UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Revision - week 2

1. See text (not an exercise).

2. For the br.csv dataset, we regress log-price on log-sqft and log-age (with an intercept). There are $N = 1080$ observations. Try with `regr = lm(log(price) log(sqft)+log(Age)`. This enables to interpret the estimated coefficients as elasticities (logarithmic derivatives). Illustrate the main estimation results.

```
> br = read.table("br.csv", sep = ",", header=T) # reads the data from a csv file
> attach(br)
> summary(lm(log(price)~log(sqft)+log(Age)))

Call:
lm(formula = log(price) ~ log(sqft) + log(Age))

Residuals:
     Min       1Q   Median       3Q      Max
-1.30120 -0.17267 -0.01204  0.18235  1.29271

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.684008   0.185793   25.21   <2e-16 ***
log(sqft)    0.952429   0.023686   40.21   <2e-16 ***
log(Age)    -0.082439   0.007066  -11.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3101 on 1077 degrees of freedom
Multiple R-squared:  0.6512,Adjusted R-squared:  0.6506
F-statistic:  1005 on 2 and 1077 DF,  p-value: < 2.2e-16
```

   The elasticity of price to sqft is estimated equal to 0.95. A 10% increase in the house dimension is expected to yield a 9.5% increase in house price. The estimated coefficients are significantly different from zero. Both variables seem to have a good explanatory power.

3. Knowing that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.3591 & -0.0455 & -0.0038 \\ -0.0455 & 0.0058 & 0.0003 \\ -0.0038 & 0.0003 & 0.0005 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} 4.6840 \\ 0.9524 \\ -0.0824 \end{bmatrix}, \sum_i e_i^2 = \mathbf{e}'\mathbf{e} = 103.5403,$$

   and that $N = 1080$, compute the residual standard error $\hat{\sigma}$ (see the slides for formula). Is it possible to compute the t-value for the coefficient β_1 from the information provided above?

   The t-value for $H_0 : \beta_2 = 0$ is -11.7 and the corresponding p-value is virtually 0; are you willing to accept the null hypothesis?

   Knowing further that the total sum of squares is $TSS = 296.8724$, what is the value of R^2? Do you think it is satisfactory?

Solution We first need to estimate σ^2:

$$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{N-p-1} = \frac{103.5403}{1080-2-1} = 0.0961.$$

Then, the $t-$value is

$$t_1 = \frac{0.9524}{\sqrt{0.0961 \cdot 0.0058}} = 40.33281$$

(this is an approximation to the value obtained from the regression output in exercise 2. The estimated standard error of $\hat{\beta}_1$ is obtained as the square root of $\hat{\sigma}^2$ times the element in position (2,2) of the matrix $\mathbf{X}'\mathbf{X}$.

It is reasonable to reject the null $H_0 : \beta_2 = 0$.

Finally, we know that $RSS = 103.5403$ and thus $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{103.5403}{296.8724} = 0.65$.

4. Regression without intercept.

$$\hat{y}_i = \frac{x_1 y_1 + x_2 y_2 + \ldots + x_i y_i + \ldots x_n y_n}{x_1^2 + x_2^2 + \ldots + x_i^2 + \ldots x_n^2} x_i.$$

Hence, $\hat{y}_i = \sum_{j=1}^n a_j y_j$, with

$$a_j = \frac{x_i x_j}{\sum_{i=1}^n x_i^2}$$

(see the discussion about leverage in point 1).

5.
```
set.seed(1)
x   = rnorm(100)
eps = rnorm(100,0,0.5)
beta0 = -1
beta1 = 0.5
y   = beta0+beta1*x+eps
plot(x,y)
#############################################
mod = lm(y~ x)
summary(mod)
plot(x, y)
abline(mod, col = 'red')
abline(beta0,beta1, col = 'blue', lty = 3)
legend(1,-1.5, c("fitted", "true"),
      col=c("red","blue"), lty =c(1,3))
mod2 = lm(y~ x+I(x^2))
summary(mod2)
```

6. Will be discussed in the classroom.