STAT 425

Model Diagnostics. Part 2

Model Diagnostics: Checking Error Assumptions

Which assumptions?

- Constant Variance
- Normality
- Uncorrelated errors

How can we check these assumptions?

- Graphical tools: Residual plots, QQ-plots

Which remedies?

- Transformations, Generalized Least-Squares, nonlinear regression

Residual Plots

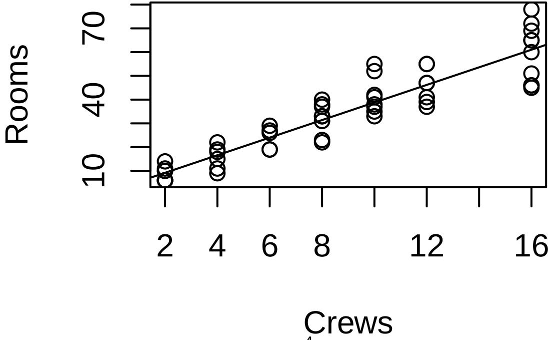
- Plot residuals r_i or studentized t_i against fitted values $\hat{y}_i$.
- Plot residuals r_i or studentized t_i against each predictor x_i.
- Plot residuals r_i or studentized t_i against an index variable such as time or case number.
- Look for systemic patterns (non-constant variance, non-linearity) and large absolute values of residuals.

Example: Cleaning data (Sheather)

```
# Read cleaning data
cleaning<-read.table("cleaning.txt",header=TRUE)

# Display basic statistical measures
summary(cleaning)

##         Case          Crews              Rooms
##  Min.   : 1   Min.   : 2.000   Min.   : 6.00
##  1st Qu.:14   1st Qu.: 4.000   1st Qu.:19.00
##  Median :27   Median : 8.000   Median :35.00
##  Mean   :27   Mean   : 8.679   Mean   :33.91
##  3rd Qu.:40   3rd Qu.:12.000   3rd Qu.:46.00
##  Max.   :53   Max.   :16.000   Max.   :78.00

modclean<-lm(Rooms ~ Crews,data=cleaning)
```

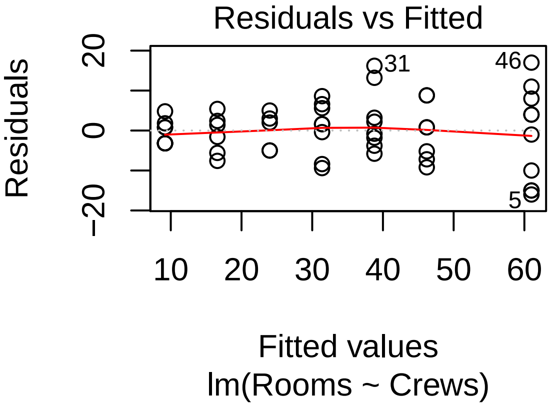
Example: Cleaning data (Sheather)

```
plot(Rooms ~ Crews,data=cleaning)
abline(modclean)
```



Example: Cleaning data (Sheather)

```
plot(modclean,which=1)
```



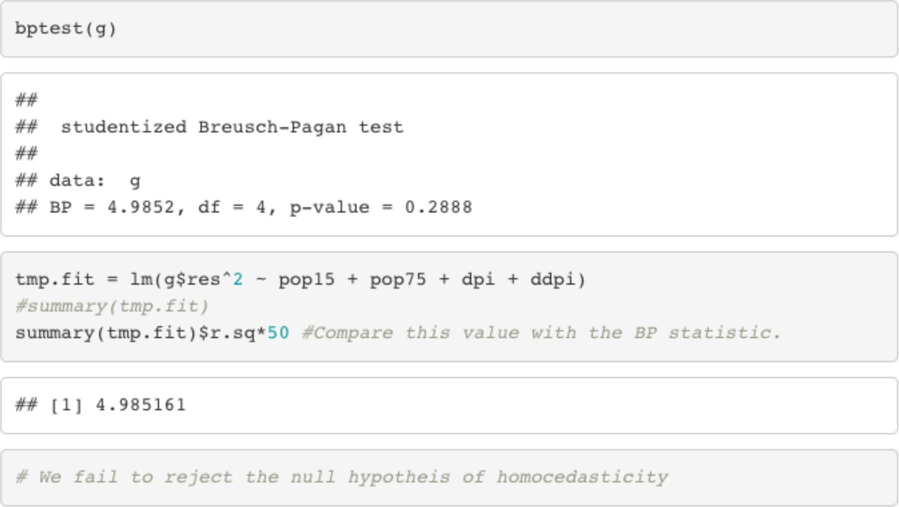
Non-Constant Variance

- Check residual plots and look for a “fan” type shape or trends
- A less rigorous but quick way:

  lm(abs(mod$residuals) ~ mod$fitted.values)
- A formal test: Breusch-Pagan Test (bptest in package lmtest)
- Remedy: transformation.

