

STAT 538 : LEC. #1 - LINEAR MODEL

outcome
dependent variable
response

(x_1, \dots, x_{p-1})

explanatory
predictors

features

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1} + \epsilon$$

covari

regress

regression

or

$$\underset{\substack{\uparrow \\ n \times 1}}{\tilde{y}} \sim N_n \left(\underset{\substack{\uparrow \\ n \times p}}{X} \underset{\substack{\uparrow \\ p \times 1}}{\tilde{\beta}}, \underset{\substack{\uparrow \\ n \times n}}{\sigma^2 I} \right)$$

i th-row is
values of $(1, x_1, x_2, \dots, x_{p-1})$
for i -th unit

- independent
- mean zero
- $N(0, \sigma^2)$

$\rightarrow E(\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1})$

Note: regression modelling of Y given x_1, \dots, x_{p-1}

Interesting running ex. in Ch. 1

Gore vs. Bush
2000

②

Units are the 159 counties of Georgia

Y is relative undercount = $\frac{\text{ballots} - \text{votes}}{\text{ballots}}$

explanatory variables include (0% to 19%)

- type of voting machinery 5 unordered categories
- economic status 3 ordered categories
- % African-American
- rural/urban, Atlanta? each binary
- votes for Gore/Bush/other

many plausible variables
X
- correlated with one another

SOMEWHAT TYPICAL REGRESSION EX. ?

Nice features of linear models?

③

① Huge generality, via ubiquity of design matrix

(line, curve, multi-groups, interactions, etc.)

i -th row could be:

• $1 \quad X_1^{(i)}$

• $1 \quad X_1^{(i)} \quad (X_1^{(i)})^2$

• $1 \quad X_1^{(i)} \quad X_2^{(i)} \quad X_1^{(i)} X_2^{(i)}$

⋮

↑

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & & & \\ 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \vdots & & & \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ \vdots & & & \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ \vdots & & & \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

ⓑ Estimation is numerically easy

$$\hat{\beta} = \underbrace{(X^T X)^{-1}}_{p \times p} \underbrace{X^T}_{p \times n} \underbrace{y}_{n \times 1}$$

// Gauss-Markov theorem

(and agreed upon, LS, MVUE, MLE, flat-prior limit of Bayes, all lead to this)

minimizer of $\|y - X\hat{\beta}\|^2 = \sum_{i=1}^n (y_i - X^{(i)}\hat{\beta})^2$

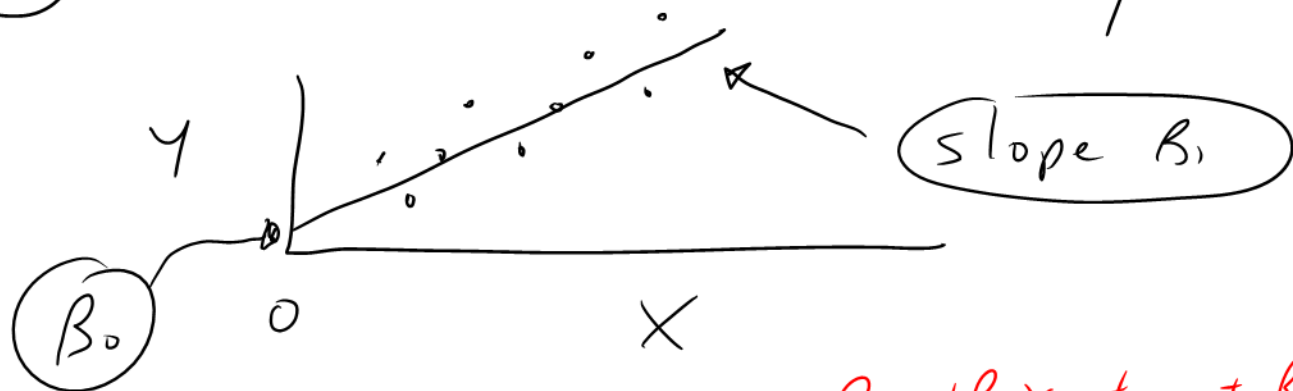
↑
ith row of the design matrix

related $\hat{\sigma}^2 = \frac{1}{n-p} \|y - X\hat{\beta}\|^2$

is (unbiased) estimate of $\sigma^2 = \text{Var}(Y | X)$

③ Parameters are easily interpreted

⑤



$$= \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

$$\beta_j = \frac{\partial}{\partial x_j} E(Y | x_1, \dots, x_{p-1})$$

provided ...

