>
<tr>
<td>Normality of error terms</td>
<td>Nonnormality of error terms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presence of outliers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Omission of important predictor variables</td>
<td></td>
</tr>
</tbody>
</table>

- Two methods for studying the appropriateness of a model
  
  (1) Graphic diagnostics;
  (2) Formal statistical tests with the null hypothesis being a basic assumption.
2.2 Residual analysis

- The \( i \)th residual is the difference between the observed value \( y_i \) and the corresponding fitted value \( \hat{y}_i \).

\[
e_i \triangleq y_i - \hat{y}_i
\]

- The \( i \)th semistudentized residual

\[
e_i^* \triangleq \frac{e_i}{\sqrt{MSE}}
\]

- The basic idea underlying residual analysis

There is an error term in the model \( Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \), which is assumed to be a random variable;

Regard \( e_i \) as the observed error;

If the model is appropriate for the data at hand, the observed residuals \( e_i \) should reflect the properties assumed for the \( \varepsilon_i \).
## Graphic diagnostics

<table>
<thead>
<tr>
<th>Departure</th>
<th>Graphic Diagnostics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlinearity of regression function</td>
<td>(1) Scatter plot</td>
</tr>
<tr>
<td></td>
<td>(2) Residual plot against the predictor variable</td>
</tr>
<tr>
<td></td>
<td>(2') Residual plot against the fitted values</td>
</tr>
<tr>
<td>Nonconstancy of error variance</td>
<td>Residual plot against the predictor variable (or fitted values)</td>
</tr>
<tr>
<td>Nonindependence of error terms</td>
<td>Sequence plot of residuals</td>
</tr>
<tr>
<td>Nonnormality of error terms</td>
<td>(1) Histogram, stem-and-leaf plot of residuals</td>
</tr>
<tr>
<td></td>
<td>(2) Normal Q-Q plot of residuals</td>
</tr>
<tr>
<td>Presence of outliers</td>
<td>(1) Boxplot, stem-and-leaf plot of residuals</td>
</tr>
<tr>
<td></td>
<td>(2) Plot semistudentized residuals</td>
</tr>
<tr>
<td>Omission of important predictor variables</td>
<td></td>
</tr>
</tbody>
</table>

## Tests involving residuals

<table>
<thead>
<tr>
<th>Departure</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomness</td>
<td>(1) Runs test</td>
</tr>
<tr>
<td></td>
<td>(2) The Durbin-Watson test (Ch.12)</td>
</tr>
<tr>
<td>Constancy of error variance</td>
<td>(1) The modified Levene test (Section 3.6)</td>
</tr>
<tr>
<td></td>
<td>(2) The Breusch-Pagan test (Section 3.6)</td>
</tr>
<tr>
<td>Normality</td>
<td>(1) Goodness of fit tests</td>
</tr>
<tr>
<td></td>
<td>(2) Correlation test (Section 3.5)</td>
</tr>
<tr>
<td>Outliers</td>
<td></td>
</tr>
</tbody>
</table>
• The modified Levene test for constancy of error variance

Divide the data set into two groups, according to the level of $X$, so that one group consists of cases where the $X$ level is comparatively low and the other group consists of cases where the $X$ level is comparatively high.

For group 1:

$n_1 \triangleq$ the sample size of group 1
$\tilde{e}_1 \triangleq$ the sample median
$e_{i1} \triangleq$ the $i$th residual
$d_{i1} \triangleq |e_{i1} - \tilde{e}_1|$ (the absolute deviation of the $i$th residual around the median of group 1)
$\bar{d}_1 \triangleq \frac{1}{n_1} \sum_{i=1}^{n_1} d_{i1}$ (the sample mean of $d_{i1}$).

For group 2:

$n_2 \triangleq$ the sample size of group 2
$\tilde{e}_2 \triangleq$ the sample median
$e_{i2} \triangleq$ the $i$th residual
$d_{i2} \triangleq |e_{i2} - \tilde{e}_2|$ (the absolute deviation of the $i$th residual around the median of group 2)
$\bar{d}_2 \triangleq \frac{1}{n_2} \sum_{i=1}^{n_2} d_{i2}$ (the sample mean of $d_{i2}$).

The null hypothesis $\mathcal{H}_0$ : the error terms have constant variance

The two-sample Student’s $t$ test

$$t^* \triangleq \frac{\bar{d}_1 - \bar{d}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p^2 \triangleq \frac{\sum_{i=1}^{n_1} (d_{i1} - \bar{d}_1)^2 + \sum_{i=2}^{n_2} (d_{i2} - \bar{d}_2)^2}{n - 2}$$

(the pooled variance)

The decision rule:

If $|\text{observed } t* \text{ value}| \leq q_t(1 - \alpha/2, n - 2)$, conclude that the error variance is constant;

If $|\text{observed } t* \text{ value}| > q_t(1 - \alpha/2, n - 2)$, conclude that the error variance is not constant.
Grade point average, page 38.

> data <- read.table("CH01PR19.DAT")
> x <- data[,2]  # entrance test score
> y <- data[,1]  # GPA at the end of the freshman year
> res <- residuals(lm(y ~ x))  # residuals

Now divide the residuals into two groups, according to \( X < 4.8 \) and \( X \geq 4.8 \).

