Diagnostics (Part II)

Lecture 9

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES
Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will discuss:

- residual plots.
  - checking constancy of variance assumption.
  - checking normality assumption.
  - checking independence assumption.
  - checking non-linearity assumption.

Model Diagnostics: Checking Error Assumptions

Model Assumptions

$$\mathbf{Y} = \beta\mathbf{X} + \varepsilon, \quad \text{where } \epsilon \sim^{IID} \mathcal{N}(0, \sigma^2)$$

- Constant Variance
- Normality
- Uncorrelated errors

How to check these assumptions?

- *Graphical tools:* Residual plots, QQ-plots

Remedial Measures?

- Transformations, Generalized Least-Squares, Nonlinear Regression

Residual Plots

Residual Plots

- Plot *plain* or *studentized residuals* (r_i or t_i) against fitted values $\hat{y}_i$.
- Plot *plain* or *studentized residuals* (r_i or t_i) against each predictor x_i.
- Plot *plain* or *studentized* residuals (r_i or 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.

Checking Constancy of Variance

Checking Constancy of Variance

– Graphical Check: *Residuals against Fitted Values*

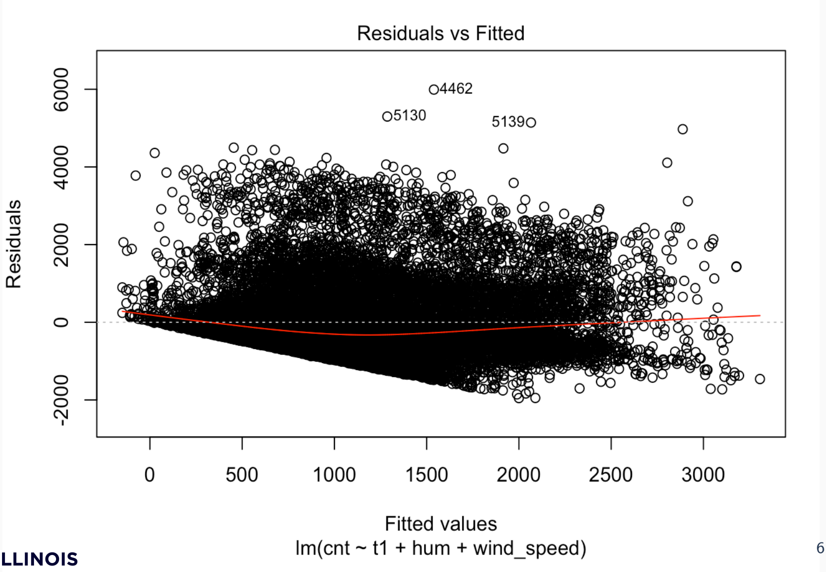
If the variance is constant, the residuals will look like a football-shaped cloud. Check residual plots and look for a "*fan*" type shape or trends.

– Statistical Test: Breusch-Pagan Test

In **R**: bptest in package lmtest

– Remedial measure: Variables' Transformation.

Residuals against Fitted Values: Bike Shares Example



Checking Constancy of Variance: Breusch-Pagan Test

- The BP test tests whether the variance of the errors from a regression is dependent on the values of the independent variables. If this is the case, we have *heteroscedasticity*.

- Under the H_0 hypothesis, the variance is constant, i.e. there is *homoscedasticity*.
  The *test statistic* is calculated as:

$$\text{BP} = nR^2,$$

  where R^2 is the coefficient of Determination between the squared residuals r_i^2 of a LS regression between Y and variables $X_1, X_2, \ldots, X_p$ (including the intercept), and the covariates (or a sub-set) $X_1, X_2, \ldots, X_p$.

- Under the H_0:

$$\text{BP} \sim \chi^2_{p-1}$$

  asymptotically.

Breusch-Pagan Test: Bike Shares Example

Regression Model (`bikeshare.mlr`): Bike Shares $\sim$ t1 + hum + wind_speed

$\Rightarrow$ Use function **bptest** from library **lmtest**

```
bptest(bikeshare.mlr)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bikeshare.mlr
## BP = 133.29, df = 3, p-value < 2.2e-16
```

Conclusion: Since the p-value is less than $\alpha = 5\%$, we reject the H_0 and conclude that the variance is not constant.

Goal:

Find a transformation of the response, $h(Y)$, to achieve constant variance.

How does it work?

