Statistics

Week 10: Regression (Chapter 10 & 11)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

Homework assignment 3 will be available on Tuesday. You can submit a hardcopy into the homework box near the entrance of the ESD offices, or submit a softcopy online.

This and next Tuesday second half: guest lectures.

This Thursday: normal and project recitations.

Outline

Multiple regression

Exercise

In the spreadsheet *regression2 – companies*, use the Data Analysis package to fit a linear model for y in terms of x_1 and x_2.

The regression model can be represented by a plane.

Dummy variables

Sometimes the data contains *categorical* variables, such as gender or seasons. We can encode them using 0's and 1's.

There are different methods of encoding. We demonstrate one method here, using the *Excel* data for triple jump distance vs year and gender.

We set gender = 0 for male and 1 for female, and use the model

$$\begin{aligned} \text{distance} &= (\beta_0 + \beta_1 \,\text{gender}) + (\beta_2 + \beta_3 \,\text{gender})\,\text{year} \\ &= \beta_0 + \beta_1 \,\text{gender} + \beta_2 \,\text{year} + \beta_3 \,\text{year} \times \text{gender}. \end{aligned}$$

One advantage of this method is that, when specializing gender to 0 or 1, we recover the least square lines for the male- or female-only data.

Dummy variables, continued

As another example, for the four seasons, we need to introduce *three dummy variables* x_1, x_2, x_3, where:

- $(x_1, x_2, x_3) = (0, 0, 0)$ for spring (chosen as the baseline),
- $(x_1, x_2, x_3) = (1, 0, 0)$ for summer,
- $(x_1, x_2, x_3) = (0, 1, 0)$ for autumn,
- $(x_1, x_2, x_3) = (0, 0, 1)$ for winter.

We do not just use indicator variables here, to avoid multicollinearity.

Again, if 'interaction' terms (in *Excel* sheet *sales1*: quarter $\times$ season) are included, then specializing the dummy variables gives the individual least square lines. This is a consequence of the underlying matrix algebra.

Outline

Set up

In simple linear regression, we can give confidence intervals for β_1 and β_0. Recall the set up

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

where ϵ_i are iid normal. We treat the x_i's as fixed, then the Y_i's are normal.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators for β_0 and β_1; in fact they are unbiased.

$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$, so as a random variable, it has distribution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{(n-1)\, s_x^2} = \frac{1}{(n-1)\, s_x^2} \sum_{i=1}^{n} (x_i - \bar{x})\, Y_i,$$

which is a linear combination of normals, and is hence normal.

Calculations

It follows that $\mathsf{E}(\hat\beta_1) =$

$$\frac{1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar x)\mathsf{E}(Y_i) = \frac{1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar x)(\beta_0+\beta_1 x_i)$$
$$= \frac{1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar x)\beta_1 x_i = \frac{\beta_1}{(n-1)\,s_x^2}\sum_{i=1}^n (x_i-\bar x)^2,$$

so $\mathsf{E}(\hat\beta_1) = \beta_1$. Likewise,

$$\mathsf{Var}(\hat\beta_1) = \frac{1}{(n-1)^2\,s_x^4}\sum_{i=1}^n (x_i-\bar x)^2\,\mathsf{Var}(Y_i) = \frac{\sigma^2}{(n-1)\,s_x^2}.$$

Similarly tedious computations show that $\hat\beta_0$ is normal, with mean β_0 and variance $\dfrac{\sigma^2}{s_x^2}\left(\dfrac{s_x^2}{n}+\dfrac{\bar x^2}{n-1}\right)$.

Confidence intervals

We estimate σ^2 by $s^2 = \mathsf{SSE}/(n-2)$, which means we will need the t-distribution.

$(1-\alpha)$-level confidence intervals for β_1 and β_0 are, respectively:

$$\hat{\beta}_1 \pm t_{n-2,\,1-\alpha/2}\,\frac{s}{s_x}\,\frac{1}{\sqrt{n-1}},$$
$$\hat{\beta}_0 \pm t_{n-2,\,1-\alpha/2}\,\frac{s}{s_x}\,\sqrt{\frac{s_x^2}{n}+\frac{\bar{x}^2}{n-1}}.$$

These CI's can also be obtained in *Excel*'s Data Analysis $\rightarrow$ Regression (check the 'confidence level' box). We will check it for the triple jump example.

Note: confidence intervals for β_i in multiple regression can be similarly derived, but involve the diagonal entries of $(\mathbf{X}^T\mathbf{X})^{-1}$ (not in the course; see textbook Section 11.4).

Correlation coefficient

Let ρ denote the true *correlation coefficient* of the random variables X and Y (from which we get the observations (x_i, y_i)). Note that r is just an estimate of ρ. We are interested in testing $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$.

If H_0 is true, then $\rho = 0 = \beta_1$, and one can check that

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1 - \beta_1}{s/(s_x\sqrt{n-1})},$$

which follows a t-distribution of $(n-2)$ degrees of freedom.

Therefore, we can reject H_0 if

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{n-2,\,1-\alpha/2}.$$

Exercise: is $r = 0.5$ always insignificant (with $\alpha = 0.05$)?

Prediction

Suppose we wish to predict the value y^* corresponding to a point x^*. Let $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

Then, it can be shown that the $(1-\alpha)$-level two-sided confidence interval for y^* is

$$\hat{y}^* \pm t_{n-2,\,1-\alpha/2}\,\frac{s}{s_x}\sqrt{\frac{s_x^2}{n} + \frac{(x^* - \bar{x})^2}{n-1}}.$$

Outline

Predictor variables and r^2

In multiple regression, increasing the number of predictor variables will increase r^2, even if random numbers are used.

This is because each extra predictor variable allows us to decrease the error (in the worst case, just set the new $\hat{\beta}$ to 0 to get the same error, but we are very likely to do better).

As an extreme example, a polynomial regression of degree $(n-1)$ achieves $r^2 = 1$.

This phenomenon of over-fitting makes r^2 no longer a good measure of how well the model fits the data.

So, how do we pick *useful* predictor variables x_i in our model, and ensure that they have an effect on y?

We first answer a weaker question: how do we know if *any* of the variables affect y?

Analysis of variance (ANOVA)

This question can be answered by ANOVA, the first step of which decomposes the total variability in y into separate components.

We have already done this for multiple (including simple) linear regression: **SST = SSE + SSR**.

Their degrees of freedom are respectively $(n-1)$, $(n-k-1)$, and k, where k is the number of predictor variables.

Explanation for the df's: n terms with 1 constraint; n terms with $(k+1)$ parameters estimated; k predictors.

Define **MSE** $=$ SSE$/(n-k-1) = s^2$, and **MSR** $=$ SSR$/k$ (mean squared regression).

Finally, define $F =$ MSR/MSE.

Hypothesis testing using F

(Intuition for SSR having 1 degree of freedom in simple linear regression: note that $\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$.)

In multiple linear regression, it can be shown that under

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

SSR/σ^2 and SSE/σ^2 are both χ^2 random variables.

Therefore, if H_0 is true, then $F =$ MSR/MSE satisfies an $F_{k,\, n-k-1}$ distribution.

If $F > f_{k,\, n-k-1,\, 1-\alpha}$, then we can reject H_0, and accept H_1 : at least one of the $\beta_i \neq 0$.

Excel can organize all this information in an ANOVA table.