Introduction

Regression modelling: rel.

$$Y = f(x_1, x_2, \ldots x_p)$$

response var. / dep. var. / out. var.

explan. var. / predictors / features / covariates.

Subset: linear RM.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \Rightarrow \text{Linear function}$$

Parts of LRM

- y: Val. of resp. variable
- β_0: Intercept: Avg. value of Y when $x_i = 0$
- x_i: Expl. var value
- $\beta_1 \ldots \beta_p$: Model params. Quantify effect of x_i on y
- ε: Random error: If only: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ Completely det. Not perfect b/c randomness. $\varepsilon \sim N(0, \sigma^2)$

Model params.

Objective: figure out model params. given data. → "Fitting a model"

Simplification: simple linear regression:

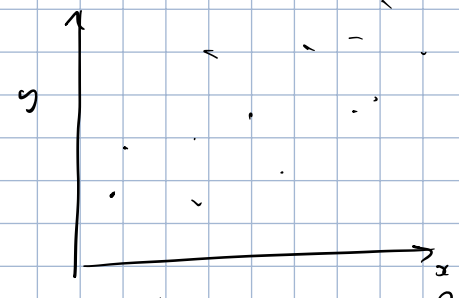
$$y = \beta_0 + \beta_1 x_1 + \varepsilon \Rightarrow 1 \text{ expl. var.}$$

Simple Linear Models

Linearity

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Check if linear rel. exists. ⇒ scatterplot



⇒ Roughly linear

Check strength of linear rel. ⇒ Pearsons corr. coef.

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}}$$

Estimate ⇒ sample corr. coef

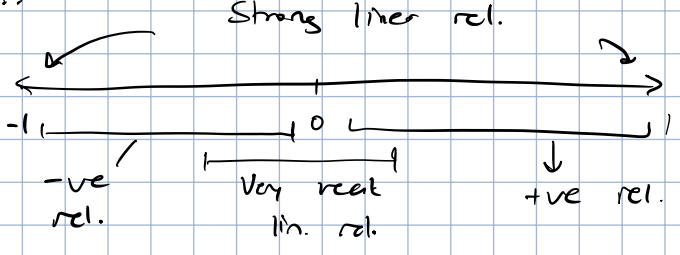
$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$S_{xy} = \sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=0}^{n} (x_i - \bar{x})^2$$

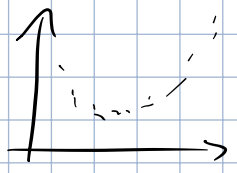
$$S_{yy} = \sum_{i=0}^{n} (y_i - \bar{y})^2$$

Interpretation:



Weaknesses:

1. Corr. coef. doesn't work in non-liner data



$\Rightarrow -1 \le r \le 1$!

2. Cannot be used for prediction.

Model Formulation

$$y = \beta_0 + \beta_1 x_1 + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

↖ Expected change / unit increase of x_i

Data:

(x_i, y_i), $i = 1, \dots n$

Each value:

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ⇒ Fitted val.

$y = \beta_0 + \beta_1 x_i + \varepsilon_i$ ⇒ Actual val.

B/C each y is $\beta_0 + \beta_1 x + \varepsilon_i$ ⇒ $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Ls proof:

$$E(y_i) = E\{\underbrace{\beta_0 + \beta_1 x_i}_{\text{constant}} + \varepsilon_i\}$$

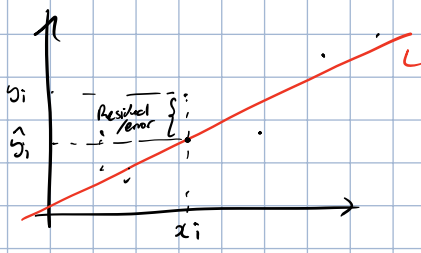
$$= \beta_0 + \beta_1 x_i + \underbrace{E[\varepsilon_i]}_{0}$$

$$= \beta_0 + \beta_1 x_i$$

$$Var(Y_i) = Var\{\underbrace{\beta_0 + \beta_1 x_i}_{\text{constant}} + \varepsilon_i\}$$

$$= \sigma^2$$

Method #1 of Estimating β_0 + β_1: Least Squares Estimation



⇒ "Best" line has least amount of error

Measure of error for all values:

$$\text{Sum of square error} = \sum_{i=0}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Get rid of -ve + weights higher error more.

Obj: minimize SSE to get least square estimators (LSEs) $\hat{\beta}_0$, $\hat{\beta}_1$

$$\underset{\beta_0, \beta_1}{\arg\min} \sum_{i=0}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

To minimize:

① Derivative:

$$\frac{\partial}{\partial \beta_0} SSE = -2 \sum (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial}{\partial \beta_1} SSE = -2 \sum (y_i - \beta_0 - \beta_1 x_i)(x_i)$$

② Zero:

a) $\sum (y_i - \beta_0 - \beta_1 x_i) = 0$

b) $\sum (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0$

③ Solve for β_0 + β_1

Start w/ a) ⇒ split out sum

Solve for β_0

Sub in b) ⇒ solve for β_1

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Remind:

$$S_{xx} = \sum (x_i - \bar{x})^2 \Rightarrow \sum (x_i - \bar{x}) = \sum x_i - n\bar{x} = \sum x_i - n \cdot \frac{\sum x_i}{n}$$

$$= \sum x_i (x_i - \bar{x}) - \bar{x} \underbrace{\sum (x_i - \bar{x})}_{0} \qquad = \sum x_i - \sum x_i = 0$$

$$= \sum x_i^2 - \bar{x} \sum x_i \quad \rightarrow \quad \sum x_i = n\bar{x}$$

$$= \sum x_i - n\bar{x}^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum x_i y_i - \bar{x} \sum y_i - \bar{y} \underbrace{\sum (x_i - \bar{x})}_{0}$$

$$= \sum x_i y_i - \bar{x} \cdot n\bar{y}$$

$$= \sum x_i y_i - n\bar{x}\bar{y}$$

$$\therefore \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Note: given data $\Rightarrow$ find $\hat{\beta}_0$ + $\hat{\beta}_1$

Ex:// $\sum x_i = a$, $\sum y_i = b$, $\sum x_i^2 = c$, $\sum x_i y_i = d$, n

Find SLR:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{d - n \cdot \frac{a}{n} \cdot \frac{b}{n}}{c - n \cdot (\frac{a}{n})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = \frac{b}{n} - \hat{\beta}_1 \cdot \frac{a}{n}$$

Residuals

$$r_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$$

Properties:

LSE:

a) $\sum (y_i - \beta_0 - \beta_1 x_i) = 0$

b) $\sum (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0$

$\therefore$

1) $\sum r_i = 0 \Rightarrow$ From a)

2) $\sum r_i x_i = 0 \Rightarrow$ From b)

3) $\sum r_i \hat{y}_i = \sum r_i (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$= \hat{\beta}_0 \sum r_i + \hat{\beta}_1 \sum r_i x_i$

$= 0$

} Model checking

Estimating σ^2:

To estimate pop. sample given $y_i \sim N(\mu, \sigma^2)$

- Sample var: $\frac{1}{n-1} \sum (y_i - \bar{y})^2$
  - $n-1$: Losing 1 d.f. by using estimate $\bar{y}$

We can do same thing for SLR:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\therefore \quad s^2 = \frac{1}{n-2} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \frac{1}{n-2} \sum r_i^2$$

Unbiased estimate of σ^2

Method #2 of Estimating $\beta_0 + \beta_1$: Maximum Likelihood

Theory:

① Maximum likelihood func.

$$L(\theta) = f(y_i; \theta) \quad \Rightarrow \text{How likely are we seeing observed data}$$

$\therefore$ For θ to maximize prob.

② Log likelihood:

$$\ell(\theta) = \log(L(\theta))$$

③ Derivative:

$$S(\theta) = \frac{\partial}{\partial \theta} L(\theta) \quad \Rightarrow \text{Score func.}$$

④ Set to 0 + solve for θ:

$$S(\theta) = 0$$

SLR:

① Likelihood function:

Each $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$L(\theta) = P(y_1 = \ldots) \cdot P(y_0 = \ldots) \ldots$$

$\theta = \{\beta_0, \beta_1, \sigma^2\}$

$$= \prod \underbrace{P(y = y_i)}_{\text{Normal}}$$

$$= \prod (2\pi\sigma^2)^{-1/2} \exp\left\{ - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}$$

② Log likelihood:

$$\ell(\theta) = \sum \left[\left(-\tfrac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) \right]$$

③ Score function:

① $\frac{\partial}{\partial \beta_0} = \frac{1}{\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i) = 0$

② $\frac{\partial}{\partial \beta_1} = \frac{1}{\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0$

③ $\frac{\partial}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2 = 0$

④ Solve:

① + ② $\Rightarrow$ same as LSE

$\therefore$ MLE = LSE

③ $\sigma^2 = \frac{1}{n} \sum r_i^2$ → This is diff.

