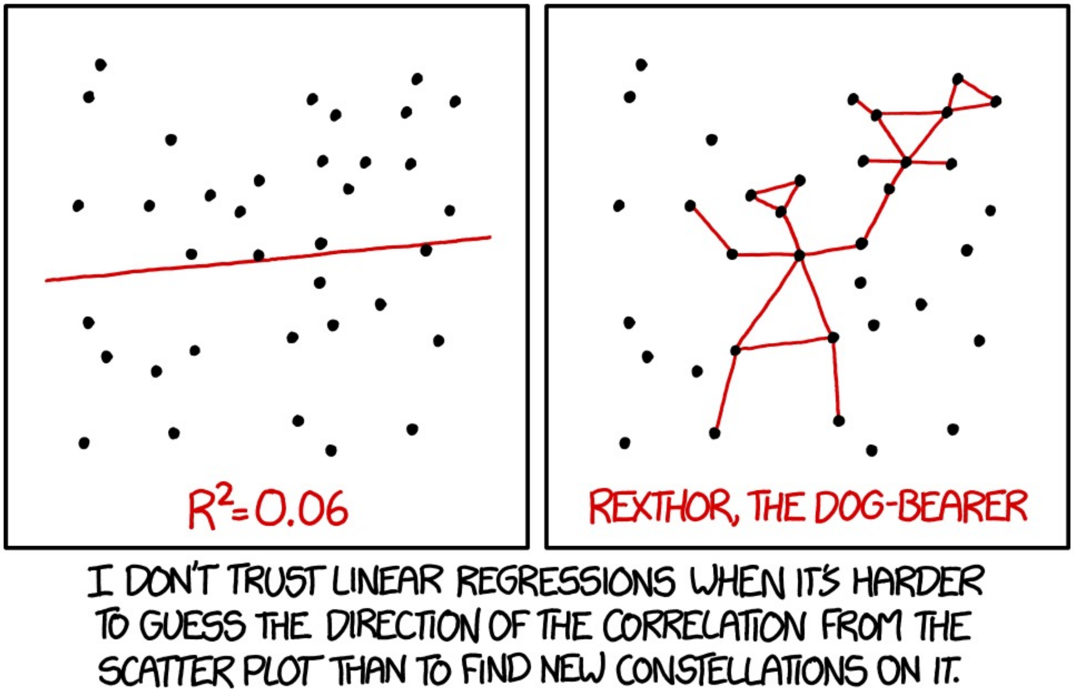
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 39 : 4/28/2021

Chapter 12

More about regression

https://xkcd.com/1725/

So far:

goes through (μ_X, μ_Y)

- $\hat{Y} = \hat{a}X + \hat{b}$ - line of "best fit" : minimizes mean squared error = $E\left[\left(Y-\hat{Y}\right)^2\right]$
- $\hat{Y}$ is called the fitted value of Y, where: $\hat{a} = \frac{r\sigma_Y}{\sigma_X}$, $\hat{b} = \mu_Y - \hat{a}\,\mu_X$
- $\hat{Y} = \hat{a}X + \hat{b} = \hat{a}X + \mu_Y - \hat{a}\mu_X = \hat{a}(X-\mu_X) + \mu_Y = \underline{\hat{a}D_X + \mu_Y}$

$E(X^* Y^*)$

- Correlation: $r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X Z_Y)$ and $-1 \le r \le 1$
- Residual $D = Y - \hat{Y}$, $E(D) = 0$, $\underline{Var(D)} = (1-r^2)\sigma_Y^2$ variance of deviations from $\hat{Y}$ (σ_Y^2 = MSE from using $\bar{Y}$ to predict Y)
- $r(D,X)$ (the residuals are uncorrelated with the predictor: why?)

