UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

Revision - week 2

1. Revision: read section 2.1, pages 15–23, of the textbook "An Introduction to Statistical Learning", by James et al. Read section 2.3.4. about reading data in R; section 2.3.5. about plots in R. For linear regression, study sections 3.1–3.2. Subsections 3.6.2–3.6.6 provide illustrations in R.

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.

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?

4. (This is not an excercise!)
   The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ plays an important role in regression analysis. The fitted values are a linear combination of the observed values, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, so that for the i-th unit

$$\hat{y}_i = h_{i1}y_1 + \cdots h_{ii}y_i + \cdots + h_{iN}y_N.$$

   The diagonal element, $h_i = h_{ii}$, is the weight that the i-th observation receives in forming the fitted value: $h_i = \frac{\partial \hat{y}_i}{\partial y_i}$, and in terms of the observations $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. It measures the leverage effect of the i-th observation, which depends on the remoteness of the i-th observation from the others in the space of the X's (think of $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ as a distance).

   The $h_i, i = 1, \ldots, N$ have the following properties:

$$\frac{1}{N} \leq h_i \leq 1, \sum_{i=1}^{N} h_i = p + 1,$$

   so that the mean is $\frac{1}{N}\sum_i h_i = \frac{p+1}{N}$. A large leverage implies that a particular observation is influential for the fit: often, values larger than twice the mean $(2(p+1)/N)$ are flagged. An index plot can be used to visualise leverage (plot h_i vs i).

   In the script `br_LinearRegression.R`, the h_i's are retrieved by the function `hatvalues()` (line 63), which applies to the output created by the function `lm()`.

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 → R Script. Copy and paste the code, select and Run. Make sure that ∼ is copied correcly).