> group1 <- res[x<4.8]
> group2 <- res[x>=4.8]

Define

> d1 <- abs(group1 - median(group1))  # "abs" is for absolute value
> d2 <- abs(group2 - median(group2))

Perform the two-sided Student's t test

> t.test(d1, d2, alternative="two.sided", var.equal=TRUE, mu=0)

    Two Sample t-test

    data:  d1 and d2
    t = 1.7187, df = 18, p-value = 0.1028

# The next part means something else. Leave it alone.
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    -0.04880672  0.48779750
sample estimates:
    mean of x  mean of y
0.4573343  0.2378389

What is the P-value? and your decision?
• Correlation test for normality

The \textit{ith} residual

\[ e_i \triangleq y_i - \hat{y}_i \]

An approximation of the expected value of the \( k \)th smallest residual under normality is

\[ \sqrt{MSE} \cdot \Phi^{-1} \left( \frac{k - 0.375}{n + 0.25} \right), \]

where \( \Phi^{-1}(t) \) is the quantile function of the standard normal distribution.

The correlation test is based on the coefficient of correlation between the ordered residuals and their expected values under normality.

\textit{Calculator maintenance}, Problem 1.20.

\begin{verbatim}
> data <- read.table("CH01PR20.DAT")
> x <- data[,2]    # the number of machines served
> y <- data[,1]    # the total number of minutes spent by the service person

The ordered residuals are obtained by \texttt{sort residuals}

> res <- residuals(lm(y ~ x))
> res.order <- sort(res)  # ordered residuals

Their expected values are

> res.exp <- sqrt(MSE)*qnorm((1:length(res)-0.375)/(length(res)+0.25))
  # MSE <- deviance(lm(y~x))/(length(y)-2)
  # deviance(lm(y~x)) is for SSE

The coefficient of correlation between the ordered residuals and their expected values under normality is

> cor(res.order, res.exp)
[1] 0.9808162
\end{verbatim}

For the level of significance \( \alpha = 5\% \), the critical value for \( n = 18 \) is 0.946 (Table B.6). Since 0.9808162 > 0.946, we conclude that the distribution of the error terms doesn't depart substantially from a normal distribution.

Compare the normal Q-Q plot of residuals with the plot of \texttt{res.order} against \texttt{res.exp}, and with the plot of \texttt{res.order} against \texttt{res.exp/sqrt(MSE)}. What do you see?
2.3 Remedial measures

<table>
<thead>
<tr>
<th>Departure</th>
<th>Remedial measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlinearity of regression function</td>
<td>Choice (1) Develop and use other models</td>
</tr>
<tr>
<td></td>
<td>Choice (2) Make a transformation</td>
</tr>
<tr>
<td>Nonconstancy of error variance</td>
<td>(1) Use the method of <em>weighted least squares</em></td>
</tr>
<tr>
<td></td>
<td>(2) Make a transformation</td>
</tr>
<tr>
<td>Nonindependence of error terms</td>
<td>Work with some models for correlated error terms</td>
</tr>
<tr>
<td>Nonnormality of error terms</td>
<td>Make a transformation</td>
</tr>
<tr>
<td>Presence of outliers</td>
<td>(1) Discard outliers</td>
</tr>
<tr>
<td></td>
<td>(2) Use some robust estimation procedure (Ch.10)</td>
</tr>
<tr>
<td>Omission of important predictor variables</td>
<td></td>
</tr>
</tbody>
</table>
• Box-Cox transformations

\[ Y' = Y^\lambda, \]

where \( \lambda \) is a parameter to be determined from the data. In particular, if \( \lambda = 0 \), then
\[ Y' = \log_e Y. \]

The plasma levels example, page 129.

> data <- read.table("CHO3TA08.DAT")
> x <- data[,1] # age
> y <- data[,2] # plasma level of a polyamine

> plot(x, y)

As is seen from the scatter plot, there is the distinct curvilinear relationship, as well as the greater variability for younger children than older ones. Look at the residual plot against the predictor variable. What does it look like?

> fit <- lm(y ~ x)
> plot(x, fit$residuals)
> abline(h=0, col=2)

Try also the residual plot against the fitted values. Is it equivalent to the residual plot against the predictor variable?

> plot(fit$fitted, fit$residuals)
> abline(h=0, col=2)

It would be helpful to make some power transformations.

Define the Box-Cox transformation as below, and then search for an appropriate lambda.

> box.cox_function(lambda, y){
  if(lambda !=0) box.cox <- y^lambda
  if(lambda==0) box.cox <- log(y)
  box.cox
}

\[ \lim_{\lambda \to 0^+} \frac{Y^\lambda - 1}{\lambda} = \log Y \]
Try $\lambda = 1/2$,

\begin{verbatim}
> plot(x, box.cox(1/2, y))
> abline(lm(box.cox(1/2, y) ~ x))
> plot(x, residuals(lm(box.cox(1/2, y) ~ x)))
> abline(h=0, col=2)
\end{verbatim}

$\lambda = -1/2$,

\begin{verbatim}
> plot(x, box.cox(-1/2, y))
> abline(lm(box.cox(-1/2, y) ~ x))
\end{verbatim}
> plot(x, residuals(lm(box.cox(-1/2, y) ~ x)))
> abline(h=0, col=2)

and $\lambda = 0$,

> plot(x, box.cox(0, y))
> abline(lm(box.cox(0, y) ~ x))
> plot(x, residuals(lm(box.cox(0, y) ~ x)))
> abline(h=0, col=2)

Which one looks better? You may compare the SSEs associated to the transformations.