STA 103 Lecture 12: Correlation and Linear Regression

Instructor: Wookyeong Song

Department of Statistics, University of California, Davis

Sep 3rd, 2025

Population Correlation

- The population correlation coefficient ρ (or ρ_{XY}), defined as

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Remember $\mathrm{Cov}(X, Y)$ shows a association between RV X and Y. If $\mathrm{Cov}(X, Y) > 0$ ($\mathrm{Cov}(X, Y) < 0$), X and Y have positive (negative) association, respectively.

- The correlation ρ_{XY} is basically scaled covariance $\mathrm{Cov}(X, Y)$ since

$$-1 \leq \rho \leq 1.$$

- **If interested**: we can prove it by using Cauchy-Schwartz Inequality.

Correlation

- Given paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$, the sample correlation coefficient is defined as:

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

- **Properties**:
  - $r \in [-1, 1]$
  - $r > 0$: Positive association; $r < 0$: Negative association
  - $r = 0$: No linear association

- This (Pearson) sample correlation coefficient is a measure of the linear relationship between two variables.

Correlation Does Not Imply Causation

It is critical to understand that correlation quantifies association, **not** causality.

- **Confounding**: A third variable may influence both X and Y.
- **Reverse Causation**: The direction of effect may be opposite to what is assumed.
- **Spurious Correlation**: A high correlation can arise purely by coincidence or due to underlying trends.

Correlation Does Not Imply Causation

- **Example (Confounding in Online Ad Campaigns)**: Suppose a data analyst observes a strong positive correlation between the number of website visits and sales revenue.
- They conclude: "**More website visits cause higher sales.**"
- However, a third variable, such as seasonal promotions, may be the true driver. This is because during promotional months, both website traffic and sales increase.
- Once promotion months are accounted for, the direct relationship between web visits and sales may vanish.

Correlation Does Not Imply Causation

- **Confounding variable**: Promotions
- **Key Insight**: Correlation between visits and revenue is real, but it is induced by a confounder. Without adjusting for promotions, the conclusion about causality is invalid.
- **(Advanced)**: Thus, statistical association must be interpreted cautiously. Establishing causality requires:
  - Randomized controlled experiments (e.g., A/B testing).
  - Controlling for confounders through stratification or multivariate regression.
  - Using causal inference frameworks (e.g., potential outcomes, DAGs).

Regression

- There is some variable Y and you want to see if another variable X can explain the variability in Y.

- Postulate a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where ε is independent of X, $\varepsilon \sim N(0, \sigma^2)$.

- For a given set of X values, $[X_1, X_2, \ldots, X_n]$, you measure the associated Y values $[Y_1, Y_2, \ldots, Y_n]$.

- These two lists are used to construct an estimate $\hat{\beta}_1$ of parameter β_1.

- Now $\hat{\beta}_1$ can be used to test $H_0 : \beta_1 = 0$, which means Y has no linear relationship with X.

Simple Linear Regression

- Simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \text{ i.i.d.}$$

- β_0: Intercept (mean value of Y when $X = 0$)
- β_1: Slope (expected change in Y per unit increase in X)
- ε_i: Random error term, representing unobserved variation

- Y is called response variable (or dependent variable) and X is predictor (or independent variable).

Estimation of Simple Linear Regression

- The least squares method finds $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the loss function:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- Once you find $\hat{\beta}_0$ and $\hat{\beta}_1$, the fitted (predicted) response is given by

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

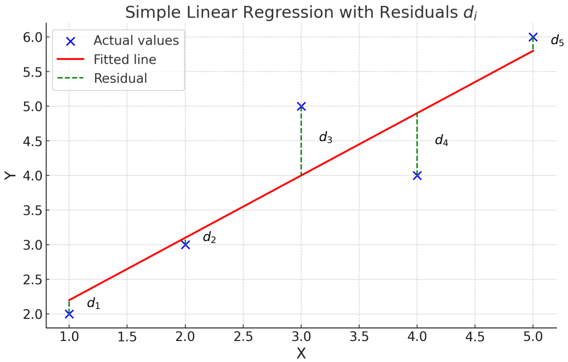
- The term

$$d_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

  is called the *residual*, the difference between the observed response Y_i and the predicted value $(\hat{\beta}_0 + \hat{\beta}_1 X_i)$.

- What we want is to make the sum of square of the difference **as small as possible** for a good fit.

Estimation of Simple Linear Regression



- **Summary**: We minimize the sum of squared errors to find the line that
  - best represents the trend in the data
  - gives the smallest overall prediction error

Point Estimator of β_0, β_1, and σ^2

- We have closed-form solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

- **Interpretation**:
  - $\hat{\beta}_1$ tells how much Y is expected to change for a one-unit increase in X.
  - $\hat{\beta}_0$ is the expected value of Y when $X = 0$ (may not always be meaningful depending on context).

