Applied Regression Analysis and Other Multivariable Methods

Chapter 4 - 5.6

Yi Zhou

2015-10-15

Outline

4-1 Preview

- Regression analysis: one of the methods in multivariate techniques.
  - Wide applicability.
  - Simplest to implement.
- Independent variables/predictors $(X_1, X_2, ...)$, dependent variable/response (Y).
- Examples
  - The relationship of Y and X_i (if when controlling for the effects of other variables).
  - The models/equations of Y and X_i.
  - Significant X_i, interactive effects of X, regression coefficients of X_i.
  - ...
- Francis Galton: regression toward the mean.

4-2 Association vs. Causality

- Bias.
- Statistically significant association does not establish a causal relationship.
- Causality cannot be established by statistical analyses. Association can be well quantified in a statistical model.
- The criteria by Bradfor Hill (1971).

4-3 Statistical vs. Deterministic Model

- Statistical model:
  - **Regression analysis**.
  - Discrimination analysis.
  - Factor analysis.
  - Analysis of variance/covariance.
  - ...
- Deterministic model assumes an ideal setting.
- Statistical model allows for the possibility of error.

5-1 Preview

- Considering the simplest form: $Y \sim X$,
  $X : X_1, X_2, ..., X_n$
  $Y : Y_1, Y_2, ..., Y_n$,
  one dependent variable and one independent variable.
- Problem: to find the curve that best fits the data, closely approximating the true relationship between X and Y.
- X and Y can yield a scatter diagram in two dimensional space (Figure 5-1).

5-2 The problem and general strategy

- Basic questions:
  - What is the appropriate mathematical model?
  - How to determine the best-fitting model?
- General strategies:
  - Forward method (most commonly used):
    Simplicity $\rightarrow$ complexity
  - Backward method:
    Complexity $\rightarrow$ simplicity
  - Details will be discussed in **Chapter 16**.

5-3 Mathematical Properties for a **Straight Line**

$$y = \beta_0 + \beta_1 x \tag{5.1}$$

- β_0: y-interception
- β_1: slope: the amount of change in y for each 1-unit change in x.
- Equation(5.1) is in the **mathematical** context, which does NOT consider y as a random variable.

5-4 Statistical Assumption for a **Straight-line Model** I

$$\mu_{(Y|X)} = \beta_0 + \beta_1 X \tag{5.2}$$

$$Y = \beta_0 + \beta_1 X + E \tag{5.3}$$

- Equation(5.2) and (5.3) are in the **statistical** context.
- Statement of assumptions: **HEIL GAUSS**
- X is a *fixed* known variable.
  Y is a *random* variable with observed value.
  $\mu_{(Y|X)}$ is the mean value of Y.
- β_0, β_1 are the parameters.

5-4 Statistical Assumption for a **Straight-line Model** II

$$
\begin{aligned}
E &= Y - (\beta_0 + \beta_1 X) \\
&= Y - \mu_{Y|X}
\end{aligned}
$$

- E is the error component with mean=0 and variance=σ^2.
  The *unobservable random* variable.
  If $E \sim N$, then $Y \sim N$.
- The point estimator $\hat{E}$(residual)

$$
\begin{aligned}
\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X + \hat{E} \\
\hat{E} &= Y - \hat{Y} \\
&= Y - (\hat{\beta}_0 + \hat{\beta}_1 X)
\end{aligned}
$$

5-5* Determining the Best-fitting Straight Line I

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{5.2.1}$$

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

- The residual sum of squares, the sum of squares due to error (SSE$=Q(\hat{\beta}_0, \hat{\beta}_1)$)
- The least-square method (or Ordinary Least Squares, OLS): $\dfrac{\partial Q}{\partial \hat{\beta}_0} = 0, \dfrac{\partial Q}{\partial \hat{\beta}_1} = 0$, to get $\hat{\beta}_0, \hat{\beta}_1$ as the minimum estimate of β_0 (eq.5.5), β_1 (eq.5.6).
- *Property 1*: $\hat{\beta}_0, \hat{\beta}_1$ are unbiased estimators.

5-5* Determining the Best-fitting Straight Line II

The centralization of eq.5.2.1 is of the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 (X_i - \bar{X}) \tag{5.2.2}$$

- $\dfrac{\partial Q}{\partial \hat{\beta}_1} = 0$, to get the minimum estimate of β_1 (eq.5.4=eq.5.6).
  $\dfrac{\partial Q}{\partial \hat{\beta}_0} = 0$, then $\hat{\beta}_0 = \bar{Y}$.
- *Property 2*: $Var(\hat{\beta}_0) = \dfrac{\sigma^2}{n}, Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$
  $X_1, X_2, \ldots$ would be better to be spread for the minimum variance of $\hat{\beta}_1$.
- *Property 3*: $Cov(\hat{\beta}_0, \hat{\beta}_1) = 0$

5-5* Determining the Best-fitting Straight Line III

- variances of $\hat{\beta}_0, \hat{\beta}_1$:
- A central property of OLS:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$
$$TSS = MSS + RSS$$

- TSS: total sum of squares
- MSS: model sum of squares
- RSS: residual sum of squares (SSE)
- A drawback of OLS is sensitivity to outliers

5-5* Determining the Best-fitting Straight Line IV

- The minimum-variance method
- Maximum likelihood
  See Chapter 12

5-6 Measure of the Quality of the Straight-line Fit and Estimate of σ^2

- Residual: $\hat{E}_i = Y_i - \hat{Y}_i$, the lower the better.
  Error: E with mean(E)=0, var(E)=σ^2.
- *Property 1*: the estimate of σ^2:
  $S^2_{Y|X} = \dfrac{RSS}{n-2} = \dfrac{SSE}{n-2} = \dfrac{\sum_{i=1}^{n} \hat{E}_i}{n-2}$ (eq.5.10) is an unbiased estimator.
- *Property 2*: given $E \sim N(0, \sigma^2)$, then $\dfrac{SSE}{\sigma^2} \sim \chi^2(n-2)$
- Residual plot: residuals on the y-axis and the independent variable on the x-axis.
  If the points in a residual plot are randomly dispersed around x-axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.