Multi-Variable Linear Regression

Tianhang Zheng

https://tianzheng4.github.io

Analyzing Least Squares Method (Unbiased)

$$E_{\epsilon_i}[\sum_i (x_i - \overline{x})] = 0$$

$$E_{\epsilon_i}\left[\frac{\sum_i (x_i - \overline{x})(\epsilon_i + \beta_1(x_i - \overline{x}))}{\sum_i (x_i - \overline{x})^2}\right]$$

$$= \frac{1}{\sum_i (x_i - \overline{x})^2} E_{\epsilon_i}\left[\sum_i (x_i - \overline{x})(\epsilon_i + \beta_1(x_i - \overline{x}))\right]$$

$$= \frac{\beta_1}{\sum_i (x_i - \overline{x})^2} E_{\epsilon_i}[\sum_i (x_i - \overline{x})^2] = \beta_1$$

Why unbiased?

Analyzing Least Squares Method

The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$SE^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2} \qquad SE^2(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}\right]$$

$$\sigma^2 = Var(\epsilon)$$

$\hat{\beta} \sim N(\beta, SE^2(\hat{\beta}))$ Why Gaussian distribution?

Confidence Level

$\hat{\beta} \sim N(\beta, SE^2(\hat{\beta}))$ means that $\beta \sim N(\hat{\beta}, SE^2(\hat{\beta}))$

A 95% confidence interval is defined as a range of values with 95% probability, and the interval for the least square method is

$$[\hat{\beta} - 2SE(\hat{\beta}), \hat{\beta} + 2SE(\hat{\beta})]$$

There is 95% probability that this interval contains the true β

Multiple Linear Regression

Linear Regression with multiple predictors (Assume the ideal model is a linear function)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

β_0 is interpreted as the average effect of one unit increase in X_i on Y

Parameter Estimation

The objective is to learn (estimate) $\beta_0, \beta_1 \ldots, \beta_p$

The estimates of $\beta_0, \beta_1 \ldots, \beta_p$ are denoted by $\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_p$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_1 x_p$$

$e = y - \hat{y}$ is the residual

Least Square Method (Solved by Software)

$$RSS = \sum_i e_i^2 = (y_i - \hat{y}_i)^2$$

$$\min_{\hat{\beta}_0 \sim \hat{\beta}_p} RSS = \sum_i e_i^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right)^2$$

Take the derivative and set it as 0 to estimate $\hat{\beta}$

Least Square Method (Solved by Matrix)

$$y = \begin{bmatrix} 6 \\ 11 \\ 4 \\ 3 \\ 5 \\ 9 \\ 10 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 4 & 5 & 4 \\ 1 & 7 & 2 & 3 \\ 1 & 2 & 6 & 4 \\ 1 & 1 & 9 & 6 \\ 1 & 3 & 4 & 5 \\ 1 & 7 & 3 & 4 \\ 1 & 8 & 2 & 5 \end{bmatrix}$$

