STAT 538 : LEC. #1 - LINEAR MODELS

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

or

$$\underset{\sim}{Y} \sim N_n\left(X\underset{\sim}{\beta}, \; \sigma^2 \underline{I} \right)$$

Note: regression modelling of Y <u>given</u> X, regardless of sampling scheme

Interesting running ex. in Ch. 1

Units are the 159 counties of Georgia

Y is relative undercount

explanatory variables include

- type of voting machinery
- economic status
- % African-American
- rural/urban, Atlanta?
- votes for Gore/Bush/other

SOMEWHAT TYPICAL REGRESSION EX.?

Nice features of linear models?

(A) Huge generality, via ubiquity of design matrix

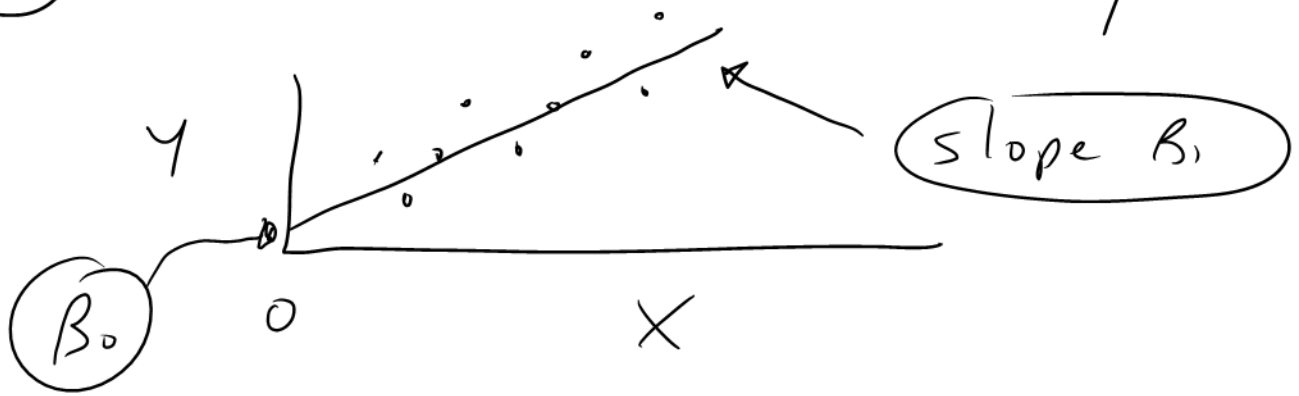
(line, curve, multi-groups, interactions, etc.)

(B) Estimation is numerically easy

$$\underset{\sim}{\hat{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{y}$$

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

(C) Parameters are easily interpreted



$$\beta_j = \frac{\partial}{\partial x_j} E(y | x_1, \ldots, x_p)$$

provided ...

(D) Model adequacy is readily assessed

FITTED VALUES

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

RESIDUALS

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

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

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

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

or generally comparing nested LMs

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

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

After much consideration and experimentation...

- form of Y
- 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:

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

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

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

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

Also categorical, count, proportion

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.