Multiple Linear Regression Models: Diagnostics

Non-Linearity: Residuals vs. Fitted-values, Residuals vs. Explanatory variables, Added Variable Plots, GAM plots

Constant Variance: Funnel plots and Weighted RSS

Outliers: Normal plots, Residuals vs. Fitted-values plots, Influence statistics

Independence: Correlogram, Durbin-Watson test

Normality of Errors: Diagnostics 5 (TODAY)
Aims of today’s lecture

- To discuss diagnostics and remedies for non-normality.
- To apply these and the other diagnostics in a case study.
Normality

- Another assumption in the regression model is that the errors are normally distributed.

- This is not so crucial, but can be important if the errors have a long-tailed distribution, since this will imply there are several outliers.

- Normality assumption important for prediction.
Detecting non-normality

The standard diagnostic is the normal plot of the residuals

\[ \text{qqnorm(residuals(\text{xyz.lm}))} \]
The Weisberg-Bingham test

- Test statistic: \( WB \) is the square of the correlation of the normal plot, measures how straight the plot is.

- \( WB \) lies between 0 and 1. Values close to 1 indicate normality.

- \textbf{R} function \texttt{WB.test} calculates \( WB \) statistic and computes \( p \)-value for test with null hypothesis of sample being normal.

- \( WB \) test is a variant of Shapiro-Wilk test.
Example: Residuals of cherry trees

\begin{verbatim}
qqnorm(residuals(cherry.cone))
WB.test(cherry.cone)

WB test statistic = 0.983
p = 0.36

Since p-value large, no evidence against normality.
\end{verbatim}
Remedies for Non-normality

- The standard remedy is to transform the response using a power transformation.

- The idea is, on the original scale the model does not fit well, but on the transformed scale it does.

- The power is obtained by means of a Box-Cox plot.

- The idea is to assume that for some power $p$, the response $y^p$ follows the regression model. The plot is a graphical way of estimating the power $p$.

- Technically, it is a plot of the profile likelihood.
Multiple Linear Regression Models: Diagnostics

Non-Linearity: Residuals vs. Fitted-values, Residuals vs. Explanatory variables, Added Variable Plots, GAM plots.

Constant Variance: Funnel plots and Weighted RSS.

Outliers: Normal plots, Residuals vs. Fitted-values plots, Influence statistics.

Independence: Correlogram, Durbin-Watson test.

Normality of Errors: QQ-plots, Weisberg-Bingham test.
1976 Chateau Margaux
NZ$250 – 800

1976 Chateau Latour
NZ$370 – 900

1976 Chateau Lafite Rothschild
NZ$550 – 1,300
Case study: The wine data

- Data on 27 vintages of Bordeaux wines.

- Variables are

  - **year**: 1952 – 1980;
  - **price**: in 1980 US$, converted to an index with 1961 = 100;
  - **temp**: average temperature during growing season (°C);
  - **h.rain**: total rainfall during harvest period, mm;
  - **w.rain**: total rainfall over preceding winter, mm;

- Data part of **R330** package, also available on course website.
Bordeaux wines: Their price

- Bordeaux wines are an iconic luxury consumer good. Many consider these to be the best wines in the world.

- The quality and the price depends on the vintage (i.e. the year the wines are made.)

- The prices are (in 1980 US$, in index form, 1961 = 100).
Bordeaux wines: Their price

Bordeaux wine price index 1952–1980

year
price
Bordeaux wines: Pairs plot

- **year**
- **temp**
- **h.rain**
- **w.rain**
- **price**

***Correlation Coefficients:***
- year: 0.29
- temp: 0.33
- h.rain: 0.27
- w.rain: 0.45
- price: 0.59
Bordeaux wines: Preliminary analysis

Call:
lm(formula = price ~ year + temp + h.rain + w.rain,
    data = wine.df)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1305.527 | 597.311    | 2.186   | 0.03977  *|
| year           | -0.82055 | 0.29140    | -2.816  | 0.01007  *|
| temp           | 19.25337 | 3.92945    | 4.900   | 6.72e-05 ***|
| h.rain         | -0.10121 | 0.03297    | -3.070  | 0.00561  **|
| w.rain         | 0.05704  | 0.01975    | 2.889   | 0.00853  **|

Residual standard error: 11.69 on 22 degrees of freedom
Multiple R-squared: 0.7369, Adjusted R-squared: 0.6891
F-statistic: 15.41 on 4 and 22 DF,  p-value: 3.806e-06
Bordeaux wines: Diagnostic plots

Residuals vs Fitted

Normal Q–Q

Scale–Location

Residuals vs Leverage

Cook's distance
Bordeaux wines: Checking normality

\[ \text{qqnorm(residuals(wine.lm))} \]
\[ \text{WB.test(wine.lm)} \]

WB test statistic = 0.957

\[ p = 0.03 \]
Bordeaux wines: Box-Cox routine

Box-Cox plot

Profile likelihood

Power = −1/3
Bordeaux wines: Transform and refit

- Use $y^{-1/3}$ as a response (reciprocal cubic root)

