Multivariate linear regression

Statistical modelling: theory and practice

Gilles Guillot

gigu@dtu.dk

September 25, 2013

Example: what explains the price paid in Italian restaurants in NYC?

Data available as text file nyc.csv on Sheather's book web site.

- EDA and simple model fit with R...
- None of the simple regression capture the whole pattern
- A model along the line of price = food + service + quality + residual would be more accurate

The multiple regression model I

Setting:

- $Y = (y_1, ..., y_n)^t$ response variable observed on n individuals
- $X_1 = x_{i1} = (x_{11}, ..., x_{n1})^t$ first explanatory variable
- $X_2 = x_{i1} = (x_{12}, ..., x_{n2})^t$ second explanatory variable

  ...
- $X_p = x_{i1} = (x_{1p}, ..., x_{np})^t$ p-th explanatory variable

All variables are quantitative, preferably continuous

The multiple regression model II

Model assumptions:

- The explanatory variables x_{ij}'s are observed and non random
- The response variables y_i's are observed and random
- y_i relates linearly to the x_{ij}'s:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$$

- The coefficients $\beta_0, ...\beta_p$ are deterministic (and unknown)
- The error terms are independent and indentically distributed (iid) as a $\mathcal{N}(0, \sigma_e^2)$
- The error variance σ_e^2 is deterministic (and unknown)

The multiple regression model III

Multiple regression equation in matrix form:

Denoting

- $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$
- $\boldsymbol{X} = (\mathbf{1}, X_1, ..., X_p)$
- $E = (\varepsilon_1, ..., \varepsilon_n)^t$

We have:

$$\begin{array}{ccc} Y & = & \boldsymbol{X}\boldsymbol{\beta} + E \\ (n \times 1) & & (n \times p + 1)(p + 1 \times 1)(n \times 1) \end{array}$$

In the above, $\boldsymbol{X}$ is known as the *design* matrix.

Stems from the situations where the scientist choses the x_{ij} values in view of maximising the amount of information obtained for a given cost (design of experiment or DoE).

Least square estimation I

What parameters?
$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$ and σ_e^2

Least square principle

We seek $\boldsymbol{\beta}$ that minimizes globally the squared errors

$$[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})]^2$$

Our estimate is

$$\hat{\boldsymbol{\beta}} = Argmax_{\boldsymbol{\beta}} \sum_i [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})]^2$$

Least square estimation II

Geometric interpretation:

- $\beta_1 \mathbf{1} + \beta_1 X_1 + ... + \beta_p X_p$ is an element of $Span(\mathbf{1}, X_1, ..., X_p)$
- $\sum_i [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})]^2 = ||Y - \boldsymbol{X}\boldsymbol{\beta}||^2$
- $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of Y on $Span(\mathbf{1}, X_1, ..., X_p)$

Deriving $\hat{\boldsymbol{\beta}}$ explictly: the normal equations

It can be shown that $\hat{\boldsymbol{\beta}}$ satisfies:

$$(\boldsymbol{X}^t \boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t y$$

which is known as the *normal equations*.

Least square estimation III

The LSE estimate of $\boldsymbol{\beta}$

If $(\boldsymbol{X}^t\boldsymbol{X})$ if of full rank, the normal equations give:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t y$$

$\hat{\boldsymbol{\beta}}$ defined above is an *unbiased* estimate of β.

Note: the LSE estimate exists if the columns of X are linearly independent.

Least square estimation IV

LSE estimate of the residual variance σ^2:

$$S^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\sum_i (\hat{y}_i - y_i)^2$$

is an unbiased estimate of σ^2.

Model fitting in R

R function `lm`:

- `lm(formula = Price  1 + Food + Service + Decor,data=nyc)`
- `lm(Price  Food + Service + Decor,data=nyc) # includes a constant term by default`
- `lm(formula = Price  0 + Food + Service + Decor,data=nyc) # no constant term`
- `lm(formula = Price ~ .,data=nyc)`
  `# include all variables in the rhs from data.frame nyc`
- `lm(formula = Price ~ .  - Decor ,data=nyc) # drop variable x2`

By default, the `lm` function does much more than just parameter estimation.

