STA302 notes

May 10, 2022

Contents

May 9

Syllabus

This is a course on linear regression. The focus is using R to do data analysis, and build the mathematical foundation for regression. We will understand how prediction works later, which is the foundation for data science.

Marking

- 2 HW - 15% each, due June 1, June 15
- Test - 25% on May 25
- Exam - 45% during June 22-27

Books J. Sheather, A Modern Approach to Regression w/ R and D. Montgomery, Linear Regression Analysis.

Review

Definition 1. A **sample space** S is the set of possible events. A **random variable** is a function $X : S \to \mathbb{R}$ assigning a number to elements of the sample space.

Constants can also be pseudo random variables. These are called **degenerate random variables** that have a **degenerate distribution** since they have infinite cdf.

Definition 2. For an event $A \subset S$, we define the **indicator function** I_A as

$$I_A(s) = \begin{cases} 1, & s \in A \\ 0, & s \notin A \end{cases}$$

These are important since we later use them to create dummy variables in linear regression. When we write an inequality involving random variables, we mean that it holds for all elements of the sample space. I.e. $X \geq Y \implies X(s) = Y(s), \forall s \in S$.

Example 1. Consider $S = \{1,2,3,4,5,6\}$. For $s \in S$, $X(s) = s$, let $Y(s) = X(s) + I_6(s)$. Then $Y = X$ for all $s \in S$ except 6, where $Y = 7$, $X = 6$.

Definition 3. **Discrete r.v.** are functions from a countable sample space, and **continuous r.v.** are functions from an uncountable sample space. There are also **mixture** random variables, which are continuous/discrete for different parts of the sample space. Random variables can be univariate and multivariate as well.

Example 2. The multinomial distribution is an example of a discrete multivariate random variable.

Definition 4. If X is a random variable, the p.d.f. is the derivative of the c.m.f. As well, $\mathscr{P}(a \leq X \leq b) = \int_a^b f(x)dx$ where $f(x)$ is pdf. Similar thing holds for discrete r.v.

Proposition 1. The expectation of two random variables is linear. For $Z = aX + bY$, X, Y r.v., then $E(Z) = aE(X) + bE(Y)$.

Definition 5. The **variance** of X is $V(X) = E(X - \mu_x)^2$. The **sample variance** $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$. Note we divide by $n-1$ so that it is an unbiased estimator (STA261).

Some properties:

- $V(X) \geq 0$
- $V(aX + b) = a^2 V(x)$
- $V(X) = E(X^2) - E(X)^2$
- $V(X) \leq E(X^2)$
- $\sigma_X = \sqrt{V(X)}$

Note: In linear regression, the variance of the predicted variable depends on the slope of regression line but not on the intercept (second property).

Let X_1, X_2, Y be r.v. and A be an event. Let $Z = aX_1 + bX_2$. Then

- $E(Z \mid A) = aE(X_1 \mid A) + bE(X_2 \mid A)$
- $E(Z \mid Y = y) = aE(X_1 \mid Y = y) + bE(X_2 \mid Y = y)$
- $E(Z \mid Y) = aE(X_1 \mid Y) + bE(X_2 \mid Y)$

Proposition 2. (Laws of Total Expecation and Variance) $E(E(Y \mid X)) = E(Y)$ and $V(X) = V(E(X \mid Y)) + E(V(X \mid Y))$.

We will see that linear regression is a conditional r.v., and the above will be very useful. For $X_1,\dots,X_n$ i.i.d. random variables, $x_1\dots x_n$ realizations, then $\overline{x}=\frac{\sum x_i}{n}$. The **sample average** $\overline{X}=\frac{\sum X_i}{n}$ is a random variable. In general, any function of n i.i.d. random variables is a random variable, and called a **sampling statistic** that follows a **sampling distribution**.

Theorem 1. (Central Limit Theorem) For $X_1,\dots,X_n$ i.i.d. $f(x,\theta)$, $E(X)$, $V(x)<\infty$, then $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\to N(0,1)$ converges in distribution for sufficiently large n.

Proof. Proof with moment generating functions. □

Example 3. In the Cauchy distribution, this does not hold since it has infinite mean and variance.

Definition 6. The **covariance** $\text{Cov}(X,Y)=E[(X-\mu_x)(Y-\mu_Y)]=E(XY)-E(X)E(Y)$. Covariance quantifies the relationship between two variables, i.e. how much one varies with the other. The **correlation** $\text{Corr}(X,Y)=\frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$.

