Statistics

Week 9: Regression (Chapter 10)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

Guest lectures:

Tuesday 21 March, 2–3pm, TT21 (lecture time).

Thursday 23 March, 1–2pm, TT21 (recitation time).

Homework **assignment 2** solutions are available on *eDimension*.

- Q5: it is not convincing to add up the powers of two one-sided tests; refer to the solution.
- Q6 (if using a CI): the CI is *for* σ, *using* s; then checking if σ_0 lies inside it. It is NOT using σ_0, then checking if s lies inside it.
- Q8: there is no way to reduce the problem to testing the variance of *one* population, since we do not know σ_1 or σ_2.

Outline

1. (Simple) linear regression
   - SSE, SST, SSR
2. Multiple (linear) regression

Introduction

Question: how can we construct a line of 'best fit' through some data points?

Set up: given n fixed x-coordinates x_i, and n corresponding y-coordinates y_i. A **regression line** is a linear model that describes their relationship.

x is called the predictor/explanatory/independent variable;
y is called the response/ outcome /dependent variable.

We should first make a scatter plot from (x_i, y_i) to check if we have a linear relationship, and if there are outliers.

If a true regression line exists, given by $y = \beta_0 + \beta_1 x$, then we estimate β_0 and β_1 using the **least square** method.

This method is used partly due to mathematical convenience. We do not explore other methods here.

Probabilistic set up

To explain why the data values (x_i, y_i) do not lie perfectly on a straight line, we can think of y_i as the observed value of a random variable Y_i, where

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

and ϵ_i is the random error arising from measurement, variables other than x, etc.

It is common to assume that ϵ_i's are iid *normal* with mean 0 and variance σ^2. This assumption will be useful later when we construct confidence intervals.

An optimization problem

To minimize

$$Q = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2,$$

we compute $\dfrac{\partial Q}{\partial \beta_0}$ and $\dfrac{\partial Q}{\partial \beta_1}$ and set them both to 0.

Denoting the solutions of these equations by $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

These two equations can be routinely solved.

Solution

Notation: let s_x, s_y be the sample standard deviations, let s_{xy} be the sample covariance (`covariance.s` in *Excel*),

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Then we can write the solutions as:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

The *least square line* is denoted by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, and is an estimate of the true regression line $y = \beta_0 + \beta_1 x$.

The fitted values are given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$;
the *residuals* are $e_i := y_i - \hat{y}_i$.

Exercise

In the spreadsheet, find the least square line for the triple jump example using the formulas, and check it against *Excel's* trendline.

Some important terms

The **sum of squared errors (SSE)** is defined to be $\sum_i e_i^2$.

The **sum of squares (total) (SST)** is $\sum_i (y_i - \bar{y})^2 = (n-1)s_y^2$.

It can be shown that

$$\begin{aligned} \textbf{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \qquad \textbf{SSE} \qquad + \qquad \textbf{SSR}. \end{aligned}$$

SSR stands for **sum of squares due to regression**.

Coefficient of determination: $r^2 = \text{SSR}/\text{SST} = 1 - \text{SSE}/\text{SST}$.

An unbiased estimator for σ^2 (of Y_i) is $s^2 = \text{SSE}/(n-2)$, also known as the **mean squared error (MSE)**.

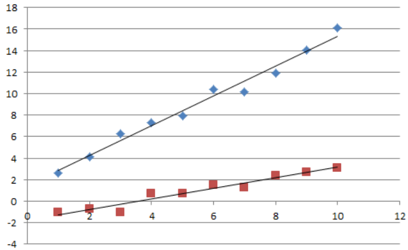
r and r^2

$r^2 \in [0, 1]$ can be interpreted as how much of the variation in y can be accounted for by the regression model.

The **correlation coefficient**, $r \in [-1, 1]$, is given by

$$r = \frac{s_{xy}}{s_x \, s_y}.$$

Its sign corresponds to the slope of the least square line. $r = \pm 1$ if and only if there is a perfect fit; $r = 0$ means no correlation.



Which least square line has larger *r* (or are they about the same)?

Residuals

Exercise

Compute r^2 for the triple jump example using the formula.

Can the least square line be used to predict the future?

A plot of the *residuals* e_i can be used to check the linearity assumption. For example, a plot which is parabolic in shape indicates the need for an x^2 term.

Rule of thumb: if $|e_i| > 2s$, then the corresponding value may be an outlier.

Example

Investigate the residual plot for the life expectancy data, using `Data Analysis` $\rightarrow$ `Regression`.

Data transformation

If there is a non-linear relationship between x and y, sometimes linear regression can still be used after appropriately transforming the data.

For example, if we suspect $y = \alpha\, x^{\beta}$, then take log of both sides. *Excel* uses `ln` for natural log.

Exercises

(1) What to do if we suspect $y = \alpha\, x^2 + \beta$? $\quad y = \alpha\, e^{\beta x}$?

(2) Interpolate a value in the spreadsheet.

Outline

Multiple regression, matrix form

When there are k independent variables, we can construct a least square regression model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

for the data values $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$.

Geometrically, this can be a curve, a surface, etc.

Set things up using matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

Question: what are $\mathbf{X}$ and $\mathbf{y}$ for the life expectancy example?

Solution

We need to minimize

$$\|\mathbf{y}-\mathbf{X}\beta\|^2 = (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) = \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta.$$

This can be done by setting the *gradient* to $\mathbf{0}$, i. e. differentiate the right hand side with respect to each of the β_i's, store the results as a column vector, then set it to the 0 vector.

After manipulation, and using the fact that $\mathbf{X}^T\mathbf{X}$ is symmetric, the result can be simplified to $-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = \mathbf{0}$.

Denoting the solution by $\hat{\beta}$, we obtain

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

A more conceptual proof of this formula can be found in the Math2 Cohort 12 and Cohort 15 slides on projections.

Some properties of multiple regression

- SSE, SST and SSR are defined the same way.
- The formula SSR + SSE = SST still holds.
- $r^2 :=$ SSR/SST.
- r is now the non-negative square root of r^2.
- For polynomial regression, just set the other independent variables as powers of the first one.
- Data transformation works the same way, e. g. for the model $y = \beta_0 \, x_1^{\beta_1} \, x_2^{\beta_2}$, take log of both sides.

Multicollinearity

Beware if some of the independent variables are almost or exactly *linearly dependent*, e. g. income, saving and expenditure. This is sometimes manifested by high correlation between the variables.

If some columns of the matrix $\mathbf{X}$ are linearly dependent, then there exists a non-zero vector $\boldsymbol{v}$ such that $\mathbf{X}\boldsymbol{v} = \mathbf{0}$, so $(\mathbf{X}^T\mathbf{X})\boldsymbol{v} = \mathbf{0}$.

This means $\mathbf{X}^T\mathbf{X}$ is not invertible, making $\hat{\beta}$ impossible to compute. Likewise, if the columns are nearly linearly dependent, then $\mathbf{X}^T\mathbf{X}$ is nearly singular, which causes numerical problems.

Solution: remove a variable that is linearly dependent on the others.