STAT 526

# Topic X
# Generalization Thoery

Dr. Qifan Song

# Generalization Error

Generalization Thoery studies how in-sample predictive performance generalizes to out-of-sample prediction task.

- In-sample performance / training error : $E_S(l_{\hat{h}}(X)) = \sum l_{\hat{h}}(X_i)/n$, where $E_S$ denotes expectation over empirical distribution, $l$ is the loss and $h$ is the estimator, $X$ denotes the a sample.
- Out-of-sample performance / testing error : $E_D(l_{\hat{h}}(X))$, where $E_D$ denotes expectation over data distribution
- $h \in \mathcal{H}$ is the parameter space
- Generalization error: $G = E_D(l_{\hat{h}}(X)) - E_S(l_{\hat{h}}(X))$
- IMPORTANT: $h$ is usually data dependent, i.e., $\hat{h} = \hat{\theta}(X_1, \ldots, X_n)$. Law of large number or Central Limit Theory doesn't trivially apply to $G$.

# Information-Based Result

- KL divergence $KL(P||Q) = E_P(\log(p/q))$.

Define a density $p^* \propto \exp[g(x)]q(x)$ for arbitrary $g$, then

$$KL(P||P^*) - KL(P||Q) = \int \log \frac{q}{p^*} p dx = -E_P h + \log E_Q(\exp(g)).$$

Thus

$$E_P[g] - KL(P||Q) = \log E_Q(\exp(g)) - KL(P||P^*) \leq \log E_Q(\exp(g))$$

where the equality holds when $\exp[g(x)]q(x) \propto p$. This implies Donsker-Varadhan representation of KL

$$\sup_g \{E_P[g] - \log E_Q(\exp(g))\} = KL(P||Q).$$

- Given joint distribution $D_{XY}$ and its marginal $D_X$ and $D_Y$, mutual information between $X$and $Y$ is $I(X;Y) = KL(D_{XY}||D_X D_Y)$
- The mutual information tells how much $X$ and $Y$ shares overlapping information. If $X$ and $Y$ are independent, then $I(X;Y) = 0$.

# Information-Based Result

$\hat{h}$ and $X$ have a joint distribution. Apply D-V to $P = P_{h,X}$, $Q = P_{\hat{h}} P_X$ leads to

$$E_{P_{\hat{h},X}} \lambda G \leq \log E_{P_{\hat{h}}} E_{P_X} \exp\{\lambda G\} + I(\hat{h}; X)$$

for any $\lambda$. The LHS is the expected generalization error.

Suppose that $l$ is bounded $[0, 1]$. For a fixed $\hat{h}$, $E_{P_X} \exp\{\lambda G\}$ is the MGF of a zero-mean r.v. $\sum_i [l_{\hat{h}}(X_i) - E l_{\hat{h}}(X)]/n$. One can show

$$\log E_{P_X} \exp\{\lambda G\} \leq \frac{\lambda^2}{8n}$$

Direct result of Hoeffding's Lemma: $Ee^{\lambda X} \leq \exp\{\lambda E(X) + \lambda^2 (b-a)^2/8\}$ if $X \in [a, b]$.

Then $E_{P_{h,X}} G \leq \lambda/8n + I(\hat{h}; X)/\lambda$, which implies

$$E_{P_{h,X}} G \leq \sqrt{\frac{I(\hat{h}; X)}{2n}}$$

# Information-Based Result

- Mutual information between $X$ and $\hat{h} = \hat{\theta}(X)$ is

$$\int \log \frac{f_X(X) I(h = \hat{\theta}(X))}{f_X(X) f_{\hat{h}}(h)} f_X(X) I(h = \hat{\theta}(X)) dX dh$$
$$= - \int \log f_{\hat{h}}(\hat{\theta}(X)) f_X(X) dX$$

- For single model (linear regression), we can compute it (Let's try).

- For most of the data science problems, $I(\hat{h}; X)$ is difficult to compute.

