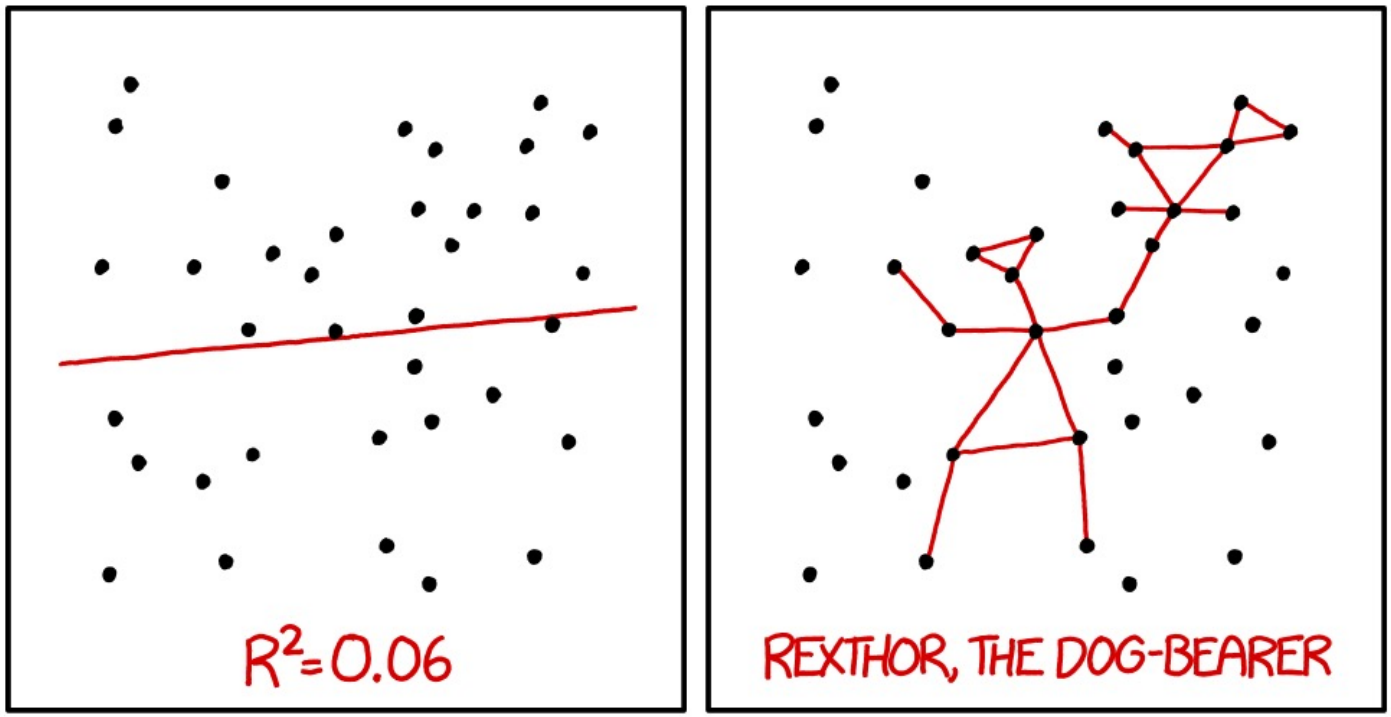
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 40 : 4/30/2021

Chapter 12

Finishing up regression

https://xkcd.com/1725/

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

- 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)

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

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

- Need to estimate σ, which we will do by using the SD of the residuals. The larger the n, the better our estimate of σ

$$\hat{\sigma} = SD(D_1, D_2, \ldots, D_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2},$$

- $D_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ (The D_i are the residuals and estimate the errors)

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

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

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*.

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,

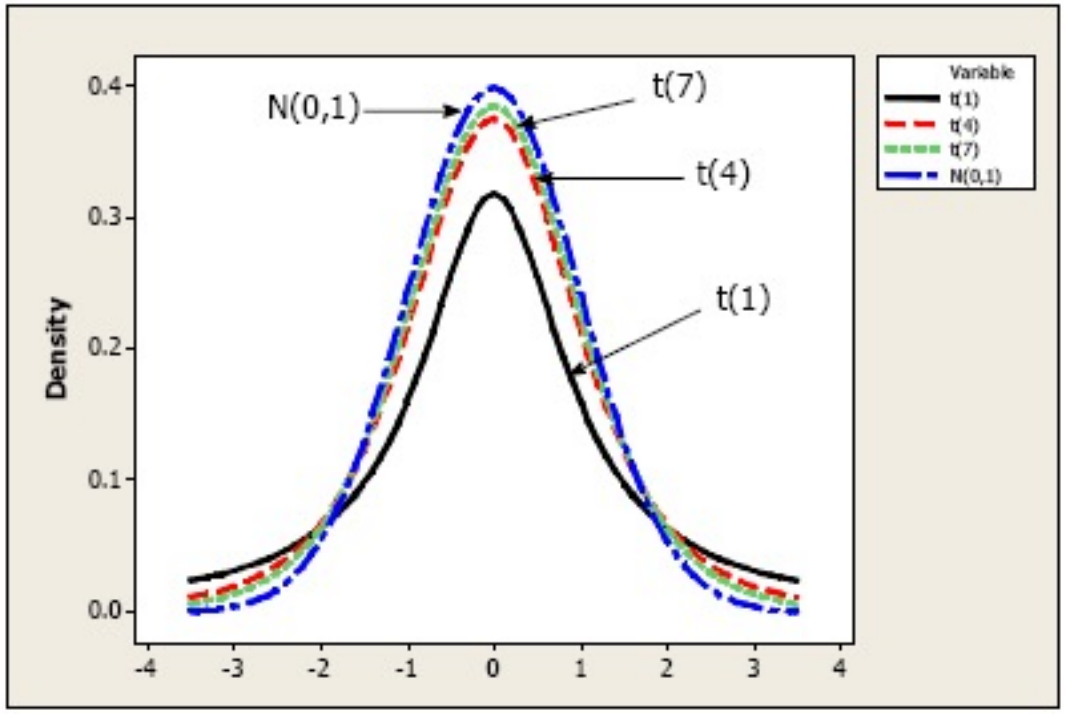
$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim N(0,1)$$

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

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.

(When the sample size n is large, so is $n-2$, so we might as well use the normal curve. When the sample is size is small, using the appropriate t curve gives more accurate answers.)



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

From onlinecourses.science.psu.edu

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.

Quick look back