STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 23
Nov 20, 2014

Victoria Liublinska

Odds and Ends

- Project update due on Monday by 5pm:
  - 1-2 pages, including outline of a poster, explanation of what you have done so far and what is left to do, and description of any challenges you have faced.
  - E-mail to stat139projects@gmail.com (one per group, CC all members).
  - Use subject line "Stat 139: Project Update".

Odds and Ends

- Next week (11/24-11/25):
  - Mon and Tue schedules are the same;
  - No sections or OHs on Wed, Thu, and Fri of next week (11/26-11/28).

- Last week of classes (12/1-12/3):
  - No OHs on Thu (12/4) and Fri (12/5).
  - Sections and last lecture (12/2) will be (partially) devoted to reviewing the material from the second half of the course.

Previous lecture: Review

- Inferential tools for multiple regression:
  - *t*-tests and CIs for coefficients and their linear combinations;

$$\frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]}}} \overset{H_0}{\sim} t_{n-(K+1)}; \qquad \hat{\beta}_j \pm t_{n-(K+1),(1-\alpha/2)}\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]}}$$

  - Confidence and prediction intervals at $X=X_0$:

$$\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}} \pm M \cdot SE$$

	C.I. for ***E(Y)***		Prediction interval for ***Y***	
	Multiplier (M)	SE	Multiplier (M)	SE
One point	$t_{n-(K+1),(1-\alpha/2)}$	$\hat{\sigma}\sqrt{\boldsymbol{X}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_0}$	$t_{n-(K+1),(1-\alpha/2)}$	$\hat{\sigma}\sqrt{1+\boldsymbol{X}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_0}$
Many points simultaneously	$\sqrt{(K+1)F_{(K+1),\,n-(K+1),\,(1-\alpha)}}$	$\hat{\sigma}\sqrt{\boldsymbol{X}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_0}$	$\sqrt{(K+1)F_{(K+1),\,n-(K+1),\,(1-\alpha)}}$	$\hat{\sigma}\sqrt{1+\boldsymbol{X}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_0}$

Previous lecture: Review

- Inferential tools for multiple regression:
  - Extra-sum-of-squares *F*-tests to compare regression models.

Reduced $H_0 : \mu(Y \mid X_1, X_2, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \ldots + \beta_M X_M$

Full (K>M) $H_a : \mu(Y \mid X_1, X_2, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \ldots + \beta_M X_M + \ldots + \beta_K X_K$

$$\boxed{R = \frac{\left(\mathrm{SSR}_{\mathrm{Reduced}} - \mathrm{SSR}_{\mathrm{Full}}\right) / \left(\mathrm{d.f.}_{\mathrm{Reduced}} - \mathrm{d.f.}_{\mathrm{Full}}\right)}{\mathrm{SSR}_{\mathrm{Full}} / \mathrm{d.f.}_{\mathrm{Full}}} \overset{H_0}{\sim} F_{\left(\mathrm{d.f.}_{\mathrm{Reduced}} - \mathrm{d.f.}_{\mathrm{Full}},\, \mathrm{d.f.}_{\mathrm{Full}}\right)}}$$

- Adjusted $R^2 = 1 - \dfrac{\hat{\sigma}^2_{\mathrm{Reg}}}{s_Y^2}$

- Strategies for variable selection.

Today's overview

- Strategies for variable selection, cont.
  - Model selection criteria
  - Automatic procedures
  - Cross-validation
- Ecological fallacy
- Collinearity between predictors

Reading:

- **Required:** Finish Ch. 12 (Ch. 12 R code), start Ch. 11
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Ch. 10 (Multicollinearity), Ch. 11 (Variable Selection)

Strategies for Variable Selection, cont.

Strategies for Variable Selection

Depend on our objective:

1. Adjusting for auxiliary explanatory variables prior to inclusion of the main variable of interest (sometimes, for the purpose of causal inference):
   - Ok to use variable selection if causal inference is not required (e.g., sex discrimination case in Ch. 12);
   - Otherwise, need more advanced techniques (e.g., subclassification or matching).

2. Looking for prediction or best set of predictors, risk-factors.
   - No interpretation needed;
   - Ok to use variable selection techniques.

