Probability & Statistics

Dongjiao Ge

djge@cityu.edu.mo

Session 11

Reading list:
http://www2.stat.duke.edu/~st118/sta250/linreg.pdf
https://web.njit.edu/~wguo/Math644_2012/Math644_Chapter%201_part4.pdf

Linear regression

In previous sessions, statistics basics, parameter estimation and hypothesis testing were introduced.

Basically, the most widely used statistical analysis is regression, where one tries to explain a **response variable** Y by an **explanatory variable** X based on paired data $(X_1, Y_1), \cdots, (X_n, Y_n)$.

We now have a look at the simple statistical model --- **linear regression.**

- **Simple linear regression**
- **Multiple linear regression** (won't be covered in this module)

Simple linear regression model

The most common way to model the dependence of Y on X is to look for a **linear relationship** with additional **noise** $\boldsymbol{\epsilon}$,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$ **---- simple linear model**

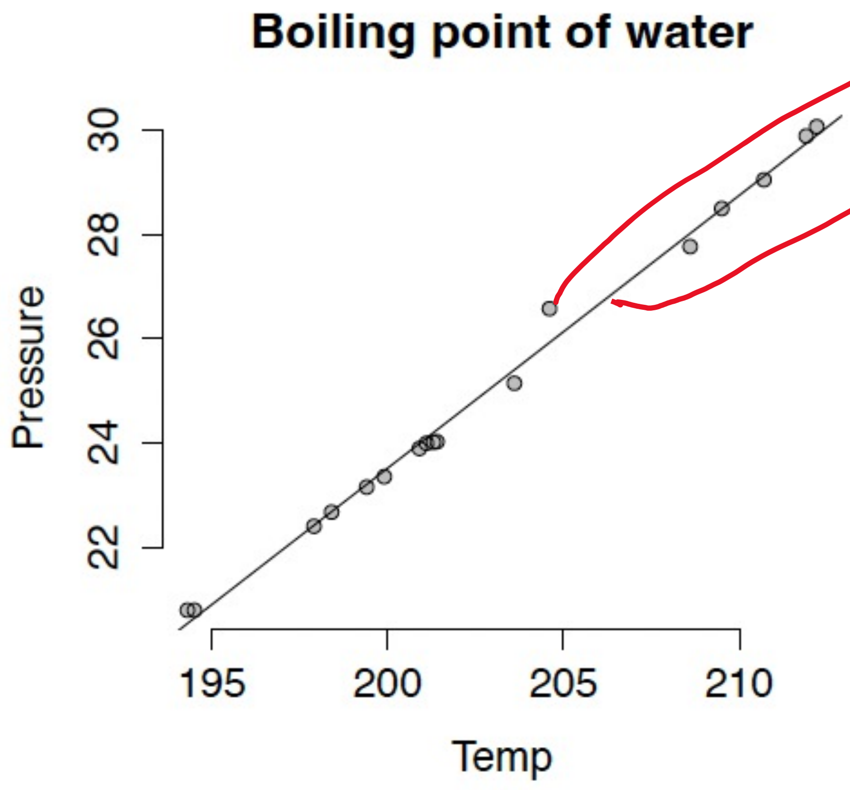
- $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are independent and identically distributed random variables with mean 0 and variance σ^2.
- The unknown parameters are $\beta_0, \beta_1, and\ \sigma^2$

Due to ϵ is the noise, so β_0 and β_1 are the main parameters that can determine the linear model.

How to estimate $\boldsymbol{\beta_0}$ *and* $\boldsymbol{\beta_1}$?

Simple linear regression model

Example: The figure below shows on the left a plot of atmospheric pressure (in inches of Mercury) against the boiling point of water (in degrees F) based on 17 pairs of observations. Although water's boiling point and atmospheric pressure should have a precise physical relationship, there would always be some deviation in actual measurements due to factors that are hard to control.



$(X_1, Y_1), \cdots, (X_n, Y_n)$

The linear could be presented as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

How to use $(X_1, Y_1), \cdots, (X_n, Y_n)$ to estimate β_0 *and* β_1?

Least squares line

The straight line in the above figure is the line that "best fits" the observed data $(x_i, y_i), i = 1, \cdots, n$. This is found as follows:

For any line $y = b_0 + b_1 x$, we can find the "residuals" $e_i = y_i - b_0 - b_1 x_i$ if we tried to explain the observed values of Y by those of X using this line.

