Emory University

QTM 220 Regression Analysis

Learning Notes

Jiuru Lyu

January 21, 2024

Contents

1 Descriptive Statistics and Binary Covariates

Definition 1.1 (Location). The *location* of the data is where it is. It is about approximating the data by a constant.

$$Y_i \approx \mu, \quad \text{for } i = 1, \ldots, n$$

Example 1.2 Different ways to summarize location: mean, median

Definition 1.3 (Spread). The *spread* of the data is how far it tends to be from is location.

Definition 1.4 (Residuals). Spread summarizes the size of the *residuals* left over after constant approximation. We use $\hat{\varepsilon}$ to denote residuals.

$$\hat{\varepsilon}_i := Y_i - \hat{\mu}.$$

Definition 1.5 (Median Absolute Deviation and Standard Deviation).

- The *median absolute deviation (MAD)* is the median size of residuals.
- The *standard deviation (sd)* is the square root of the mean squared size of residuals.

  Remark 1.1 *The standard deviation is a sort of average in which big residuals count more than smaller ones.*

Definition 1.6 (Distribution). We use *histograms* to summarize the *distribution* of the data.

Remark 1.2 *Distribution of the data tells us more information than location and spread, but less than dot plot.* For example, in this context, dot plot also include the identities of the individuals in addition to the number of people having salary in the range.

Definition 1.7 (Binary Data). *Binary data* only have two options, and we usually denote those two options as 1's and 0's.

Corollary 1.8 : Hence, when drawing a dot plot, everyone falls into either of the two lines representing 1 and 0.

Theorem 1.9 Location of Binary Data
The median is whichever outcome is the most common, and the mean is the proportion of 1's in the data.

Remark 1.3 *Hence, a histogram tells us no more information than* $\hat{\mu}$.

Theorem 1.10 Spread of Binary Data

- Median absolute deviation will always be 0 in a binary case.
- The standard deviation is the square root of the mean squared distance from the mean, and
$$\text{sd} = \sqrt{\hat{\mu}(1 - \hat{\mu})}.$$

Proof 1. The claim concerning MAD is trivial. *Hint: there's only two possible values in the data, so median and MAD should always be the same.*

Now, let's consider the claim on standard deviation.

$$
\begin{aligned}
\mathrm{sd}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\mu})^2 \\
&= \frac{1}{n}\sum_{y:\{0,1\}}\sum_{i:Y_i=y}(Y_i-\hat{\mu})^2 \\
&= \frac{1}{n}\big\{N_1\big(1-\hat{\mu}^2\big)+(n-N_1)\big(0-\hat{\mu}^2\big)\big\} && [N_1 = \text{number of 1's}] \\
&= \frac{1}{n}\big\{N_1\big(1-2\hat{\mu}+\hat{\mu}^2\big)+(n-N_1)\hat{\mu}^2\big\} \\
&= \frac{1}{n}\big\{N_1-2N_1\hat{\mu}+n\hat{\mu}^2\big\} \\
&= \frac{1}{n}\big\{n\hat{\mu}-2n\hat{\mu}\cdot\hat{\mu}+n\hat{\mu}^2\big\} && [N_1 = n\hat{\mu}] \\
&= \frac{1}{n}\big\{n\hat{\mu}-n\hat{\mu}^2\big\} \\
&= \hat{\mu}-\hat{\mu}^2 = \hat{\mu}(1-\hat{\mu}).
\end{aligned}
$$

Therefore, we know

$$
\mathrm{sd} = \sqrt{\hat{\mu}(1-\hat{\mu})}.
$$

■

Remark 1.4 *In binary data, knowing the mean $\equiv$ knowing everything else.*