# A more conservative bound

- Instead of studying $G = E_D(l_{\widehat{h}}(X)) - E_S(l_{\widehat{h}}(X))$ for data-dependent estimation $h$, we try to find a uniform bound

$$\sup_{h \in \mathcal{H}} E_D(l_h(X)) - E_S(l_h(X))$$

- Apparently, $\sup_{h \in \mathcal{H}} E_D(l_h(X)) - E_S(l_h(X)) \geq G$
- Intuitively, the complexity $\mathcal{H}$ shall impact this bound, in the sense that more complexity modeling, i.e., larger $\mathcal{H}$, lead to worse bound.
- This usually refers to the trade-off among model complexity, model fitting power and model predictive performance

# PAC Bayes analysis

```
> summary(gavote)
   equip        econ         perAA          rural
 LEVER:74   middle:69   Min.   :0.0000   rural:117
 OS-CC:44   poor  :72   1st Qu.:0.1115   urban: 42
 OS-PC:22   rich  :18   Median :0.2330
 PAPER: 2               Mean   :0.2430
 PUNCH:17               3rd Qu.:0.3480
                        Max.   :0.7650
      atlanta          gore              bush
 Atlanta   : 15   Min.   :   249   Min.   :   271
 notAtlanta:144   1st Qu.:  1386   1st Qu.:  1804
                  Median :  2326   Median :  3597
                  Mean   :  7020   Mean   :  8929
                  3rd Qu.:  4430   3rd Qu.:  7468
                  Max.   :154509   Max.   :140494
    other             votes            ballots
 Min.   :   5.0   Min.   :   832   Min.   :   881
 1st Qu.:  30.0   1st Qu.:  3506   1st Qu.:  3694
 Median :  86.0   Median :  6299   Median :  6712
 Mean   : 381.7   Mean   : 16331   Mean   : 16927
 3rd Qu.: 210.0   3rd Qu.: 11846   3rd Qu.: 12251
 Max.   :7920.0   Max.   :263211   Max.   :280975
```

max perAA influential?

All vote count distributions are highly skewed

Will consider models for percent undercount

# Distribution of Percent Undercount

```
> percunder <- (gavote$ballots - gavote$votes)/gavote$ballots
> hist(percunder,xlab="Percent",las=1,main="Undercount")
```

# Distribution of Percent Undercount

```
> plot(density(percunder),main="Undercount",las=1)
> rug(percunder)
```

# Pairwise Relationships

```
> pergore = gavote$gore/gavote$votes
> perbush = gavote$bush/gavote$votes
> pairs(~percunder+gavote$perAA+pergore+perbush,pch=20)
```

# Pairwise Relationships

```
> plot(percunder~rural,gavote,las=1,ylab="Percent")
> plot(percunder~equip,gavote,las=1,ylab="Percent")
```

# Multiple Linear Regression

- Use multiple predictors to explain variation in response $Y$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \text{ for } i = 1, \ldots, n \\ &\downarrow \\ \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned}$$

- Have $p-1$ predictors $\longrightarrow p$ coefficients

- Use least squares to estimate $\boldsymbol{\beta} \rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$

- Gauss-Markov theorem: These OLS estimates are

  - unbiased
  - minimum variance among all unbiased linear estimators
  - BLUEs = best linear unbiased estimators

# Normal Error Regression Model

- Gauss-Markov results assume errors are uncorrelated with mean 0 and constant variance. No assumed distribution.
- Assuming errors are uncorrelated (i.e., independent) $N(0, \sigma^2)$ random variables greatly simplifies inference
  - $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}) \rightarrow \mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
  - Maximum likelihood estimates of $\boldsymbol{\beta}$ the same as OLS
  - MLE estimate of $\boldsymbol{\sigma^2}$ is biased (downward)
  - Sampling distributions follow known dists (e.g., $\boldsymbol{t}, \boldsymbol{F}$)
  - Inference robust to moderate deviations from Normality
    - Variations of the central limit theorem

