Hypothesis testing about β_j

21.02.2017
h11

Previously: $Y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I)$

$\hat{\beta}_j$ is the jth element of $\hat{\beta} = (X^T X)^{-1} X^T Y$

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

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\varepsilon}^T\hat{\varepsilon} = \frac{SSE}{n-p}$$

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}$$

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t_{n-p}$$

c_{jj}: jth diagonal element of $(X^T X)^{-1}$

n: # obs, p: # param. we estimate

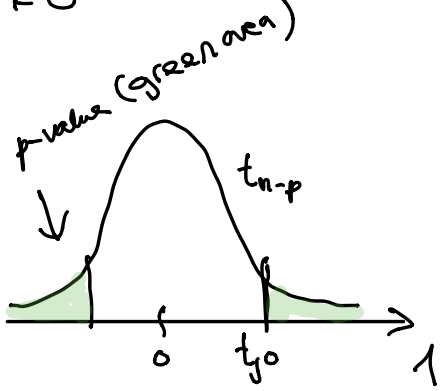
Test for association (linear) between response Y and x_j:

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

When H_0 is true:

$$T_{j0} = \frac{\hat{\beta}_j - 0}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t_{n-p}$$

T_{j0}: test statistic



p-value: $P(|T_{j0}| \geq |t_{0j}|, H_0 \text{ true})$ (t_{j0})

$= 2 \cdot P(T_{j0} \geq |t_{0j}|)$ (t_{j0})

T_{j0}: $n-p$

Reject H_0 when $|t_{j0}| > t_{\frac{\alpha}{2}, n-p}$ sign. level α.

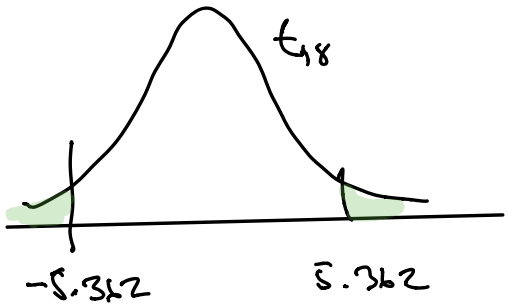
Ex: Acid rain: linear association between SO_4 and pH

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

From summery of lm in R (slide)

$t_{10} = -5,362$

$n-p = 18$



p-value =

$2 \cdot P(T_{18} > 5.362) = 4.3 \cdot 10^{-5}$

R: `2*(1-pt(5.362, 18))`

↑ area to left

Reject H_0 for all $\alpha > 4.3 \cdot 10^{-5}$.

Residuals (again)

$$\hat{\varepsilon} = Y - \hat{Y}, \quad \hat{\varepsilon} \sim N_n(0, \sigma^2(I - H))$$

R: residuals(fit)

The residuals have heteroscedastic variances

$\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ and $\text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(0 - h_{ij})$

can in general be $\neq 0$, but in most cases experience shows that ≈ 0

Standardized residuals:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

will be (approx.) homoscedastic.

R: rstandard(fit)

Studentized residuals: Fitting the model to all obs. except i to make r_i^*. ← see slide

⇓

use studentized!

R: rstudent(fit)

See example on slide for r_i vs $\hat{\varepsilon}_i$

Analysis of variance decomposition and R^2

$y_1, \dots, y_n$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

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

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

Sums of squared total — SST

$(y_i - \hat{y}_i) = \hat{\varepsilon}_i$

Sums of squared errors — SSE

Sums of sq regression — SSR

↑ explained by regression

With vectors and matrices. $\hat{\varepsilon}^T\hat{\varepsilon}$

$$Y^T(I - \tfrac{1}{n}11^T)Y = Y^T(I-H)Y + Y^T(H - \tfrac{1}{n}11^T)Y$$

SST SSE SSR

This is used to define:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

↑ coefficient of determination

↖ relative proportion of total variability explained by the regression

Ex: Acid rain, full model (all covariates available)

$R^2 = 0.93$, 93%

Is the regression significant?

