1. Probability, regression review

Friday, January 7, 2022 5:19 PM

Probability Review

often written $\underline{X}$ or $\underline{\underline{X}}$ by hand to make it clear

Definition/Convention A capital letter, like X, is usually a random variable

$\mathbb{P}(X \geq .3)$ or $\text{Prob}(X \geq .3)$ or $P(X \geq .3)$ mean the probability that $X \geq .3$

A lowercase letter, like x, is usually a realization

eg., $x = 0.3$, $\mathbb{P}(X \geq x) = .5$

Definition A time series X_t or (X_t) or $X(t)$ is a sequence of random variables

If $t \in \mathbb{N} := \{0, 1, 2, 3, \ldots\}$ or $t \in \mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$ it's a discrete time series ← our main focus for this class

If $t \in \mathbb{R}$ or $t \in [a, b] \subseteq \mathbb{R}$ it's a continuous time series.

A realization of a time series is x_t or (x_t) or $x(t)$, just a sequence.

We often call (x_t) a "time series" also

Definition Independence $X_1, X_2, \ldots, X_n$ are independent if their cumulative distribution function (CDF) factors, namely

$$\underbrace{F(x_1, x_2, \ldots, x_n)}_{:= \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)} = F_1(x_1) \cdot F_2(x_2) \cdot \ldots \cdot F_n(x_n)$$

(joint CDF; CDF for X_1)

Equivalently, the pdf / pmf factors

⚠ Independent variables are uncorrelated

but uncorrelated variables need not be independent*

*unless jointly normal.

Def Expectation, $\mathbb{E}$. Let X be a r.v.
and Variance, Var

- Continuous r.v. Let f_X be the pdf (prob. density function) of X, then

$$\mathbb{E}[X] := \int x \cdot f_X(x)\, dx \qquad \text{aka the mean, } \mu_X$$

(often written $\mathbb{E}X$ or $E[X]$ or EX)

- Discrete r.v. Let f_X be the pmf (prob. mass function) of X, then

$$\mathbb{E}[X] := \sum_x x\, f_X(x) = \mu_X$$

- Variance: in all cases,

$$\mathrm{Var}[X] := \mathbb{E}[(X-\mu_X)^2] = \sigma_X^2$$

Note: not all distributions are cts or discrete (e.g., a mixture of the two). Similar def'n apply

- In all cases,

① $\mathbb{E}$ is linear, $\mathbb{E}(aX+Y) = a\cdot\mathbb{E}(X) + \mathbb{E}(Y)$

② $\sigma_X^2 = \mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ very handy

$\sigma^2 \geq 0$ via Jensen's ineq.

③ $\mathbb{E}[X\cdot Y] = \mathbb{E}[X]\cdot\mathbb{E}[Y]$ only if X, Y are uncorrelated

(in particular, this is true if they're independent)

④ $\mathrm{Var}[aX+b] = a^2\cdot\mathrm{Var}[X]$ (not linear)

$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ only if X, Y are uncorrelated

⑤ $\sigma = \sqrt{\mathrm{Var}[X]}$ is the "standard deviation"

Def Covariance

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[X\cdot Y] - \underbrace{\mathbb{E}[X]}_{\mu_X}\cdot\underbrace{\mathbb{E}[Y]}_{\mu_Y}$$

(over joint distribution)

i.e. X, Y uncorrelated iff $\mathrm{Cov}(X,Y) = 0$

Properties

① $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

② $\mathrm{Cov}(aX+b, cY+d) = a\cdot c\cdot\mathrm{Cov}(X,Y)$

③ $\mathrm{Cov}(X+Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$

Remark Intro probability classes focus on calculating $\mathbb{E}$, Var, Cov... using pdf/pmf f_X and/or cdf F_X.

In more advanced usage, try to avoid using f_X and F_X directly until you really have to. Exploit properties (eg., linearity of $\mathbb{E}$)

More prob. review

See Appendix A in Brockwell + Davis 3rd ed.

