**Homework 2**

1. (10% credit) What does the low p-value of the regression coefficient tells us while considering a regression of Y vs X (select all that applies):

1. that relation between Y and X is likely to be linear
2. that Y is likely to be independent of X
3. that a hypothesis of a regression coefficient being zero (X has no linear effect on Y) should be rejected
4. that a fraction of variation of Y that could be explained through linear regression on X is low
5. that a fraction of variation of Y that could be explained through linear regression on X is high

Answer: c)

2. (10% credit) What is the meaning of R2 in a linear regression of Y vs X (select all that applies):

a) degree of linearity of dependence of Y on X

b) Percentage of response variable’s (Y) variation explained by a considered linear model

c) statistical significance of the hypothesis of linear independence of Y vs X

Answer: b)

3. (20% credit). Using attached data Compute correlation coefficient of Y and X. Visualize a scatter plot.

Answer:

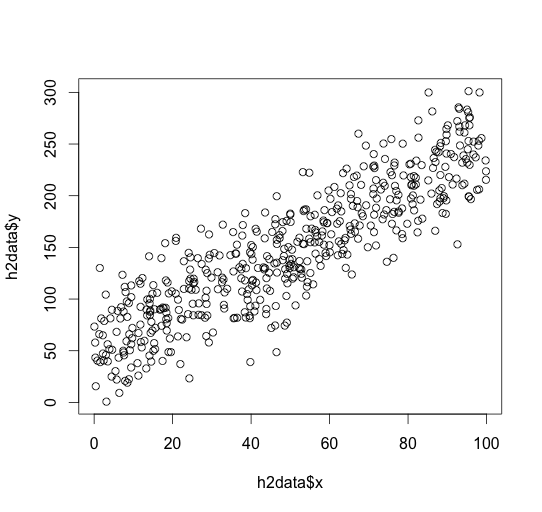
0.8804123

par(mfrow=c(1,1))

h2data<-read.csv('/Users/stansobolevsky/Desktop/NYU/lectures/L2/codes/data/h2data.csv')

plot(h2data$x,h2data$y)

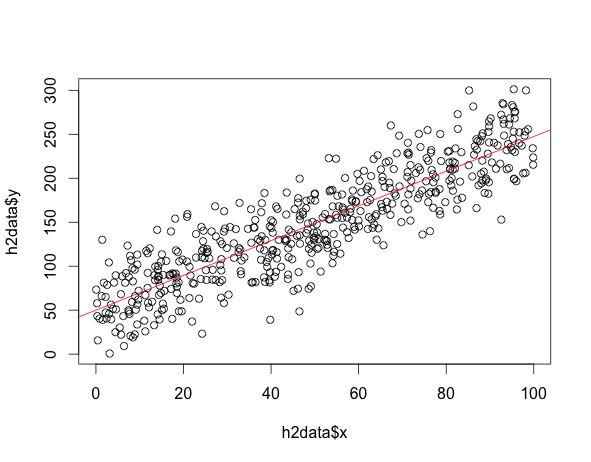
print(cor(h2data$y,h2data$x))



4. (35% credit). For the data from the problem 3 perform and visualize a linear regression of Y vs X. Print regression statistics (R2, p-value of the X-coefficient, confidence intervals of the both coefficients).

Answer:

Y=1.975795\*X+50.02628



lmfit = lm( h2data$y ~ h2data$x )

abline(lmfit, col='red')

lmfit$coefficients

summary(lmfit)

summary(lmfit)$r.squared

summary(lmfit)$coefficients[2,4]

(Intercept) h2data$x

50.026280 1.975795

> summary(lmfit)$r.squared

[1] 0.7751258

> summary(lmfit)$coefficients[2,4]

[1] 1.742289e-163

> summary(lmfit)

Call:

lm(formula = h2data$y ~ h2data$x)

Residuals:

Min 1Q Median 3Q Max

-93.238 -19.718 -1.129 21.770 81.443

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 50.02628 2.71110 18.45 <2e-16 \*\*\*

h2data$x 1.97579 0.04769 41.43 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.34 on 498 degrees of freedom

Multiple R-squared: 0.7751, Adjusted R-squared: 0.7747

F-statistic: 1717 on 1 and 498 DF, p-value: < 2.2e-16

5 (25% credit). For the regression from the problem 4, visualize the histogram and cumulative distribution of errors. Estimate parameters of this empirical distribution and graphically compare it with the normal distribution. Does this distribution seem to be consistent with the classical regression assumptions?

mu=mean(lmfit$residuals)

sigma=sd(lmfit$residuals)

print(mu)

print(sigma)

par(mfrow=c(2,1))

hist(lmfit$residuals, freq=FALSE)

x<-seq(mu-2\*sigma,mu+2\*sigma,sigma/100) #scale of x

y<-dnorm(x, mean=mu, sd=sigma, log=FALSE) #produce pdf of the normal distribution with the same parameters

lines(x,y,col='red')

plot(ecdf(lmfit$residuals))

y<-pnorm(x, mean=mu, sd=sigma, log=FALSE) #produce pdf of the normal distribution with the same parameters

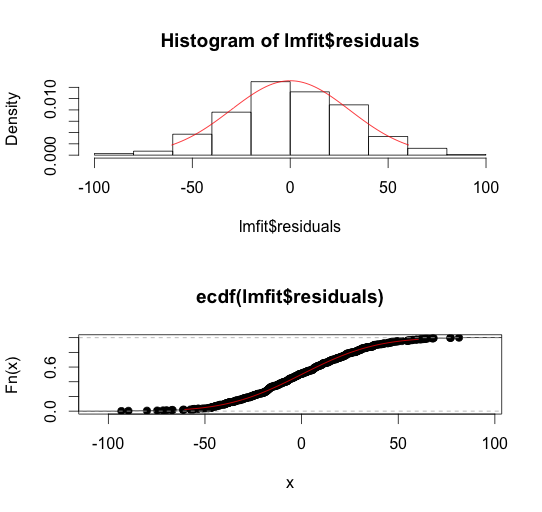
lines(x,y,col='red')

> print(mu)

[1] -1.6257e-15

> print(sigma)

[1] 30.31319



Yes, distribution looks close to normal with mean close to 0, which is consistent with the classical regression assumptions.

**Extra credit assignment**

(30% of additional credit to be applied for this or further homeworks):

For the real estate prices of individual houses (data from the lab also attached to the homework together with latitude-longitude coordinates of the US zip codes), estimate and visualize (as a scatter plot map, where values for each zip codes are shown by color – the higher the brighter) correlation coefficients and power-law scaling exponents of house price vs gross square footage for all the zip codes, having over 20 individual houses for sale.