The total deviation can be measured by the **sum of squares of the residuals**

$$d(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Now we can find b_0 and b_1 by minimizing $d(b_0, b_1)$. It can be done by solving the following equations:

$$\begin{cases} \dfrac{\partial}{\partial b_0} d(b_0, b_1) = 0 \\ \dfrac{\partial}{\partial b_1} d(b_0, b_1) = 0 \end{cases}$$

Least squares line

Therefore, we have

$$0 = \frac{\partial}{\partial b_0} d(b_0, b_1) = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-1) = -2n(\bar{y} - b_0 - b_1 \bar{x})$$

$$0 = \frac{\partial}{\partial b_1} d(b_0, b_1) = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i) = -2(\sum_{i=1}^{n} x_i y_i - n b_0 \bar{x} - b_1 \sum_{i=1}^{n} x_i^2).$$

These are two linear equations in two unknowns b_0, b_1. The solutions are:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

Least squares line

Using the following notations:

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

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

we can write the least squares solution as

$$\hat{b}_0 = \bar{y} - s_{xy}\bar{x}/s_x^2,\ \hat{b}_1 = s_{xy}/s_x^2$$

- The method of least squares was used by physicists working on astronomical measurements in the early 18th century.
- A statistical framework was developed much later. The main import of the statistical development, as usual, has been to incorporate a notion of uncertainty.

Statistical analysis of simple linear regression

To put the linear regression relationship $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ into a statistical model, we need a distribution on the ϵ_i's. The most common choice is a normal distribution $N(0, \sigma^2)$.

This can be justified as follows:
The additional factors that give rise to the noise term are many in number and act independently of each other, each making a little contribution. By the central limit theorem, the aggregate of such numerous, independent, small contributions should behave like a normal variable. The mean is fixed at zero because any non-zero mean can be absorbed in the intercept β_0.

Statistical analysis of simple linear regression

So, usually, our statistical model of simple linear regression is presented as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i are i.i.d r.v.s; the error terms ϵ_i are independent of the explanatory variables X_i (because the errors account for additional factors beyond the explanatory variables).

We can estimate the model parameters using MLE.
We have $[y_i - (\beta_0 + \beta_1 x_i)] \sim N(0, \sigma^2)$The log-likelihood function is

$$l = \log\{\prod_{i=1}^{n}\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\right)\}\}$$

$$l = \log\{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{-\sum_{i=1}^{n}\frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}\}\}$$

Statistical analysis of simple linear regression

$$l = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

The MLEs of β_0, β_1 and σ^2 can be obtained by solving:

$$\begin{cases} \dfrac{\partial}{\partial \beta_0} l = 0 \\ \dfrac{\partial}{\partial \beta_1} l = 0 \\ \dfrac{\partial}{\partial \sigma^2} l = 0 \end{cases} \Rightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma^2} = 0 \\ \displaystyle\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i) x_i}{\sigma^2} = 0 \\ \displaystyle -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2(\sigma^2)^2} = 0 \end{cases}$$

Statistical analysis of simple linear regression

$$\hat{\beta}_{0,MLE} = \bar{y} - \hat{\beta}_{1,MLE}\bar{x}$$

$$\hat{\beta}_{1,MLE} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_{0,MLE} - \hat{\beta}_{1,MLE}x_i\right)^2$$

It is more common to estimate σ^2 by

$$\hat{\sigma}^2 = s_{y|x}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

where $n-2$ indicates that two unknown quantities (β_0 and β_1) were to be estimated to define the residuals.

Analysis of Variance (ANOVA) approach to regression analysis

Recall the model again

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n$$

The observations can be written as

obs	Y	X
1	Y_1	X_1
2	Y_2	X_2
$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_n

The deviation of each Y_i from the mean $\bar{Y}$,

$$Y_i - \bar{Y}$$

Analysis of Variance (ANOVA) approach to regression analysis

The fitted $\hat{Y}_i = b_0 + b_1 X_i, i = 1, ..., n$ are from the regression and determined by X_i.