- The variance σ^2 is estimated by

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2.$$

Inference on β_1

- What we are interested in the most is to do inference on the regression effect parameter β_1:

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

- Sampling variability of $\hat{\beta}_1$ given $X_1, \ldots, X_n$ are fixed. We expect it to be

$$\hat{\beta}_1 \sim N(E(\hat{\beta}_1), \mathrm{Var}(\hat{\beta}_1)).$$

- The mean $E(\hat{\beta}_1)$ is

$$E(\hat{\beta}_1) = \beta_1, \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \approx \frac{s^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Inference on β_1

- **Example**: Suppose $Y = \beta_0 + \beta_1 X + \varepsilon$, ε and X are independent, and $\varepsilon \sim N(0, \sigma^2)$. Based on random samples $(X_1, Y_1), \ldots, (X_{25}, Y_{25})$, one obtained the following statistics:

$$\bar{X} = 0.0335, \quad \bar{Y} = -0.7713,$$
$$\sum_{i=1}^{25}(X_i - \bar{X})^2 = 27.3, \quad \sum_{i=1}^{25}(X_i - \bar{X})Y_i = -269.2$$

- **Question 1**: Find $\hat{\beta}_1$ and $\hat{\beta}_0$:

- **Answer**:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \frac{-269.2}{27.3} = -9.86.$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = -0.7713 - (-9.86) \times 0.0335 = -0.44.$$

Inference on β_1

- **Question 2**: If $\hat{\sigma}^2 = s^2 = 5.7711$, approximate $\mathrm{SE}(\hat{\beta}_1)$.

- **Answer**:

$$\mathrm{SE}(\hat{\beta}_1) = \mathrm{sd}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum(X_i - \bar{X})^2}} \approx \sqrt{\frac{s^2}{\sum(X_i - \bar{X})^2}}$$
$$= \sqrt{\frac{5.7711}{27.3}} = 0.46.$$

Inference on β_1

- **Question 3**: Find an approx 95% confidence interval for β_1.

- **Answer**: For the confidence interval,

$$\beta_1 \in \hat{\beta}_1 \pm z_{0.975} \times \mathrm{SE}(\hat{\beta}_1).$$

In Question 1, we found that $\hat{\beta}_1 = -9.86$. In Question 2, we found that $\mathrm{SE}(\hat{\beta}_1) = 0.46$. Then,

$$\beta_1 \in -9.86 \pm 1.96 \times 0.46 = [-10.78, -8.94].$$

Inference on β_1

- **Question 4**: Is there evidence that the slope is negative $\beta_1 < 0$ with significance level $\alpha = 0.01$.

- **Answer**: Formalize the hypothesis testing:

$$H_0 : \beta_1 \geq 0, \quad H_a : \beta_1 < 0.$$

- The Z-score of point estimator $\hat{\beta}_1$ is

$$Z = \frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\mathrm{sd}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{0.46} = -21.43.$$

- Thus, if $\beta_1 = 0$, $\hat{\beta}_1$ would be -21.43 $\mathrm{sd}(\hat{\beta}_1)$'s away from what we expect. This is basically impossible. Formally, compare p-value $P(Z < -21.43) < 10^{-12}$ and significance level $\alpha = 0.01$. We reject the null H_0 and take the alternative H_a.

Inference on β_1

- **Question 5**: If I found another sample (X_{26}, Y_{26}) and told you $X_{26} = 0.05$. What is your best guess for Y_{26}?

- **Answer**: The prediction of new sample (X_{26}, Y_{26}) based on previous observations, $(X_1, Y_1), \ldots, (X_{25}, Y_{25})$, is

$$\hat{Y}_{26} = \hat{\beta}_0 + \hat{\beta}_1 X_{26} = -0.44 - 9.86 \times 0.05 = -0.933.$$

Inference on β_1

- **Real Case Example (Advertising and Sales)**: A marketing analyst regresses monthly sales (in $\$1000s$, Y) on advertising spending (in $\$1000s$, X), based on data from 12 months. The fitted model is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 80 + 5.2X.$$

- We have the following information:

$$\hat{\beta}_1 = 5.2, \quad \mathrm{SE}(\hat{\beta}_1) = 1.4.$$

- **Question 1**: Conduct hypothesis testing that advertising spending is associated with sales at significance level $\alpha = 0.05$, i.e.,

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

Inference on β_1

- **Answer**: The Z-score for $\hat{\beta}_1$ is

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{5.2}{1.4} = 3.71.$$

- Since p-value $P(|Z| > 3.71) < 0.05$, then we reject the H_0. There is strong evidence that advertising spending is associated with sales.

Inference on β_1

- **Question 2**: Find the 95% confidence interval for β_1.

- **Answer**:

$$\beta_1 \in \hat{\beta}_1 \pm z_{0.975} \times \mathrm{SE}(\hat{\beta}_1) = 5.2 \pm 1.96 \times 1.4 = [2.456, 7.944].$$

Takeaway

- A statistically significant slope suggests a non-random relationship, but effect size and precision matter.
- Confidence intervals provide useful bounds for expected change—not just a yes/no decision.
- Always interpret estimates in real-world units to communicate value to stakeholders.