BS1820: Maths and Statistics Foundations for Analytics

Statistics 3

Zhe Liu

Imperial College Business School

Email: zhe.liu@imperial.ac.uk

Outline

3.1 Regression Analysis

Regression model:

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon$$

- $\mathbf{X}$: the **independent variables** (predictors, explanatory variables)
- Y: the **dependent variables** (outcome, response)
- $\boldsymbol{\beta}$: the **unknown parameters** (coefficients)
- ϵ: the **error term** (residual)

Key assumptions:

1. Errors have mean zero: $\mathsf{E}(\epsilon_i) = 0$
2. Errors have constant finite variance (homoscedastic): $\mathsf{Var}(\epsilon_i) = \sigma^2 < \infty$
3. Errors are uncorrelated: $\mathsf{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$
4. Independent variables are linearly independent

Remark: Assumptions 1–3 form the Gauss–Markov theorem: OLS estimator is the MVUE for $\boldsymbol{\beta}$.

3.1 Regression Analysis

Regression is a supervised learning problem where the goal is to estimate

$$\mathsf{E}[Y|\mathbf{X}]$$

from training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ with $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$.

E.g.

- Predicting sales from advertising costs.
- Predicting stock returns from fundamental and stock-specific factors.
- Predicting e-commerce revenue based on customer features, cookies, etc.
- Predicting crime rates in a neighborhood from various socioeconomic factors.

Common regression models:

1. Linear regression
2. Logistic regression
3. Nonlinear regression
4. Nonparametric models

3.2 Linear Regression

In a linear regression model the dependent variable Y is a RV that satisfies

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

where $\mathbf{X} = (X_1, \ldots, X_p)$ are the independent variables and ϵ is the error term.

Linear model, therefore, implicitly assumes $\mathbb{E}[Y|\mathbf{X}]$ is approximately linear in $\mathbf{X}$.

The independent variables are numerical inputs

- or possibly transformations (e.g. product, log, square root, $\phi(x)$) of the "original" numerical inputs
- the ability to transform provides considerable flexibility

The X_i's can also be used as 0–1 dummy variables that encode the levels of categorical inputs

- an input with K levels would require $K-1$ dummy variables, $X_1, \ldots, X_{K-1}$

3.3 Ordinary Least Squares

Given training data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ with $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, the **ordinary least squares (OLS) estimator** $\hat{\beta}$ minimizes the residual sum of squares (RSS):

$$\min \sum_{i=1}^{n} \epsilon_i^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 ,$$

where

$$\mathbf{y} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} := \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

The solution is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

3.3 Ordinary Least Squares

The geometry of OLS:

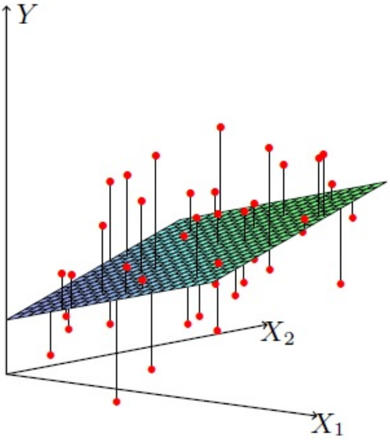
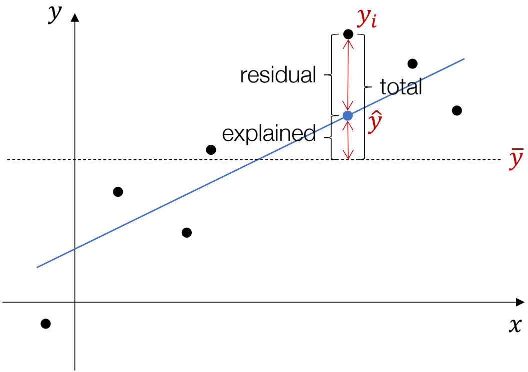


Figure 3.1 from *The Elements of Statistical Learning*: Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.

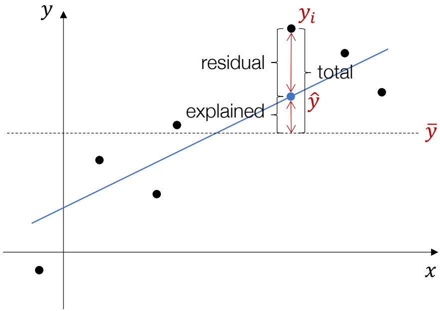
3.4 Analysis of Variance in OLS

Partition of sums of squares: (Proof is beyond scope).

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{total sum of squares (TSS)}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{residual sum of squares (RSS)}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{explained sum of squares (ESS)}}$$



3.4 Analysis of Variance in OLS



$$\mathsf{TSS} = \mathsf{RSS} + \mathsf{ESS}$$

The $\mathbf{R}^2$ **statistic** is a measure of the linear relationship between $\mathbf{X}$ and Y:

$$R^2 := \frac{\mathsf{ESS}}{\mathsf{TSS}} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}.$$

R^2 always lies in the interval $[0, 1]$ with values closer to 1 being "better"

- in physical science applications we look for R^2 close to 1
- in social science an $R^2 \approx 0.1$ might be deemed good

3.4 Analysis of Variance in OLS

Analysis of Variance (ANOVA):

Source	df	Sum of Squares (SS)	Mean Square (MS)
Model	p	ESS $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	ESS$/p$
Residual	$n-p-1$	RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	RSS$/(n-p-1)$
Total	$n-1$	TSS $= \sum_{i=1}^{n}(y_i - \bar{y})^2$	TSS$/(n-1)$

Remark: Alternative notation:

- SSR (sum of squares due to regression) for ESS
- SSE (sum of squares due to error) for RSS
- MSE (mean squared error) for RSS$/n-p-1$

(Unbiased) estimator of error variance $\sigma^2 = \mathbf{Var}(\epsilon)$

$$\hat{\sigma}^2 = \text{MSE} := \frac{\text{RSS}}{n-p-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}.$$

3.4 Hypothesis Testing: Significance of Regression

This test checks the significance of the whole regression model:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_1 : \text{at least one } \beta_i \neq 0$$

Assume $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ (simultaneously satisfies the Gauss-Markov Assumptions), we can compute the **F-statistic**

$$F = \frac{\mathsf{ESS}/p}{\mathsf{RSS}/(n-p-1)} > 0$$

which follows an $F_{p,n-p-1}$ distribution under H_0. Hence

- large values of F constitute evidence against H_0
- we can compute the p-value $= \mathsf{Prob}(F_{p,n-p-1} \geq F)$

3.5 Hypothesis Testing: Individual Coefficients

This test checks the significance of an individual regression coefficient:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Assume $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ (simultaneously satisfies the Gauss-Markov Assumptions), we can compute the T-**statistic**

$$T = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}}$$

where $\hat{\sigma}_{\hat{\beta}_i}$ is the standard error estimate of $\hat{\beta}_i$ given by $\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma}\sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{ii}}$.

Since T follows a t_{n-p-1} distribution under H_0,

- the p-value $= 2\ \mathrm{Prob}(t_{n-p-1} \geq |T|)$
- a $100(1-\alpha)\%$ CI $= \hat{\beta}_i \pm t^{\alpha/2}_{n-p-1} \hat{\sigma}_{\hat{\beta}_i}$

3.6 Hypothesis Testing: Example

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```