The other is unbiased $\Rightarrow$ This MLE is biased (don't use)

Properties of LSE

Note: 1. estimator vs. estimate → Observed value

⇅ r.v.

$\hat{\beta}_0, \hat{\beta}_1 \leftarrow f(y_i) \overset{N}{\Rightarrow} \hat{\beta}_0, \hat{\beta}_1 \sim N$

2. Assumptn: $\varepsilon_i \sim N(0, \sigma^2)$

Property #1: LSE are unbiased

a) $E(\hat{\beta}_1) = \beta_1$ ← true value

$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$ → Constants

$E(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x}) E(y_i)}{\sum (x_i - \bar{x})^2}$

b) $E(\hat{\beta}_0) = \beta_0$

$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x})$ (r.v. ✓, r.v. ✓)

$= \frac{1}{n} \sum E(y_i) - \bar{x} E(\hat{\beta}_1)$

$= \beta_0 + \beta_1 \bar{x} - \bar{x} \beta_1$

$= \beta_0$

$$= \frac{\sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\beta_1 \left[\sum (x_i - \bar{x}) x_i\right]}{(x_i - \bar{x})^2} \quad \sum (x_i - \bar{x})^2$$

$$= \beta_1$$

Property # 2: Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \qquad Var(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

a)
$$Var(\hat{\beta}_1) = Var\left(\frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}\right)$$

$$= \left(\frac{1}{S_{xx}}\right)^2 Var\left(\sum (x_i - \bar{x}) y_i\right)$$

$$= \left(\frac{1}{S_{xx}}\right)^2 \underbrace{\sum (x_i - \bar{x})^2}_{S_{xx}} \cdot Var(y_i)$$

$$= \frac{\sigma^2}{S_{xx}}$$

$Var(\sum y_i) = \sum Var(y_i) + \underbrace{\sum_{i<j} Cov(y_i, y_j)}_{0 \text{ if each } ij \text{ indep.}}$

Assume indep. sample $\Rightarrow$

b)
$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x})$$

$$= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}\, Cov(\bar{y}, \hat{\beta}_1)$$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{1}{n}\sum y_i, \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}\right)$$

Cov of sure variable!

$$= 0$$

$$\therefore Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}}$$

$$= \vdots$$

$$= \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$$

Property # 3: Covariance of $\hat{\beta}_0$ + $\hat{\beta}_1$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}$$

Property # 4: Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

y_i are iid N & $\hat{\beta}_0 + \hat{\beta}_1$ are linear combos of y_i

$$\therefore \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{n\sum(x_i-\bar{x})^2}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$$

Issue: all params can be calc'd except σ^2

↳ Recall: $\hat{\sigma}^2 = \frac{1}{n-2}\sum r_i^2 = s^2 \Rightarrow E(s^2) = \sigma^2$

↳ MSE: mean squared error

Since this is calculable $\Rightarrow$ use this in our distn (sub σ^2 for s^2)

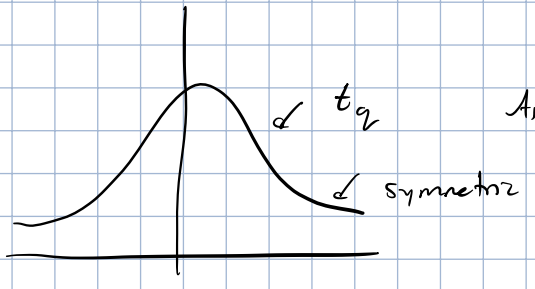
Standard error:

standard dev. if using σ^2 instead

$$se(\hat{\beta}_0) = \sqrt{Var(\hat{\beta}_0)} = \sqrt{\frac{s^2 \sum x_i^2}{\sum(x_i-\bar{x})^2}} \qquad Se(\hat{\beta}_1) = \sqrt{Var(\beta_1)} = \sqrt{\frac{s^2}{S_{xx}}}$$

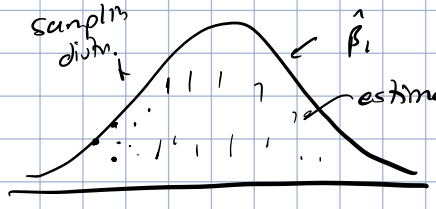
If we use S.E. $\Rightarrow$ $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2}$, $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_0)} \sim t_{n-2}$

Recall:



As $q \uparrow$, $t_q \to N$

Inference for Regression Parameters.

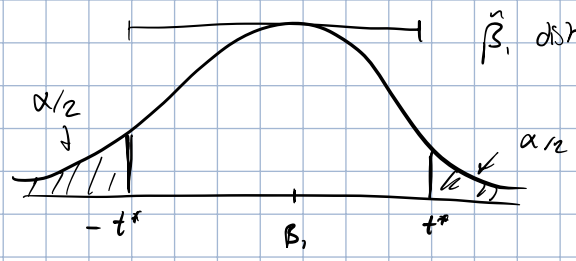


→ Inference

#1: C.I

α level

$1-\alpha$



$$\therefore \quad P\left(-t^* < \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} < t^*\right) = 1-\alpha$$

$$P\left(\hat{\beta}_1 - t^* se(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t^* se(\hat{\beta}_1)\right) = 1-\alpha$$

(β_1: true pop param.)

$$C.I. = \hat{\beta}_1 \pm t^* se(\hat{\beta}_1)$$

$$\hat{\beta}_1 \pm t^* se(\hat{\beta}_1)$$

} Range of pop param in 95% confidence.

t^* → t-table via $\alpha/2$ + d.f. $(n-2)$

If estimate ∉ C.I ⇒ can reject H_0

#2: Hypothesis Testing

Obj: the value of param $=/>/<$ some estimate.

① Hypothesis:

H_0 (status quo) vs. H_a (Not status quo)

You assume H_0 is T → if test gets crazy result → reject H_0

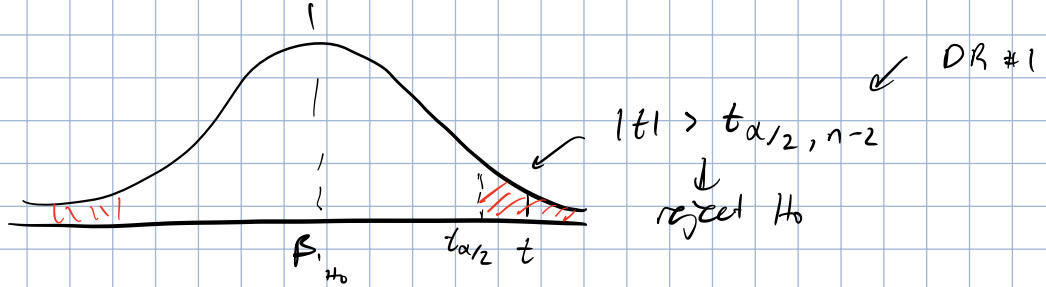
② Decision rule: what will make you reject H_0

③ Test statistic ⇒ decision.

Test stat: t-stat

$$\frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} \sim t_{n-2} \quad \text{(if } H_0 \text{ is true)}$$

If your data is off ⇒ t-stat is quite far



Other decision rule: p-val

$P(|T| > t)$ => prob of seeing data if H_0 is true.

If p-val $< \alpha$ => reject H_0 . (DR ≠ 2)

Common test:

- $\hat{\beta}_1$ test of significance: is there a linear rel.

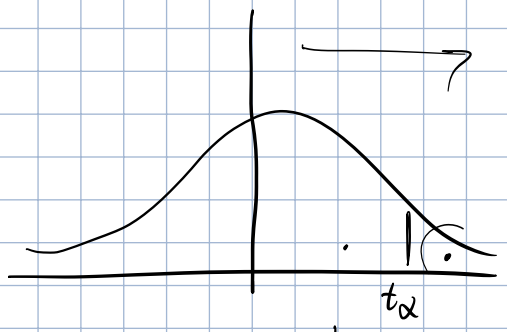
$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

Test stat: $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$ => t-ratio

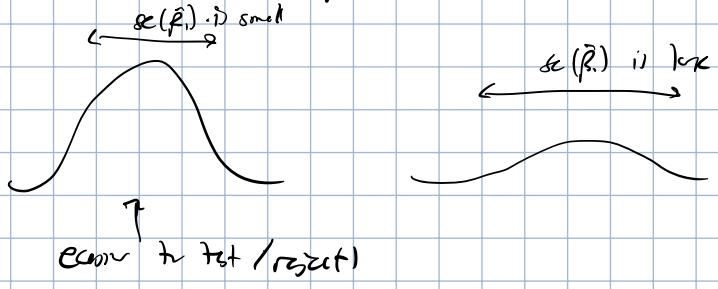
Reject in H_0 & acceptance of H_a.