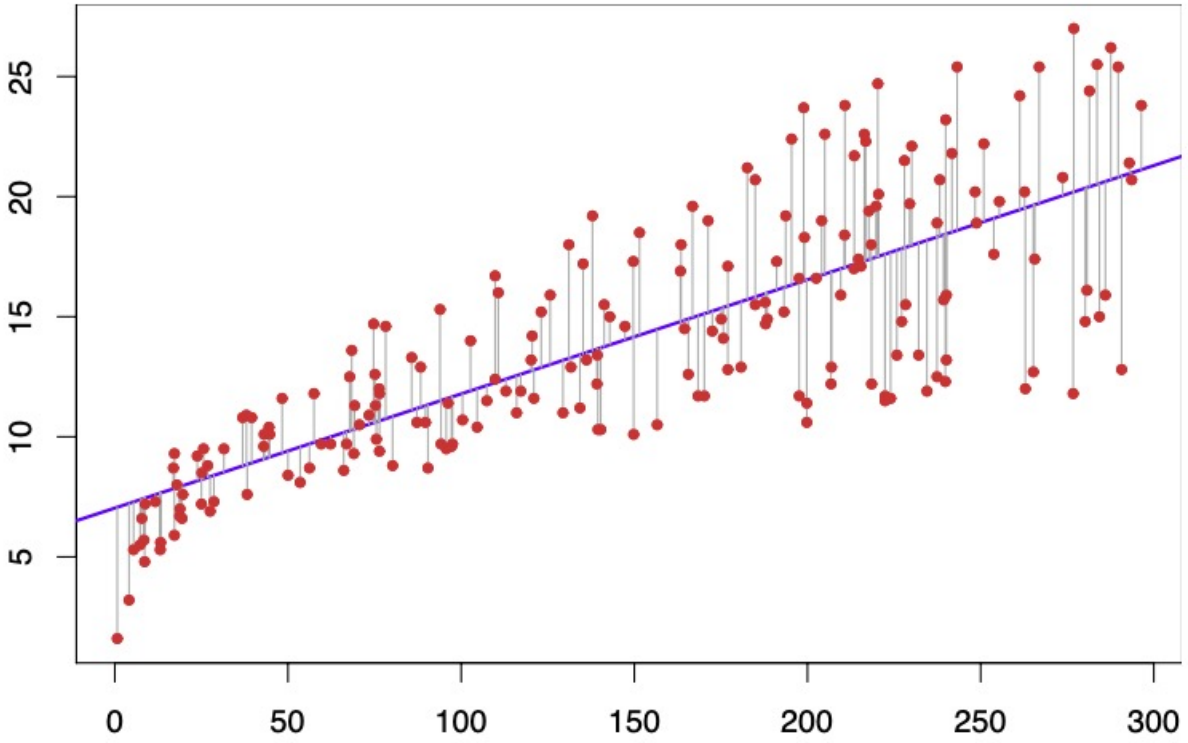
Estimation:

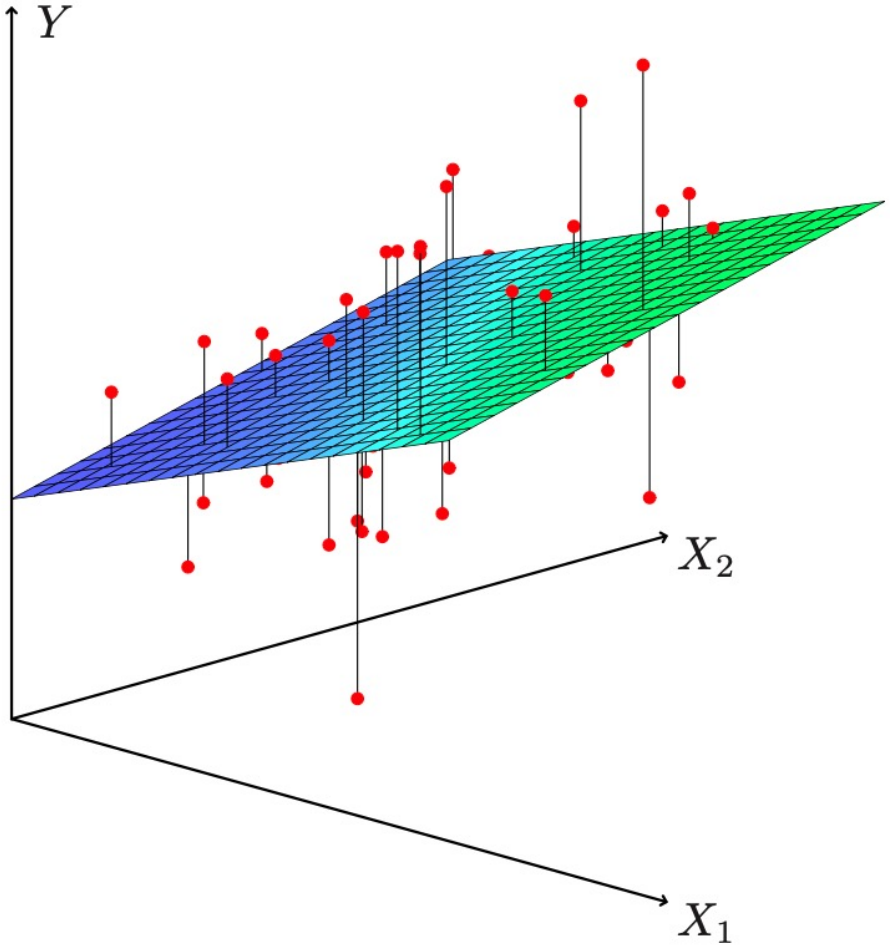
$$b = (X'X)^{-1}X'y$$

$$b = [\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_3]$$

$$y = Xb + e$$

Visualization





Hypothesis Testing

$$H_0: \beta_j = 0$$

We can use t-statistics $$t = \frac{\hat{\beta}_j - 0}{\widehat{SE}(\hat{\beta}_j)}$$

$$\widehat{\boldsymbol{\beta}} = \left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathbf{t}\mathbf{X})^{-1})$$

P value

A p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis (H_0) is correct.

$$\text{P-value} = P[T > |t|]$$

If p-value is large, we tend to accept H_0. Otherwise, we tend to reject it.

Hypothesis Testing

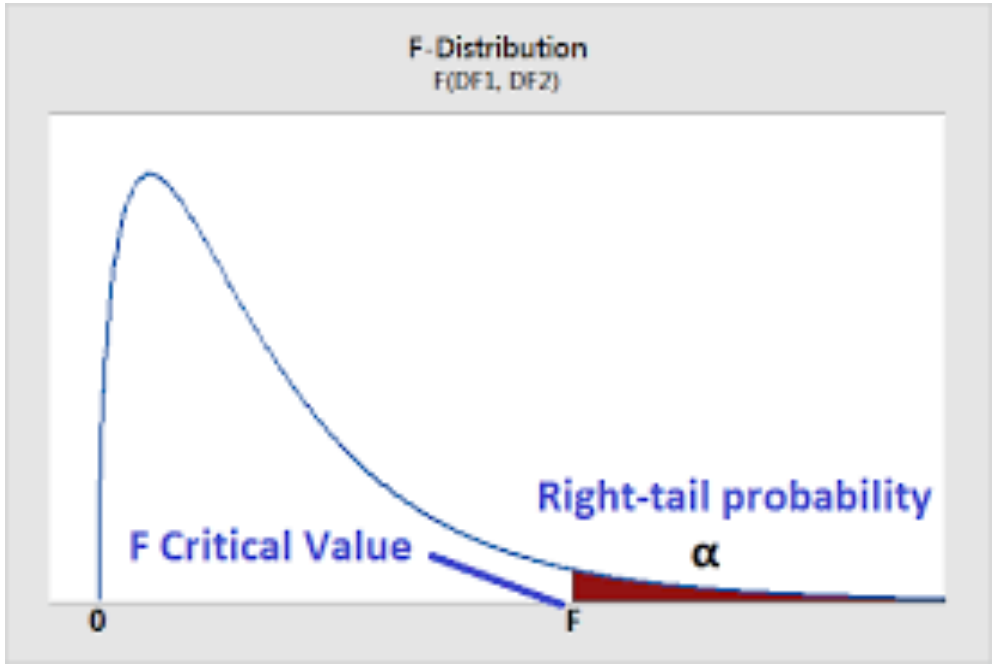
$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

We need to use F-statistics

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Hypothesis Testing

$$df_1 = p \quad df_2 = n - p - 1$$



	df_1=1	2	3	4	5	6	7	8
df_2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85

Variable Selection

Forward Selection

Backward Selection

Colinearity

Forward Selection

Begin with the null model — a model that contains an intercept but no predictors.

Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.

Add to that model the variable that results in the lowest RSS amongst all two-variable models. (Continue until some stopping rule is satisfied)

Backward Selection

Start with all variables in the model.

Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.

Continue to fit and remove until a stopping rule is reached

Colinearity

Two or more variables are exactly correlated.

The parameters are not fixed and will be affected by small changes in the training data

Increase the difficulty for interpretation

Variable Interaction / Nonlinear Effects

Consider interaction between X_1 and X_2.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Consider nonlinear effects of X_1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

Next

Friday: First Assignment---Linear Regression

Next Week: Classification and Regression

Q & A