Breusch-Pagan test example: saving data set

Suppose $g : sr \sim pop15 + pop75 + dpi + ddpi$
We use function **bptest** from library **lmtest**



```
bptest(g)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  g
## BP = 4.9852, df = 4, p-value = 0.2888
```

```
tmp.fit = lm(g$res^2 ~ pop15 + pop75 + dpi + ddpi)
#summary(tmp.fit)
summary(tmp.fit)$r.sq*50 #Compare this value with the BP statistic.
```

```
## [1] 4.985161
```

```
# We fail to reject the null hypotheis of homocedasticity
```

Variance stabilizing transformations

The goal is to find a transformation of the response $h(Y)$ to achieve constant variance. The method for finding these transformations is based on the following. Suppose h is a smooth function. Then by the Taylor's theorem, the following expansion of $h(Y)$ around the $E[Y]$ holds:

$$h(Y) = h(E[Y]) + h'(E[Y])(Y - E[Y]) + \dots$$

where the dots $\dots$ represent the reminder of this approximation. This reminder is assumed small with high probability and we can ignore it. Then we have:

$$var[h(Y)] \approx (h'(E[Y]))^2 var[Y]$$

We want to choose a transformation h such that:

$$var[h(Y)] \approx (h'(E[Y]))^2 var[Y]$$

is approximately constant.
For example, suppose that the variance of Y is proportional to the mean of Y, i.e., $Var(Y) \propto E[Y]$, then is we select h such that:

$$h'(z) = \frac{1}{\sqrt{z}}$$

$$\Rightarrow h(z) \propto \sqrt{z}$$

When plugging-in the value of $h'(z)$ evaluated in $E[Y]$ in the variance of $h(Y)$ equation, the variance of $h(Y)$ will be approximately constant.

Another example:
Suppose $var(Y) \propto E[Y]^2$, then

$$h'(z) = \frac{1}{z} \Rightarrow h(z) = \log(z)$$

Residual plots can give an idea of the relationship between $var(Y) = var(e)$ (Residual variance) and the estimated $E[Y]$ (fitted values).

A summary a variance stabilizing transformations:

- When $var(e) \propto E[Y], \;\; h(Y) = \sqrt{Y}$. Suitable for counts from the Poisson distribution.
- When $var(e) \propto E[Y]^2, \;\; h(Y) = \log(Y)$ or $\log(Y+1)$. Suitable for data whose range of Y is very broad, e.g., from 1 to several thousands; suitable for estimating percentage effect ($Y \propto CX^{\alpha}$.)
- When $var(e) \propto E[Y]^4, \;\; h(Y) = 1/Y$ or $1/(Y+1)$. Suitable for data where Y measures the waiting time or survival time. Taking reciprocals changes the scale from time (time per response) to rate (response per unit time).

Assessing Normality

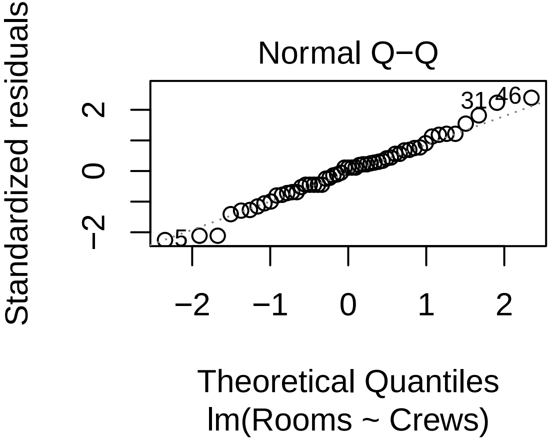
Suppose that we have a sample $z_1, z_2, \ldots, z_n$, and we wish to examine the hypothesis that the z's are a sample from a normal distribution with mean μ and variance σ^2. A standard graphical method for inspecting the normal assumption is the **QQ-plot**. It is calculated as follows:

1. Order the z's: $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$
2. Compute $u_i = \Phi^{-1}(\frac{i}{n+1})$, where Φ is the cdf of the $N(0,1)$ and i is the order if the data $(i = 1, 2, \ldots, n)$
3. Plot $z_{(i)}$ against u_i. If the z's are normal, the plot should be approximately a straight line

A more formal way to test normality: Shapiro-Wilks test.

Example: Cleaning data (Sheather)

```
plot(modclean,which=2)
```



Shapiro-Wilks test

H_0: The residuals follow a Normal distribution

```
# Shapiro-Wilks test
shapiro.test(residuals(modclean))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(modclean)
## W = 0.98681, p-value = 0.822
```

Since the p-value is large, we fail to reject the Null hypothesis

Correlated Errors

- Correlation is normally present when we have data with temporal, or spatial predictors
- We can plot residuals against time or other index, such as case number and look whether data above or below the mean tend to be followed by data above or below the mean
- To detect correlation: use formal tests like the Durbin-Watson test (dwtest in package lmtest)

Checking Model Structure Assumptions (Non-linearity)

How do we check that the linearity assumption $E[y] = \mathbf{X}\boldsymbol{\beta}$ is correct?