Tips from G&H on Building Models for Prediction

- Consider including interactions for predictors with large and significant main effects;

- In general, keep the variable if:
  - Sign makes sense (significant or not);
  - Significant but sign is unexpected. Think hard why:
    - Additional interactions?
    - Unobserved confounders?
    - Ecological fallacy?

- In general, remove the variable if:
  - Insignificant and sign is unexpected

Strategies for Variable Selection: Prediction

General principles

Our general principles for building regression models for prediction are as follows:

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors—for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
3. For inputs that have large effects, consider including their interactions as well.

Gelman and Hill, p. 69

4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance (typically measured at the 5% level; that is, a coefficient is "statistically significant" if its estimate is more than 2 standard errors from zero):

   (a) If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.

   (b) If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).

   (c) If a predictor *is* statistically significant and does not have the expected sign, then think hard if it makes sense. (For example, perhaps this is a country such as India in which incumbents are generally unpopular; see Linden, 2006.) Try to gather data on potential lurking variables and include them in the analysis.

   (d) If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.

Gelman and Hill, p. 69

Strategies for Variable Selection

Depend on our objective:

3. "Fishing" for explanation of the outcome (i.e., what are the important X's?):
   - Most common "method" in observational studies!
   - Use automatic selection techniques with great caution! – chosen variables are not necessarily *special*.
   - Interpretation of coefficients is extremely difficult if explanatory variables are correlated (multicollinearity).
     - Principal Component Analysis (PCA) finds uncorrelated linear combination of predictors that explain the outcome.

Strategies for Variable Selection

Depend on our objective:

3. "Fishing" for explanation of the outcome (i.e., what are the important X's?):

- When explanatory variables are related, it may be impossible to "hold all other variables fixed" while changing one of them – it is "outside of the experience provided by the data".
- Causal inference may not be possible due to unobserved confounders.
- **Best attitude:** use this analysis for hypothesis generation.

Strategies for Variable Selection

Techniques for variable selection:

1. Fixed set by design (treatment indicator + background variables);

2. Fit all possible subsets of models and find the one that fits the best according to some criterion:
   - E.g., Adj-R^2, *Cp* statistic, AIC, or BIC

3. Sequential: forward / backward / stepwise selection;

See Sections 12.2 - 12.4.

Model Selection Criteria

If two models have the same number of parameters, choose the one with smaller residual variance.

