STAT 420: Methods of Applied Statistics

Model Diagnostics — Normality

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign
March 7, 2017

Model Diagnostics

- We talked about how to fit linear models, and how to perform hypothesis testing problems.
- However, there are several key assumptions that we are relying on. It is important to check them when fitting a linear model.
- In the next few weeks, we will talk about checking those conditions.

Model Diagnostics

- Normal i.i.d. errors
- Constant error variance
- Absence of influential cases
- Linear relationship between predictors and outcome variable
- Collinearity

Normality of Residuals

- Throughout our previous derivations of the $\widehat{\beta}$ distribution, and hypothesis testings, we assumed that the residuals were normally distributed

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n})$$

- What if the data do not satisfy this assumption?
- Severe violations of the normally assumption may cause the confidence intervals to be too narrow or too wide.
- We need approaches to testing this assumption. A first step is creating graphs to evaluate potential deviations from normality such as boxplots or histograms.

Normality of Residuals

- Lets use the AT&T data (on our course website) as an example.
- A description of the data can be found at https://ww2.amstat.org/publications/jse/datasets/aptness.txt
- The aim is to model the number of work hours using the function points as a predictor.
- We fit a simple linear regression and investigate the residuals

Example: investigating residuals

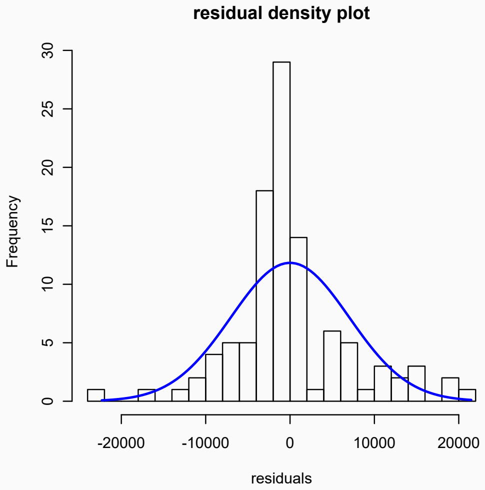
```r
ATT = read.table("ATT.txt", header = FALSE)
colnames(ATT) = c("FPC", "Work", "OS", "DMS", "Lang")
fit = lm(Work ~ FPC, data = ATT)
res = fit$residuals

# histogram and density plot

h = hist(res, main = "residual density plot", xlab = "residuals"
    , breaks = 20)
xgrid <- seq(min(res),max(res),length=100)
yden <- dnorm(xgrid,mean=0,sd=7047)
yden <- yden*length(res)*diff(h$mids[1:2])
lines(xgrid, yden, col="blue", lwd=2)
```

Example: investigating residuals

Is the distribution of the residual normal?



QQ plot

- The residual distribution looks more peaked, and deviates quite a lot from the matched (with mean and sd) normal density.
- There are many approaches for testing normality of a variable
- QQ plot graph e_i against a quantile that assumes e_i is normally distributed.
- Intuition: If a certain cut-off value corresponds to the $q \times 100\%$ th percentile of the normal distribution, then we would expect approximately $q \times n$ number of residuals fall below that cut-off.

QQ plot

- Formally, we identify "theoretical values" of the residual at each percentile, and compare that with the correspond observed residual values at the same percentile.
- We first get the percentile $f(e_i)$ corresponding to each e_i:

$$f(e_i) = \frac{\texttt{rank}(e_i) - 0.5}{n}$$

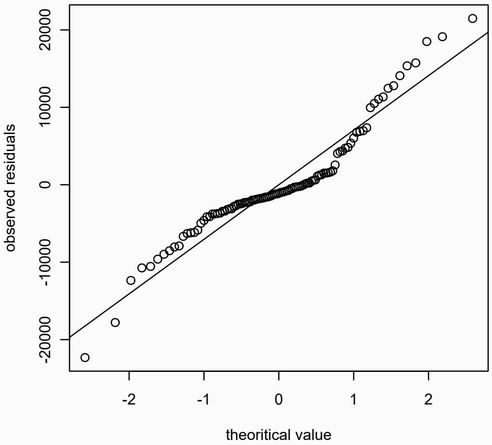
- Then we match this percentile to the standard normal distribution and get the "theoretical values":

$$\Phi^{-1}\big(f(e_i)\big)$$

  where Φ is the cdf function of the standard normal.
- Plot $\Phi^{-1}\big(f(e_i)\big)$ against the observed residual values e_i

QQ plot

$\Phi^{-1}\big(f(e_i)\big)$ against the observed residual values e_i



QQ plot

- In the ideal case, i.e., when e_i's actually come from a normal distribution, we would expect the dots to line up with the theoretical value
- Lets try a simulation study to see what is a “good looking” QQ plot
- In the previous plot, e_i deviates quite a lot from the theoretical line
- Note: how to generate the theoretical line is a bit tricky. There are different ways to do it, see our R code. The package car provides a nice function qqPlot
- Using graphs is nice and intuitive, but we should use more rigorous criteria