# Fitted/Predicted Values

- The fitted values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ called *hat matrix*
- Can equivalently write $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$
- $\mathbf{H}$ symmetric and idempotent ($\mathbf{HH} = \mathbf{H}$)
- Matrix $\mathbf{H}$ used in diagnostics along with residuals
- Large $h_{ii}$ (diagonals) implies fit being forced close to $Y_i$

# Residuals

- Residual vector (random)

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y}\end{aligned}$$

- Because it is random

$$\begin{aligned}E(\mathbf{e}) &= (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} \\ &= 0 \\ \boldsymbol{\sigma}^2(\mathbf{e}) &= (\mathbf{I} - \mathbf{H})\boldsymbol{\sigma}^2(\mathbf{Y})(\mathbf{I} - \mathbf{H})' \\ &= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' \\ &= (\mathbf{I} - \mathbf{H})\sigma^2\end{aligned}$$

# ANOVA Table

- Quadratic form defined as

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_i \sum_j a_{ij} Y_i Y_j$$

  where $\mathbf{A}$ is symmetric $n \times n$ matrix

- Sums of squares can all be expressed as quadratic forms

- Quadratic forms play significant role in the theory of linear models when errors are Normally distributed

# Overall F-test

| Source of Variation | df | SS | MS | F |
|---|---|---|---|---|
| Regression | $p-1$ | SSR | SSR/$(p-1)$ | MSR/MSE |
| Error | $n-p$ | SSE | SSE/$(n-p)$ | |
| Total | $n-1$ | SSTO | | |

- Do predictors *collectively* explain the variation in $Y$
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
  - $H_a$ : at least one $\beta_k \neq 0$

- $F$ also scaled ratio of explained to unexplained variation

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

$$\text{where } R^2 = \frac{\text{SSR}}{\text{SSTO}}$$

- Note that $\sqrt{R^2} = Corr(\hat{Y}, Y)$

# Inference on $\hat{\beta}$

- Vector $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}\mathbf{Y}$
- The mean and variance are

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \\ \boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}}) &= \mathbf{A}\boldsymbol{\sigma}^2(\mathbf{Y})\mathbf{A}' \\ &= \mathbf{A}\sigma^2\mathbf{I}\mathbf{A}' \\ &= \sigma^2\mathbf{A}\mathbf{A}' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Thus, $\hat{\boldsymbol{\beta}}$ is $N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

# Testing an Individual Coefficient

- Overall $F$ test gives no direct conclusion regarding individual predictor's contribution
- Because $\widehat{\beta}_k \sim N(\beta_k, \sigma^2(\widehat{\beta}_k))$, perform a $t$ test

$$t = \frac{\widehat{\beta}_k - 0}{s(\widehat{\beta}_k)}$$

- Under $H_0 : \beta_k = 0$, this is $t$ distributed with $n - p$ df
- Assesses contribution of predictor **after accounting for all other predictors in the model**

# Testing a Set of Coefficients

- Consider two models, one nested within the other
  - Full Model :

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ji} + \varepsilon_i$$

  - Reduced Model ($\beta_1 = \beta_2 = \ldots = \beta_q = 0$):

$$Y_i = \beta_0 + \sum_{j=q+1}^{p-1} \beta_j X_{ji} + \varepsilon_i$$

- Can compute

$$F = \frac{(\text{SSE(R)} - \text{SSE(F)})/q}{\text{SSE(F)}/(n-p)}$$

- Assesses significance of the set of predictors given the other predictors are already in the model

# Inference on Prediction

- Consider vector $X_h = [1\, x_{h1}\, x_{h2} \cdots x_{h(p-1)}]$
- Mean response $\hat{\mu}_h = X_h\hat{\boldsymbol{\beta}}$

  $E(\hat{\mu}_h) = X_h\boldsymbol{\beta}$

  $\text{Var}(\hat{\mu}_h) = X_h\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}})X_h' = \sigma^2 X_h(\mathbf{X}'\mathbf{X})^{-1}X_h'$
- Prediction $\hat{Y}_h = X_h\hat{\boldsymbol{\beta}}$