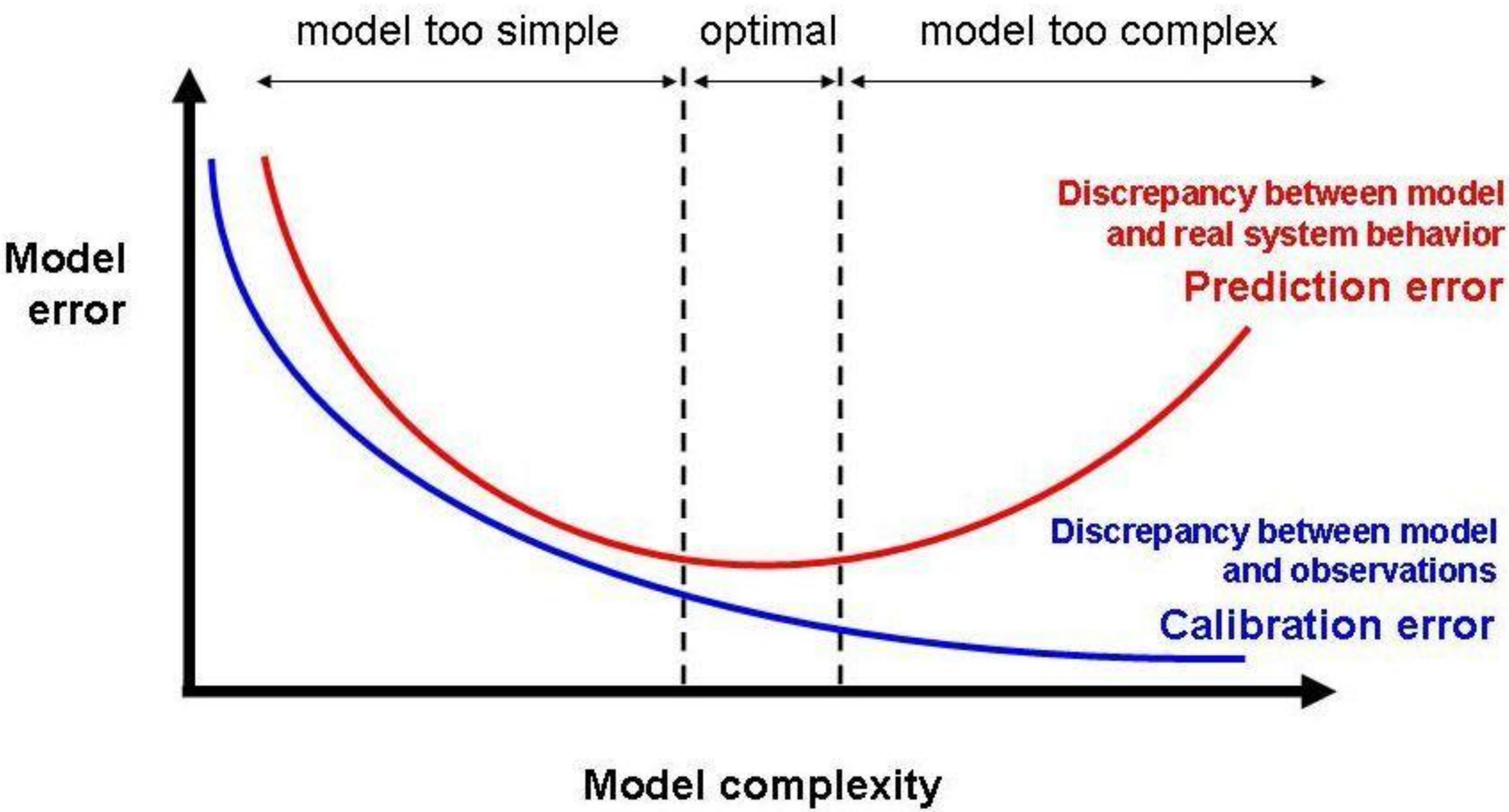
More generally: take into account both estimate of residual variance and number of parameters.

General form of any criterion:

$$f(\hat{\sigma}^2) + g(p)$$

***p*=*K*+1, where *K* is the number of predictors.**

Goal of Model Selection



Model Selection Criteria

- Maximize: $\text{Adj. } R^2 = 1 - \hat{\sigma}^2 / s_Y^2$

always chooses the model with smallest residual variance, doesn't penalize for p (or K) as much as (others below).

- Mallow's C_p statistic,

Minimize: $C_P = (n-p)\dfrac{\hat{\sigma}^2 - \hat{\sigma}^2_{full}}{\hat{\sigma}^2_{full}} + p$

Trade-off between bias due to excluding important variables and extra variance due to including too many.

Mallow's C_P Statistic

- $C_P = p$ for the full model; also, we assume that the full model has *no bias*.
- Let

$$MSE(\hat{Y}_i) = \underbrace{\left(\hat{Y}_i - E(\hat{Y}_i)\right)}_{\text{BIAS}}^2 + Var(\hat{Y}_i)$$

- For any other model, the statistic estimates the total mean squared error (TMSE), scaled by σ^2, where

$$TMSE = \sum_{i=1}^{n} MSE(\hat{Y}_i).$$

- Sometimes it is advised to only consider models with $C_P \le p$.

Model Selection Criteria

- Akaike's Information Criterion (AIC),

<u>Minimize:</u> $\mathrm{AIC} = n\log(\mathrm{SSRes}/n) \boxed{+2p}$

- Bayes Information Criterion (BIC),

<u>Minimize:</u> $\mathrm{BIC} = n\log(\mathrm{SSRes}/n) + \boxed{p\log(n)}$

•2nd and 3rd ed. have different definitions!

•Use the ones defined in class.

Criterion	SSR part	Extra penalty for too many terms
Adj-R^2	$1-\hat{\sigma}^2/s_Y^2$	0
C_P	$\mathrm{SSRes}/\hat{\sigma}^2_{full} - n$	$2p$
AIC	$n\log(\mathrm{SSRes}/n)$	$2p$
BIC	$n\log(\mathrm{SSRes}/n)$	$\log(n)\cdot p$

Check out "Notes on AIC and BIC" by Matteo Bonvini under Readings.