$$r(D,X) = E\left(\left(\frac{D-0}{\sigma_D}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right) = \frac{1}{\sigma_D\sigma_X}E(D\cdot D_X) = \frac{1}{\sigma_D\sigma_X}E\left(\left(D_Y - \hat{a}D_X\right)D_X\right)$$

$$\boxed{D = Y - \hat{Y} = Y - \hat{a}D_X - \mu_Y}$$

$$= \frac{1}{\sigma_D\sigma_X}E\left(D_Y D_X - \hat{a}D_X^2\right) = \frac{1}{\sigma_D\sigma_X}\left[E(D_Y D_X) - \hat{a}E(D_X^2)\right]$$

$$= \frac{1}{\sigma_D\sigma_X}\left[r\sigma_Y\sigma_X - r\frac{\sigma_Y}{\sigma_X}\cdot\sigma_X^2\right] = 0$$

$$\left(\begin{array}{l} E(D_Y D_X) = r\sigma_Y\sigma_X \\ E(D_X^2) = Var(X) \end{array}\right)$$

The Simple Linear Regression Model

- Regression model from data 8
- Model has two variables: response (Y) & (x) predictor/covariate/feature variable

or explanatory variable.

- **Assumptions**: response is a linear function of the predictor (signal) + random error (noise), where the noise has a **normal** distribution, centered at 0. The signal is not random, but the response is, because the noise is random:

$$\text{response} = \text{signal} + \text{noise}$$

- In mathematical language:

For individuals $i = 1, 2, \dots n$ $\quad (x_1, Y_1), (x_2, Y_2) \dots (x_n, Y_n)$

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{} + \varepsilon_i$$

x_i not random

β_0, β_1 constants

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + 0$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \beta_0 + \beta_1 \bar{x}$$

$$\mathrm{Var}(\bar{Y}) = \sigma^2/n$$

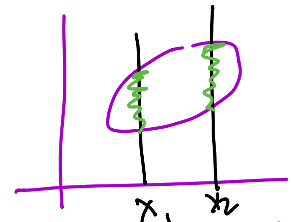
Data generation

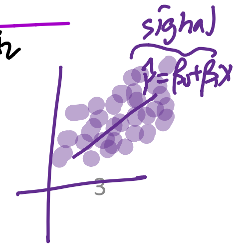
X population $\rightarrow Y = \beta_0 + \beta_1 X + \varepsilon$

$\varepsilon_i \sim N(0, \sigma^2)$

n samples $X_1 \dots X_n$

obs. values $x_1, \dots, x_n$





① $\beta_0, \beta_1, \sigma^2$: constant parameters that we do <u>not</u> observe.

② x_i is the observed value of the predictor for individual i & we assume it is not random

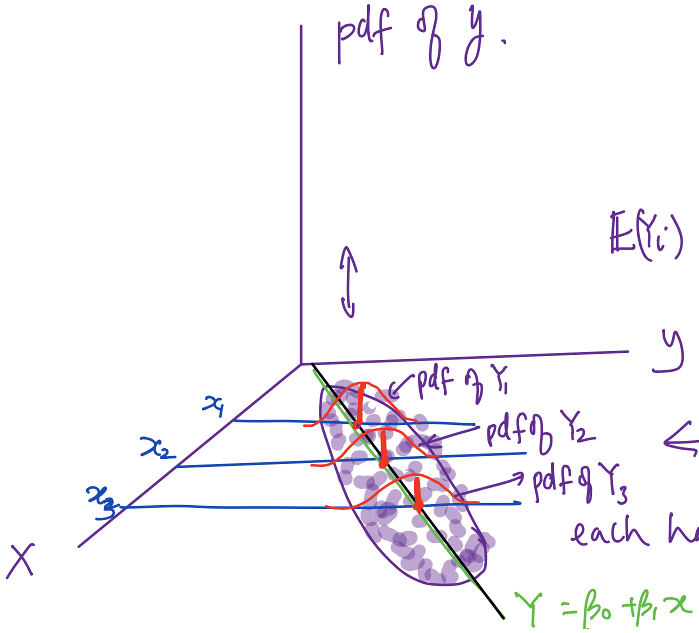
③ ε_i : i^{th} error : $\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n$ are assumed to be iid $N(0, \sigma^2)$

$\beta_0, \beta_1, \sigma^2$ are called unobservable constant parameters.