  $E(\hat{Y}_h) = X_h\boldsymbol{\beta}$

  $\text{Var}(\hat{Y}_h) = \sigma^2(1 + X_h(\mathbf{X}'\mathbf{X})^{-1}X_h')$

  Relies heavily on the Normal distribution assumption

# Back to Georgia Undercount Data

- For these data, we are interested mostly in finding important predictors ($\beta$)
- Variables of interest include
  - atlanta: Indicator of county in Atlanta
  - rural: Indicator that county is rural
  - perAA: percent of African Americans in county
  - econ: economic status of county (poor, middle, rich)
  - equip: voting equipment used (five types)
  - pergore: percent of votes for Gore
- Will consider various linear models to investigate and describe additional linear model concepts

# Fitting a Linear Model in R

```
> model1 = lm(percunder ~ pergore + perAA, gavote)
> summary(model1)

Call:
lm(formula = percunder ~ pergore + perAA, data = gavote)

Residuals:
      Min         1Q    Median        3Q       Max
-0.046013 -0.014995 -0.003539  0.011784  0.142436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)        Neither variable
(Intercept)  0.03238    0.01276   2.537   0.0122 *       significant
pergore      0.01098    0.04692   0.234   0.8153         but overall F
perAA        0.02853    0.03074   0.928   0.3547           test is
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02445 on 156 degrees of freedom
Multiple R-squared:  0.05309,Adjusted R-squared:  0.04095
F-statistic: 4.373 on 2 and 156 DF,  p-value: 0.01419
```

