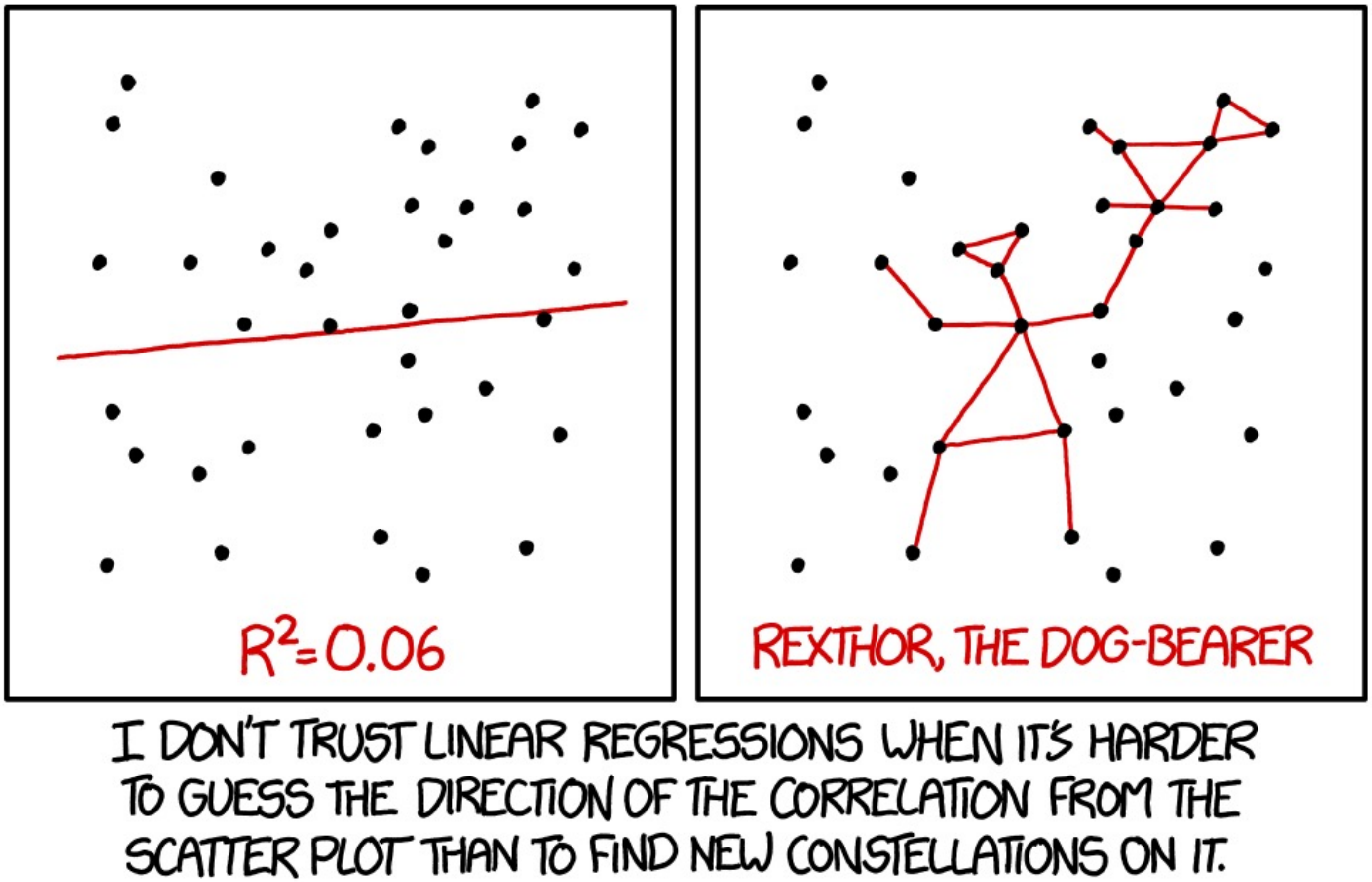
Stat 88: Probability & Mathematical Statistics in Data Science



https://xkcd.com/1725/

Lecture 27 : 4/28/2022

Finishing Chapter 11, and some of chapter 12

Correlation, Regression

Mathematical derivation of the formulas for a and b OPTIONAL

- As usual, $E(X) = \mu_X, SD(X) = \sigma_X; E(Y) = \mu_Y, SD(Y) = \sigma_Y$
- (X, Y) are our random variables, that we *think* are related by a linear function, perhaps with some error: $Y = aX + b + error$
- We want to estimate the equation of the line, that is, find $\hat{Y}$ such that $\hat{Y} = aX + b$
- Find the a and b by minimizing the mean square error, where error is the difference between our estimate $\hat{Y}$ and the original random variable Y.
- Notice that the mean squared error will be a function of a and b:

$$MSE(a,b) = E\left((Y - \hat{Y})^2\right) = E\left(\left(Y - (aX + b)\right)^2\right)$$

- First, we can look for the best intercept for some fixed slope:

Mathematical derivation of the formulas for a and b OPTIONAL

- Looking for the best intercept for some fixed slope, that is, fix a, and then see, for this *given* value of a, what would be the b that minimizes the MSE?

