UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

1 Revision Week 2

This week we have covered the linear regression model. Study Chapter 3 of your textbook. Sections 3.1–3.2 deal with specification, estimation and testing. Sections 3.3.1-3.3.2 deals with qualitative input variables, polynomial terms and interactions. Subsection 3.3.5 deals with leverage. Subsection 3.3.6 deals with multicollinearity. Subsections 3.6.2–3.6.6 provide illustrations in R.

- Replication of the illustrations of linear regression using the Advertising data presented in Ch. 3 of the textbook are available in Lab1.

- (*This is not an exercise, but a discussion of leverage*)
  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: $h_i = \frac{\partial \hat{y}_i}{\partial y_i}$.

  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).

  It can be shown that for $p = 1$ (simple linear regression), since $\mathbf{x}_i' = [1, x_i]$ and
  $$\mathbf{X}'\mathbf{X} = \left[\begin{array}{cc} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{array}\right],$$

it follows that

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}.$$

We see from this expression that h_i measures how far away is x_i from the average of the xs, relative to the deviance of x. In the script `br_LinearRegression.R`, the h_i's are retrieved by the function `hatvalues()`, which applies to the output created by the function `lm()`.

The h_i's contribute to the variance of the OLS residuals and fitted values as $\text{Var}(e_i|\mathbf{X}) = \sigma^2(1 - h_i)$ and $\text{Var}(\hat{y}_i|\mathbf{X}) = \sigma^2 h_i$.

2 Exercises

1. 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.

2. 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?

3. Solve Exercise n. 5, page 122 of the textbook.

4. Solve Exercise n. 13, points (a)-(g) page 126.

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

6. *Exam preparation: This question was worth 10 (2+2+2+2+2) points out of 70 in the last written exam.*
The following table summarizes the results of estimating a linear regression model for total sales (logarithms) from a training sample of $N = 300$ observations:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.00315    0.27603  28.993   <2e-16
log(nfull)  -0.06970    0.06286  -1.109   0.2684
log(nown)   -0.17284    0.08320  -2.077   0.0386
log(npart)   0.12427    0.07764   ****    0.1105
log(hoursw)  1.16220    0.06277  18.515   <2e-16
---
Residual standard error: 0.4514 on 295 degrees of freedom
Multiple R-squared:  0.589,Adjusted R-squared:  0.5834
```

(a) What is the interpretation of the p-value 0.2684 associated to the explanatory variable `log(nfull)` (log of number of full-time workers)?

(b) Is `log(nfull)` significant at the 10% level?

(c) Why is there a difference between `Multiple R-squared` and `Adjusted R-squared`? Describe the nature of the adjustment.

(d) Obtain the missing `t value` for `log(npart)`.

(e) Construct an approximate 95% confidence interval for the coefficient β_4 of the variable `log(hoursw)`, assuming $\hat{\beta}_4 - \beta_4 \sim \mathrm{N}(0, 0.06277^2)$ and recalling that for $Z \sim \mathrm{N}(0,1)$, $P(-1.96 < Z < 1.96) = 0.95$.