Review of a few Probability facts and linear regressions

Xin Bing

Department of Statistical Sciences
University of Toronto

Mathematical notations

- Vector norm: for a vector $v \in \mathbb{R}^d$, its ℓ_p norm , for $0 \leq p \leq \infty$ is defined as

$$||v||_p = \left(\sum_{j=1}^{d} |v_j|^p \right)^{1/p}.$$

We mainly use $||v||_1$ and $||v||_2$.

- Inner-product between vectors $v_1, v_2 \in \mathbb{R}^d$:

$$v_1^\top v_2 = \sum_{j=1}^{d} v_{1j} v_{2j}.$$

Mathematical notations

- For any square matrix $M \in \mathbb{R}^{d \times d}$, the trace of M is defined as

$$\operatorname{Tr}(M) = \sum_{j=1}^{d} M_{jj}$$

In particular, for any vectors $v_1, v_2 \in \mathbb{R}^d$,

$$v_1^\top v_2 = \operatorname{Tr}\left(v_1 v_2^\top\right).$$

Let X and Y be two random variables.

- $$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- More generally, for any function f,

$$\mathrm{Var}(f(X)) = \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right] = \mathbb{E}[(f(X))^2] - (\mathbb{E}[f(X)])^2.$$

- X is said to be uncorrelated with Y if

$$\mathrm{Cov}(X, Y) = 0.$$

In particular, the fact that X is independent of Y implies that $\mathrm{Cov}(X, Y) = 0$.

- For any constants a, b,

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y).$$

In particular, if X is uncorrelated with Y, then

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y).$$

- For any function f and g, if X is independent of Y, then

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)],$$

  and

$$\mathbb{E}[f(X) \mid Y] = \mathbb{E}[f(X)].$$

- For any function h,

$$\begin{aligned}\mathbb{E}[h(X, Y)] &= \mathbb{E}_X\left[\mathbb{E}_{Y|X}[g(X, Y) \mid X]\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_{X|Y}[g(X, Y) \mid Y]\right]\end{aligned}$$

  where $\mathbb{E}_X$ is the expectation w.r.t. the randomness of X whereas $\mathbb{E}_{Y|X}$ is w.r.t. the randomness of $Y \mid X$.

Simple Linear Regression

- We assume a model
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
where β_0 and β_1 are two unknown constants that represent the **intercept** and **slope**, also known as **coefficients** or **parameters**, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict response at $X = x$ as
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Least Square Estimates

- Training data: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Least square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are minimizers of RSS, given by

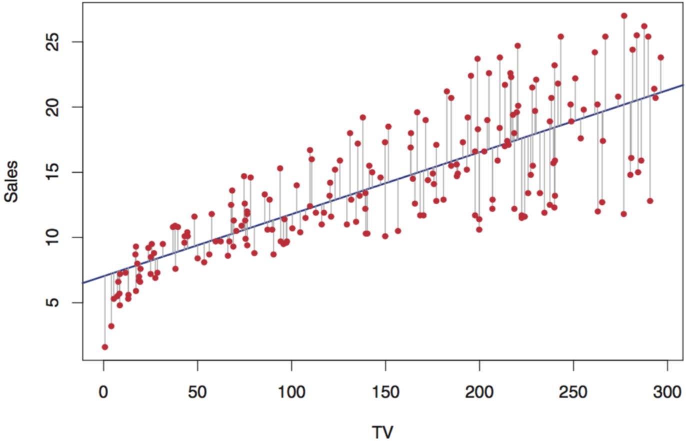
$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

They have the following closed-form solution

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

where $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ and $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ are the sample means.

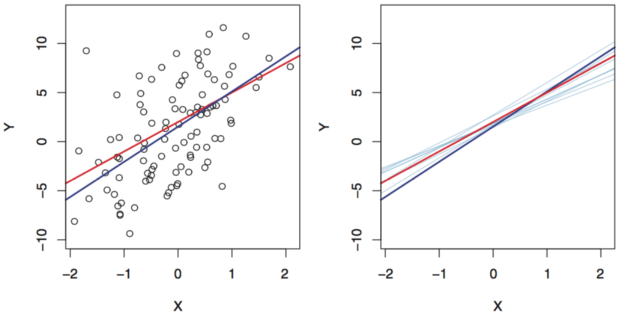
Advertising Data



Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. A linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Understand the randomness in $\hat{\beta}_0$ and $\hat{\beta}_1$

We cannot hope $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$, because they depend on the observed data which is random.



Left: The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares fit based on the observed data.
Right: The light blue lines represent ten least squares fits. Each one is computed on the basis of a different training set.
The fitted least squares lines are different, but their average is quite close to the true regression line.

Derivation of the OLS formula

Recall that

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

Taking the derivative with respect to $\boldsymbol{\beta}$ and setting it equal to zero yield

$$-\frac{2}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) = 0.$$

The solution $\hat{\boldsymbol{\beta}}$ has to satisfy the above equation.

When $\mathbf{X}$ has full column rank such that $\mathbf{X}^\top\mathbf{X}$ is invertible, there exists a unique solution, i.e.

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