- We can write out the MSE as a function of b, take the derivative, and set it equal to 0, and look for the best b.

Equation of the regression line

- $\hat{Y} = \hat{a}X + \hat{b}$
- $\hat{Y}$ is called the fitted value of Y, $\hat{a}$ is the slope, $\hat{b}$ is the intercept where:
- $\hat{a} = \frac{r\sigma_Y}{\sigma_X}, r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X \times Z_Y)$
- $\hat{b} = \mu_Y - \hat{a}\,\mu_X$

Correlation

- The expected product of the deviations of X and Y, $E(D_X D_Y)$ is called the ***covariance*** of X and Y.
- The problem with using covariance is that the units are multiplied *and* the value depends on the units
- Can get rid of this problem by dividing each deviation by the SD of the corresponding SD, that is, put it in standard units. The resulting quantity is called the ***correlation coefficient*** of X and Y:
- $r(X, Y) =$
- Note that it is a pure number with no units, and now we will prove that it is always between -1 and 1.

Bounds on correlation

- $r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X Z_Y)$
- (Note that this implies that $E(D_X D_Y) = r\sigma_X\sigma_Y$. We will use this later.)

Correlation as a measure of linear association

- $D = Y - \hat{Y},\ E(D) = 0,\ Var(D) = (1 - r^2)\sigma_Y^2$
- What if the correlation is very close to 1 or -1? What does this tell you about X & Y?

- What about if the correlation is close to 0? What does this tell you about X & Y?

Residual is uncorrelated with X

- What about $r(D, X), D = Y - \hat{Y}$?
- Intuitively, what should this be? Why?
- What should your residual (diagnostic) plot look like?

The Simple Linear Regression Model

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

response = signal + noise

- In mathematical language:

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.

- 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 ?

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

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

- $Var(\bar{Y}) =$

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

- Y_i are normal with expectation $\beta_0 + \beta_1 x_i$ and variance σ^2
- Note that the individual responses are independent of each other.

- Let $\bar{Y}$ be the average response.
- $E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ (the expected average response is the *signal* at the average value of the predictor variable)

- $Var(\bar{Y}) = \frac{\sigma^2}{n}$ (only involves the error variance since the randomness in the Y_i's comes only from the errors or noise)

- Since $\bar{Y}$ is a linear combination of independent normally distributed random variables, it is also normal.

The estimated slope β_1

- Recall the slope we derived in the previous chapter

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

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

$$\hat{\beta}_1 =$$

- Notice that $\hat{\beta}_1$ is random (because of the Y_i). How would we find its distribution?
- Note that $E(Y_i - \bar{Y}) = \beta_1(x_i - \bar{x})$
- $E\left(\hat{\beta}_1\right) =$

The estimated slope β_1

- 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).
- Also, since Y_i is normal, and $\bar{Y}$ is normal, so is $Y_i - \bar{Y}$, therefore $\hat{\beta}_1$ *is also normally distributed*

- $E(Y_i - \bar{Y}) = \beta_1(x_i - \bar{x})$

- $E(\hat{\beta}_1) = \beta_1$, so $\hat{\beta}_1$ is an ***unbiased*** estimator of β_1

- $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$(to be taken as fact, proof beyond the scope of this class)

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 ________ estimator of β_1

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

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

- 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.

Confidence intervals for β_1

- $SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, for large n, $SE(\hat{\beta}_1) \to SD(\hat{\beta}_1)$

Therefore, for large n, the distribution of $\hat{\beta}_1$, standardized, is approximately standard normal.