Their mean is

$$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Thus the deviation of $\hat{Y}_i$ from its mean is

$$\hat{Y}_i - \bar{Y}$$

The residuals $e_i = Y_i - \hat{Y}_i$, with mean is

$$\bar{e} = 0$$

Thus the deviation of e_i from its mean is

$$e_i = Y_i - \hat{Y}_i$$

Analysis of Variance (ANOVA) approach to regression analysis

Write

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation}\\ \text{due the regression}}} + \underbrace{e_i}_{\substack{\text{Deviation}\\ \text{due to the error}}}$$

obs	deviation of Y_i	deviation of $\hat{Y}_i = b_0 + b_1 X_i$	deviation of $e_i = Y_i - \hat{Y}_i$
1	$Y_1 - \bar{Y}$	$\hat{Y}_1 - \bar{Y}$	$e_1 - \bar{e} = e_1$
2	$Y_2 - \bar{Y}$	$\hat{Y}_2 - \bar{Y}$	$e_2 - \bar{e} = e_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Y_n - \bar{Y}$	$\hat{Y}_n - \bar{Y}$	$e_n - \bar{e} = e_n$
Sum of squares	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ Total Sum of squares (SST)	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ Sum of squares due to regression (SSR)	$\sum_{i=1}^{n} e_i^2$ Sum of squares of error/residuals (SSE)

We have

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{SSE}}$$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n}\{(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\} \\
&= SSR + SSE + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \\
&= SSR + SSE + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})e_i \\
&= SSR + SSE + 2\sum_{i=1}^{n}(b_0 + b_1 X_i - \bar{Y})e_i \\
&= SSR + SSE + 2b_0\sum_{i=1}^{n} e_i + 2b_1\sum_{i=1}^{n} X_i e_i - 2\bar{Y}\sum_{i=1}^{n} e_i \\
&= SSR + SSE
\end{aligned}$$

β_1 is obtained by $\sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = 0$

It is also easy to check

$$SSR = \sum_{i=1}^{n}(b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2 = b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1}$$

Breakdown of the degree of freedom

The degrees of freedom for SST is $n-1$: noticing that

$$Y_1 - \bar{Y}, \ldots.., Y_n - \bar{Y}$$

have one constraint $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$

The degrees of freedom for SSR is 1: noticing that

$$\hat{Y}_i = b_0 + b_1 X_i$$

The degrees of freedom for SSE is $n-2$: noticing that

$$e_1, ..., e_n$$

have TWO constraints $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} X_i e_i = 0$ (i.e., the normal equation).

Mean (of) Squares

$$MSR = SSR/1 \qquad \text{called } \textbf{regression mean square}$$

$$MSE = SSE/(n-2) \qquad \text{called } \textbf{error mean square}$$

Analysis of variance (ANOVA) table Based on the break-down, we write it as a table

Source of variation	**SS**	**df**	**MS**	**F-value**	$P(> F)$
Regression	SSR $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	MSR $= \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$	p-value
Error	SSE $=\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	n-2	MSE $= \frac{SSE}{n-2}$		
Total	SST $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	n-1			

F-test of $H_0 : \beta_1 = 0$

Consider the hypothesis test

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

Note that $\hat{Y}_i = b_0 + b_1 X_i$ and

$$SSR = b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

If $b_1 = 0$ then $SSR = 0$

Thus we can test $\beta_1 = 0$ based on SSR. i.e. under H_0, SSR or MSR should be "small".

We consider the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

Under H_0,

$$F \sim F(1, n-2)$$

For a given significant level α, our criterion is

$$\text{If } F^* \leq F(1-\alpha, 1, n-2) \text{ (i.e. indeed small), accept } H_0$$

$$\text{If } F^* > F(1-\alpha, 1, n-2)\text{(i.e. not small), reject } H_0$$

where $F(1-\alpha, 1, n-2)$ is the $(1-\alpha)$ quantile of the F distribution.

We can also do the test based on the p-value $= P(F > F^*)$,

$$\text{If p-value} \geq \alpha, \text{ accept } H_0$$

$$\text{If p-value} < \alpha, \text{ reject } H_0$$

Exercises:

1. We assume X presenting the pizza cost, and Y presenting the delivery fee. X and Y follows an linear relationship $Y = \beta_0 + \beta_1 X$. We now have four observations of $\{(x_i, y_i)\}_{i=1}^{4}$ shown in the Table below. Please estimate the values of β_0 and β_1.

Pizza cost (MOP) x_i	Delivery fee (MOP) y_i
50	8
60	9
100	12
120	13

Preparing for the exam:

1. Axioms
2. Bayesian, total probability
3. Independent (what does it mean for independent?)
4. How to compute the cdf and pdf? How to compute the mean and variance?
5. Joint distribution (how to compute the marginal p.m.f/density)
6. The sample mean and variance
7. MLE (how to find the MLE) **参数估**计及其原理
8. Hypothesis testing
9. Linear regression (Least square and MLE)

中心极限定理，大数定律