One-sided testing:

$H_0: \sim > 0$, $H_a: \sim \leq$ --



This is based on $se(\hat{\beta}_1) \propto 1/s_{xx}$



If s_{xx} ↑ => se ↓ => reject / accept easier!

Spread out x-vals in LR

Ex:// Sig. rel. b/w prop. age & bonding strength

① Hypothesis:

$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

② Decision rule:

If $\left| \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \right| > t_{\alpha/2, n-2}$ → reject H_0.

③ Math:

$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = -12.86$, $t_{\alpha/2, n-2} = 2.101$.

qt(0.975, n-2)

$\therefore$ Reject H_0. $\Rightarrow$ Interpret.

Estimation of Mean Response

Mean response var. val. at $x = x_p$

① Random var:

$$\mu = \beta_0 + \beta_1 x_p$$

② Estimate

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

③ Unbiased?

$$E(\hat{\mu}) = E(\hat{\beta}_0) + x_p E(\hat{\beta}_1)$$
$$= \beta_0 + x_p \beta_1$$

④ Variance?

$$Var(\hat{\mu}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_p)$$

$\bar{y}, \hat{\beta}_1$

$$= Var(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_p)$$
$$= Var(\bar{y} + \hat{\beta}_1 (\bar{x} + x_p))$$
$$= Var(\bar{y}) + 2(\bar{x} + x_p) \underbrace{Cov(\bar{y}, \hat{\beta}_1)}_{\to 0} + (x_p - \bar{x})^2 Var(\hat{\beta}_1)$$
$$= \frac{\sigma^2}{n} + (x_p - \bar{x})^2 \cdot \sigma^2 / S_{xx}$$
$$= \sigma^2 \left[\frac{1}{n} + (x_p - \bar{x})^2 \cdot \frac{1}{S_{xx}} \right]$$

$\uparrow \sigma^2 / S_{xx}$

$$Var\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n^2} \sum Var(y_i) \quad \downarrow Cov(y_i, y_j) = 0 \text{ if } i \neq j$$
$$= \frac{1}{n^2} \sum \sigma^2$$
$$= \sigma^2 / n$$

$$Var\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n^2} \cdot \sum Var(y_i)$$
$$= \frac{1}{n^2} \cdot n\sigma^2$$
$$= \sigma^2 / n$$

⑤ Sampling distn.

$$\frac{\hat{\mu} - \mu}{\sqrt{Var(\hat{\mu})}} \sim N(0, 1)$$

$\Downarrow$

$\sigma^2 \to s^2$

$$\frac{\hat{\mu} - \mu}{se(\hat{\mu})} \sim t_{n-2}$$

$s^2 = \frac{1}{n-2} \sum r_i^2$

Run C.I. & inference tests!

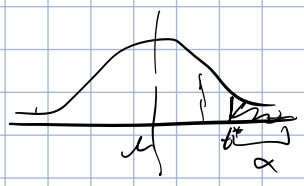
Ex.// Is avg. strength of obj at $x = 16 < 2000$?

① Estimate:

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 (16) = 2031.805$$

② Hypothesis:

$H_0: \mu \leq 2000$ $\quad$ $H_a: \mu > 2000$



③ Decision rule:

If $t > t_{\alpha, n-2} \Rightarrow$ reject H_0

④ Math

$$\frac{\hat{\mu} - 2000}{se(\hat{\mu})} = 1.462, \quad t_{\alpha, n-2} = 1.734$$

⑤ Concl.

Cannot reject $H_0 \Rightarrow$ might be wrong.

Prediction

True value of y_p from a new subject $x = x_p \Rightarrow$ not part of data

① R.V.

$$y_p = \beta_0 + \beta_1 x_p + \varepsilon$$ ← future random error.

② Estimate

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

③ Prediction:

$y_p - \hat{y}_p \Rightarrow$ No idea what y_p is.

Sampling distn.:

a) Mean:

$$E\{y_p - \hat{y}_p\} = (\beta_0 + \beta_1 x_p) - (\beta_0 + \beta_1 x_p)$$
$$= 0$$

b) Var:

$$Var\{y_p - \hat{y}_p\} = Var\{y_p - \hat{\beta}_0 - \hat{\beta}_1 x_p\}$$
$$= Var(\varepsilon) + Var(\hat{\beta}_0 + \hat{\beta}_1 x_p)$$
$$= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$
$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$

i) Greater than expected response.

ii) Dep. on $(x_p - \bar{x})$

c)

$$\frac{y_p - \hat{y}_p}{se(y_p - \hat{y}_p)} \sim t_{n-2} \Rightarrow \text{C.I. / tests.}$$

ANOVA

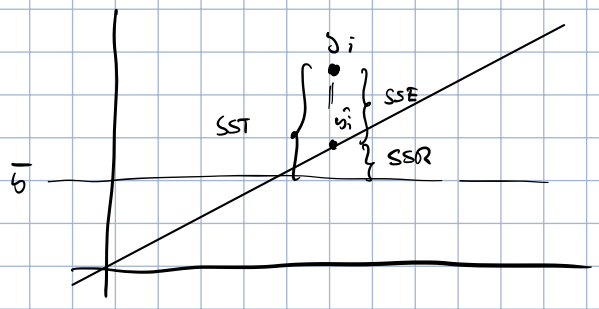
$$SST = SSE + SSR$$

Total sum of squares: $\sum (y_i - \bar{y})^2$

Error sum of sq. $\sum (y_i - \hat{y}_i)^2 = \sum r_i^2$

Regression sum of sq. $\sum (\hat{y}_i - \bar{y})^2$

Variance of y!



Proof:

$$SST = \sum (y_i - \bar{y})^2$$

$$= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum (y_i - \hat{y}_i) + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2$$

$$= \quad \| \quad + \underbrace{2\left(\sum r_i \hat{y}_i - 2\bar{y} \sum r_i\right)}_{0} + \quad \|$$

$$= SSE + SSR$$

Degrees of freedom:

- SST: $n-1$
- SSE: $n-2$
- SSR: 1

Mean square:

- MST: $SST/d.f. = SST/n-1$
- MSE: $SSE/d.f. = SSE/n-2 \rightarrow \frac{1}{n-2} \sum r_i^2 \Rightarrow$ unbiased est. of σ^2
- MSR: $SSR/d.f. = SSR$

MSR:

- $E[MSR] = \sigma^2 + \beta_1^2 S_{xx}$
  - Proof:

$$MSR = \sum (\hat{y}_i - \bar{y})^2 \qquad \hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$$

$$= \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$$

$$= \hat{\beta}_1^2 \cdot S_{xx}$$

$$E(MSR) = E(\hat{\beta}_1^2 \cdot S_{xx})$$
$$= S_{xx} \cdot (Var(\beta_1) + E(\beta_1)^2)$$
$$= \sigma^2 + \beta_1^2 S_{xx}$$

Inference:

① $MSR >> MSE \Rightarrow \beta_1 \neq 0$ (error can be explain better w/ regression line)

② $\frac{MSR}{MSE} \sim F_{1, n-2}$

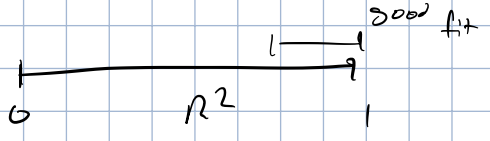
↳ F-test

$H_0: \beta_1 = 0, \quad H_a = \beta_1 \neq 0$

} High → more error expl. by SLR

Coefficient of determination:

$$R^2 = \frac{SSR}{SST} \Rightarrow \text{prop. of error expl. by reg. line.}$$



MULTIPLE LINEAR REGRESSION

Random vectors & Multivariate Normal Distribution

Random vector:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \rightarrow \text{r.v.}$$

- Mean:

$$E[Y] = \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_n) \end{bmatrix} = \mu$$

- Variance-covariance matrix:

$$Var(Y) = \begin{bmatrix} Var(y_1) & Cov(y_1, y_2) & \cdots & Cov(y_1, y_n) \\ & Var(y_2) & & \\ & & \ddots & \\ Cov(y_n, y_1) & & & \end{bmatrix}$$

→ Symmetric to lower triangle.