- Suppose h is a smooth function.
- Using *Taylor's theorem*, the expansion of $h(Y)$ around $\mathbf{E}(Y)$ is:

$$h(Y) = h\Big(\mathbf{E}(Y)\Big) + h'\Big(\mathbf{E}(Y)\Big)\Big(Y - \mathbf{E}(Y)\Big) + \text{ Remainder}$$

- The remainder is assumed *small with high probability* and we can ignore it:

$$\mathrm{Var}\Big(\mathrm{h}(\mathrm{Y})\Big) \approx \left(\mathrm{h}'\Big(\mathbf{E}(\mathrm{Y})\Big)\right)^2 \mathrm{Var}(\mathrm{Y})$$

- We want to **choose** a transformation h such that $\mathrm{Var}\Big(\mathrm{h}(\mathrm{Y})\Big)$ is *approximately constant*.

Example

- Suppose that the variance of Y *is proportional to the mean of* Y, i.e., $\mathrm{Var(Y)} \propto \mathbf{E}(\mathrm{Y})$.

- 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 at $\mathbf{E}(Y)$ in the variance of $h(Y)$ equation, the variance of $h(Y)$ will be approximately constant. Indeed,

$$\mathrm{Var}\left(\sqrt{\mathrm{Y}}\right) \approx \left(\frac{1}{\sqrt{\mathbf{E}(\mathrm{Y})}}\right)^2 \mathrm{Var(Y)} = \frac{\mathrm{Var(Y)}}{\mathbf{E}(\mathrm{Y})} \approx \text{const.}$$

Another Example

- Suppose that the variance of Y *is proportional to the squared mean of* Y, i.e., $\mathrm{Var(Y)} \propto \Big(\mathbf{E}(\mathrm{Y})\Big)^2$.
- Select h such that:

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

- Then,

$$\mathrm{Var}\Big(\log \mathrm{Y}\Big) \approx \frac{1}{\Big(\mathbf{E}(\mathrm{Y})\Big)^2}\mathrm{Var(Y)} \approx \mathrm{const.}$$

In Practice

- How can we get an idea of the relationship between the *Residual Variance*, i.e. $\text{Var(Y)} = \text{Var}(\varepsilon)$ and the *Fitted Values*, i.e the estimated $\mathbb{E}(Y)$?
- Using residual plots.

Summary

A summary of variance stabilizing transformations:

- When $\mathrm{Var}(\varepsilon) \propto \mathbf{E}(\mathrm{Y})$, then $h(Y) = \sqrt{Y}$.
  Suitable for counts from the Poisson distribution.
- When $\mathrm{Var}(\varepsilon) \propto \Big(\mathbf{E}(\mathrm{Y})\Big)^2$, then $h(Y) = \log(Y)$ 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 $\mathrm{Var}(\varepsilon) \propto \Big(\mathbf{E}(\mathrm{Y})\Big)^4$, then $h(Y) = 1/Y$ 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).

Checking Normality

Assessing Normality: Graphical Test

– Suppose that we have a sample $z_1, z_2, \ldots, z_n$.
– We wish to examine the hypothesis that the z's are a sample from a normal distribution with mean μ and variance σ^2.

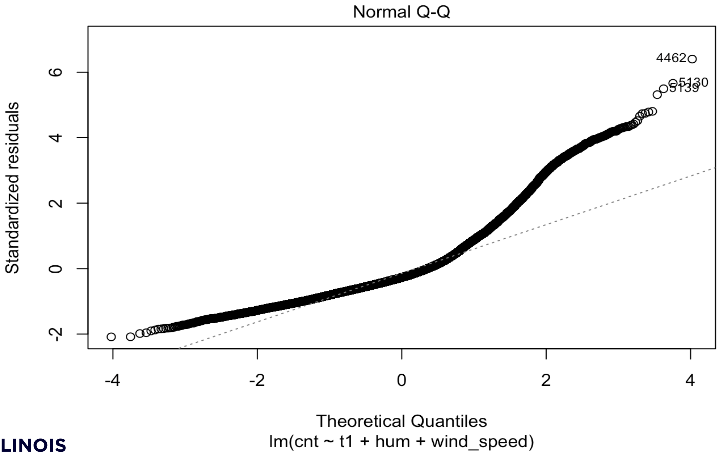
QQ-Plot

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.

$\Rightarrow$ If the z's are normal, the plot should be approximately a straight line.

