ST430: Introduction to Regression Analysis, Chapter 8, Sections 1-4

Luo Xiao

November 2, 2015
Residual Analysis
Inferences about a regression model are valid only under assumptions about the random errors in the observations.

Objectives:
- Show how residuals reveal departures from assumptions;
- Suggest procedures for coping with such departures.
Regression residuals

The random errors $\epsilon$ satisfy

$$Y = E(Y) + \epsilon, \text{ or } \epsilon = Y - E(Y).$$

We observe $Y$, but we do not know $E(Y)$, so we cannot calculate $\epsilon$.

We estimate $E(Y)$ by $\hat{Y}$, the predicted (or fitted) value.

We approximate the random errors by regression residuals:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots, n.$$
Properties of residuals

If the model contains an intercept, the sum of the residuals, and also their mean, is zero:

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0, \text{ and so } \bar{\hat{\epsilon}} = 0.$$  

The covariance of the residuals and any term in the regression model is zero:

$$\sum_{i=1}^{n} \hat{\epsilon}_i X_{i,j} = 0, \quad j = 1, 2, \ldots, k.$$  

Detecting lack of fit

A *misspecified* model is one that leaves out a relevant predictor.

The residuals from a misspecified model do not have mean zero.

Example: serum cholesterol \((Y)\) and dietary fat \((X)\) in Olympic athletes.

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("OLYMPIC.RData")
pairs(OLYMPIC)
```
Scatter plot
Suppose we ignore the graph, and fit a first-order model (see “output1.txt”):

```r
l1 <- lm(CHOLES ~ FAT, OLYMPIC)
summary(l1)
plot(OLYMPIC$FAT, residuals(l1),pch=20) #X versus residuals
abline(h=0,col="gray") #0 horizontal line
```

- The summary of the fitted model looks reasonable.
- But the graph of the residuals against $X$ show that the assumption $E(\epsilon) = 0$ is violated.
- Because this is a straight-line model, this graph is effectively the same as the “residuals versus fitted value” graph from `plot(l1)`.
Residual plot
Because of the curvature, we could fit the second-order (quadratic) model (see “output2.txt”):

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("OLYMPIC.RData")

l2 <- lm(CHOLES ~ FAT + I(FAT^2), OLYMPIC)
summary(l2)
plot(OLYMPIC$FAT, residuals(l2),pch=20)
abline(h=0,col="gray")
```

- The residual plot suggests that the model is satisfactory.
- The quadratic term is highly significant.
Residual plot
Partial residuals

Sometimes the effect of an independent variable is better described by a transformed version: \(\log(X), 1/X\), etc.

The *partial residual* plot can help identify the transformation:

- The partial residuals for independent variable \(X_j\) are

\[
\hat{\epsilon}^* = \hat{\epsilon} + \hat{\beta}_j X_j
\]

- Plot \(\hat{\epsilon}^*\) against \(X_j\).

Also known as a “Component + Residual” plot.
Example

Effect of price ($p$) and advertising ($X_2$) on demand ($Y$) for coffee.

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("COFFEE2.RData")
pairs(COFFEE2,pch=20)
```

Try a first-order model:

```r
l1 <- lm(DEMAND ~ PRICE + X, COFFEE2)
summary(l1)
plot(COFFEE2$PRICE, residuals(l1),pch=20)
abline(h=0,col="gray")
```

The residual plot shows misspecification.
Scatter plot

Residual Analysis
Residual plot
The Component + Residual plot:

```r
library(car)
crPlot(l1, variable = "PRICE", pch=20)
```

- Curve suggests either adding 'PRICE2', or transforming to log(PRICE) or 1/PRICE.
- $R^2$ and $R_a^2$ are highest for 1/PRICE.
- Note: the partial regression plot is different.
Component + Residual plot

Residual Analysis
Detecting unequal variances

**Homoscedasticity versus heteroscedasticity.**

That is, constant variance versus varying variance.

When the variance is not constant, it is most often related to the mean.

For Poisson-distributed data (counts), \( \text{var}(Y) = E(Y) \).

When errors are multiplicative, \( Y = E(Y) \times (1 + \epsilon) \), and \( \text{var}(Y) \propto E(Y)^2 \).
Sometimes the variance can be made constant by transforming \( Y \).

For example, with multiplicative errors,

\[
\log(Y) = \log[E(Y) \times (1 + \epsilon)] \\
= \log[E(Y)] + \log[1 + \epsilon] \\
\approx \log[E(Y)] + \epsilon.
\]

So \( \text{var}[\log Y] \) is (approximately) constant.

Sometimes variance can be made constant by transformation, but a different method may be better than using a transformation.

For example, with Poisson-distributed counts, \( \sqrt{Y} \) has approximately constant variance, but a generalized linear model may be more satisfactory.
Example

Salary and experience for social workers.

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("SOCWORK.RData")
pairs(SOCWORK,pch=20)

Try a second-order model:

```r
l2 <- lm(SALARY ~ EXP + I(EXP^2), SOCWORK)
summary(l2)
plot(fitted(l2), residuals(l2), pch=20)
abline(h=0, col="gray")
```
Scatter plot

EXP

SALARY

LNSALARY
Residual plot

Residual Analysis
The “Residuals vs Fitted” plot shows a fan-shaped scatter, and the “Scale-Location” plot shows an upward trend (try ’plot(l2’)).

It suggests

\[ \text{std dev}(Y) \propto E(Y), \]

hence

\[ \text{var}(Y) \propto E(Y)^2, \]

so try logarithms.
Second-order model for \( \log(\text{SALARY}) \):

\[
\text{lLog2} \leftarrow \text{lm}(\log(\text{SALARY}) \sim \text{EXP} + \text{I}(\text{EXP}^2), \text{SOCWORK})
\]
\[
\text{summary}(\text{lLog2})
\]

The quadratic term is not significant, so try a first-order model:

\[
\text{lLog1} \leftarrow \text{lm}(\log(\text{SALARY}) \sim \text{EXP}, \text{SOCWORK})
\]
\[
\text{summary}(\text{lLog1})
\]
\[
\text{plot(fitted(lLog1)}, \text{residuals(lLog1),pch=20)}
\]
\[
\text{abline(h=0,col="gray")}
\]

The residual plots are more satisfactory.
Residual plot

Residual Analysis
Simple test for heteroscedasticity

Divide the data set in two, for instance low fitted values versus high fitted values.

Fit the model separately to each part, and compare the MSEs (Mean Square Errors).

Under $H_0$: variance is constant,

$$F^* = \frac{MSE_1}{MSE_2}$$

has the $F$-distribution with $\nu_1 = n_1 - (k + 1)$ and $\nu_2 = n_2 - (k + 1)$ degrees of freedom.
This is usually a two-sided test; $H_a$: variance is not constant.

Reject $H_0$ at level $\alpha$ if $F^*$ differs too far from 1 in either direction; that is, if

- $F^* < F_{1-\alpha/2}(\nu_1, \nu_2)$, the lower $\alpha/2$-point of the distribution, or
- $F^* > F_{\alpha/2}(\nu_1, \nu_2)$, the upper $\alpha/2$-point of the distribution.
Note: $F_{1-\alpha/2}(\nu_1, \nu_2) = 1/F_{\alpha/2}(\nu_2, \nu_1)$, so an equivalent method is based on

\[ F = \frac{\text{Larger MSE}}{\text{Smaller MSE}} = \max \left( F^*, \frac{1}{F^*} \right). \]

Then we reject $H_0$ if

\[ F > F_{\alpha/2}(\nu_{\text{Larger}}, \nu_{\text{Smaller}}). \]