MTH499/599 Lecture Notes 05

Donghui Yan

Department of Math, Umass Dartmouth

Mar 04, 2015

Outline

- Overview of regression diagnosis
- Linear model and normality assumption
- Constant variance assumption
- Leverage and influence

Regression diagnosis

- Is the linear model appropriate (goodness-of-fit)
- Normality assumption
- Homoscedasticity (constant variance)
- Leverages of individual data points.

OLS output for the toy example

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2500 -0.6875 -0.0625  0.7812  1.2500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.3750     0.6847  19.535 1.17e-06 ***
x            -1.1250     0.1976  -5.692  0.00127 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9682 on 6 degrees of freedom
Multiple R-squared:  0.8437,	Adjusted R-squared:  0.8177 
F-statistic:  32.4 on 1 and 6 DF,  p-value: 0.001269
```

The R^2 statistic for goodness-of-fit

Recall that

$$R^2 = SSR/SST = SSR/(SSR + SSE)$$

- Amount of variance explained by the linear model
  - A natural statistic for testing goodness-of-fit

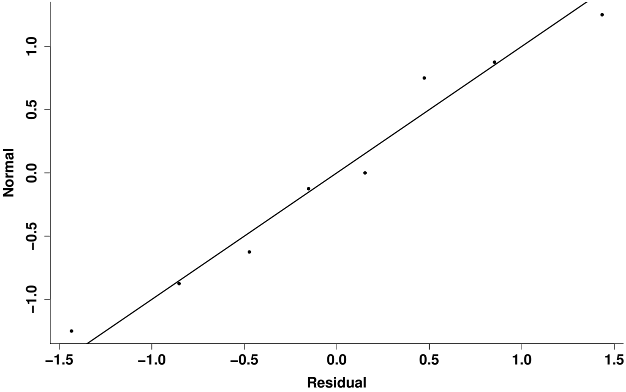
$$F = \frac{SSR/(p-1)}{SSE/(n-p)} \sim F_{(p-1),(n-p)}.$$

  - In the toy example, $n = 8, p = 2$, $SSE = 5.625, SSR = 30.375$

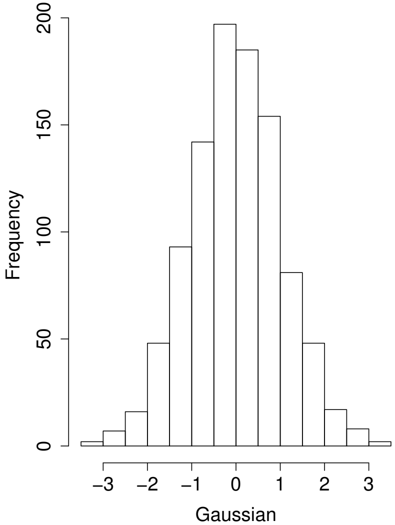
$$F = \frac{30.375/(2-1)}{5.625/(8-2)} = 32.4.$$

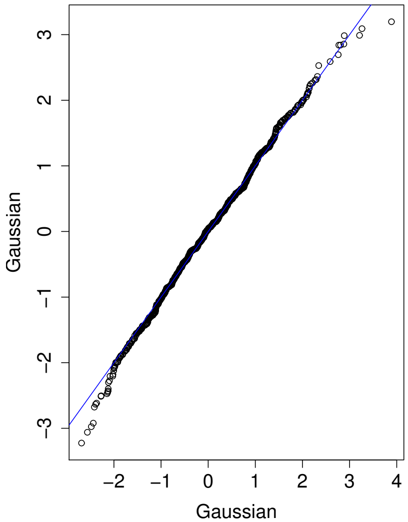
The normality assumption

- Visual inspection by a normal Q-Q plot
  - Q-Q plot close to the $y = x$ line

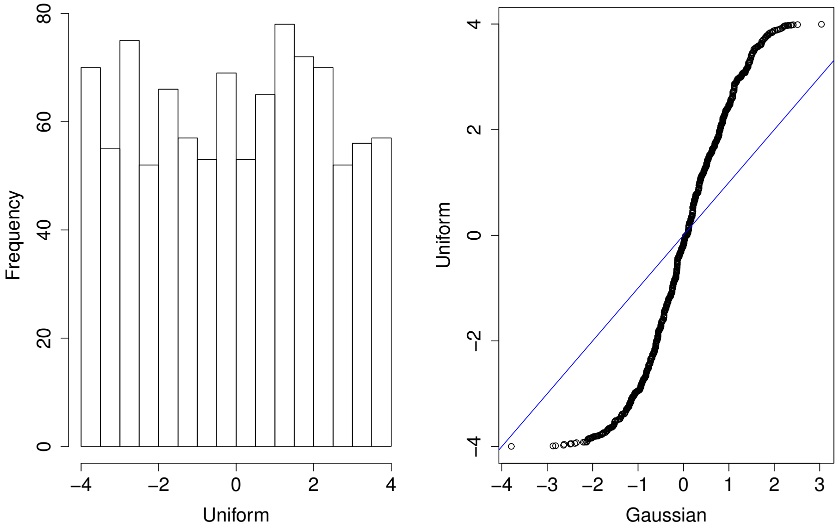


Q-Q plot of normal Vs normal



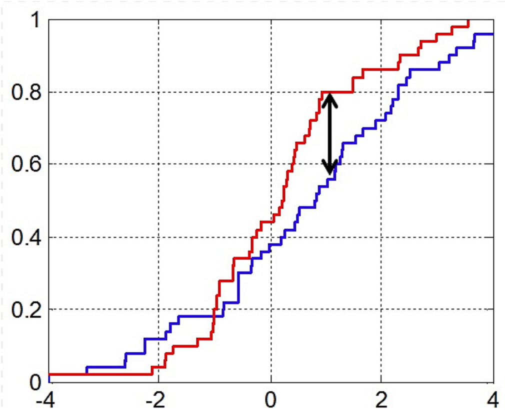


Q-Q plot of normal Vs Uniform



The Kolmogorov-Smirnov test of normality

- Testing statistic $D = \sup_x \mid F_m(x) - G_n(x) \mid$
- *ks.test(data1, data2)*



The Kolmogorov-Smirnov test of normality (continued)

- *ks.test(x, rnorm(100,0,1))*