- We can apply the Lack-of-fit test when replicates are available (will be discussed later)
- Use partial regression plots
- Use partial residual plots
- Remedies to lack of linearity: Apply transformations, nonlinear regression (will be discussed later)

Partial Regression PLot (added Variable Plot)

- We want to know the relationship between the response Y and a predictor X_k after the effect of the other predictors has been removed.
- To remove the effect of the other predictors, run the following two regression models:

  $$Y \sim X_1 + \ldots + X_{k-1} + X_{k+1} + \ldots \tag{1}$$
  $$X_k \sim X_1 + \ldots + X_{k-1} + X_{k+1} + \ldots \tag{2}$$

  Get the following residuals:

  $$\mathbf{r}_y = \text{residuals from (1)}$$
  $$\mathbf{r}_k^X = \text{residuals from (2)}$$

- Plot $\mathbf{r}_y$ vs. $\mathbf{r}_k^X$: For a valid model, then the added-variable plot should produce points randomly scattered around a line through the origin with slope $\hat{\beta}_k$. This is also a useful plot to detect high influential data points.

Using transformations to overcome non-linearity

Examples of linearizing transformations:

- Use $\log(Y)$ vs. $\log(X)$ (apply logarithm to the response and the predictors): Suitable when $E[Y] = \alpha X_1^{\beta_1} \dots X_p^{\beta_p}$
- $\log(Y)$ vs. X (apply logarithm to the response only): Suitable when $E[Y] = \alpha \exp \sum_j X_j \beta_j$
- $1/Y$ vs. X (Take the inverse of the response): Suitable when $E[Y] = \frac{1}{\alpha + \sum_j X_j \beta_j}$

Box-Cox transformation of the Y variable

- Box and Cox (1964) suggested a family of transformations (for positive response) designed to reduce non-normality of the errors. It turns out that in doing this, it often reduces non-linearity as well.
- Suppose each $y_i > 0$, and consider the following transformation: [1]:

$$g_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

[1] The transformation for $\lambda = 0$ is justified because $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y)$

The aim is to choose λ that maximizes the likelihood of the data, under the normal assumption that the transformed data $g_\lambda(\mathbf{y})$ has a normal distribution:

$$g_\lambda(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$$

- The maximum log-likelihood function for $\lambda \neq 0$ is:

  $$L(\lambda) = -\frac{n}{2}\log(RSS_\lambda/n) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i)$$

  where RSS_λ is the residual sum of squares when $g_\lambda(\mathbf{y})$ is the response, and for $\lambda = 0$ is:

  $$L(0) = -\frac{n}{2}\log(RSS_0/n) - \sum_{i=1}^{n}\log(y_i)$$

  The second term in these log-likelihood function corresponds to the Jacobian of the transformation.
- Note that it doesn't make sense to simply pick λ that minimizes RSS_λ since for each λ, the residual sum of squares are measured in a different scale.

- In **R**, we can graph the log-likelihood as a function of λ ($L(\lambda)$) versus $\lambda \in (-2, 2)$ [2] and then pick the maximizer $\hat{\lambda}$.
- It is common to round $\hat{\lambda}$ to a nearby value like:

$$-1, -0.5, 0, 0.5, \quad \text{or} \quad 1$$

then the transformation defined by $\hat{\lambda}$ is easier to interpret.

[2] The method tends to work well for λ in this range

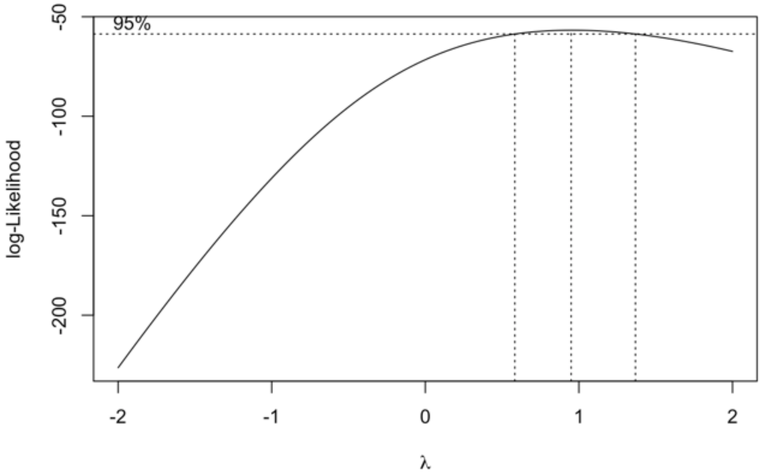
- To answer the question whether we really need the transformation g_λ, we can do hypothesis testing $(H_0 : \lambda = 1)$, or equivalently construct a Confidence Interval for λ as follows[3]:

$$\{\lambda : L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi_1^2(1-\alpha)\}$$

[3] This is based on the result that $2(L(\hat{\lambda}) - L(\lambda_0)) \sim \chi_1^2$ under H_0

Box-cox transformation example

```
boxcox(g,plotit=T) # plotit=T is the default setting
```


log-Likelihood
λ