2017-03-01-lec04

Today's topic is mainly about **simple linear regression models.**

Suppose lots of data in a population with random variables X and Y. (There is an obvious trend going through the data points.) The size of the population is N, which is hard to know, because we always sample the population to inference.

We have $E[Y|X] = \beta_0 + \beta_1 X_i$ as the mean of the model.

For data point (X_i, Y_i), what we really care for is that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, 2, \ldots, N$$

where for $i = 1, 2, \ldots, N, \epsilon_i$ is some random variation, the vertical distance from (X_i, Y_i) to the line. And in the population we call it the **error**.

But in practice, we never have the chance to draw a line like this because that is the population. In fact, hopefully after a representative sampling process, we can draw a sample picture instead.

What's changed? Axises becomes x, y and the number of data points deceases (because of sample). Then we are gonna fit a model into the data we have as estimation. We wish it could reflect the true model in population.

Note that β_0 was the intercept of linear line to x-axis, β_1 was the slope of the line. And we are gonna estimate those two:

$$\hat{\beta}_0 = b_0, \hat{\beta}_1 = b_1$$

where b_0 is the intercept of the sample line, b_1 is the slope of the sample line.

And for each data point in the sample $(x_i, y_i), i = 1, 2, \ldots, n$, n is the sample size, there is a corresponding point on the linear line $(x_i, \hat{Y}_i)$, and the vertical difference between them is $e_i = y_i - \hat{Y}_i$ which is called the **residual**.

So the line should be:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \ldots, n$$

Ok, we've all set up. Back to the big question: **How do we estimate $\hat{\beta}_0 = b_0, \hat{\beta}_1 = b_1$?**

We use Gauss' method of least squares (check textbook): find b_0, b_1 that minimize the sum of squares of the errors.

$$
\begin{aligned}
\text{population} \sum_{i=1}^{N} \epsilon_i^2 &= \sum_{i=1}^{N} (Y_i - \hat{Y}_i) \\
\text{or sample} \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2, \\
&\hat{\beta}_0 = b_0, \hat{\beta}_1 = b_1 \text{ are the estimates that minimize this!}
\end{aligned}
$$

To calculate b_0, b_1 in practice we need means and variances of the x, y sample variables and we also need the covariance of X, Y:

To estimate this in the sample we use sum of products of the deviation from x to its mean and the deviation of y to its mean (over degrees of freedom):

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

then

$$
\begin{aligned}
\hat{\beta}_1 &= b_1 = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\
\hat{\beta}_0 &= b_0 = \bar{y} - b_1 \bar{x}
\end{aligned}
$$

Two competing models

model 1: population is $Y = \beta_0 + \epsilon$, the estimated version (sample) is $\hat{Y} = \bar{y}$

model 2: population $Y = \beta_0 + \beta_1 X + \epsilon$, sample $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = b_0 + b_1 x$

What's the difference between the two models? The term $\beta_1 X$. The first does not assume that X has any effect.

Do wee need this term?

- if we don't then $\beta_1 = 0$ and we have model 1.
- if we are convinced that a positive linear trend is a better fit then $\beta_1 > 0$ and we have model 2.

We should do hypothesis test of $H_0 : \beta_1 = 0$ v.s. $H_A : \beta_1 > 0$.

By definition, the standard error of β_1 is the standard deviation of the sampling distribution of β_1. So how do we work this out?

Assumptions underlying a simple linear regression (SLR) model

1. General assumptions (applicable to most statistical models)

    (a) that the sample is representative of the population of interest

    (b) that the explanatory (X) variables are measured without error (or at least minimal error of Y) $\rightarrow$ all the error is in the Y direction (vertical on the earlier plots)

    (c) that a model of the proposed form (e.g. a linear model) is appropriate

2. Model-specific assumptions (most regression-type models including SLR)

    (a) (population) $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, 2, \ldots, N$ where $\beta_0 + \beta_1 X_i$ is the deterministic model for the mean $E[Y_i|X] = \beta_0 + \beta_1 X_i$, and ϵ_i is the stochastic model for the variance. The assumptions, specific to this model, are about ϵ_i

    $$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

    That erros (ϵ_i) are independent and identically (normally) distributed with mean 0 and constant variance σ^2. This in a nutshell is the variance model.