Positive definite

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix} \Rightarrow \Sigma_{n \times n}$$

If Y is indep. $\Rightarrow$ $diag\{\sigma_1^2 \ldots \sigma_n^2\}$

Properties:

① Linear combination on expectation:

$$E[AY + b] = AE[Y] + b$$

② Variance of vector + constant

$$Var[Y + b] = Var[Y]$$

③ Linear comb. on variance:

$$Var[AY + b] = A\,Var(Y)A' \Rightarrow \text{Square a matrix/vector: } AA'$$

Multivariate normal distn.

$$Y \sim MVN(\mu, \Sigma) \Leftrightarrow f(y_1, \ldots, y_n) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\tfrac{1}{2}(Y-\mu)^T \Sigma^{-1} (Y-\mu)\right)$$

- Properties:

① Linear comb.

$$Y \sim MVN(\mu, \Sigma) \Leftrightarrow AY + b \sim MVN(A\mu + b, A\Sigma A')$$

② Indiv. normality

$$Y \sim MVN(\mu, \Sigma) \Rightarrow y_i \text{ in } Y \sim N(\mu_i, \sigma_i^2)$$

③ Partioning:

$$Y = \left[\begin{array}{c} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{array}\right] = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \Rightarrow \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{bmatrix}$$

$$\hookrightarrow \quad Y_1, Y_2 \sim MVN(\mu_i, \Sigma_i)$$

④ Independence of indv. vars:

$$y_1, \ldots, y_n \overset{iid}{\sim} N(\mu_i, \sigma_i^2) \Rightarrow Y \sim MVN(\mu, \Sigma), \text{ diag}\{\sigma_1^2 \ldots \sigma_n^2\}$$

⑤ Variance & independence:

Not true for other distn.

$$Var(Y) = \Sigma \text{ is a diag.} \Leftrightarrow \text{independence of } y_i$$

⑥ Independence of scaled MVN

$$Y \sim MVN(\mu, \Sigma), \quad V = AY, \quad W = BY$$

$\hookrightarrow \quad V \perp W \iff A \Sigma B' = 0$

Multiple Linear Models

1 observation: y response, p explanatory

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

impact of each expl var. (β); i^{th} obs. of n^{th} explanatory (x); $\varepsilon_i \sim N(0, \sigma^2)$

Matrix:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(first column of 1s ← β_0)

$$Y = X\beta + \varepsilon$$

X ← design matrix

Interpretation:

β_0: expected resp. val. if $x_1 \dots = x_p = 0$

$\beta_i \ (i \neq 0)$: change in resp. variable by 1 unit inc if all other vars are fixed.

Distn:

Assumptn: $\varepsilon_i \sim N(0, \sigma^2) \Rightarrow \varepsilon = MVN(0, \Sigma)$ Σ = diag$\{\sigma^2 \dots \sigma^2\} = \sigma^2 I$

$\therefore$ Y is a linear combo of $\varepsilon \Rightarrow$ $\boxed{Y \sim MVN(X\beta, \sigma^2 I)}$

Parameter Estimation

Least squares:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip})\right)^2$$

(the bracketed sum is $X\beta$)

$$= \underset{\beta}{\text{argmin}} \left[(Y - X\beta)'(Y - X\beta)\right]$$

Take derivative:

$$\frac{\partial}{\partial \beta} \hat{\beta} = \frac{\partial}{\partial \beta} \left[Y'Y - Y'X\beta - X'\beta'Y + \beta' X' X \beta \right]$$

Aside:

$$z = a'\beta \Rightarrow \frac{\partial}{\partial \beta} z = a$$

$$z = \beta' a \Rightarrow \frac{\partial}{\partial \beta} z = a$$

$$z = \beta' A \beta \Rightarrow \frac{\partial}{\partial \beta} z = (A + A')\beta$$

$$\frac{\partial}{\partial \beta}\left[\underbrace{Y'Y}_{\text{no } \beta} - \underbrace{Y'X\beta}_{1} - \underbrace{X\beta'Y}_{1} + \beta' X' X \beta\right] = -(Y'X)' - (XY')' + (X + X')\beta$$

$$= -YX' - X'Y + 2X'X\beta$$

$$= -2X'Y + 2X'X\beta$$

Set to 0:

$$-2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

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

Properties:

① $E(\hat{\beta}) = \beta$

$$E(\hat{\beta}) = E\left[(X'Y)(X'X)^{-1}\right]$$

$$= (X'X)^{-1} X' E[Y]$$

$$= \underbrace{(X'X)^{-1}}_{\to I} \underbrace{X'X}_{I} \beta$$

$$= \beta$$

② $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

$$Var(\hat{\beta}) = Var\left[(X'Y)(X'X)^{-1}\right]$$

$$= \left[(X'X)^{-1}X'\right] Var(Y) \left[(X'X)^{-1}X'\right]'$$

$$= \left((X'X)^{-1}X'\right) Var(Y) \times (X'X)^{-1}$$

$Var(Y) \to \sigma^2$; $[(X'X)^{-1}]' = (A')^{-1}$

$$= (X'X)^{-1}\sigma^2$$

③ $\hat{\beta} \sim MVN(\beta, \sigma^2 (X'X)^{-1})$ b/c $\hat{\beta}$ is a linear comb of Y

④ $\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X'X)^{-1}_{jj})$

Fitted Values & Residuals

SLR $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

MLR: $\hat{Y} = X\hat{\beta} = X\left[(X'X)^{-1}X'Y\right]$

$= \underbrace{\boxed{X(X'X)^{-1}X'}}_{H} Y$

Hat matrix: $H = X(X'X)^{-1}X'$

① Symmetric:

Show: $H' = H$

$H' = \left[X(X'X)^{-1}X\right]'$

$= X\left[X(X'X)^{-1}\right]'$

$= X'(X'X)^{-1}X'$

$= H$

② Idempotent:

Show: $HH = H$

$HH = \left(X(X'X)^{-1}X'\right)\left(X(X'X)^{-1}X'\right)$ → Cancel.

$= X(X'X)^{-1}X'$

$= H$

③ $\hat{Y}$ is a L.T. of Y $\Rightarrow$ $\hat{Y} \sim MVN$

Residuals:

SLR: $r_i = y_i - \hat{y}$

MLR: $r = Y - \hat{Y}$

$= Y - HY$

$= (I - H)Y$

① r is a L.T. of Y $\Rightarrow$ $r \sim MVN$

② $E[r] = 0$

Proof: $E[r] = E[(I-H)Y]$

$$= (I - H) X\beta$$

$$= X\beta - HX\beta$$

$$= X\beta - (X\underbrace{(X'X)^{-1}}\underbrace{X'})X\beta$$

$$= X\beta - X\beta$$

$$= 0$$

(?) $\text{Var}[r] = \sigma^2(I - H)$

$$\text{Var}[r] = \text{Var}((I-H)Y)$$

$$= (I-H)\,\text{Var}(Y)\,(I-H)'$$

$$= \sigma^2 (I-H)(I-H)'$$

$$= \sigma^2 (I-H)(I-H)$$

$$= \sigma^2 (I-H)$$

$I - H$ is symmetric.

↓ Idempotent

$$\therefore \quad r \sim MVN(0, \sigma^2(I-H))$$

④ $\sum_{i=1}^{n} r_i = 0$

⑤ $\sum_{i=1}^{n} r_i x_{i1} = 0, \quad \sum_{i=1}^{n} r_i x_{i2} = 0, \ \ldots, \ \sum_{i=1}^{n} r_i x_{ip} = 0$

⑥ $\sum_{i=1}^{n} r_i \hat{y} = 0$

Proofs of ④, ⑤, ⑥:

Note that $X'r$ will give us ④ & ⑤:

$$\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np} \end{bmatrix}' \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} \sum r_i \\ \sum x_{i1} r_i \\ \vdots \end{bmatrix}$$

$X'r = 0 \Rightarrow$ done:

$$X'r = X'(Y - \hat{Y})$$

$$= X'Y - X'\hat{Y}$$

$$= X'Y - X'(X(X'X)^{-1}X'Y)$$

$$= X'Y - X'Y$$

$= 0$

Estimation of σ^2 in MLR

SLR: $\hat{MSE} = \frac{1}{n-2} \sum r_i^2$

MLR: $\frac{SSE}{n-(p+1)} \Rightarrow$ unbiased estimator of σ^2

Inference in MLR

Recall:

a) $\hat{\beta} \sim MVN(\beta, \sigma^2 (X'X)^{-1})$

b) $\hat{\beta}_j \sim N(\beta_j, [\sigma^2 (X'X)^{-1}]_{jj})$

$\therefore$ $$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 [(X'X)^{-1}]_{jj}}} \sim N(0,1) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 [(X'X)^{-1}]_{jj}}} \sim t_{n-(p+1)}$$

C.I.:

$$\hat{\beta}_j \sim t_{\alpha/2,\, n-(p+1)} \times se(\hat{\beta}_j)$$

→ Exact same mechanics.

Hypothesis:

$$|t| = \left| \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \right| > t_{\alpha/2,\, n-(p+1)}$$

Difference: interpretation

C.I. & test of $\hat{\beta}_j$ is accounting for relation with other $\hat{\beta}_j$ included in model.