Goal is to get as close as we can to the signal $(\beta_0 + \beta_1 x)$

Estimate of signal is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{\beta}_0, \hat{\beta}_1$ estimated from the sample & are therefore random variables (change when a different sample is used.)



$E(Y_i) = \beta_0 + \beta_1 x_i$ ← we don't know

estimate it as

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

each has the same variance σ^2

The regression line

- For each i, we want to get as close as we can to the *signal* $\beta_0 + \beta_1 x_i$
- There is some "true" regression line $\beta_0 + \beta_1 x$ that we cannot observe since there is noise. We estimate this line by minimizing the squared observed error.
- Estimate of the line given the data is $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope, respectively, given the data.
- We will investigate the distribution of the slope estimate (why is it random?) after looking at the individual and average response.

The individual response Y_i and the average response $\bar{Y}$

- For any fixed i, Y_i is the sum of the signal and the noise.

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}}$$

- The signal is not random, but the noise is random with $\epsilon_i \sim N(0, \sigma^2)$
- Therefore what is the distribution of the Y_i ?

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

- What can you say about the independence and distribution of each of the Y_i ? Are they iid?

Y_i are indep b/c ε_i are indep.
Y_i are not identically distributed.

- Let $\bar{Y}$ be the average response. What would be its distribution?
- $E(\bar{Y}) =$ $\beta_0 + \beta_1 \bar{x}$

- $Var(\bar{Y}) =$ $\frac{\sigma^2}{n}$, $Var(\varepsilon_i) = \sigma^2$

The estimated slope β_1

$$\hat{a} = r\frac{\sigma_Y}{\sigma_X}$$

- Recall the slope we derived in the previous chapter

$$\hat{a} = \frac{E(D_X D_Y)}{\sigma_X^2} = \frac{r\sigma_Y\sigma_X}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}$$

- Now we have data, so we need to use the empirical distribution
- The least squares estimate of the true slope β_1 is:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

- Notice that $\hat{\beta}_1$ is random (because of the Y_i). How would we find its distribution? $\hat{\beta}_1$ has normal dsn. b/c Y_i are normal.
- Note that $E(Y_i - \bar{Y}) = \beta_1(x_i - \bar{x})$
- $E(\hat{\beta}_1) = \mathbb{E}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\mathbb{E}(Y_i-\bar{Y}))}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}\beta_1(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$= \beta_1$

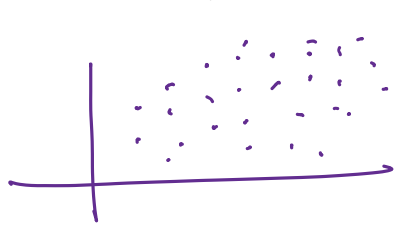
Distribution of $\hat{\beta}_1$

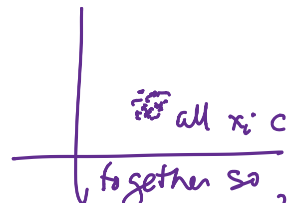
- From the formula of $\hat{\beta}_1$, we see that it is a linear combination of the independent normal rvs $Y_1, Y_2, \ldots, Y_n$ and therefore $\hat{\beta}_1$ is also normal.

- $E(\hat{\beta}_1) = \beta_1$ indicating that $\hat{\beta}_1$ is an unbiased estimator of β_1

- Recall that the common variance of the errors ϵ_i is σ^2

- FACT: $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$





all x_i close together so $\Sigma(x_i - \bar{x})^2$

- What you want to note is that the numerator is constant, so as we have more terms, the denominator gets larger, and our estimated slope gets closer to the true slope.