x_j doesn't appear in any interaction terms, or any nonlinear terms

i.e. no $x_j x_1$, no x_j^2 , etc, etc,

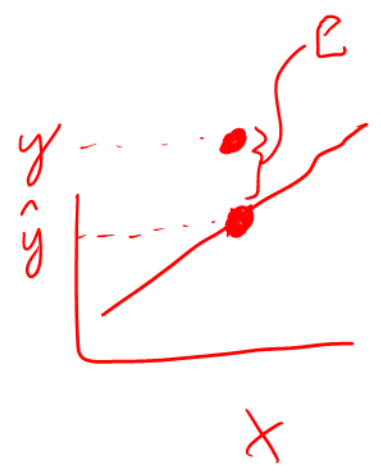
④ Model adequacy is readily assessed ⑥

FITTED VALUES

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

RESIDUALS

$$e = y - \hat{y}$$

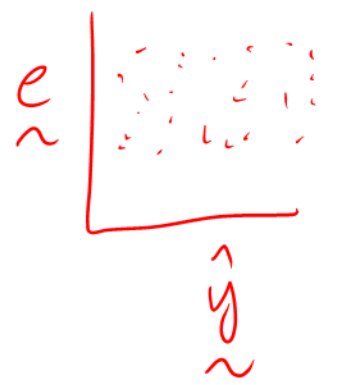


$$= X (X^T X)^{-1} X^T y$$

H "hat matrix"

- idempotent
- trace(H) = p

connection to n-p on prev. slide



(E) Distributional theory for hypothesis tests and CIs is well developed (N-t-F)

$$E(\hat{\beta}) = \beta, \quad \text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

w.r.t. dist of (Y|X), not dist of (Y,X)

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_{n-p} \quad \text{gives CI}$$

so SE($\hat{\beta}_j$) is j-th diagonal element of $\sigma^2 (X^T X)^{-1}$

or generally comparing nested LMS

$$RSS_{red} = \| \underline{y} - X_{red} \hat{\beta}_{red} \|^2$$

$$RSS_{full} = \| \underline{y} - X_{full} \hat{\beta}_{full} \|^2$$

If reduced model is true,

$$\frac{(p-q)^{-1} \{ RSS_{red} - RSS_{full} \}}{(n-p)^{-1} RSS_{full}} \sim F_{p-q, n-p}$$

Back to Georgia voting

(8)

After much consideration and experimentation...

- form of Y ^{relative} ^{undercount}
- which X 's to include
 - substantive
 - testing
 - stepwise selection AIC
- how to include
 - transformations
 - interactions
- whether to weight units
- ... a 'final' model is reported with some guidance on interpretation

LM inextricably linked to:

$$Y = E(Y | X_1, \dots, X_p) + \text{"noise"}$$

Many important applications are hard/impossible to shove into this framework

BINARY

$$Y = \begin{cases} 1 & \text{patient dies} \\ 0 & \text{patient lives} \end{cases}$$

can't transform problem away for any g

$$Y \neq \Pr(Y=1 | X_1, \dots, X_p) + \text{noise}$$

$$\text{(and } g(Y) \neq E[g(Y) | X_1, \dots, X_p] + \text{noise)}$$

Also categorical, count, proportion

$$Y = \begin{cases} \text{Brand A} \\ \text{Brand B} \\ \text{Brand C} \end{cases}$$

$$Y = \begin{cases} \text{complete recovery} \\ \text{partial recovery} \\ \text{no recovery} \end{cases}$$

$$Y = \# \text{ hospital admissions on day } i$$

SO THIS COURSE HAS A RAISON D'ÊTRE !