Test for normality

- We are going to introduce several tests.
  - Shapiro-Wilk
  - Kolmogorov-Smirnov
  - Anderson-Darling
  - Correlation test
- There is no “best” test theoretically, however, based on a 2011 paper by Razali and Wah, Shapiro-Wilk is the best test, while Anderson-Darling performs almost the same. Kolmogorov-Smirnov is more general, and can be applied to any distribution. Correlation test is conceptually simple.
- We will not derive these test statistics, but only focus on their intuitions.

Shapiro-Wilk test

- The test steatitic for the normality of a set of samples $\{x_1, \ldots, x_n\}$ is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

  where $x_{(i)}$ is the i-th order statistics, i.e., the i-th smallest number in the sample, and a_i's are constants derived from the distribution of the order statistics.
- The Null hypothesis is $\mathrm{H}_0 : x \sim \mathcal{N}(\mu, \sigma^2)$ vs. $\mathrm{H}_1 : \mathrm{H}_0$ is false.
- In R, the test can be performed by shapiro.test

Shapiro-Wilk test

- Example: Use the Shapiro-Wilk test on the gala data.
- Do we reject the normality test?

A : Yes B : No C : Maybe?

Kolmogorov-Smirnov test

- The Kolmogorov-Smirnov test compares the empirical distribution function F_n for a set of n samples $\{x_1, \ldots, x_n\}$ with its true distribution (normal), and calculate the largest discrepancy across the entire domain of x:

  $$D = \sup_t |F_n(t) - F(t)|$$

  where F is the cdf of a normal distribution.
- What is an empirical distribution?
- In R, the test can be performed by ks.test

Anderson-Darling test

- The Anderson-Darling test, instead of looking at the maximum discrepancy, uses the integrated square discrepancies:

$$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

  where F is again the cdf of a normal distribution.

- In R, the test can be performed by ad.test in the nortest package

Correlation test

- Intuition: if the residuals perfectly line up with the theoretical value, we would expect a perfect correlation between the theoretical value and the observed values.
- In this test, Looney & Gulledge (1985) propose to use the theoretical value

$$z_i = \Phi^{-1}\left(\frac{\mathtt{rank}(e_i) - 0.375}{n + 0.25}\right)$$

  where Φ is the cdf function of the standard normal. Note that the constants used here are not the ones we used in the QQ plot.
- Hence we are testing whether the values z_i's and the values e_i's have a perfect correlation or not.

$$\text{H}_0 : \text{Corr}(\mathbf{z}, \mathbf{e}) = 1 \quad \text{vs.} \quad \text{H}_1 : \text{Corr}(\mathbf{z}, \mathbf{e}) < 1$$

Correlation test

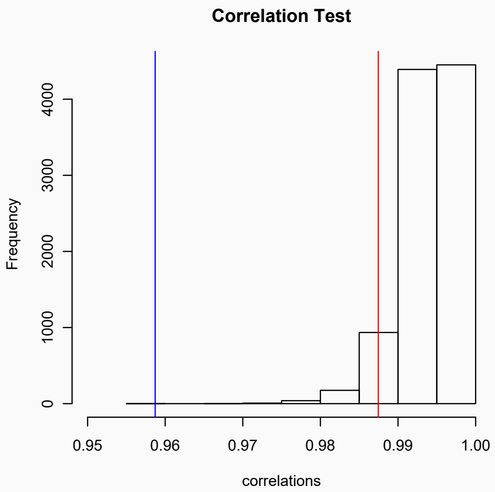
- Let's first estimate the correlation, and use a simulation study to approximate the p-value

```
> lg.test <- function(x) {
+   z <- qnorm((rank(x) - 0.375)/(length(x) + 0.25))
+   c(cor(x, z), length(x))
+ }
>
> lg.test(res)
[1]   0.9587105 104.0000000
```

Correlation test

```
> scores = rep(NA, 10000)
>
> for (i in 1:10000)
+ {
+   x = rnorm(length(res))
+   scores[i] = lg.test(x)[1]
+ }
>
> lgcrit = quantile(scores, prob = 0.05)
> hist(scores, xlim = c(0.95, 1))
> abline( v = lgcrit, col = "red")
> abline( v = lg.test(res)[1], col = "blue")
```

Correlation test



Solutions for Non-normal Residuals

- Use bootstrap resampling to perform inferences
- Use a rank transformation of the original data and use the approach proposed by

  Conover, W.J. & Iman, R.L. (1981). Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *The American Statistician*, 35, 124-129.
- Model the data with a different non-normal distribution
- Employ other nonparametric regression techniques