Multiple linear regression: focus on inference

GLM
M2W2
2018

The GLM way ⇒ likelihood / score / Fisher information

$X^TX = \sum_{i=1}^n x_i x_i^T$

$X^TY = \sum_{i=1}^n x_i y_i$

1. $Y_i \sim N(\mu_i, \sigma^2)$
2. $\eta_i = x_i^T \beta$
3. $\mu_i = \eta_i$

(x_i, Y_i) $i = 1, .., n$ independent pairs

$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$ full rank

$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ $\quad \hat\beta = (X^TX)^{-1} X^T Y$

$$\hat\beta \sim N_p(\beta, \sigma^2 (X^TX)^{-1})$$

$$\hat\sigma^2 = \frac{1}{n-p}(Y - X\hat\beta)^T(Y - X\hat\beta) = \frac{SSE}{n-p} = \frac{\sum_{i=1}^n (Y_i - x_i^T\hat\beta)^2}{n-p}$$

$$T_j = \frac{\hat\beta_j - \beta_j}{\hat{SD}(\hat\beta_j)} \sim t_{n-p} \overset{\text{asymptotic}}{\underset{n \to \infty}{\approx}} N(0,1)$$

$\hat{SD}(\hat\beta_j) = \hat\sigma \cdot \sqrt{c_{jj}}$

c_{jj} elem (j,j) of $(X^TX)^{-1}$

Statistical inference

1) Confidence interval for β $(1-\alpha)\cdot 100\%$

$$\left[\hat\beta_j \pm t_{\frac{\alpha}{2}, n-p} \sqrt{c_{jj}}\, \hat\sigma\right]$$

2) Prediction interval for new observation Y_0

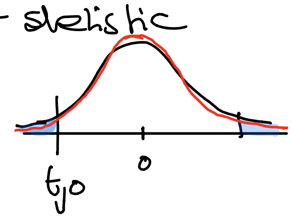
[not use so much in this course - focus on interpretation]

3) Single hypothesis testing

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

$$T_{j0} = \frac{\hat\beta_j - 0}{\hat{SD}(\hat\beta_j)} \quad \text{test statistic}$$

p-value: $2 \cdot P(T_{n-p} > |t_{j0}|)$



4) Testing linear hypotheses

$$H_0: \underbrace{C\beta}_{r\times 1} = d \quad \text{vs} \quad H_1: \overset{r\times p}{C}\underset{p\times 1}{\beta} \neq \overset{r\times 1}{d}$$

(Why $C\beta = d$?)

A: large model (H_1) and B: small model (H_0 true)

"nested" within A

Motivation $\quad C\hat{\beta} \sim N(C\beta, \; C\,\sigma^2(X^TX)^{-1}C^T)$

$Y \sim N_p(\mu, \Sigma)$

$(Y-\mu)^T\Sigma^{-1}(Y-\mu) \sim \chi^2_p \quad \rightarrow \quad (C\hat{\beta}-d)^T\left[C\,\sigma^2(X^TX)^{-1}C^T\right]^{-1}(C\hat{\beta}-d) \sim \chi^2_r$

but $\hat{\sigma}$ estimate inserted

and $\dfrac{\hat{\sigma}\cdot(n-p)}{\sigma^2} \sim \chi^2_{n-p}$

Test statistic:

$$F_{obs} = \frac{1}{r}\left[C\hat{\beta}-d\right]\frac{1}{\hat{\sigma}^2}\left[C(X^TX)^{-1}C^T\right]^{-1}(C\hat{\beta}-d) \sim F_{r,n-p}$$

Asymptotic p-value: $\quad r\cdot F_{r,n-p} \xrightarrow{n\to\infty} \chi^2_r$

Using $r\cdot F_{obs} \sim \chi^2_r$ is called a Wald test.

Alternative way of writing the Fobs:

$$F_{obs} = \frac{\frac{1}{r}(SSE_{H_0} - SSE)}{\frac{SSE}{n-p}}$$

SSE_{H_0} : from smaller model B

SSE : larger A

Is the regression significant?

A : our (candidate) model $\rightarrow$ SSE

B : a model with only intercept $\rightarrow$ SST

$\hat{\varepsilon} = (Y_i - \hat{Y}_i) = (Y - \bar{Y})$

$$F_{obs} = \frac{\frac{1}{r}(SST - SSE)}{\frac{SSE}{n-p}}$$

5) The likelihood ratio test (but

AIM: compare two (nested) models

A: large H_1

B: small — nested within large H_0

$L(\hat\beta_A, \tilde\sigma_A)$: (maximum) likelihood at $\hat\beta_A, \tilde\sigma_A$ under model A

maximum likelihood estimates under model 0

$L(\hat\beta_B, \tilde\sigma_B)$: (maximum) likelihood at $\hat\beta_B, \tilde\sigma_B$

because

$$\Big(-\frac{1}{2\tilde\sigma^2}\underbrace{(Y-X\hat\beta)^T(Y-X\hat\beta)}_{\tilde\sigma^2 n}\Big)$$