Regression Review

Working example: Sea Surface Temperature ("SST") data

Data: $(t, SST(t))$

we'll do ordinary least squares (OLS) regression

→ means there is not expected to be error in the independent variable (in this case, "year")

First, choose the independent variables aka covariates aka regressor

- The obvious choice is "year", t. What else?
  We'll require our model to be linear in these covariates, so nonlinearity has to be explicitly added.

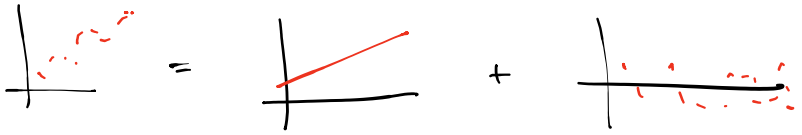
  So could include t^2 or $\sin(t)$, for example

- Let's stick with t and an offset (a constant).

  So model:

$$SST(t) = \underbrace{\text{linear function of time}}_{\text{technically it's an "affine" function since we include the offset}} + \text{error}$$

$$= \beta_0 + \beta_1 t + \varepsilon(t)$$



θ is another common notation for β

X as independent, Y as dependent is a universal convention.

to find the coefficients $\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ we write

our model in vector form

matrix multiplication

$$\begin{pmatrix} SST(1850) \\ \vdots \\ SST(2012) \end{pmatrix} = \begin{pmatrix} 1 & 1850 \\ 1 & 1851 \\ \vdots & \vdots \\ 1 & 2012 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon(1850) \\ \vdots \\ \varepsilon(2012) \end{pmatrix}$$

(163 years)

$$(*) \quad \vec{Y} = X \cdot \vec{\beta} + \vec{\varepsilon}$$ ← "residual"

now we're using the convention that capitals mean matrices, not random vectors

OLS in this case makes sense if we have the following assumptions:

1. (*) is the true model, and
2. $\varepsilon(1850), \varepsilon(1851), \dots, \varepsilon(2012)$ are mutually independent
3. $\varepsilon(t) \sim N(0, \sigma^2)$ normally distributed (and "centered" aka mean 0)

then OLS is the maximum likelihood estimate (MLE) for $\vec{\beta}$:

common convention: a hat ^ is used for estimators (something based on data)

$$\hat{\vec{\beta}}_{OLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \underbrace{\| \vec{Y} - X \cdot \vec{\beta} \|_2^2}$$

$$= (\vec{Y} - X\vec{\beta})^T (\vec{Y} - X\vec{\beta})$$

$$= \sum_{t=1850}^{2012} (SST(t) - \beta_0 - \beta_1 t)^2$$

Solvable in closed form by solving the normal equations

$$X^T \cdot X \cdot \vec{\beta} = X^T \vec{y}$$

so if X has full column rank, $\vec{\beta} = (X^T X)^{-1} X^T y$

you should almost never do this formula on a computer

and we can estimate σ^2 also:

$$\hat{\sigma}^2 = \frac{1}{n-2} \| \vec{Y} - X \hat{\vec{\beta}}_{OLS} \|_2^2$$

n = #observations; $\hat{\vec{\beta}}_{OLS}$ = size of $\vec{\beta}$; "RSS" = Residual Sum of Squares

$$R^2 = 1 - \frac{RSS}{SS_{tot}} \quad \leftarrow \sum (y_i - \bar{y})^2$$

$R^2 \approx 1$ means most variability has been explained

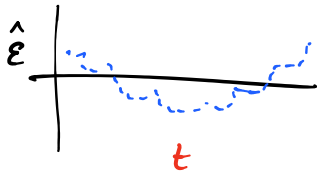
and define $\widehat{SST}(t) = \hat{\beta}_0 + \hat{\beta}_1 \cdot t$

and estimate residuals

$$\hat{\varepsilon}(t) = SST(t) - \widehat{SST}(t)$$

Always a very good idea to plot residuals

we assumed true residuals were independent



← doesn't look independent

We could model

$$SST(t) = \beta_0 + \beta_1 t + \underbrace{\beta_2 t^2}_{\text{new term}}$$

though if you add too many more terms, you're at risk of overfitting

In time series, we're going to discuss relaxing the assumption that $\varepsilon(t)$ is independent

(both in a regression context and in other contexts)