- We will need to estimate σ^2 since it is an unknown parameter.

SD of the estimated slope $\hat{\beta}_1$

- $SD(\hat{\beta}_1) =$

- Need to estimate σ, which we will do by using the SD of the residuals. Since we are estimating the SD from the data, we will call it *standard error* of the estimator.
- That is, we will denote this estimated $SD(\hat{\beta}_1)$ by $\boldsymbol{SE(\hat{\beta}_1)}$**.**

- The larger the n, the better our estimate of σ

$$\hat{\sigma} = SD(D_1, D_2, \dots, D_n) = \sqrt{\frac{1}{n}}$$

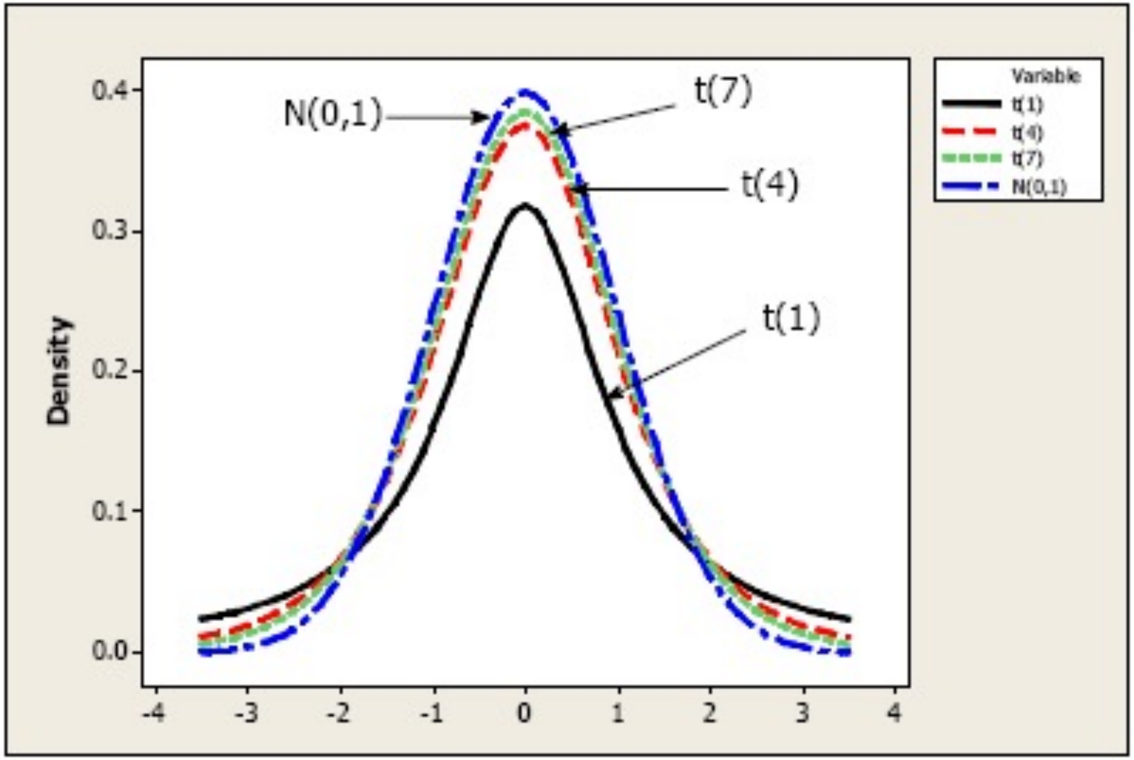
- A 95% CI for β_1 is given by $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$
- For large n, the distribution of $\hat{\beta}_1$, standardized, is approximately standard normal.

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$$

- Let's look at the example from the text on pulse rates.

The t-distribution

Rather than a normal curve, a t-curve is used. For regression, "degrees of freedom" for T equals $n - 2$. For large enough n, use the normal curve.



$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

From onlinecourses.science.psu.edu