Unit III Coda

Regression Example,
Regression Issues

A Regression Example

Quiz 1

Question: Does extra preparation time improve performance on a quiz?

Population: MIT 2.086 students (present, future).

Experiment: Administer same quiz to
m^1 students at t^1,
m_2 students at t^2,
m_3 students at t^3.

Model $n = 2$

$t \equiv$ time $\qquad t^1 \equiv 1$; $t^2 \equiv 3$; $t^3 \equiv 5$
(M) (W) (F)

$y \equiv$ grade $\qquad 0 \le y \le 100$

$$Y_i = \underbrace{\beta_0^{\text{true}} + \beta_1^{\text{true}}\, t_i}_{Y_{\text{model}}(t_i;\, \beta^{\text{true}})} + \epsilon_i , \quad 1 \le i \le m$$

ϵ_i: N1, N2, N3

grade of i^{th} student

day on which i^{th} student takes quiz ($t_i \in \{1,3,5\}$)

$$\Downarrow$$

$$E(Y_i) = Y_{\text{model}}(t_i;\, \beta^{\text{true}})$$

An engineering context: Heat Treatment

$t \equiv$ time part removed from furnace (duration)

$Y \equiv$ toughness of part

$$Y_i = \beta_0^{true} + \beta_1^{true} t_i + \epsilon_i \quad \sim N1, N2, N3 \qquad ?$$

toughness of i^{th} part

time at which i^{th} part removed

$$\Downarrow$$

$$E(Y_i) = \beta_0^{true} + \beta_1^{true} t_i$$

Data

M_grades $m^1 \times 1$

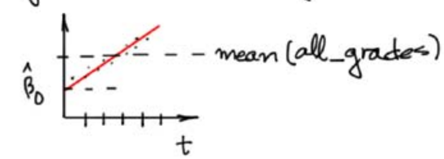
W_grades $m^2 \times 1$

F_grades $m^3 \times 1$

all_grades = [M_grades; W_grades; F_grades] Y

$m = m^1 + m^2 + m^3$ (= length(all_grades))

Note $\hat{\beta}_0 \neq$ in general mean(all_grades):



Regression Matrix X . . .

$$(X\beta)_i = Y_{model}(t_i; \beta), \quad 1 \le i \le m$$

$$= \beta_0 + \beta_1 t_i, \quad 1 \le i \le m$$

$$= \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad (\text{any } \beta)$$

$X \quad m \times n^{(2)}$

. . . Regression Matrix X

$$(X\beta)_i = Y_{model}(t_i; \beta), \quad 1 \le i \le m$$

$$= \beta_0 + \beta_1 t_i, \quad 1 \le i \le m$$

$$= \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad (\text{any } \beta)$$

$X \quad m \times n^{(2)}$

[ones(m,1); [ones(m^1,1), 3*ones(m^2,1); 5*ones(m^3,1)]]

Regression

```
beta_hat = X \ Y ;

sigma_hat = sqrt(1/(m-2)) * norm(Y - X*beta_hat) ;

XTXinv = inv(X'*X) ;

ci_0 = ...          ;

ci_1 = [beta_hat(2) -
     sigma_hat * sqrt(XTXinv(2,2)) * sqrt(finv(0.95,2,m-2)),
         beta_hat(2) +
     sigma_hat * sqrt(XTXinv(2,2)) * sqrt(finv(0.95,2,m-2))] ;
```

Y_hat (X*beta_hat)

(minimum residual)

... Results DEMO

"Issues" in Regression Analysis

Setting

x independent, y dependent

$$Y_i = 0 + 1 \cdot x_i + \alpha x_i^2 + \epsilon_i \, , \quad 1 \le i \le m$$

$$Y_{model}(x;\beta) = \sum_{j=0}^{n-1} \beta_j x^{j-1}$$

$$= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \beta_{n-1} x^{n-1}$$

If $\alpha = 0$, $n \ge 2$, and ϵ_i satisfies N1, N2, N3

$$\beta^{true} = (0 \; 1 \; 0 \cdots 0)^T ;$$

if $\alpha > 0$, $n \ge 3$,

$$\beta^{true} = (0 \; 1 \; \alpha \cdots 0)^T .$$

What can go wrong, and how can we at least detect the issue? and perhaps remedy the problem?

Overfitting: $m \approx n$

Design of Experiment: choice of x_i, $1 \le i \le m$ $\Rightarrow X$

Underfitting: no β^{true} $\Rightarrow$ bias

Correlated Noise: (N1, N2, or) N3 not satisfied

Repetition: challenge assumptions

⋮

DEMO

2.086 Numerical Computation for Mechanical Engineers
Spring 2013