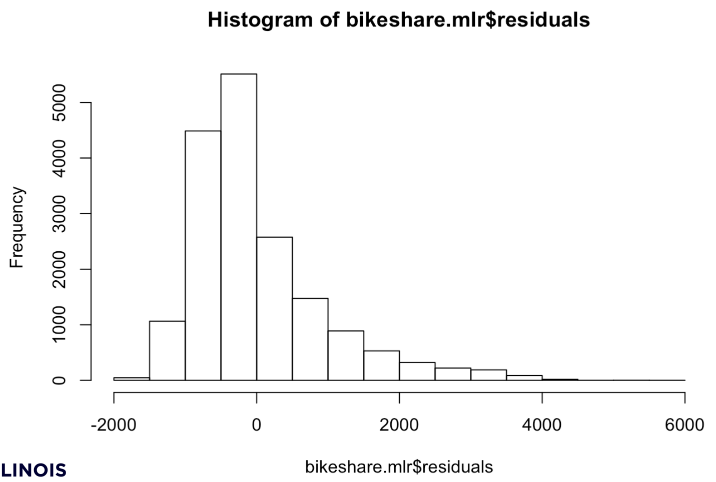
QQ-Plot: Bike Shares Example

```
plot(bikeshare.mlr, which=2)
```



Histogram: Bike Shares Example

```
hist(bikeshare.mlr$residuals)
```



- **Shapiro-Wilk Test (good for $n \leq 50$):**
  It tests the null hypothesis that a sample came from a normally distributed population:

$$W = \frac{\left(\sum_{i=1}^{n} a_i r_{(i)}\right)^2}{\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

  where $r_{(i)}$ is the ith largest value of the r_i's and the a_i terms are are calculated using the means, variances, and covariances of the r_is.
  Small values of W will lead to rejection of the null hypothesis.

- **Kolmogorov-Smirnov Test (good for $n > 50$):**

$$D_n = \sup_x \left| F_n(x) - \Phi(x) \right|$$

  where $\Phi(x)$ is the cdf of the Normal and F_n the empirical distribution function F_n for n i.i.d. ordered observations X_i is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[-\infty, x]}(X_i)$$

Kolmogorov-Smirnov Test: Bike Shares Example

H_0: The residuals follow a Normal distribution

```
ks.test(residuals(bikeshare.mlr), y=pnorm)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuals(bikeshare.mlr)
## D = 0.63627, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

$\Rightarrow$ Since the p-value is small, *we reject the Null hypothesis*.

Checking Serial Dependence

Sequence Plot

- 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)

Durbin-Watson Test

- **Durbin-Watson statistic:**

$$DW = \frac{\sum_{k=1}^{n-1} (r_k - r_{k+1})^2}{\sum_{k=1}^{n} r_k^2}$$

If **DW** $<$ **2**, then there is evidence for *positive serial dependence*.

Checking Non-Linearity

Checking Model Structure Assumptions (Non-linearity)

How do we check that the linearity assumption $\mathbb{E}(y) = \mathbf{X}\beta$ is correct?

We can use:

- *Partial Regression* plots.
- *Partial Residual* plots.
- *Lack-of-Fit* tests when replicates are available (will be discussed later)
- Remedial Measures to lack of linearity:
  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:

$$Y \sim X_1 + \ldots + X_{i-1} + X_{i+1} + \ldots \quad (1)$$

$$X_i \sim X_1 + \ldots + X_{i-1} + X_{i+1} + \ldots \quad (2)$$

  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, 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.*

Examples of linearizing transformations

- Use $\log(Y)$ vs. $\log(X)$, i.e. apply logarithm to the response and the predictors.
  Suitable when $\mathbb{E}(Y) = \alpha X_1^{\beta_1} \dots X_p^{\beta_p}$.

- $\log(Y)$ vs. X, i.e. apply logarithm to the response only.
  Suitable when $\mathbb{E}(Y) = \alpha \exp \sum_j X_j \beta_j$.

- $1/Y$ vs. X, i.e. Take the inverse of the response.
  Suitable when $\mathbb{E}(Y) = \frac{1}{\alpha + \sum_j X_j \beta_j}$.

Box-Cox Transformations

Box-Cox Transformation of the Y Variable

- Box and Cox (1964) suggested a family of transformations (for *strictly* 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 $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y)$

Box-Cox Transformation of the Y Variable

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

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

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

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

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

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

The second term in these log-likelihood function corresponds to the Jacobian of the transformation.

Box-Cox Transformation of the Y Variable

- 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

Box-Cox Transformation of the Y Variable

– 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]:

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

[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: Bike Shares Example