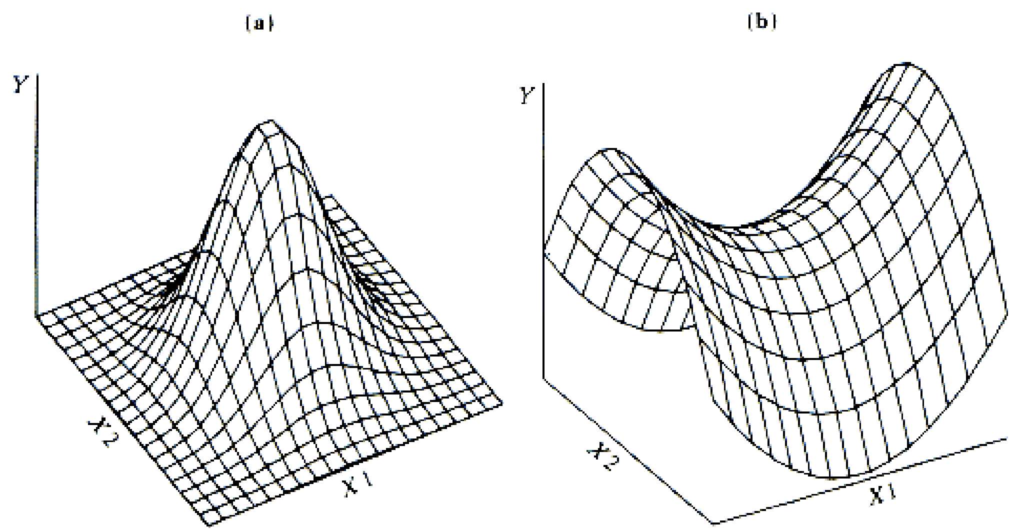
Criterion-Based Model Selection

- Fit of *all possible models* and compare them using Adj-R^2, C_P statistic, AIC, or BIC.

- For example, a saturated second-order model (SSOM) includes:
  - All main effects
  - All quadratic terms
  - All paired interactions

- SSOM describes

$\mu(Y|X_1, .., X_K)$ as an *arbitrary parabolic surface*.



Saturated Second-Order Model (SSOM)

- Given K predictors, SSOM contains $K(K+3)/2+1$ parameters.

- Total number of *p*-parameter hierarchical models (not including residual variance):

$$\sum_{j=0}^{K} C_j^K \times C_{p-1-j}^{C_2^{j+1}}, \text{where } C_m^n = \frac{n!}{m!(n-m)!}$$

$$\text{and } C_m^n = 0 \text{ if } n < m \text{ or } m < 0, \text{and } C_0^n = C_n^n = 1.$$

  - For example, if K=4, we have 1,337 models with p from 1 to 15.

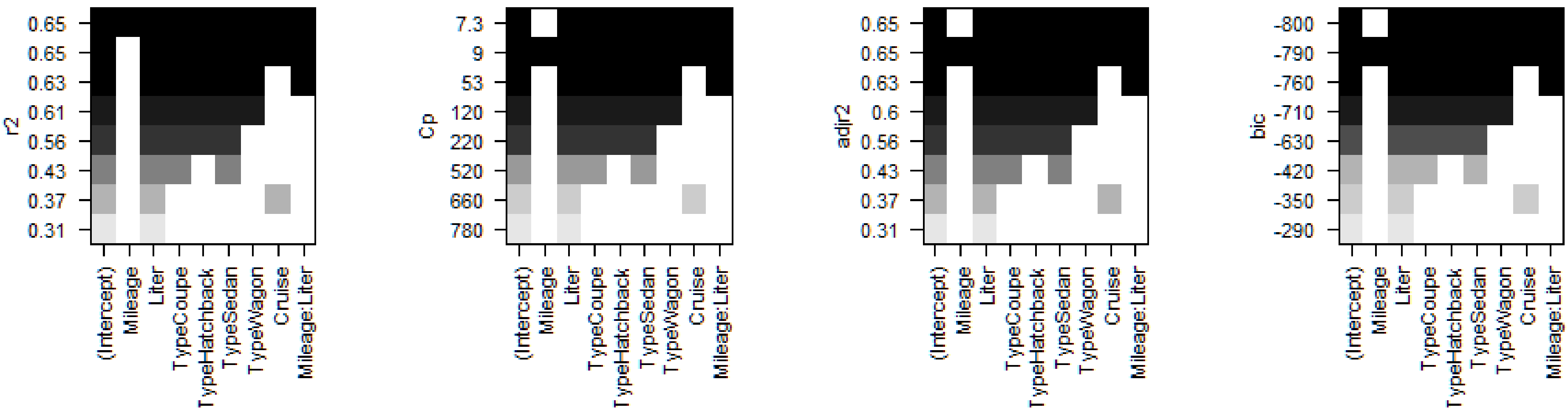
Criterion-Based Model Selection: Example

```
library(leaps)
leaps <- regsubsets(Price ~ Mileage*Liter + Cruise + Type,
                    data = CarData, nbest=1)
#nbest means the max number of optimal models
#for each size
#view results
summary(leaps)

# plot a table of models showing variables in each model.
# models are ordered by the selection statistic.
plot(leaps,scale="r2")
plot(leaps,scale="Cp")
plot(leaps,scale="adjr2")
plot(leaps,scale="bic")

?plot.regsubsets #Learn more about the function plot
```