# ANOVA table in R

```
> anova(model1)                                    ***Type I SS

Analysis of Variance Table
Response: percunder
           Df   Sum Sq   Mean Sq F value   Pr(>F)
pergore     1 0.004713 0.0047129  7.8845 0.005623 **
perAA       1 0.000515 0.0005151  0.8617 0.354701
Residuals 156 0.093249 0.0005978
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(car)
> Anova(model1, type=3)

Anova Table (Type III tests)
Response: percunder
              Sum Sq  Df F value  Pr(>F)
(Intercept) 0.003847   1  6.4366 0.01216 *       Neither significant
pergore     0.000033   1  0.0547 0.81531         when fitted last
perAA       0.000515   1  0.8617 0.35470
Residuals   0.093249 156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Multicollinearity

- Numerical analysis issue: The matrix $\mathbf{X}'\mathbf{X}$ is almost singular (linear dependent columns - no inverse exists)
- Unless singular, inverse computed well with current algorithms
- Statistical inference issue: Very high correlation among the explanatory/predictor variables
- Although inverse exists, regression coefficients unstable
  - Increased uncertainty / variance
  - Spurious coefficient estimates
- Predicted values and $R^2$ largely unaffected

# Multicollinearity Example

- Consider a two predictor model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

- The estimate of $\beta_1$ in model above is

$$\widehat{\beta}_1 = \frac{\widehat{\beta}_1' - \sqrt{\frac{s_Y^2}{s_{X_1}^2}} r_{12} r_{Y2}}{1 - r_{12}^2}$$

  where $\widehat{\beta}_1'$ is the estimate fitting $Y$ vs $X_1$

# Extreme Cases

- Extreme case #1: Consider $X_1$ and $X_2$ are uncorrelated
  - $r_{12} = 0 \rightarrow \widehat{\beta}_1 = \widehat{\beta}_1'$
  - Estimator $\widehat{\beta}_1$ does not depend on $X_2$
  - The contribution of each predictor is the same regardless of whether or not the other predictor is in the model
- Extreme case #2: Consider $X_1 = a + bX_2$
  - $r_{12} = \pm 1 \rightarrow$ estimator $\widehat{\beta}_1$ does not exist
  - Type III SS are zero
  - There is no contribution of the predictor if the other predictor is already in the model

# Diagnostics

- Procedures to determine appropriateness of the model and to check conditions needed for standard inference
- If there are violations, inference and model may not be reasonable thereby resulting in faulty conclusions
- Always check assumptions before any inference!!!!!!!!
- Residuals used in many diagnostics so fit model first
- Procedures involve both graphical methods and formal statistical tests
- I prefer focus on visual assessments. Often a formal test is less robust to model conditions than the regression model itself

# Individual Variable Assessments

- Can look at marginal distribution of each $X$
  - Recall model does **not** state $X \sim$ Normal
  - Can provide information regarding outlying values that could be influential
- Careful looking at marginal distribution of $Y$
  - If mean of $Y$ depends on $X$, looking at $Y$ alone may be deceiving (i.e., mixture of normal dists)
  - Better to look at residuals $\rightarrow$ have adjusted for diff means so you can look at all residuals together

# Residuals

- Standard residual $e_i = Y_i - \widehat{Y}_i$
- Studentized residual

$$r_i = \frac{e_i}{\sqrt{\mathrm{M}SE(1-h_{ii})}}$$

- Studentized deleted residual

$$\begin{aligned} t_i &= \frac{Y_i - \widehat{Y}_{i(i)}}{\sqrt{\mathrm{MSE}_{(i)}/(1-h_{ii})}} \sim t(n-p-1) \\ &= e_i \left[\frac{n-p-1}{\mathrm{S}SE(1-h_{ii}) - e_i^2}\right]^{1/2} \end{aligned}$$

# Residual Plots

- Used to visually assess if
  - Model is reasonable (all models are wrong)
  - Errors are Normally distributed
  - Errors have constant variance
  - Errors are independent
- Plot $e$ vs $\widehat{Y}$ (overall)
- Plot $e$ vs $X_j$ (with respect to $X_j$)
- Plot $e$ vs non-included variable (e.g., $X_j X_k$)
- Normal QQplot of residuals
- Histogram of residuals

# Other Measures

- Hat matrix diagonals – How much $Y_i$ is contributing to $\widehat{Y}_i$. Large $h_{ii}$ means case is far away from center of $X$'s.
- Cook's Distance – Influence of case $i$ on all the predicted values (need large $r_i$ and $h_{ii}$)

$$D_i = \frac{\sum (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p\mathrm{M}SE} = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}}$$

- DFBETAS - Influence of case $i$ on each of the regression coefficients. A standardized difference between regression coefficient computed with and without case $i$

$$\mathrm{D}FBETAS_{k(i)} = \frac{\widehat{\beta}_k - \widehat{\beta}_{k(i)}}{\sqrt{\mathrm{M}SE_{(i)} c_{kk}}}$$

where $c_{kk}$ is from $(\mathbf{X}'\mathbf{X})^{-1}$

# Diagnostics

# Model Selection

- Model selection criteria used to balance trade-off between predictive power and model complexity
- Adjusted $R^2$ (maximize)

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SSTO}} = 1 - \frac{\text{MSE}}{\text{MSTO}}$$

- AIC - (minimize)

$$-2\log(L) + 2p = n\log\left(\frac{\text{SSE}_p}{n}\right) + 2p$$

- BIC (minimize)

$$-2\log(L) + p\log(n) = n\log\left(\frac{\text{SSE}_p}{n}\right) + p\log(n)$$

# Returning to Georgia Undercount Data

- Given six predictors, there are 63 possible main-effect models.
- If we include interactions or higher-order terms, the number of models explodes.
- There are automated search tools for looking at models but when interested in coefficients, need to proceed with caution.

# Alternative Model

```
> cpergore = pergore - mean(pergore)
> cperAA = gavote$perAA - mean(gavote$perAA)
> model2 = lm(percunder ~ cperAA+cpergore*rural+equip, gavote)