Inference for Linear Combos of Parameters

Constant vector:

$$c = (1, x_1, \ldots, x_p)'$$

Estimated response: $\hat{\mu}_c = c' \hat{\beta}$

Distn: $\hat{\mu}_c \sim MVN\left(c'\beta,\ \sigma^2 c'(X'X)^{-1}c\right)$

$$\therefore \quad \frac{\hat{\mu}_c - \mu_c}{\sqrt{s^2 c'(X'X)^{-1}c}} \sim t_{n-(p+1)}$$

C.I. & test based off this.

Prediction Interval.

True value: $y_p = c'\beta + \varepsilon_p$

Estimate: $\hat{y}_p = c'\beta$

Prediction error: $y_p - \hat{y}_p$

Distn?

① $E[y_p - \hat{y}_p] = 0$

$$E[\underbrace{c'\beta + \varepsilon_p}_{y_p} - \underbrace{c'\hat{\beta}}_{\hat{y}_p}] = c'\beta + \cancel{E[\varepsilon_p]}^{0} - c'E[\hat{\beta}]$$

$$= c'\beta - c'\beta$$

$$= 0$$

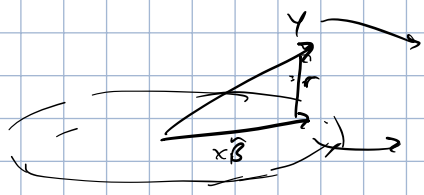
② $Var[y_p - \hat{y}_p] = \sigma^2(1 + c'(X'X)^{-1}c)$

$$\therefore \quad \frac{y_p - \hat{y}_p}{se(y_p - \hat{y}_p)} \sim t_{n-(p+1)}$$

$\hookrightarrow \sqrt{s^2(1 + c'(X'X)^{-1}c)}$

Geometric Interpretation

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underline{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}, \quad x_j = \underbrace{\begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}_{n\times 1}}_{\text{All obsv. of } j^{th} \text{ variable}}, \quad X = \underbrace{[\underline{1}\ x_1 \dots x_j]}_{1\times(p+1) \text{ row vector}}$$



particular linear combination.

$L(x)$ = subspace of all L.C. of X

To minimize $X\hat{\beta}$ & Y:

① Def $r = Y - X\hat{\beta}$

② Orthogonal to all X:

$$X'(Y - X\hat{\beta}) = 0$$

$$X'Y - X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

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

Hat matrix:

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

$$= \underbrace{X(X'X)^{-1}X'}_{H}Y$$

H: projects Y on $L(X)$

Also:

$X'(r) = \hat{Y}'(r) = 0 \Rightarrow$ b/c orthogonal.

ANOVA in MLR

Source	S.S.	d.f.	M.S.
Regression	$SSR = \sum(\hat{y}_i - \bar{y})^2$	p	$MSR = SSR/p$
Error	$SSE = \sum(y_i - \hat{y}_i)^2$	$n-(p+1)$	$MSE = \frac{1}{n-(p+1)} \cdot SSE$
Total	$SST = \sum(y_i - \bar{y})^2$	$n-1$	

Still adds up

F-test of overall significance:

① Hypothesis:

$H_0: \beta_1 = \dots = \beta_p = 0$. H_a: one of them is non-zero

② Stat:

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/n-(p+1)} \sim F_{p,\, n-(p+1)}$$

③ Conclusion:

Choose a critical value & test

R^2:

$$R^2 = \frac{SSR}{SST} \nwarrow \propto \#\text{ of predictors}$$

Doesn't work for MLR ⇒ adjusted R^2

$$\text{Adj. } R^2 = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$$

Interpretation diff.

SPECIFICATION ISSUES

Categorical Variables

Cat. Variable

- Binary: 0/1
- Ordered: e.g. mild, moderate, severe
- Non-ordered: es. R, G, B

Indicator variables:

① # of categories

② Create binary indicator for $n-1$ categories.

Es:

$$x_{i2} = \begin{cases} 1, & \text{if "type"} = t_1 \\ 0 & \text{o.w.} \end{cases}, \quad x_{i3} = \begin{cases} 1, & \text{if "type"} = t_2 \\ 0, & \text{o.w.} \end{cases} \quad \dots$$

If all indicator variables = 0 ⇒ encode into for last cat.

↳ Inclusion: $(X'X)^{-1}$ not possible → not L.I. → not invertible

Numerical method:

1 variable for all categories

$$x_{i2} = \begin{cases} 0 \\ \vdots \\ n \end{cases}, \text{ if type} = t_1, \quad ;$$

→ Assumes linear regression to x_{i2} → not ind.

Model: (Categorical indicator variable)

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots$$

Interpretation of coef. for cat. variables?

① Take mean response for all categories type:

type = "bc" → $y_i = \beta_0 + \beta_1 x_1$ — ①

= "prof" → $y_i = \beta_0 + \beta_1 x_1 + \beta_2$ — ②

= "wc" → $y_i = \beta_0 + \beta_1 x_1 + \beta_3$ — ③

② Take diff!

Avg. diff. in respon b/w "prof" & "bc": ② − ① = β_2

" "wc" & "bc": ③ − ① = β_3

" "wc" & "prof": ③ − ② = $\beta_3 - \beta_2$

Testing:

Funct: is respon stat. diff b/w cat 1 & cat 2:

$H_0: \beta_{2/3-} = 0$ vs. $H_2: \beta_{-} \neq 0$ → t-test

coef. of indicator variable of interest

Special case:

To test $\beta_i - \beta_j = 0$

① Estimate

② Variance: (covariance matrix)

$$Var(\beta_i - \beta_j) = Var(\beta_i) + Var(\beta_j) - 2Cov(\beta_i, \beta_j)$$

③ Standard error: $\sqrt{Var(\beta_i - \beta_j}$

④ t-test

Interaction Effects

Obj: include effect of $x_1 \leftrightarrow x_2$

Create interaction model: (int. effect)

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_i$$

Interpretation: ← main model

Change in unit incr of x_1: $\beta_1 + \beta_3 x_2$

" x_2: $\beta_2 + \beta_3 x_1$

For binary variables:

Coefficient of interaction → if slope is different b/w categ.

↓

Test:

Obj: is relationship between b/w y & x_1, $x_2 = 0$ diff compared to rel. b/w y & x_1, $x_2 = 1$

H_0: interaction coef. = 0 vs. IC $\neq 0$

↓

can drop interaction effect.

GENERAL LINEAR HYPOTHESES

Tests

- Single variable is significant → t-test
- All variables are non-sig. → F-test for sig.
- Specific combo of variable sign. → General linear hyp.

Test:

$H_0: A\beta = \vec{0}$ $H_a: A\beta \neq 0$

A: matrix of $l \times (p+1)$

↑ # of constraints on β

Ex:// $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$H_0: \beta_2 = \beta_3 = 0$ H_a: o.w.

↓

① # of liner constraints

$l = 2$

② Write out matrix

$$\underbrace{\begin{bmatrix} a & b & c & d \\ e & f & g & h \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}}_{\beta} = \begin{bmatrix} a\beta_0 + b\beta_1 + c\beta_2 + d\beta_3 \\ -\beta_3 - \beta_1 \quad \beta_2 - \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

∴ Put 0's in cols of variable I don't care about

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ex:// $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$H_0: \beta_1 = 0, \quad \beta_2 = \beta_3, \quad H_a: \text{o.w.}$

① Make R.H.S of constr. to be 0:

$\beta_1 = 0, \quad \beta_2 - \beta_3 = 0$

② Make A:

$$A = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \beta = \begin{bmatrix} \beta_1 \\ \beta_2 - \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Q: What is actual test?

1. Reduced model: model if $A\beta = 0$
2. Given reduc model → find fitted value $(\hat{Y}_A)$
3. Find SSE of $\hat{Y}_A \Rightarrow \sum (y_i - \hat{y}_{i,A})^2 = SSE_A$
4. Test statistic:

$$\frac{\overbrace{(SSE_A - SSE)}^{\text{full model}} / \ell}{SSE / n-p-1} \sim F_{\ell,\, n-p-1}$$

Principle:

$SSE - SSE_A$ is "extra" error explained by including full variab

↳ we hope that $SSE - SSE_A \approx SSE$ if reduced model is good!

↳ we hope test statistic is large → reject H_0

Ex:// Prestige example:

Full model: $y_i = \beta_0 + \underbrace{\beta_1 x_{i1}}_{\text{educ}} + \underbrace{\beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{"type" indicators}} + \underbrace{\beta_4 x_1 x_2 + \beta_5 x_1 x_3}_{\text{Interaction effects}}$

Question: is the effect of education the same for diff. types?