Criterion-Based Model Selection: Example



Caution: These procedures do not take into account the hierarchy of the predictors:

- Add and drop factors *as a group.*
- Hierarchical structure: include main effects if interactions or squared terms are in the model.

Sequential Variable Selection: Forward

1. (May) start with an intercept-only model, $E(Y|X)=\beta_0$.
2. Consider all models with one more "term".*
3. For each, calculate some statistic (*F*-statistic, AIC, BIC, Cp, adj-R^2)
4. Include a new term with
   - The largest *F*-statistic (e.g., if > 4); or
   - The smallest AIC/BIC/*Cp* (if lower than in the current model); or
   - The largest adj-R^2 (if larger than in the current model).
5. Iterate steps 2-4 until no more variables can be added.

In R: `step(RegModel, direction = “forward“, k=2)) #k is the d.f. multiple, use k=log(n) for BIC, add test="F“ for F-statistic`

Sequential Variable Selection: Backward

1. Start with all predictors in the model (possibly, with interactions and polynomial terms).

2. Consider all models with one "term" removed*.

3. For each, calculate some test-statistic (*F*-statistic, AIC, BIC, Cp, adj-R^2)

4. Remove the term with

   - The smallest *F*-statistics (e.g., if < 4); or
   - The smallest AIC/BIC/Cp (if lower than in the current model); or
   - The largest adj-R^2 (if larger than in the current model).

5. Iterate until no more variables can be removed.

```
step(RegModel, direction = “backward")
```

Sequential Variable Selection: Stepwise

1. *May* start with an intercept-only model;
2. Do one step of <u>forward</u> selection;
3. Do one step of <u>backward</u> elimination;
4. Iterate.

```
step(RegModel, direction = “both")
```

Sequential Variable Selection: Caution

- Forward, backward, and stepwise may lead to different final models!
  - Inclusion/exclusion depends on correlation between the new variable and the ones that are already in the model.
  - Different initial models may produce difference final ones.
  - It is a form of data snooping!

- Think not: "~~here is the best model~~." Think instead: "here is one, possibly useful model."
  - Variable selection is a means to an end and not an end itself.

- Variable selection methods are sensitive to outliers and influential points.

Sequential Variable Selection: Example

```
> RegModel <- lm(Price ~ Mileage * Liter * Cruise + Type ,
                    data = CarData)
> slm1 <- step(RegModel, direction = "backward")

Start:  AIC=13958.29
Price ~ Mileage * Liter * Cruise + Type

                        Df  Sum of Sq        RSS   AIC
- Mileage:Liter:Cruise  1 6.5671e+06 2.7052e+10 13956
<none>                               2.7046e+10 13958
-  Type                  4 2.0066e+10 4.7112e+10 14396
```