- Has the fit improved?
  - Are the errors more normal? (normal plot and Weisberg-Bingham test)
  - Has $R^2$ increased?
  - Would further transformations improve normality?
  - Minimum at $p = 1$ means fit cannot be improved by further transformation.
Bordeaux wines: Re-Checking normality

qqnorm(residuals(wine.upd))
WB.test(wine.upd)

WB test statistic = 0.988
p = 0.65
Bordeaux wines: Re-Checking normality

Call:
lm(formula = price^(-1/3) ~ ., data = wine.df)
---
Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| -3.666e+00 | 1.613e+00  | -2.273  | 0.03317  * |
| year       | 2.639e-03  | 7.870e-04  | 3.353   | 0.00288  ** |
| temp       | -7.051e-02 | 1.061e-02  | -6.644  | 1.11e-06  *** |
| h.rain     | 4.423e-04  | 8.905e-05  | 4.967   | 5.71e-05  *** |
| w.rain     | -1.157e-04 | 5.333e-05  | -2.170  | 0.04110  * |
---
Residual standard error: 0.03156 on 22 degrees of freedom
Multiple R-squared: 0.8331, Adjusted R-squared: 0.8028
F-statistic: 27.46 on 4 and 22 DF, p-value: 2.841e-08
Bordeaux wines: Box-Cox revisited

Box–Cox plot

Profile likelihood

Power = 1
Bordeaux wines: Conclusions on Normality

- Transformation has been spectacularly successful in improving the fit!

- What about other aspects of the fit?
  - residuals/fitted values?
  - Pairs
  - Added variable plots.
Bordeaux wines: Diagnostics plots revisited

Residuals vs Fitted

Normal Q–Q

Scale–Location

Residuals vs Leverage

ALL GOOD
Bordeaux wines: GAM plots

library(mgcv)
plot(gam(price^(-1/3)~temp+s(h.rain)+s(w.rain)+year,
data=wine.df))
Bordeaux wines: Polynomial fit

Call:
\( \text{lm(formula = price}^{(-1/3)} \sim \text{temp + h.rain + year}
\quad + \text{poly(w.rain, 4), data = wine.df)} \)

---

Coefficients:

|                | Estimate  | Std. Error | t value | Pr(>|t|) |
|----------------|-----------|------------|---------|----------|
| (Intercept)    | -2.974e+00 | 1.532e+00 | -1.942  | 0.06715  .|
| temp           | -7.478e-02 | 1.048e-02 | -7.137  | 8.75e-07 ***|
| h.rain         | 4.869e-04  | 8.662e-05 | 5.622   | 2.02e-05 ***|
| year           | 2.284e-03  | 7.459e-04 | 3.062   | 0.00642  **|
| poly(w.rain, 4)1 | -7.561e-02 | 3.263e-02 | -2.317  | 0.03180  * |
| poly(w.rain, 4)2 | 4.469e-02  | 3.294e-02 | 1.357   | 0.19079  |
| poly(w.rain, 4)3 | -2.153e-02 | 2.945e-02 | -0.731  | 0.47374  |
| poly(w.rain, 4)4 | 6.130e-02  | 2.956e-02 | 2.074   | 0.05194  .|

---

Residual standard error: 0.02931 on 19 degrees of freedom
Multiple R-squared:  0.8757, Adjusted R-squared:  0.8299
F-statistic: 19.13 on 7 and 19 DF,  p-value: 2.352e-07
Bordeaux wines: Final fit

Call:
lm(formula = price^(-1/3) ~ ., data = wine.df)
---
Coefficients:

                     Estimate  Std. Error  t value  Pr(>|t|)  
(Intercept)   -3.666e+00  1.613e+00   -2.273  0.03317 *  
   year          2.639e-03  7.870e-04    3.353  0.00288 ** 
    temp         -7.051e-02  1.061e-02   -6.644  1.11e-06 ***
   h.rain        4.423e-04  8.905e-05    4.967  5.71e-05 ***
  w.rain        -1.157e-04  5.333e-05   -2.170  0.04110 *  
---
Residual standard error: 0.03156 on 22 degrees of freedom
Multiple R-squared:  0.8331, Adjusted R-squared:  0.8028
F-statistic: 27.46 on 4 and 22 DF,  p-value: 2.841e-08
Bordeaux wines: Conclusions

- Model using $\text{price}^{-1/3}$ as response fits very well.

- Use this model for prediction, understanding relationships:
  - Coefficient of year positive, so transformed response increases with year (i.e. older vintages are more valuable).
  - Coefficient of temp negative, so high temperatures decrease transformed response (i.e. increase price).
  - Coefficient of h.rain positive, so high harvest rain increases transformed response (i.e. decreases price).
  - Coefficient of w.rain negative, so high winter rain decreases transformed response (i.e. increases price).
Other uses for Box-Cox plots

- The transformation indicated by the Box-Cox plot is also useful in fixing up non-planar data, and unequal variances, as we saw in Lecture 10.

- If the GAM plots indicate several independent variables need transforming, then transforming the response using the Box-Cox power often works.
Thank you

http://www.savagechickens.com/