Predictive Analytics
Regression and Classification
Lecture 2 : Part 2

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2020

Check the Model Assumptions

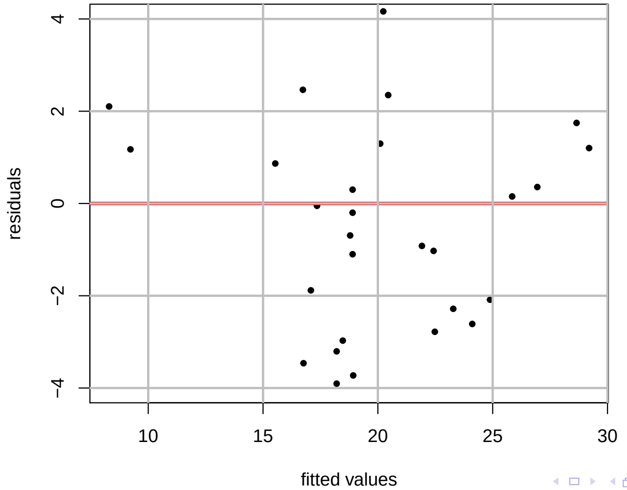
- Consider the regression model:
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$

- **Assumptions**:
  1. **Linearity**: Data is in linear hyper-plane.
  2. **Independence**: $\mathbb{C}ov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j = 1, 2, \cdots, n$ (randomness !!!)
  3. **Homoskedasticity**: $\mathbb{V}ar(\epsilon_i) = \sigma^2$ for all $i = 1, 2, \cdots, n$
  4. **Gaussian distribution**: ϵ follows Gaussian distribution

How to check Linearity

visualization : plot fitted vs residual

`corr between e & y-hat =  -1.584386e-16`



Test for Randomness to check independence

- $H_0 : \{\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_n\}$ are random numbers
  vs.
  $H_a : \{\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_n\}$ are not random.

  1. Bartels Rank Test (aka. Bartlet's Ratio Test) (1984)
  2. Mann-Kendall rank test of randomness. (1945)
  3. Wald-Wolfowitz Runs Test for randomness. (1940)

- In R, we have a package called `randtests` which implements several nonparametric randomness tests of hypothesis.

Homoskedasticity vs Heteroskedasticity

- When the variance for all residuals are equal, we call that **homoskedasticity**. $H_0 : \mathbb{V}ar(\epsilon_i) = \sigma^2, \quad i = 1, 2, \cdots, n$

- When the variance for residuals are different for at least one case, we call that **heteroskedasticity**.
$H_a : \mathbb{V}ar(\epsilon_i) \neq \sigma^2, \quad \text{at leat for one } i = 1, 2, \cdots, n$

- Equal variances across populations is called **homoscedasticity** or **homogeneity** of variances.

- If equal variances does not hold then known as **heteroskedasticity** or **heterogenity**.

Homoskedasticity

- How to check the **homoskedasticity**?
- Breusch-Pagan Test
- Bartlett's test
- Box's M test for homoskedasticity in multivariate data or equal covariance

Breusch-Pagan Test

- Consider general form of the variance function:

$$\mathbb{V}ar(y_i) = \mathbb{E}(e_i^2) = g(\gamma_1 + \gamma_2 z_2 + \cdots + \gamma_q z_q)$$

- To test:

  $H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_q = 0$

  H_a : At least one $\gamma_i \neq 0$

- Note that $z_2, z_3, \cdots, z_q$ could be same or different from $x_1, x_2, \cdots, x_p$

- The dependent variable e_i^2 are unobservable. Substitute with its least squares estimate $\hat{e}_i^2$

- $\hat{e}_i^2 = \gamma_1 + \gamma_2 z_2 + \cdots + \gamma_q z_q + \nu_i$

Breusch-Pagan Test

- Consider general form of the variance function:

$$\mathbb{V}ar(y_i) = \mathbb{E}(e_i^2) = g(\gamma_1 + \gamma_2 z_2 + \cdots + \gamma_q z_q)$$

- To test:

  $H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_q = 0$

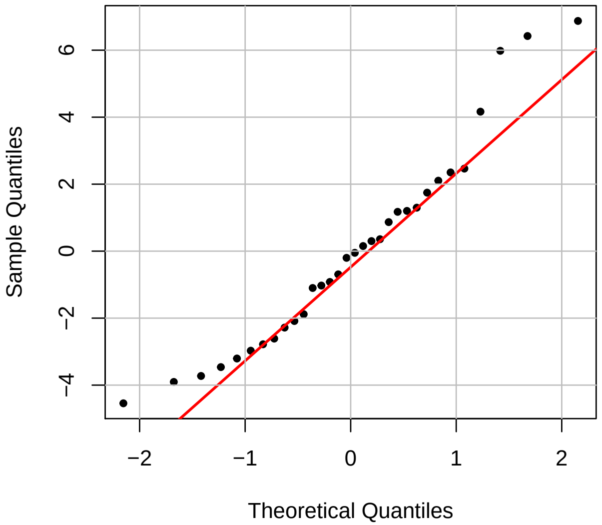
  H_a : At least one $\gamma_i \neq 0$

- Test Statistics under H_0: $\chi^2 = n \times R^2 \sim \chi^2_{q-1}$

- In `R`, the package `lmtest` contains the function `bptest`.

- In `Python`, in `statsmodels` module, you have `statsmodels.stats.diagnostic.het_breuschpagan`.

Check Normality with Q-Q Plot



Check Normality with Statistical Test

- **Kolmogorov-Smirnov test**:

$$H_0: \; e \sim N(0, \sigma^2) \quad \text{vs} \quad H_a: \; e \nsim N(0, \sigma^2)$$

Kolmogorov-Smirnov statistic is defined as

$$K_n = \sqrt{n} D_n = \sqrt{n} \sup_x |F_n(e) - \Phi_\sigma(e)|,$$

where

$$F_n(e) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{(-\infty, e)}(e_i),$$

$\mathbf{I}_{(-\infty, e)}(e_i)$ is the indicator function, equal to 1 if $e_i \leq e$ and equal to 0 otherwise.

- It rejects null hypothesis at lavel α if

$$K_n > K_\alpha,$$

where K_α is

$$\mathbb{P}(K_n \leq K_\alpha | \text{under } H_0) = 1 - \alpha$$

Check Normality with Statistical Test

- **Kolmogorov-Smirnov test for normality**:

$$H_0 : \; e \sim N(0, \sigma^2) \quad \text{vs} \quad H_a : \; e \nsim N(0, \sigma^2)$$

- In `R`, you can use '`ks.test`' from `stats` package to run the Kolmogorov-Smirnov test.

- In `Python`, you can use the '`scipy.stats.kstest`' to run the Kolmogorov-Smirnov test.

Check Normality with Statistical Test

- Kolmogorov-Smirnov test
- Anderson-Darling test
- Shapiro-Wilk test

Discussion

- What happened if any one of the model assumptions is not true?
  - You should not use that model further.
- What happened if all three assumptions of the model are good?
  - Great... you overcome one hurdle. Now check, how good is your prediction accuracy?

In the next part of this lecture...

- We will discuss, how do you compare the performance of two models and different choices of model selection criteria.