Sequential Variable Selection: Example

```
Step:  AIC=13956.48
Price ~ Mileage + Liter + Cruise + Type + Mileage:Liter +
  Mileage:Cruise +
    Liter:Cruise
                 Df  Sum of Sq        RSS   AIC
- Mileage:Cruise  1 9.3950e+06 2.7062e+10 13955
- Liter:Cruise    1 3.9139e+07 2.7092e+10 13956
<none>                         2.7052e+10 13956
- Mileage:Liter   1 2.9993e+08 2.7352e+10 13963
- Type            4 2.0097e+10 4.7150e+10 14395

Step:  AIC=13954.76
Price ~ Mileage + Liter + Cruise + Type + Mileage:Liter +
  Liter:Cruise
                Df  Sum of Sq        RSS   AIC
- Liter:Cruise   1 4.2687e+07 2.7105e+10 13954
<none>                        2.7062e+10 13955
- Mileage:Liter  1 3.9205e+08 2.7454e+10 13964
-  Type          4 2.0088e+10 4.7150e+10 14393
```

Sequential Variable Selection: Example

```
Step:  AIC=13954.03
Price ~ Mileage + Liter + Cruise + Type + Mileage:Liter

                Df  Sum of Sq        RSS   AIC
<none>                        2.7105e+10 13954
- Mileage:Liter  1 3.9263e+08 2.7497e+10 13964
- Cruise         1 1.6138e+09 2.8718e+10 13998
- Type           4 2.0262e+10 4.7366e+10 14395
```

```
> summary(slm1)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.119e+04  1.830e+03  11.582  < 2e-16 ***
Mileage        3.941e-02  7.193e-02   0.548 0.583904
Liter          6.057e+03  4.815e+02  12.581  < 2e-16 ***
Cruise         3.691e+03  5.365e+02   6.880 1.21e-11 ***
TypeCoupe     -2.135e+04  9.733e+02 -21.931  < 2e-16 ***
TypeHatchback -2.093e+04  1.166e+03 -17.949  < 2e-16 ***
TypeSedan     -1.831e+04  8.745e+02 -20.935  < 2e-16 ***
TypeWagon     -1.106e+04  1.131e+03  -9.776  < 2e-16 ***
Mileage:Liter -7.470e-02  2.201e-02  -3.394 0.000724 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5839 on 795 degrees of freedom
Multiple R-squared:  0.6545,	Adjusted R-squared:  0.6511
F-statistic: 188.3 on 8 and 795 DF,  p-value: < 2.2e-16
```

Computer-assisted variable selection

Best:

- Compare all possible (hierarchical) subsets of models using either *Cp*, AIC, or BIC; find some model with a fairly small value.

Next best (e.g., if too many models to compare):

- Use sequential variable selection, like forward, backward, or stepwise regression.

- If there are multiple candidate models, consider:
  - 1. Do the models have similar qualitative consequences?
  - 2. Do they make similar predictions? (SSR)
  - 3. Which has the best diagnostics?
  - 4. What is the cost of measuring the predictors?