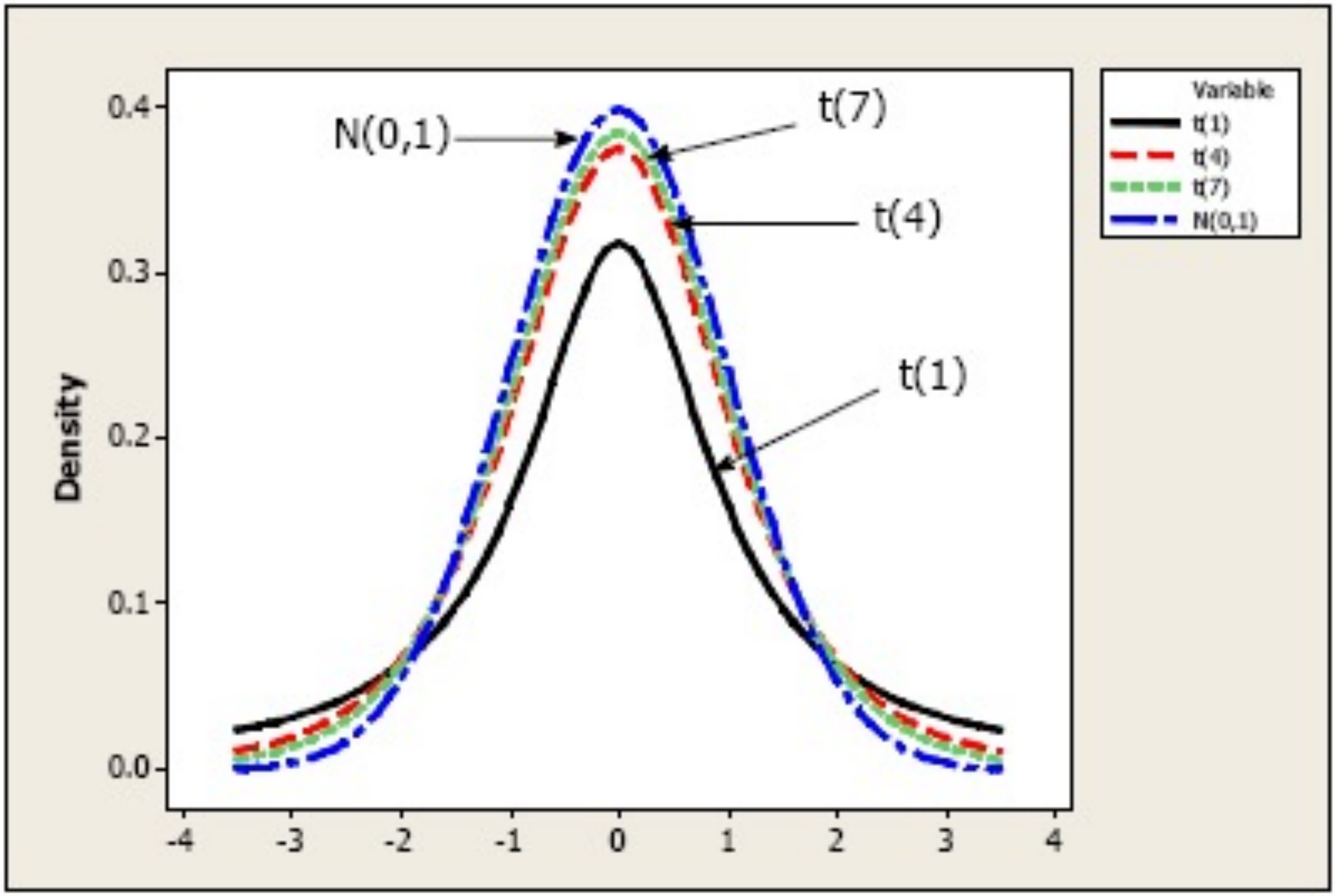
$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$$

- A 95% CI for β_1 is given by $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$

- Note that if the sample size is not large enough, the distribution of T is not necessarily normal, since the assumption that $SE(\hat{\beta}_1) \approx SD(\hat{\beta}_1)$ may not hold.
- In this situation, we model the distribution of T using a family of bell-shaped distributions, called the *t-distributions*.

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

Hypothesis tests to test $\beta_1 = 0$

- $\beta_1 = 0$ is a very important question: is there any linear relationship at all?

- A 95% CI for β_1 is given by $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$: we can use this CI. If 0 is not in this interval, then we reject the null hypothesis of the slope being 0 at the 5% significance level.
- We can set up a test: $H_0: \beta_1 = 0$ vs $H_1:\ \beta_1 \neq 0$ and use the fact that under the null hypothesis,

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

- Let's look at the example from the text on pulse rates after looking at the t-distribution

Example (12.4.3)

slope, intercept, r, p, se_slope=

```
(1.142879681904831,
 13.182572776013345,
 0.6041870881060092,
 1.7861044071652305e-24,
 0.09938884436389145)
```

```
mean_active, sd_active
```

```
(91.29741379310344, 18.779629284683832)
```

```
mean_resting, sd_resting
```

```
(68.34913793103448, 9.927912546587986)
```

c) Find the SD of the residuals.