Module 7: Linear regression

Yuan Tian

07/20/2023

Outline

In this module, we will review linear regression.

Linear regression

- Model:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

- Equivalently:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$$

Linear regression

- Model:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

- Equivalently:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$$

- Standard assumptions
  - y_i independent (equivalently ϵ_i independent)
  - $\mathbb{E}(\epsilon_i) = 0$
  - $\mathrm{var}(\epsilon_i) = \sigma^2$, constant
  - x_i known, β to be estimated

Linear regression

- Model:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

- Equivalently:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$$

- Standard assumptions
  - y_i independent (equivalently ϵ_i independent)
  - $\mathbb{E}\left(\epsilon_i\right) = 0$
  - var $\left(\epsilon_i\right) = \sigma^2$, constant
  - x_i known, β to be estimated

- More concisely:

$$\mathbb{E}(Y \mid X) = X\beta, \quad \mathrm{var}(Y \mid X) = \sigma^2 I$$

Interpretation of β_j

- Effect on the expected response of a unit change in jth explanatory variable, all other variables held fixed

Least squares estimation

- Definition (minimize the residuals)

$$\hat{\beta}_{\mathrm{LS}} := \min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^{\mathrm{T}} \beta \right)^2$$

- Equivalently,

$$\hat{\beta}_{LS} := \min_{\beta} (y - X\beta)^{\mathrm{T}} (y - X\beta)$$

- Equivalently (L2 distance),

$$\hat{\beta}_{\mathrm{LS}} := \min_{\beta} \|\mathrm{y} - X\beta\|_2^2$$

- Equivalently, $\hat{\beta}$ is the solution of the score equation

$$X^{\mathrm{T}} (y - X\beta) = 0$$

- Solution

$$\hat{\beta}_{\mathrm{LS}} = \left(X^{\mathrm{T}} X \right)^{-1} \left(X^{\mathrm{T}} \boldsymbol{y} \right)$$

Another interpretation: the projection of Y onto the linear subspace spanned by the columns of $\mathbf{X}$

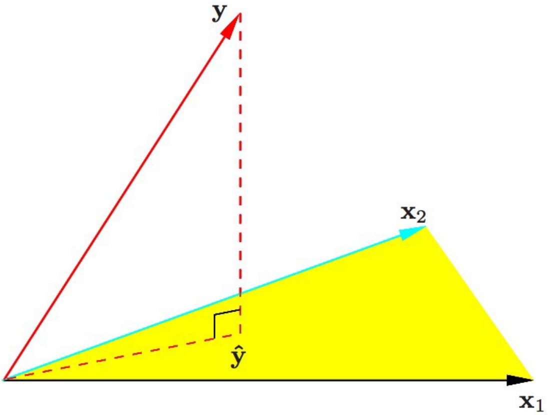

FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Ele

Least squares estimation (cont'd)

Assume X is fixed,

- Expected value

$$\mathbb{E}\left(\hat{\beta}_{\mathrm{LS}}\right) = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\mathbb{E}(y) = \left(X^{\mathrm{T}}X\right)^{-1}\left(X^{\mathrm{T}}X\right)\beta = \beta$$

- Variance

$$\begin{aligned} \mathrm{var}\left(\hat{\beta}_{LS}\right) &= \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\,\mathrm{var}(y) X \left(X^{\mathrm{T}}X\right)^{-1} \\ &= \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\sigma^2 I X \left(X^{\mathrm{T}}X\right)^{-1} \\ &= \sigma^2 \left(X^{\mathrm{T}}X\right)^{-1} \end{aligned}$$

Assumptions for ordinary least squares

- **Linearity**: the expectation of Y is linear in $X_1 \ldots X_p$
- **Independence**: the ϵ_i are independent
- **Mean zero errors**: the ϵ_i have mean zero, i.e. $E\left[\epsilon_i\right] = 0$
- **Equal variance (homoscedasticity)**: the ϵ_i have the same variance, i.e. $\text{Var}\left[\epsilon_i\right] = \sigma^2$

What about normal distribution?

- If we further assume $\epsilon_i \sim N\left(0, \sigma^2\right)$ (and independent across i), then
- $y \mid X \sim N\left(X\beta, \sigma^2 I\right)$, and
- likelihood function is

$$L\left(\beta, \sigma^2; y\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}$$

- log-likelihood function is

$$\ell\left(\beta, \sigma^2; y\right) = -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)$$

- maximum likelihood estimate of β is

$$\hat{\beta}_{ML} = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}\boldsymbol{y} = \hat{\beta}_{\mathrm{LS}}$$

What about normal distribution? (cont'd)

- distribution of $\hat{\beta}$ is normal

$$\hat{\beta} \sim N_p\left(\beta, \sigma^2\left(X^{\mathrm{T}}X\right)^{-1}\right)$$

- distribution of $\hat{\beta}_j$ is

$$N\left(\beta_j, \sigma^2\left(X^{\mathrm{T}}X\right)^{-1}_{jj}\right), \quad j = 1, \ldots, p$$

- maximum likelihood estimate of σ^2 is

$$\frac{1}{n}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

- but we use

$$\tilde{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

Maximum likelihood estiamtion vs. OLS

- We did not place any distributional assumptions on the outcome,
  - We only required that $E[\epsilon_i] = 0$ with constant variance
  - In other words, OLS is a semiparametric method