> sumary(model2)
                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)      0.0394149  0.0039232 10.0466 < 2.2e-16
cperAA           0.0282641  0.0310921  0.9090   0.36479
cpergore         0.0038371  0.0454202  0.0845   0.93279
rural1           0.0093183  0.0023241  4.0095 9.564e-05
equip1          -0.0054359  0.0043605 -1.2466   0.21448
equip2           0.0010465  0.0048155  0.2173   0.82825
equip3           0.0102037  0.0054419  1.8750   0.06273
equip4          -0.0145280  0.0135759 -1.0701   0.28628
cpergore:rural1  0.0043997  0.0193581  0.2273   0.82051

n = 159, p = 9, Residual SE = 0.02335, R-Squared = 0.17
```

# Model Summary

- The use of "A*B" is shorthand for "A+B+A:B"
- Indicators for rural and equip are coded differently than in the textbook because I switched the option from "contr.treatment" to "contr.sum"
- Our intercept no longer represents a particular type of county. Instead can be thought of as the "average" over all types.
- To get textbook's intercept (lever, rural)

$$\hat{Y} = 0.0394149 + 0.0093183 + (-0.0054359) = 0.04330$$

- Do additional variables contribute to explaining $Y$?

```
> anova(model1,model2)
Model 1: percunder ~ pergore + perAA
Model 2: percunder ~ cperAA + cpergore * rural + equip
  Res.Df      RSS Df Sum of Sq      F   Pr(>F)
1    156 0.093249
2    150 0.081775  6  0.011474 3.5077 0.002823 **
```

# Continuing Model Search

```
> drop1(model2,test="F")

Single term deletions

Model:
percunder ~ cperAA + cpergore * rural + equip
               Df Sum of Sq      RSS     AIC F value  Pr(>F)
<none>                      0.081775 -1186.1
cperAA          1 0.0004505 0.082226 -1187.2  0.8264 0.36479
equip           4 0.0054438 0.087219 -1183.8  2.4964 0.04521 *
cpergore:rural  1 0.0000282 0.081804 -1188.0  0.0517 0.82051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(1) Procedure maintains hierarchy in that it does not test main effects of rural and cpergore
(2) Suggests cperAA and the interaction could be removed
(3) We already knew cperAA and cpergore were highly correlated

# New Model

```
> model3 = lm(percunder ~ cpergore+rural+equip, gavote)
> sumary(model3)
               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  0.03950062  0.00385811 10.2383 < 2.2e-16
cpergore     0.04390787  0.01826155  2.4044   0.01740
rural1       0.00934797  0.00227384  4.1111 6.423e-05
equip1      -0.00515432  0.00433736 -1.1884   0.23655
equip2       0.00049161  0.00476869  0.1031   0.91803
equip3       0.00971968  0.00538615  1.8046   0.07312
equip4      -0.01420887  0.01345025 -1.0564   0.29246

n = 159, p = 7, Residual SE = 0.02328, R-Squared = 0.16

> anova(model2,model3)
Analysis of Variance Table

Model 1: percunder ~ cperAA + cpergore * rural + equip
Model 2: percunder ~ cpergore + rural + equip
  Res.Df      RSS Df   Sum of Sq      F Pr(>F)
1    150 0.081775
2    152 0.082344 -2 -0.00056858 0.5215 0.5947
```

# Automated Search

```
> modelmax = lm(percunder ~ (equip+econ+rural+atlanta)^2 +
(equip+econ+rural+atlanta)*(pergore+perAA), gavote)
> modelbetter = step(modelmax,trace=FALSE)
> summary(modelbetter)

Call:
lm(formula = percunder ~ equip + econ + rural + perAA + equip:econ +
    equip:perAA + rural:perAA, data = gavote)

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0290149  0.0079612   3.645 0.000377 ***
equip1        0.0149377  0.0088743   1.683 0.094570 .
equip3        0.0288363  0.0113918   2.531 0.012476 *
econ2         0.0188711  0.0041043   4.598 9.50e-06 ***
equip3:econ1 -0.0083037  0.0049666  -1.672 0.096792 .
equip2:econ2 -0.0108228  0.0054450  -1.988 0.048814 *
equip3:econ2  0.0303931  0.0073457   4.138 6.05e-05 ***
equip1:perAA -0.0801429  0.0252661  -3.172 0.001864 **

Residual standard error: 0.01964 on 139 degrees of freedom
Multiple R-squared:  0.4554,Adjusted R-squared:  0.381
F-statistic: 6.118 on 19 and 139 DF,  p-value: 4.37e-11
```