Cross-Validation

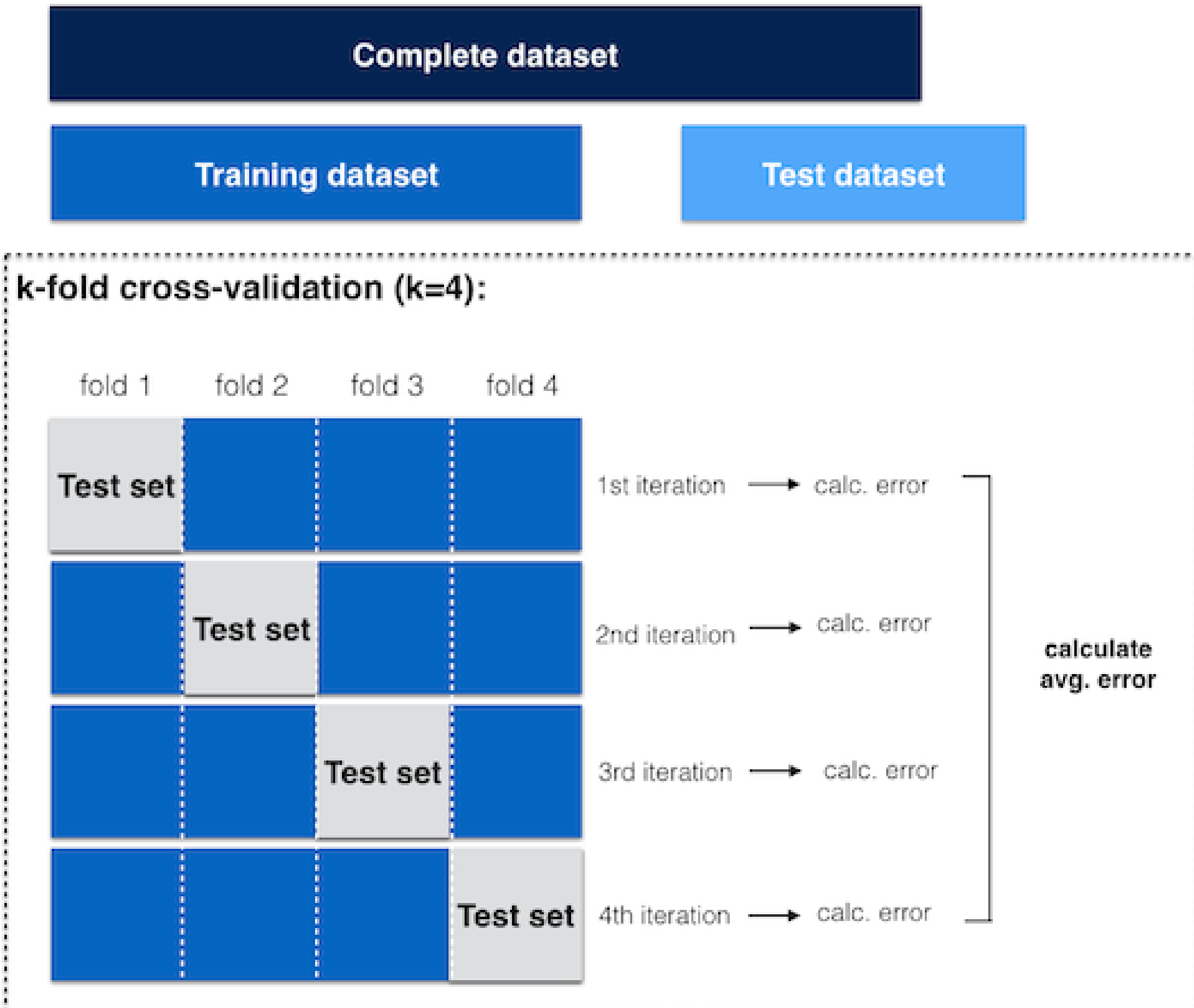
- If n is large, one may try cross validation:

A method of assessing the accuracy (i.e., bias) and precision (i.e., residual error) of a statistical model.

Train	Test	Validation

1. The available data set is divided into two (or three) random parts.
2. Training set is used to a the model.
3. Test set is used to check the predictive capability (e.g., SSR) and refine the model.
4. Validation set is used <u>once</u> to estimate the model's true error.

K-Fold Cross-Validation



Ecological Fallacy

Ecological Fallacy

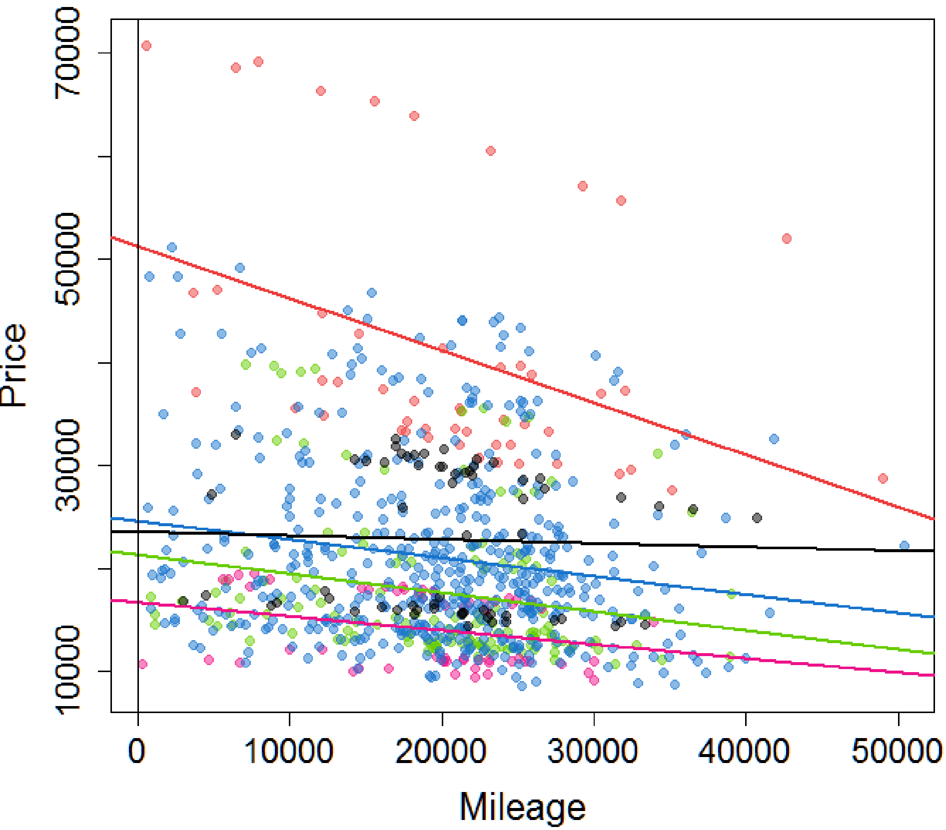
- Occurs when one makes conclusions about individuals based only on analyses of aggregated group data.
  - Simpson’s paradox
  - Group-level correlations vs. individual correlations
  - Group average vs. individual likelihood