- Covariance is an inner product, variance is norm.
- $V(X+Y)=V(X)+V(Y)+2\text{Cov}(X,Y)$.
- If $X\perp Y$, $V(X+Y)=V(X)+V(Y)$.
- In general, $V(\sum_i X_i)=\sum_i V(X_i)+2\sum_{i<j}\text{Cov}(X_i,X_j)$.

These will be useful in regression, where we try to identify relationships between r.v.s.

Definitions in statistics

In probability, we are given a mathematical model to work with. In statistics, we infer properties of a mathematical model. The steps of data analysis are: state the problem, identify what data is needed, decide on a model and collect data, clean data, estimate parameters of the model, and carry out appropriate tests, draw conclusions.

Start of Course

Definition 7. The **corelation coefficient**

$$\rho_{X,Y}=\frac{\sum_i(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i(x_i-\overline{x})^2}\sqrt{\sum_i(y_i-\overline{y})^2}}=\frac{\text{Cov}(X,Y)}{s_x s_y}$$

The above value is somewhat like the $\cos(\theta)$ between the vectors X,Y; recall dot product. When we discuss corelation, we talk about linear relations only; the linear association between X,Y. We can see this by considering X and $Y=X^2$. Corelation is symmetric, it does not indicate the direction of the symmetry (which causes which/causation). Corelation only says the influence on the change of one variable when the other changes; think about moving along non-orthogonal vectors and projecting.

Galton investigated the effect of fathers heights on their sons height. Galton termed **regression** as a 'regression' of heights towards the mean; on average, heights of sons move towards the mean, so the average height across generations is the same.

In a linear regression, we assume there is a linear relation $Y = \beta_0 + \beta_1 X + \epsilon$ between the random variables X, Y where ϵ is an error random variable. The deviation not captured by linearity is incorporated to ϵ. Given two values of X, it is not guaranteed that the value of Y is the same. But for a unique X we get **unique average** Y. We want $E(Y \mid X = x) = \beta_0 + \beta_1 X$; the relationship between the mean of Y and a specific value of X is linear. Note $E(\epsilon) = 0$. We call X the **explanatory, predictor, independent** variable and Y as the **response, outcome, dependent** variable. Suppose we are given paired data $(x_1, y_1), \ldots, (x_n, y_n)$. We try to fit a linear regression to model the relationship between X and Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \text{ and want } E(Y \mid X = x) = \beta_0 + \beta_1 X$$

The values of β_0, β_1 are not yet known and need to be estimated. In the sample, the error e_i replaces ϵ_i. The line best predicting Y as X changes should minimize the squares of the errors $e_i = y_i - \hat{y}_i$ where $\hat{y}_i = b_0 + b_1 x_i$ where b_0, b_1 are the intercept and slope of the regression line. We minimize the squares $\sum_i e_i^2$. The e_i are referred to as **residuals**; minimize residual sums squared. Note

$$RSS(b_0, b_1) = \sum_i e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

Aside: What value of a minimizes (1) $\sum |x_i - a|$, and which minimizes (2) $\sum (x_i - a)^2$? Answer: (1) $a = \text{Med}(X)$, (2) $a = \overline{x}$. We do not minimize the sum of the residuals, since this must always be 0. We minimize the RSS with respect to b_0, b_1.

$$\frac{\partial RSS}{\partial b_0} = -2 \sum_i (y_i - b_0 - b_1 x_i), \quad \frac{\partial RSS}{\partial b_1} = -2 \sum_i x_i (y_i - b_0 - b_1 x_i)$$

so setting these to 0, we get the **normal equations**

$$\sum_i y_i = b_0 n + b_1 \sum_i x_i, \quad \sum_i x_i y_i b_0 \sum_i x_i + b_1 \sum_i x_i^2$$

Solving these, we get

$$\hat{\beta}_0 = b_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \qquad \hat{\beta}_1 = b_1 = \frac{\sum_i x_i y_i - n \overline{x}\,\overline{y}}{\sum x_i^2 - n \overline{x}^2} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{X,Y}}{S_X}$$

The intercept is the average value of the response when $X = 0$.

Afterthoughts

When the errors have $E(\epsilon = 0)$, then $V(\epsilon) = E(\epsilon^2) - E(\epsilon)^2 = E(\epsilon^2)$. By minimizing this in the sample, we minimize the variance of the errors (?)

May 11