Example of output

Printing the ouput of `lm`

```
> res = lm(formula = Price ~ 1 + Food + Service + Decor,data=ny
> res

Call:
lm(formula = Price ~ 1 + Food + Service + Decor, data = nyc)

Coefficients:
(Intercept)         Food      Service        Decor
    -24.641        1.556        0.135        1.847
```

Example of output, cont'

Summary of output of `lm`:

```
> summary(res)

Call:
lm(formula = Price ~ 1 + Food + Service + Decor, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8440  -3.7039  -0.1525   3.6218  19.0576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Service       0.1350     0.3957   0.341    0.733
Decor         1.8473     0.2176   8.491 1.17e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared: 0.617, Adjusted R-squared:  0.61
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

Testing global significance I

We are interested in testing whether the response y is pure noise and does not relate linearly to any of the putative explanatory variables.
This is a global test of significance and answers to the question "does the linear equation explain anything at all?"

Null hypothesis in global significance testing

The null hypthesis is $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ i.e $H_0 : Y = \beta_0 + E$

Testing global significance II

We define

- $SST = \sum_i (y_i - \bar{y})^2$
- $SSM = \sum_i (\hat{y}_i - \bar{y})^2$
- $SSE = \sum_i (\hat{y}_i - y_i)^2$

If one of the parameters β_j is $\neq 0$, the model should "do good" with $SSM \approx SST$ and $SSE \approx 0$.

Test statistic in global significance testing

Under H_0, $F = \dfrac{SSM/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$.

A large value of F suggests that the model "explains something" and leads to reject H_0.

Testing global significance III

In R, the p-value for the global F test can be found at the last line of `summary(lm(...))`.

Testing individual coefficients I

If H_0 is rejected in the global test above, we want to identify wich coefficient β_j is non zero.
Mor generally, we can be interested in testing $H_0 : \beta_j = \beta_j^0$ for some specific β_j^0 values.

Test for β_j

Under $H_0 : \beta_j = \beta_j^0$,

$$T_j = \frac{\beta_j - \beta_j^0}{\widehat{sd(\hat{\beta}_j)}} \sim St_{n-p-1}$$

Testing individual coefficients II

The previous test of $\beta_j = 0$ is performed under the assumption that other parameters are non zero. It measures the improvement brought by X_j in a model containing the other variables.

The result of the test of $\beta_j = 0$ depends on which other variables are included in the model.

Testing individual coefficients, cont'

```
Call:
lm(formula = Price ~ 1 + Food + Service + Decor, data = nyc)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Service       0.1350     0.3957   0.341    0.733
Decor         1.8473     0.2176   8.491 1.17e-14 ***

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared: 0.617, Adjusted R-squared:  0.61
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16

##############
Call:
lm(formula = Price ~ 1 + Food + Service, data = nyc)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.1586     5.6651  -3.735 0.000258 ***
Food          1.4954     0.4462   3.351 0.000997 ***
Service       1.7041     0.4185   4.072 7.22e-05 ***

Residual standard error: 6.942 on 165 degrees of freedom
Multiple R-squared: 0.4486, Adjusted R-squared: 0.4419
F-statistic: 67.12 on 2 and 165 DF,  p-value: < 2.2e-16
```

Model assessment and variable selection I

Coefficient of determination R^2

Same definition as for simple linear regression:

$$R^2 = 1 - \frac{SSE}{SST}$$

An important word of caution:

$$R^2_{X_1,\ldots,X_p} \leq R^2_{X_1,\ldots,X_p,X_{p+1}}$$

Including an extra variable will always increase *automatically* R^2.

Model assessment and variable selection II

Adjusted coefficient of determination R^2_{adj}:

A coefficient that includes a penalty term in p

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

R^2_{adj} makes it possible (in principle) to compare models of different dimensions.

Other criteria for model selection include Akaike Information Criterion (AIC) and Bayeisan Information Criterion (BIC). To be defined after later in this course.

Suggested reading

- Sheather, chapter 5. Multiple linear regression.