We see that the rural:perAA can be dropped. A subsequent test reveals that rural can also be removed. This gives us a final model of:

```
> finalm <- lm(undercount~equip + econ + perAA + equip:econ
+ equip:perAA, gavote)
```

```
> summary(finalm)
```

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.04187	0.00503	8.33	6.5e-14
equipOS-CC	-0.01133	0.00737	-1.54	0.12670
equipOS-PC	0.00858	0.01118	0.77	0.44429
equipPAPER	-0.05843	0.03701	-1.58	0.11669
equipPUNCH	-0.01575	0.01875	-0.84	0.40218
econpoor	0.02027	0.00553	3.67	0.00035
econrich	-0.01697	0.01239	-1.37	0.17313
perAA	-0.04204	0.01659	-2.53	0.01239
equipOS-CC:econpoor	-0.01096	0.00988	-1.11	0.26922
equipOS-PC:econpoor	0.04838	0.01380	3.51	0.00061
equipPAPER:econpoor	NA	NA	NA	NA
equipPUNCH:econpoor	-0.00356	0.01243	-0.29	0.77492
equipOS-CC:econrich	0.00228	0.01538	0.15	0.88246
equipOS-PC:econrich	-0.01332	0.01705	-0.78	0.43615
equipPAPER:econrich	NA	NA	NA	NA

equipPUNCH:econrich	0.02003	0.02200	0.91	0.36405
equipOS-CC:perAA	0.10725	0.03286	3.26	0.00138
equipOS-PC:perAA	-0.00591	0.04341	-0.14	0.89198
equipPAPER:perAA	0.12914	0.08181	1.58	0.11668
equipPUNCH:perAA	0.08685	0.04650	1.87	0.06388

Residual standard error: 0.02 on 141 degrees of freedom

Multiple R-Squared: 0.428, Adjusted R-squared: 0.359

F-statistic: 6.2 on 17 and 141 DF, p-value: 1.32e-10

Because there are only two paper-using counties, there is insufficient data to estimate the interaction terms involving paper. This model output is difficult to interpret because of the interaction terms.

Conclusion: Let's attempt an interpretation of this final model. Certainly we should explore more models and check more diagnostics, so our conclusions can only be tentative. The reader is invited to investigate other possibilities.

To interpret interactions, it is often helpful to construct predictions for all the levels of the variables involved. Here we generate all combinations of `equip` and `econ` for a median proportion of `perAA`:

```
> pdf <- data.frame(econ=rep(levels(gavote$econ),5),  
  equip=rep(levels(gavote$equip),rep(3,5)),perAA=0.233)
```

We now compute the predicted undercount for all 15 combinations and display the result in a table:

```
> pp <- predict(finalm,new=pdf)  
> xtabs(round(pp,3) ~ econ + equip, pdf)
```

	equip				
econ	LEVER	OS-CC	OS-PC	PAPER	PUNCH
middle	0.032	0.046	0.039	0.004	0.037
poor	0.052	0.055	0.108	0.024	0.053
rich	0.015	0.031	0.009	-0.013	0.040

We can see that the undercount is lower in richer counties and higher in poorer counties. The amount of difference depends on the voting system. Of the three most com-

We create a three-level factor for the three levels of `perAA` to aid the construction of the table:

```
> propAA <- gl(3,1,15,labels=c("low", "medium", "high"))
> xtabs(round(pp,3) ~ propAA + equip,pdf)
      equip
propAA  LEVER OS-CC OS-PC PAPER PUNCH
low     0.037 0.038 0.045 -0.007 0.031
medium  0.032 0.046 0.039 0.003 0.036
high    0.027 0.053 0.034 0.014 0.042
```

We see that the effect of the proportion of African Americans on the undercount is mixed. High proportions are associated with higher undercounts for OS-CC and PUNCH and associated with lower undercounts for LEVER and OS-PC.