① Hypotheses

$H_0: \beta_4 = \beta_5 = 0$ vs. $H_a: \neg H_0$

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

② Create test statistic

A: Reduced model ⇒ assume H_0 is the model?

$$y_r = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

B: Fit both full & reduced model

C: Find SSE for each model

In R output ⇒ residual sum of sq. $= \sqrt{MSE}$

$\therefore$ $SSE = (\text{res. ss})^2 \times \text{d.f.}$

D: Create F-stat

$$\frac{(SSE_A - SSE)/\ell}{SSE/(n-(p+1))} = 0.844$$

③ Compare against critical val.

Is $F > F_{\alpha, \ell, n-(p+1)}$? ⇒ No it wasn't, don't reject H_0

ANOVA in R:

① anova (model$_{\text{reduced}}$, model)

↳ Model 1: ~ —

Model 2: ~ — + —

	Res. d.f.	RSS	d.f.	Sum of sq.	F	prval
1	model 1 d.f.	$SSE_{reduced}$				
2	" 2 "	SSE_{full}	Δ d.f. $= \ell$	$SSE_A - SSE$	f-stat	$P(F > f\text{-val})$

Model Assumptions And Residual Analysis

Intro:

Assumptions:

① Linear rel. b/w y & $x_1 \dots x_p$ — most important

② $E[\varepsilon_i] = 0$

③ $\text{Var}[\varepsilon_i] = \sigma^2 \quad \forall i$ (constant variance)

④ $\varepsilon_i \sim N$

⑤ $\varepsilon_i \perp \varepsilon_j \quad \forall i \neq j$
↳ OK if random sampling — least important

If assumptions not true → analysis is faulty

Monitoring ε_i is hard ← random error from theoretical model
↳ Use residuals instead.

Claim: Residuals behave very similar to random error:

Proof:

$$\begin{aligned} r &= Y - \hat{Y} \\ &= (I - H)Y \\ &= (I - H)(X\beta + \varepsilon) \\ &= (X\beta - HX\beta) + (I - H)\varepsilon \\ &= (X\beta - X\beta) + (I - H)\varepsilon \\ &= (I - H)\varepsilon \end{aligned}$$

$\Rightarrow r$ is a linear transform of ε!

$X\beta - HX\beta$
$= X\beta - X(X'X)^{-1}X'X\beta$ ($(X'X)^{-1}X'X = I$)
$= X\beta - X\beta$
$= 0$

$$\therefore r \sim MVN(0, (I-H)'\sigma^2 I (I-H))$$
$$r \sim MVN(0, \sigma^2(I-H))$$

Analysis: If off-diag of $I - H \neq 0 \Rightarrow r_i$ is not indep.

But, if $H \ll I \Rightarrow r_i \doteq \varepsilon_i$

Residual Analysis

Types of residuals:

① Raw

$$r_i = y_i - \hat{y}_i$$

② Standardized:

$$d_i = \frac{r_i}{\hat{\sigma}} \sim N(0,1)$$
$\hat{\sigma} \longrightarrow$ estimate $\sqrt{\text{Var}(r_i)}$

Effectively standardizes b/c $\text{Var}(r_i) \approx \sigma^2$ if $H \ll I$

③ Studentized:

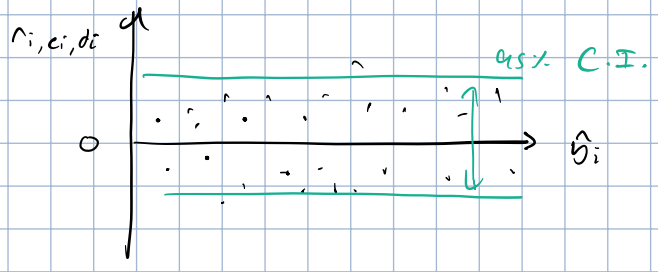
$$e_i = \frac{r_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}} \hookrightarrow \text{Actual variance.}$$

Departure of assumptions should reflect in residuals. We plot residual!

Plots:

① Residuals vs. fitted value

Good:

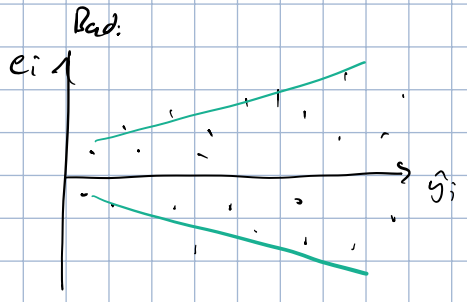


Random scatter around x-axis within horizontal band.

Why: $\hat{Y} \perp r$

Conclusion: no visible model defect.

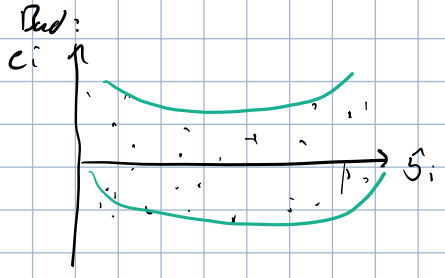
Bad:



Violation of common variance

Hetero scedasitivity

Bad:



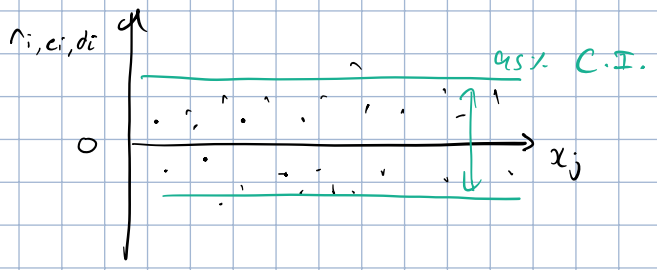
Linear model is not adequate

↳ Non-linear rel. btw x

↳ Missing variable

② Residual vs. explanatory

Good:

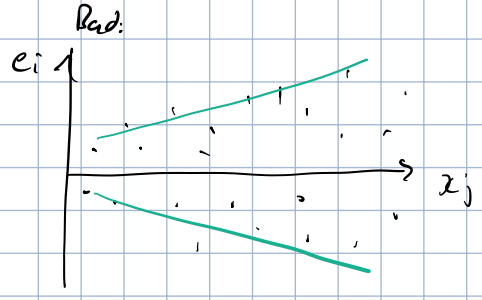


Random scatter around x-axis within horizontal band.

Why: $x_j \perp r$

Conclusion: no visible model defect.

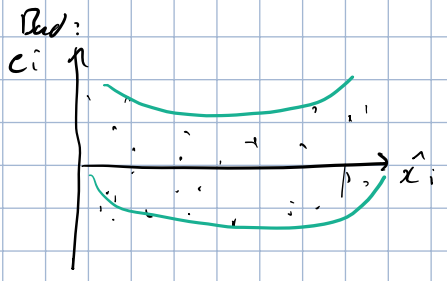
Bad:



Violation of common variance

Hetero scedasitivity

Bad:



Linear model is not adequate
↳ Non-linear wrt x_j (use x_j^2?)
↳ Missing variable

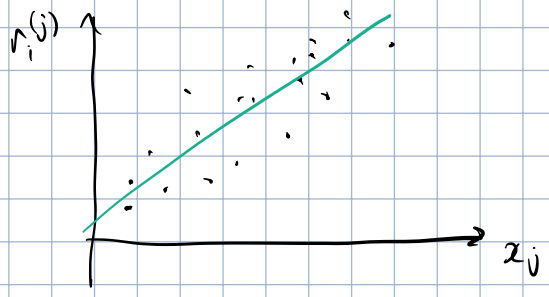
③ Partial residual plot

Partial residual of x_j: $r_i^{(j)} = r_i + \hat{\beta}_j x_{ij}$

r_i: full model residuals

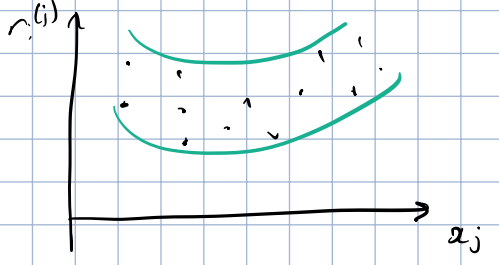
$\hat{\beta}_j x_{ij}$: effect of x_j

Good:



x_j is linear

Bad:



Non-linear → higher order term might help.