$$L(\hat\beta_A, \tilde\sigma_A) = \left(\frac{1}{2\pi\tilde\sigma_A^2}\right)^{n/2} e^{-\frac{n}{2}}$$

$$L(\hat\beta_B, \tilde\sigma_B) = \left(\frac{1}{2\pi\tilde\sigma_B^2}\right)^{n/2} e^{-\frac{n}{2}}$$

since

$$L(\beta,\sigma) = \left(\prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}}\right)\cdot \exp\Big(-\frac{1}{2\sigma^2}\underbrace{(y-X\beta)^T(y-X\beta)}_{\sum_{i=1}^{n}(y_i-x_i^T\beta)^2}\Big)$$

$$\left.\begin{array}{l}\sum_{i=1}^{n}(y_i-x_i^T\hat\beta)^2 = SSE \\ \tilde\sigma^2 = \frac{SSE}{n}\end{array}\right\} L(\hat\beta,\tilde\sigma) = \left(\frac{1}{2\pi\tilde\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\tilde\sigma^2}\cdot n\tilde\sigma^2\right)$$

The likelihood ratio test statistic

$$-2\ln\lambda = -2\Big(\ln L(\underbrace{\hat\beta_B, \tilde\sigma_B}_{\text{small}}) - \ln L(\underbrace{\hat\beta_A, \tilde\sigma_A}_{\text{large}})\Big)$$

mild regularity conditions $\longrightarrow \approx \chi^2_{df}$

df: (number of parameters in A − number of parameters in B)

Reject H_0 when (prefer A (large) to B (small)) p-value $< \alpha$: $P(\chi^2_{df} > -2\ln\lambda)$

Comment: can be shown that

$$-2\ln\lambda = n\left(\ln\tilde\sigma_B^2 - \ln\hat\sigma_A^2\right) \quad\text{and}\quad \lambda = \left(1 + \frac{p_A - p_0}{n-p_A} F_{obs}\right)^{-n/2}$$

(Ex: munich rent index)

6) Deviance ← used for model assessment and to replace ANOVariance ...

Assume all n observations have different $\underbrace{\text{covariate patterns}}_{\text{comb. of the } k \text{ covariates}}$.

Candidate model: the model we fit and want to assess.

Saturated model: the model that would give the best fit to the data: $\hat{\mu}_i = y_i$

and we have $\eta_i = \hat{\mu}_i$ (so not $\eta_i = x_i^T\beta$) (unless $x_i^T = 0,1$ and we have n β's)

The log likelihood of the saturated model is

$$\ln\left(\prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(y_i-\mu_i\right)^2\right)\right) \text{ and let } \hat{\mu}_i = y_i$$

$$-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-y_i)^2 = -\frac{n}{2}\ln(2\pi\sigma^2)$$

The loglikelihood for the candidate model:

$$-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$$

The deviance is LRT-statistic for saturated vs candidate

$$D = -2\left(\underbrace{-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - x_i^T\hat{\beta})^2}_{\text{cand.}} - \left(-\frac{n}{2}\ln(2\pi\sigma^2)\right)\right)$$

$$= \frac{1}{\sigma^2}\sum(y_i - x_i^T\hat{\beta})^2 = \frac{SSE}{\sigma^2}$$ "scaled deviance"

unscaled deviance $= \phi D = \sigma^2 D = SSE$

⇒ For MLR the (unscaled) deviance is SSE —
and for GLM's we use the deviance for model assessment
(not SSE)

will not be = SSE for
bin, Pois, .. etc GLM.

7) <u>ANOVA tables</u> = sequential table of
reductions in SSE as each term in the linear
predictor is added in turn.
For this: β_0 is always in the model

$SSE(\beta_1, \beta_2, \dots, \beta_k)$ is SSE of our candidate model with
k predictors

$SSE(\beta_1)$ = SSE when only $\eta = \beta_0 + \beta_1 x_1$

$SSE(\beta_1, \beta_2)$ = SSE when $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

NEW: $SSE(\beta_2 | \beta_1) = SSE(\beta_1, \beta_2) - SSE(\beta_1)$

added effect of model with β_1 and β_2 as
compared to model with only β_1.

$$F = \frac{\frac{1}{r} SSE(\beta_2 | \beta_1)}{\left\{ \frac{SSE(\beta_1, \dots, \beta_k)}{n-p} \right.}$$

ALWAYS candidate model

r = 1 here
||
difference in number
of parameters
between model with
(β_1, β_2) and (β_1)

for covariates with
dummy coding of 3
levels β_2 is really $(\beta_{2.1}, \beta_{2.2})$
so r = 2

5

H_0: no added effect of β_2 compared to model with β_1
when "candidate model is the reference point"

H_1: effect

$F \sim F_{r, n-p}$, p-value upper tail also presented

"Next": ANODevience not ANOVariance.. in "similar way"