```
Two-sample Kolmogorov-Smirnov test

data:  mylm$residuals and datNorm
D = 0.125, p-value = 1
alternative hypothesis: two-sided
```

- *ks.test(x, rnorm(100,0,1), runif(100,-4,4))*

```
Two-sample Kolmogorov-Smirnov test

data:  runif(100, -4, 4) and rnorm(100, 0, 1)
D = 0.34, p-value = 1.908e-05
alternative hypothesis: two-sided
```

The Shapiro-Wilk test of normality

- The testing statistic given by

$$W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x_i - \overline{x})^2}$$

  - $x_{(i)}$'s are order statistics
  - a_i expected value of order statistics of data from normal distribution, normalized by covariance matrix
  - Roughly

W as 'correlation' of order statistics by the data and by $\mathcal{N}$.

The Shapiro-Wilk test of normality (continued)

- *shapiro.test(data)*

```
Shapiro-Wilk normality test

data:  mylm$residuals
W = 0.9519, p-value = 0.7308
```

- *shapiro.test(runif(100,-4,4))*

```
Shapiro-Wilk normality test

data:  runif(100, -4, 4)
W = 0.9354, p-value = 0.0001022
```

Quiz

- Given a sample $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_n, Y_n)$, and linear model $Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$.
  1). What is the OLS estimate for $\boldsymbol{\beta}$?
  2). What is the hat matrix?
  3). What can you say about the OLS estimate of $\boldsymbol{\beta}$?