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs

H_1: at least one $\beta_j \neq 0$ $\quad j = 1, \ldots, k$

Test statistics: $F = \dfrac{SSR/k}{SSE/(n-p)} \sim F_{k, n-p}$ ($SST - SSE$)

↑ prove this in Part 3 in a general setting

Ex: Acid rain:

F-observed: 34.15

$k = 7$, $n - p = 18$ (26 − 8)

$P(F_{7,18} > 34.15) = 3.9 \cdot 10^{-9}$ (p-value)

Ex: Volume of tree and the lumberjack

Big model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

Small model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

$R^2_{Big} \geq R^2_{small}$ ⇐ since $\hat{\beta}_3$ is found to minimize $SSE = \hat{\varepsilon}^T\hat{\varepsilon}$, thus maximize $R^2 = 1 - \frac{SSE}{SST}$

$R^2_{Big} = R^2_{small}$ if $\hat{\beta}_3 = 0$

if $\hat{\beta}_3 \neq 0$ the $SSE_{Big} < SSE_{small}$ and $R^2_{big} > R^2_{small}$.

R^2 will always increase (or stay unchanged) when a new covariate is added to the model.

Next: more on choosing a good model, and then

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$ ← penalizing adding many covariates

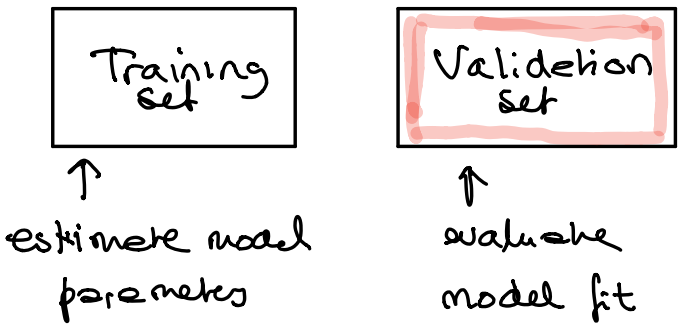
is one criterion to use instead of R^2 for model selection.

Model choice and variable selection [ISL 3.4]

Question 1: Is a full model (all available covariates fitted) the best model?

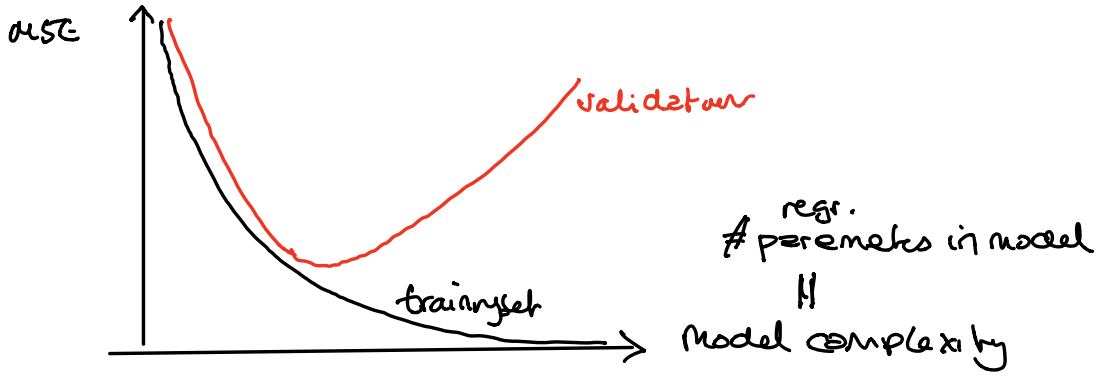
good interpretability ← → good for future predictions

Data is divided into



$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Now calculate the MSE on the training and on the validation set and plot as a function of model complexity:



Answer 1: No, this may lead to overfitting = fitting the trend + the noise!

⇒ so, what can we do instead?