# Continued Refinement

```
> drop1(modelbetter,test="F")
Single term deletions

Model:
percunder ~ equip + econ + rural + perAA + equip:econ + equip:perAA +
    rural:perAA
             Df Sum of Sq      RSS     AIC F value   Pr(>F)
<none>                    0.053627 -1231.1
equip:econ    6 0.0075232 0.061150 -1222.3  3.2500 0.005084 **
equip:perAA   4 0.0068439 0.060471 -1220.0  4.4348 0.002101 **
rural:perAA   1 0.0010214 0.054649 -1230.1  2.6474 0.105984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Continued Refinement

```
> modelbetter2 = lm(percunder ~ equip+econ+rural+perAA+equip:econ+
      equip:perAA, gavote)
> drop1(modelbetter2,test="F")
Single term deletions

Model:
percunder ~ equip + econ + rural + perAA + equip:econ + equip:perAA
            Df Sum of Sq      RSS     AIC F value   Pr(>F)
<none>                    0.054649 -1230.1
rural        1 0.0017234 0.056372 -1227.2  4.4151 0.037414 *
equip:econ   6 0.0075465 0.062195 -1221.6  3.2221 0.005384 **
equip:perAA  4 0.0060162 0.060665 -1221.5  3.8531 0.005307 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Nothing additional should be dropped
(b) This final model differs from textbook in inclusion of rural

# Final Model

```
> summary(modelbetter2)

Call:
lm(formula = percunder ~ equip + econ + rural + perAA + equip:econ +
    equip:perAA, data = gavote)

Coefficients: (2 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0278686  0.0079765   3.494 0.000638 ***
equip3        0.0302132  0.0114269   2.644 0.009127 **
equip4       -0.0488104  0.0293275  -1.664 0.098284 .
econ2         0.0197774  0.0040901   4.835 3.45e-06 ***
rural1        0.0048738  0.0023195   2.101 0.037414 *
equip2:econ2 -0.0121172  0.0054182  -2.236 0.026907 *
equip3:econ2  0.0311425  0.0073743   4.223 4.31e-05 ***
equip1:perAA -0.0688544  0.0244374  -2.818 0.005539 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01976 on 140 degrees of freedom
Multiple R-squared:  0.4451,Adjusted R-squared:  0.3737
F-statistic: 6.238 on 18 and 140 DF,  p-value: 5.233e-11
```

# Results

```
> pdf <- data.note(econ=rep(levels(gavote$econ),5),equip=rep(levels(
  gavote$equip),rep(3,5)), perAA=0.233, rural="rural")
> ppr = predict(modelbetter2,new=pdf)
> xtabs(round(ppr,3)~econ+equip,pdf)
        equip
econ      LEVER  OS-CC  OS-PC  PAPER  PUNCH
  middle  0.035  0.049  0.043  0.000  0.046
  poor    0.053  0.056  0.107  0.024  0.054
  rich    0.022  0.040  0.017 -0.011  0.050
> pdf <- data.note(econ=rep(levels(gavote$econ),5),equip=rep(levels(
  gavote$equip), rep(3,5)), perAA=0.233, rural="urban")
> ppu = predict(modelbetter2,new=pdf)
> xtabs(round(ppu,3)~econ+equip,pdf)
        equip
econ      LEVER  OS-CC  OS-PC  PAPER  PUNCH
  middle  0.025  0.039  0.034 -0.010  0.036
  poor    0.043  0.046  0.097  0.014  0.044
  rich    0.012  0.030  0.008 -0.021  0.040
```

# Results

- Undercount higher in poorer regions
- Undercount higher in rural regions
- Effect of proportion African American varies across equipment. Sometimes it goes up, sometimes it goes down
- Lever seems to be the best equipment. There the undercount goes down with perAA increases