④ Q-Q plots

Check normality of r_i / d_i / e_i

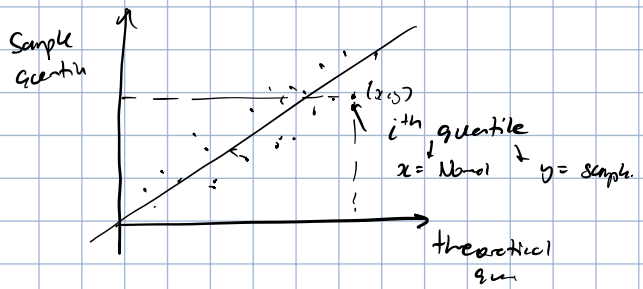
Steps:

① Order standardized residuals $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n)}$

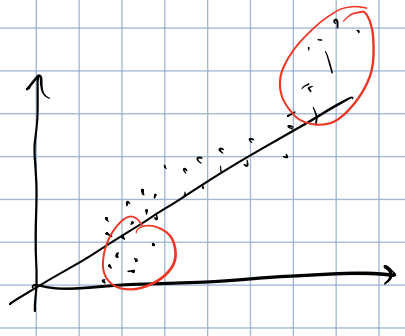
② Find quantile of data

③ Plot against quantile of $N(0,1)$

Good:

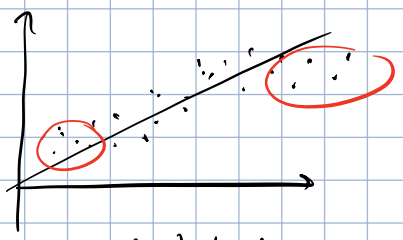


$e_i \sim N(0,1)$



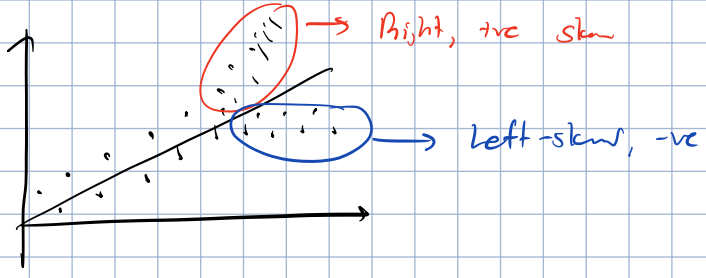
Heavy-tailed ⇒ outliers.

Hard to fix



Light-tailed

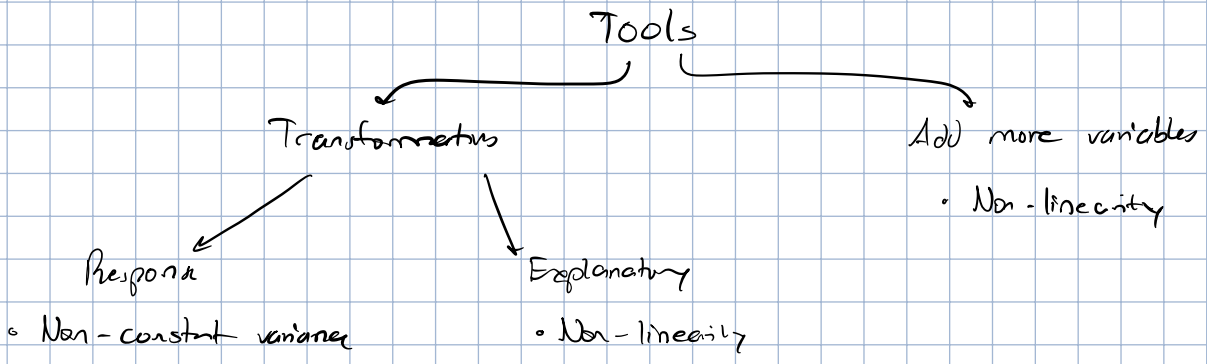
Fine



Skewed

Transform → make normal

Addressing Model Assumption Problems



Transforming Response Variables

Obj: constant variance → apply transform $g(y)$ → fit model $g(y) = \dots$

Assumption: variance of y is func. of mean μ_i

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\varepsilon_i) \\ &= h(\mu_i)\,\sigma^2 \end{aligned}$$ } Assume $h > 0$

Back off goal:

$$\text{Var}(g(y_i)) = \sigma^2$$

Taylor expansion at μ_i:

$$g(y_i) \doteq g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$$

$$\begin{aligned} \text{Var}(g(y_i)) &\doteq g'(\mu_i)^2 \ \text{Var}(y_i) \\ &\doteq g'(\mu_i)^2 \cdot h(\mu_i)\,\sigma^2 \\ \sigma^2 &= g'(\mu_i)^2 \cdot h(\mu_i)\,\sigma^2 \end{aligned}$$

$$\boxed{g'(\mu_i)^2 \propto \frac{1}{h(\mu_i)}}$$

In practice: make assumption of $h(\,)$ → use formula to find $g(\cdot)$

Ex:// a) $h(\mu_i) = \mu_i \Rightarrow \text{Var}(y_i) = \sigma^2 \mu_i$

$$g'(\mu_i)^2 \propto \frac{1}{\mu_i}$$

$$g'(\mu_i) \propto \frac{1}{\sqrt{\mu_i}}$$

$$g(\mu_i) \propto \int \frac{1}{\sqrt{\mu_i}} = \sqrt{\mu_i} + C$$

$\therefore$ g should be sqrt transform.

Box-Cox Transform

$$g(y_i) = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log y_i, & \lambda = 0 \end{cases}$$

Q: How to choose λ?

Try many λ → choose one of highest log likelihood.

Special case:

$\lambda = 1/2$ → sqrt

$\lambda = 0$ ⇒ log

$\lambda = 1$ → no transform

$\lambda = -1$ → reciprocal transform.

Interpretation changes w/ transform ⇒ $\hat{\beta}_j$ is change in $g(y_i)$ if x_j ↑

Ex:// y_i is log transform.

$$\log y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Original:

$$y = e^{\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p}$$

$\therefore$ 1 unit chng in $\hat{\beta}_j$ ⇒ $100(e^{\hat{\beta}_j} - 1)$ % change in y

Transform explanatory vars. / add more vars.

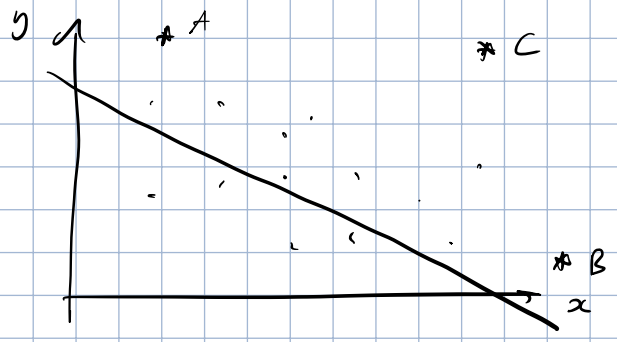
① Power transform of variable

② Polynomial term additive:

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Effects of Individual Observations

Intro



Outliers:
- A: outlier in respon
- B: outlier in expl.
- C: outlier in respon & expl.

Outliers have big impact on regression

Q: Why do outliers occur?

1. Human error
2. Sampling issue (sample from wrong pop.)
3. Natural variation

Q: What do we do w/ outliers:

Human error → correct/remove. O.W. → don't!

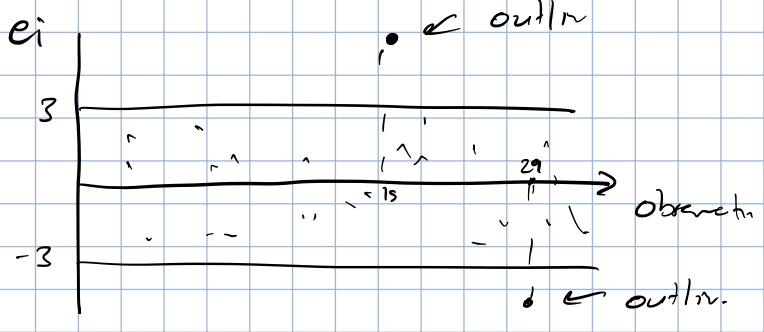
Studentized Residuals - Outliers in Response

Studentized residual:

$$e_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \sim N(0,1)$$

If $|e_i| > 3 \Rightarrow$ outlier in response (v. extreme in standard normal distn.)

Plot:



Leverage - Outliers in Explanatory

Theorem: h_{ii} indicates rel. b/w $\hat{y}_i$ & y_i.

