H. Regression

1. Setup
   a. One response variable \( Y \)
   b. One or more explanatory variables \( X, W, \ldots, U \)
      i. Presume that you set these
      ii. Consider the distribution of \( Y \) without making any allowance for randomness in \( X, W, \ldots, U \).
   c. Related by \( Y = f(X, W, \ldots, U) + \text{error} \)

2. Observe \( n \) times with different values for \( X, W \), etc.
   a. \( X_1, X_2, \ldots, X_n \)
   b. \( W_1, W_2, \ldots, W_n \)
   c. \( Y_1, Y_2, \ldots, Y_n \)

3. Objective: figure out \( f \)

4. Linear case: \( f(X, W, \ldots, U) = \alpha + \beta X + \cdots + \gamma U \)

5. Simple Linear case: \( f(X) = \alpha + \beta X \)
   a. Interpretation: \( \beta \) is average change in \( Y \) for unit change in \( X \)
   b. Error behavior:
      i. Population mean is zero
Lecture 21

ii. Constant spread $\sigma$

iii. Independent

iv. Approximately Normal

c. Behavior of $Y$ knowing $X$
   
   i. Population mean is $\alpha + \beta X$

   ii. SD $\sigma$

   iii. Approximately normal

d. Recall estimates:
   
   i. $b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

   ii. $a = \bar{Y} - b\bar{X}$

   iii. $s_e^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{(n - 2)}$

   - Note in this case, the numerator is $n - 2$ rather than $n - 1$.

   - For example, if $n = 2$, then every data set may be fit exactly by a line, and so the numerator of $s_e^2$ will always be zero.

   - So the denominator should also be zero, to show that we don’t have any information about the variation of the $Y$’s about the regression line.

e. Objective: Learn about $\alpha, \beta$: Tests, CI’s.

   i. Sampling distribution for $b$:
• Mean

  ▶ Population means combine linearly

  ▶ Population mean of

\[
\frac{\sum_{i=1}^{n} (X_i - \bar{X})Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(\alpha + \beta \bar{X} + \beta(X_i - \bar{X}))}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(\alpha + \beta \bar{X}) + \beta \sum_{i=1}^{n} (X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = 0 + \beta
\]

• SD

  ▶ Population variance of \(\frac{\sum_{i=1}^{n} (X_i - \bar{X})Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2\sigma^2/\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^2 = \frac{\sigma^2/\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\)

  ▶ Standard error \(s_e/\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}\)

• Shape

  ▶ If all errors start out normal, \(b\) is exactly normal, since sums of (multiples of) normals are still normal regardless of \(x\)’s.
If all \((X_i - \bar{X})\)'s were same, for large samples, \(b\) approx normal: central limit theorem

Normality also holds in inbetween case, with approximately normal errors and moderately similar \(X\)'s.

ii. Statistic:

- \(\frac{(b - \beta)}{(\sigma/\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\) approximately standard normal
- \(\frac{(b - \beta)}{(s_e/\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\) approximately \(t(n - 2)\).
- Usual construction for confidence intervals and tests.

iii. Example: Fig. 37/ shows association between daily changes in two financial indices.

- Fig. 38/ shows box plots for one of these, with data grouped by rounding other.
- We often determine where to take samples, so we determined \(X\)

iv. Example: Fig. 39/ shows measurements of the oxygen content of steel in a bar as a function of the distance along its length.

- Fig. 40/ shows regression lines when new data are randomly generated.
**Fig. 37: Changes in Dow Jones Average and Standard and Poor’s 500 Index**

- Lines through means
- Line of Symmetry (SD line)
- Regression Line

![Graph showing the correlation between changes in Dow Jones Average and Standard and Poor’s 500 Index.](image)

Correlation = 0.958

- When we don’t determine $X$, a more sophisticated analysis might be able to squeeze more out of data set.
Fig. 38: Distribution of Changes in Dow Jones Average, Conditional on Changes in Standard and Poor’s 500 Index
Fig. 39: Oxygen Content of a Steel Bar as a Function of Distance from the End

Regression Line

Correlation = 0.204
Fig. 40: Potential Regression Lines for Steel Example

Lines formed by adding random error to observations