Maximum likelihood estiamtion vs. OLS

- We did not place any distributional assumptions on the outcome,
  - We only required that $E\left[\epsilon_i\right] = 0$ with constant variance
  - In other words, OLS is a semiparametric method
- Sometimes, people assume that $\epsilon_i \sim N\left(0, \sigma^2\right)$, which means

$$Y_i \sim N\left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_1 X_{ip}, \sigma^2\right)$$

  - If this additional assumption is made, then we can instead use maximum likelihood estimation for β
  - This connects to a whole other class of models called generalized linear models (GLMs)
  - Interestingly, in this case, you will end up with the same estimates for β

`lm` function in R

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance

Usage

```
lm(formula, data, subset, weights, na.action, method
= "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Check out utility functions: `summary`, `residuals`, `fitted`, `deviance`, `coef`, `...`

Example

```r
catF <- lm(y~x)
summary(catF)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00871 -0.68599 -0.04506  0.79583  2.21858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.4855   2.007 0.050785 .
## x             2.6364     0.6254   4.215 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.162 on 45 degrees of freedom
## Multiple R-squared:  0.2831, Adjusted R-squared:  0.2671
## F-statistic: 17.77 on 1 and 45 DF,  p-value: 0.0001186
```

Decomposition of sum of squares

- Total sum of squares (SS_{total}): $\|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- Explained sum of squares(SS_{model}): $\|\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}\|^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- Residual sum of squares, RSS (also denoted as SS_{error}): $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$
- The above equation decomposes SS_{total} into two parts: explained due to the LM and unexplained:

$$SS_{\text{total}} = SS_{\text{model}} + SS_{\text{error}}$$

ANOVA table

Source	SS	d.f.	MS	F
model	SS_{model}	$\mathrm{p} - 1$	MS_{model}	MS_{model} / MSE
error	SS_{error}	$\mathrm{n} - \mathrm{p}$	MSE	
total	SS_{total}	$\mathrm{n} - 1$		

```
anova(catF)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 24.002 24.0020  17.768 0.0001186 ***
## Residuals 45 60.788  1.3508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Goodness-of-fit

- It is useful to know how well a LM fits the data. One obvious measure of of goodness-of-fit is the RSS.
- A measure of goodness-of-fit is the `coefficient of determination`, or R^2:

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{total}}} = 1 - \frac{SS_{\text{error}}}{SS_{\text{total}}}$$

  It gives the proportion of the variation in the response explained by the LM
- R^2 is the square of the `multiple correlation coefficient` which is defined as the sample correlation coefficient between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$

Adjusted R^2

- Adjusted R^2 is a modification of R^2 that adjusts for the number of independent variables in a model:

$$\bar{R}^2 = 1 - \frac{SS_{\text{error}} /(n-p)}{SS_{\text{total}} /(n-1)}$$

- When a variable is added to the model, R^2 always increases while $\bar{R}^2$ can increase or decrease
- Unlike R^2, $\bar{R}^2$ increases only if the new term improves the model more than would be expected by chance. $\bar{R}^2$ can be negative, and will always be less than or equal to R^2
- $\bar{R}^2$ does not have the same interpretation as R^2. As such, care must be taken in interpreting and reporting this statistic
- $\bar{R}^2$ is useful in the variable selection stage of model building. R^2 is not useful for variable selection

Diagnostics: Under-fitting

Suppose the true model is

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\gamma + \boldsymbol{\epsilon}$$

and we fit the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

That is, covariates in Z are missed. Consequences are

- $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \left(X^T X\right)^{-1} X^T Z\gamma$. Therefore, $\hat{\boldsymbol{\beta}}$ is biased if $X^T Z \neq 0$.
- $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(X^T X\right)^{-1}$, unchanged
- $E\left(\hat{\sigma}^2\right) \geq \sigma^2$. That is, $\hat{\sigma}^2$ is inflated since it includes variation due to Z which is uncounted for by the fitted model

Lurking variables

- Lurking (confounding) variables are factors (often “hidden”) may effect the relationship between the response and the covariates but are not measured or considered
- They can make it seem like there is a relationship when there’s not or they can hide an existing relationship
- For example,we will observe a positive relationship between the height and reading ability among elementary school students. This may be driven by the lurking variable - age
- Some designed experiments makes Z orthogonal to X, that is, $X^T Z = 0$, then β is unbiased
- Randomization helps to reduce the effects of lurking variables
- Matching and/or stratification

Over-fitting

Suppose the true model is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and we fit the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\gamma + \boldsymbol{\epsilon}$$

Consequences are

- $E(\hat{\beta}) = \beta$ and $E(\hat{\gamma}) = \gamma$ that is, both are unbiased
- $\text{Var}(\hat{\beta}) \geq \sigma^2 \left(X^T X\right)^{-1}$, that is,lose precision due to the need to estimate more parameters
- $E\left(\hat{\sigma}^2\right) = \sigma^2$, unchanged but with less `df`

Correlation and non-constant variance

So far we have assumed that $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$. Suppose in reality $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$.

Consequences are

- $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, unchanged
- $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X}\right) \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \neq \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$
- $E\left(\hat{\sigma}^2\right) \neq \sigma^2$, biased
- Correlation is a more serious violation which could severely bias inference. We need to model correlation or apply robust procedures when correlation is present

Resources

This tutorial is based on

- Linear Regression Analysis, George A.F.Seber,Alan J.Lee
- Harvard's Biostatistics Preparatory Course Methods [links].