AID-521 Mathematics for Data Science

Module: Statistics | Lecture: 7

LINEAR REGRESSION MODELS

simple regression, multiple regression (matrix notation), OLS estimator, parameter inference, prediction, diagnostics

Introduction

- $\rightarrow$ Understanding relationships between Y and X
- $\rightarrow$ Three step procedure of (linear) regression modeling:-
  - $\rightarrow$ Specify a model $Y = f(X) + \epsilon$
  - $\rightarrow$ Estimate the model, perform model diagnostics. Select the best model.
  - $\rightarrow$ Use the model for parameter inference and prediction.

Simple Regression & Correlation Analysis

...

Multiple Regression

A multiple linear regression model relating a random response/dependent variable Y to a set of predictor/independent variables $X_1, ..., X_k$ is an equation of the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

where $\beta_0, \beta_1, ..., \beta_k$ are unknown parameters, and ϵ is the r.v. representing an error term.

→ $E[\epsilon] = 0$

→ $Var(\epsilon) = \sigma^2$

→ $\epsilon \sim \mathcal{N}$

Multiple Regression -- Visualization

... linear, iid, exogeneity, homoskedastic, normal, invertible

Multiple Regression

→ Estimators $\hat{Y}$, and $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$

$$\hat{Y} \;=\; \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

→ The sum of squares for errors (SSE) or sum of squares of the residuals for all of the n data points is

$$SSE \;=\; \sum_{i=1}^{n} \hat{e}^2 \;=\; \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)]^2$$

→ The least-squares line is a line of the form $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ for which the error sum of squares SSE is minimum.

Least Squares Estimator

$$Y = \beta X + \epsilon$$

is the matrix notation for n observations

$$Y_i = \beta_0 + \beta_1 X_{1,\,i} + \cdots + \beta_k X_{k,\,i} + \epsilon_i, \quad i = 1, \ldots, n.$$

The estimator of β vector is

$$\hat{\beta} \mid X = (X'X)^{-1}X'Y$$

Properties of Least Squares Estimator

Theorem 8.2.1 *Let $Y = \beta_0 + \beta_1 x + \varepsilon$ be a simple linear regression model with $\varepsilon \sim N(0, \sigma^2)$, and let the errors ε_i associated with different observations $y_i (i = 1, \ldots, N)$ be independent. Then*

(a) *$\hat{\beta}_0$ and $\hat{\beta}_1$ have normal distributions.*

(b) *The mean and variance are given by*

$$E\left(\hat{\beta}_0\right) = \beta_0, \quad Var\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2,$$

and

$$E\left(\hat{\beta}_1\right) = \beta_1, \quad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}},$$

where $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$. In particular, the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively.

Gauss-Markov Theorem: The least-squares estimators are best linear unbiased estimators.

Estimation of Error Variance σ^2

→ The greater the variance σ^2 of the random error ϵ, the larger will be the errors in the estimation of model parameters.

→ An unbiased estimator of the error variance σ^2 is

$$\hat{\sigma^2} = \frac{SSE}{n-2} := MSE.$$

Inferences on Least Squares Estimator

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1)$$

$$t_{\beta_1} = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\frac{SSE}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim T_{(n-2)}$$

- → Hypothesis tests for β_1 (and similarly for β_0) can be obtained accordingly.
- → Confidence intervals for β_1 (and similarly for β_0) can be obtained accordingly.

Inferences using ANOVA

It can be verified that (see Exercise 8.3.7)

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 .$$

Denoting

$$SST=\sum_{i=1}^{n}(y_i-\overline{y})^2 ,\quad SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 ,\ \text{and } SSR=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 ,$$

the foregoing equation can be written as

$$SST=SSR+SSE.$$

Note that the total sum of squares (SST) is a measure of the variation of y_i's around the mean $\overline{y}$, and SSE is the residual or error sum of squares that measures the lack of fit of the regression model. Hence, SSR (sum of squares of regression or model) measures the variation that can be explained by the regression model.

Inferences using ANOVA

Table 8.1 ANOVA Table for Simple Regression

Source of variation	**Degrees of freedom**	**Sum of squares**	**Mean sum of squares**	***F*-ratio**
Regression (model)	1	SSR	$MSR = \frac{SSR}{d.f.}$	$\frac{MSR}{MSE}$
Error (residuals)	$n-2$	SSE	$\frac{SSE}{d.f.}$	
Total	$n-1$	SST		

To test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, we could use the statistic

$$\frac{MSR}{MSE} \sim F(1, n-2).$$

Confidence Interval for Prediction of Y

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

A $(1-\alpha)100\%$ prediction interval for Y is

$$\hat{Y} \pm t_{\frac{\alpha}{2},\, n-2}\, S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

where $S^2 = \frac{SSE}{n-2}$.

Regression Diagnostics

- → Linearity
- → I.I.D. errors
- → Exogeneity
- → Homoskedasticity
- → Normal distribution (small sample size)
- → Multicollinearity (Full rank)