UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

Revision - week 2

1.

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. (This is not an excercise!)
   The hat matrix etc.

5. Let $\mathbf{x}_1$ denote a vector of $N = 10$ temperatures in degrees Celsius, generated as $x_1 \sim \mathrm{N}(20, 3)$ and let $\mathbf{y}$ be the corresponding consumption of beer (in cans), that is linearly related to temperature. Here, $\mathrm{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

   We create a new variable, $\mathbf{x}_2 = 32 + 1.8\mathbf{x}_1$, which represents temperatures converted to degrees Fahrenheit.

   We regress Y on X_1 and X_2. However, as we can see from the code below, something went wrong. What, in particular?

```
x1 = rnorm(10, 20, 3)
y =  round(20 * x1 + rnorm(10,0,15))
plot(x1,y)
x2 = 32+1.8 * x1
summary(lm(y ~ x1+x2))
cor(x1,x2)
```

   (To execute the above code, open RStudio; from the menu File select New $\rightarrow$ R Script. Copy and paste the code, select and Run. Make sure that $\sim$ is copied correcly).

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5526  -5.0158  -0.6911   8.0913   9.5560

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.118     20.851   1.444    0.187
x1            18.837      1.063  17.727 1.05e-07 ***
x2                NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 8.975 on 8 degrees of freedom
Multiple R-squared:  0.9752,Adjusted R-squared:  0.9721
F-statistic: 314.2 on 1 and 8 DF,  p-value: 1.049e-07
```

The coefficient for the second variable is not estimable as the two regressors are perfectly collinear.