↳
$$\hat{y} = X\hat{\beta}$$
$$\hat{y} = Hy$$

Consider $\hat{y}_i$:

$$\hat{y}_i = \begin{bmatrix} h_{i1} & h_{i2} & \cdots & h_{ip} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= h_{ii} y_i + \sum_{i \neq j} h_{ij} y_j$$

Properties of h_{ii}:

① $0 < h_{ii} \leq 1$

Proof:

From idempotency ($HH = H$)

$$h_{ii} = \begin{bmatrix} h_{i1} & \cdots & h_{ip} \end{bmatrix} \begin{bmatrix} h_{1i} \\ \vdots \\ h_{ni} \end{bmatrix}$$

$$= \sum_j h_{ij} h_{ji}$$

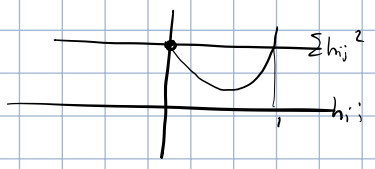
From symmetry: $h_{ij} = h_{ji} \quad \forall i, j$

$$\therefore \sum h_{ij} h_{ji} = \sum h_{ij}^2$$

$$= h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

$$\therefore h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

$$h_{ii}(1 - h_{ii}) = \boxed{\sum h_{ij}^2} \geq 0$$



$\therefore h_{ii} \in (0, 1]$

② As $h_{ii} \to 1$, off diagonal $\to 0$

Proof:

$$\underbrace{h_{ii}(1 - h_{ii})}_{\to 0 \text{ if } h_{ii} \to 1} = \underbrace{\sum h_{ij}^2}_{\to 0}$$

Impl:

$$\hat{y}_i = y_i h_{ii} + \cancel{\sum h_{ij} y_j} \quad 0 \text{ if } h_{ii} \to 1$$

$$\therefore \hat{y}_i \approx y_i$$

③ If $h_{ii} >> h_{ij} \Rightarrow \hat{y}_i$ is det. by y_i

$\therefore$ If i was an outlier $\to$ fitted value in model will be skewed.

$\therefore$ i has high leverage!

④ If obs. $(x_{i1} \cdots x_{ip})$ is close to centroid of expl. $(\bar{x}_1, \ldots, \bar{x}_p)$

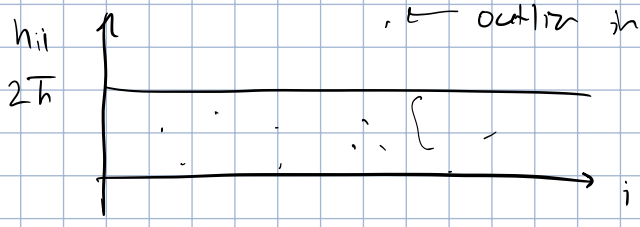
$\Leftarrow$ h_{ii} is small.

Rule of thumb:

$h_{ii} > 2\bar{h} = \frac{2(p+1)}{n} \Rightarrow$ i^{th} obs. has high leverage.

Corollary: From prop. 4, outlier in explanatory.

Plot:



Cook's Distance - Influential Datapoints

Defn of infl. d.p: removing point causes big change in $\hat{\beta}$

Note: $\hat{\beta}^{(i)}$ is fitted $\hat{\beta}$ after removing i^{th} obs.

To find distance b/w $\hat{\beta}$ & $\hat{\beta}^{(i)} \Rightarrow$ Cook's distance.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{(i)})' (X'X) (\hat{\beta} - \hat{\beta}^{(i)})}{\hat{\sigma}^2 (p+1)}$$

$$= e_i^2 \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{p+1}$$

$\therefore D_i$ is high if obs. is influential in both expl. & response.

Just b/c outlier → not always influential

Rule of thumb:

$$D_i = \begin{cases} > 0.5 \Rightarrow \text{might be influential} \\ > 1 \text{ / sig. diff from other data pnt} \Rightarrow \text{influential} \end{cases}$$

MODEL SELECTION

Goal: find optimal # of predictors

Selection Criteria

① R^2_{adj}

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-k-1}\cdot(1-R^2)$$

k = # of predictors

Will penalize if $k \uparrow$

If we just use R^2, model with more predictors will have bigger R^2

② Akaike Information Criteria (AIC)

$$AIC = 2q - 2\ln(L(\hat{\theta})) \qquad -\frac{n}{2}\log\left(\frac{SSE}{n} + \text{constant}\right)$$

q = # of model params = $k+2$ (β_0, σ^2)

The lower AIC, better the model

AIC penalizes high predictor models

③ Bayesian Info. Criteria (BIC)

$$BIC = q\ln(n) - 2\ln(L(\hat{\theta}))$$

Penalizes high predictor models more & incl. sample size.

None of these interpretable.

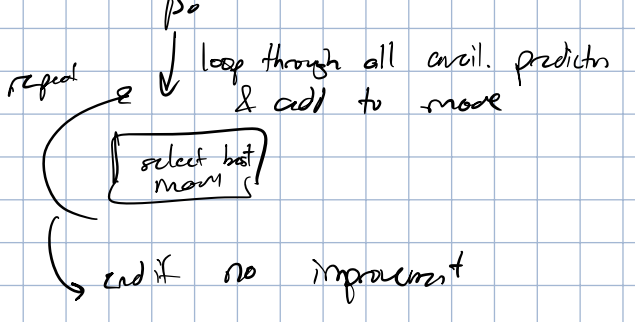
Search Strategy

① All subset regression

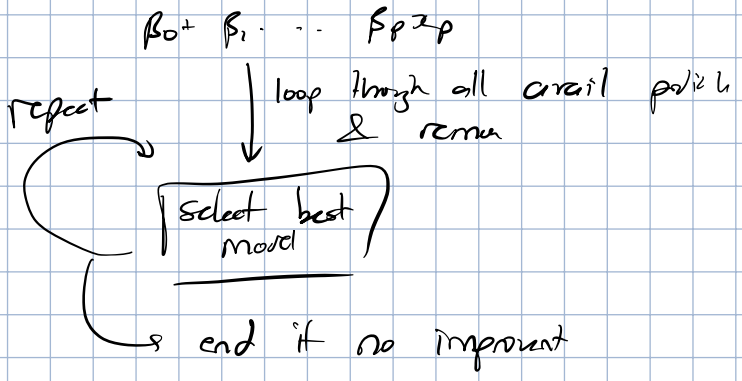
Find all subsets of predictors → fit → choose best

Very expensive (2^p models)

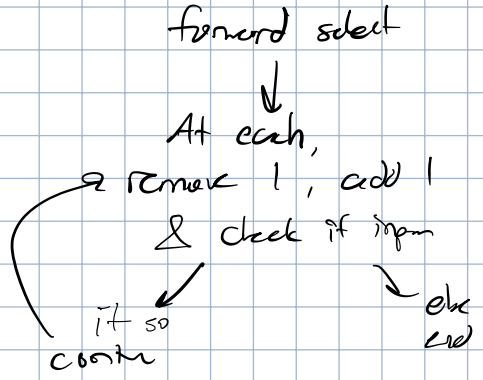
② Forward selection



③ Backward selection:



④ Forward-backward stepwise:



BUILDING PREDICTIVE MODELS

Metrics to evaluate model on new data (on testing/validation data)

① Mean squared prediction error

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

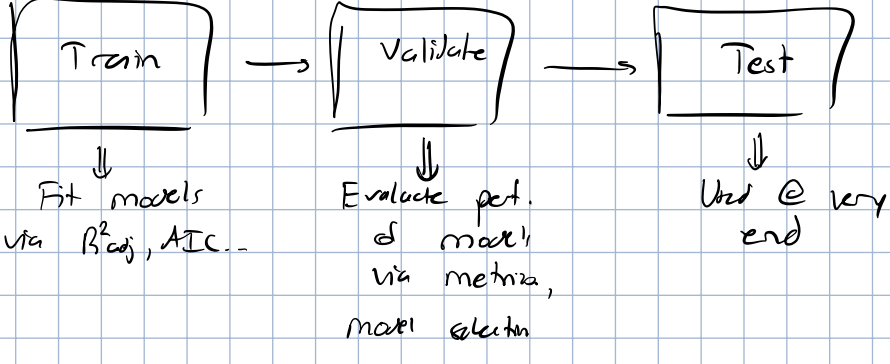
② Root mean squared error

$$RMSE = \sqrt{MSPE}$$

③ Mean absolute error:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Process of building:

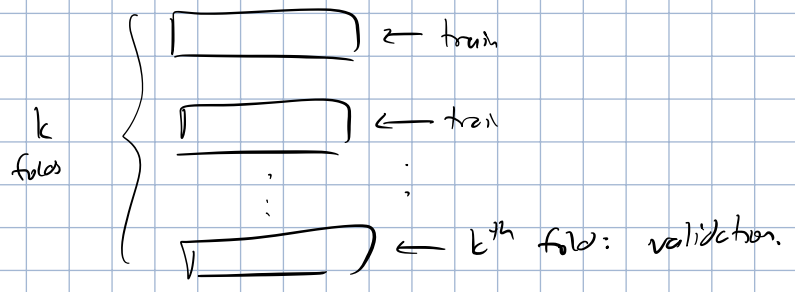


How to split data?

① Randomly
- \+ Simple
- − Variation in diff. splits ← randomness

② K-Fold C.V.

1. Divide data into k folds



2. Train model on k-1 fold, eval on k^{th} fold
3. Repeat so each fold becomes eval fold once
4. Take avg. of error

$$\frac{1}{k} \sum_{k=1}^{K} MSPE_k$$