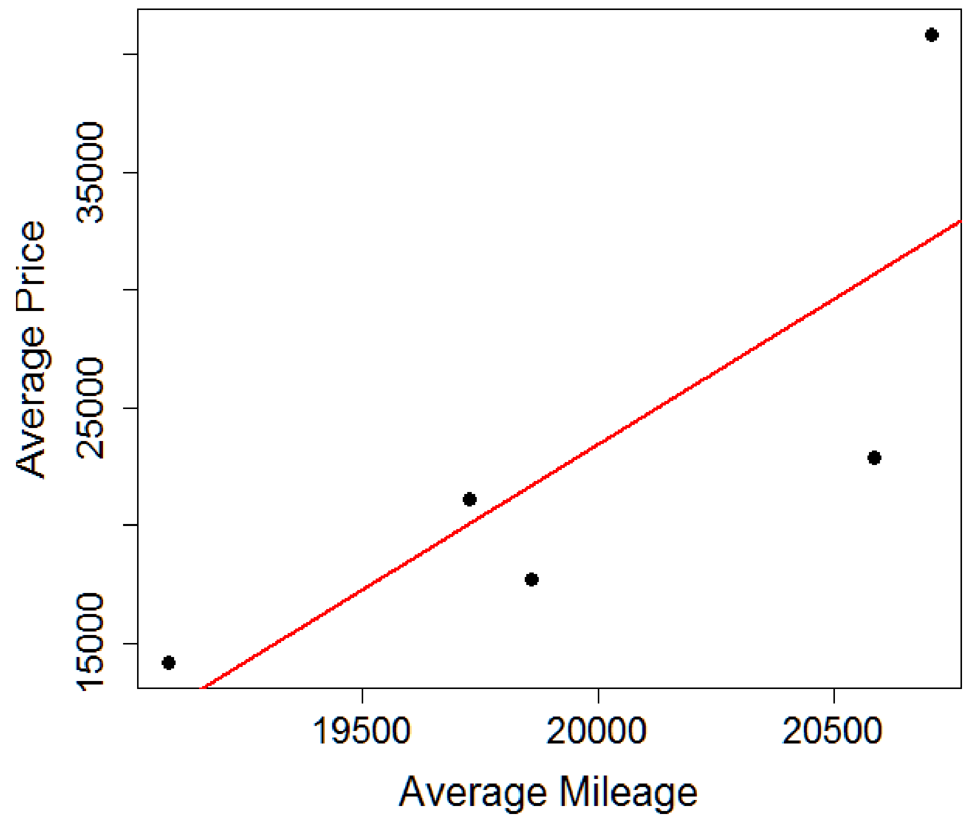
- Example: "Red State, Blue State" by A. Gelman: Rich *people* within particular states tend to vote more Republican than poor people, but higher-income *states* tend to vote more Democratic than poor states.

Issue: Unequal distribution of rich vs. poor within states.

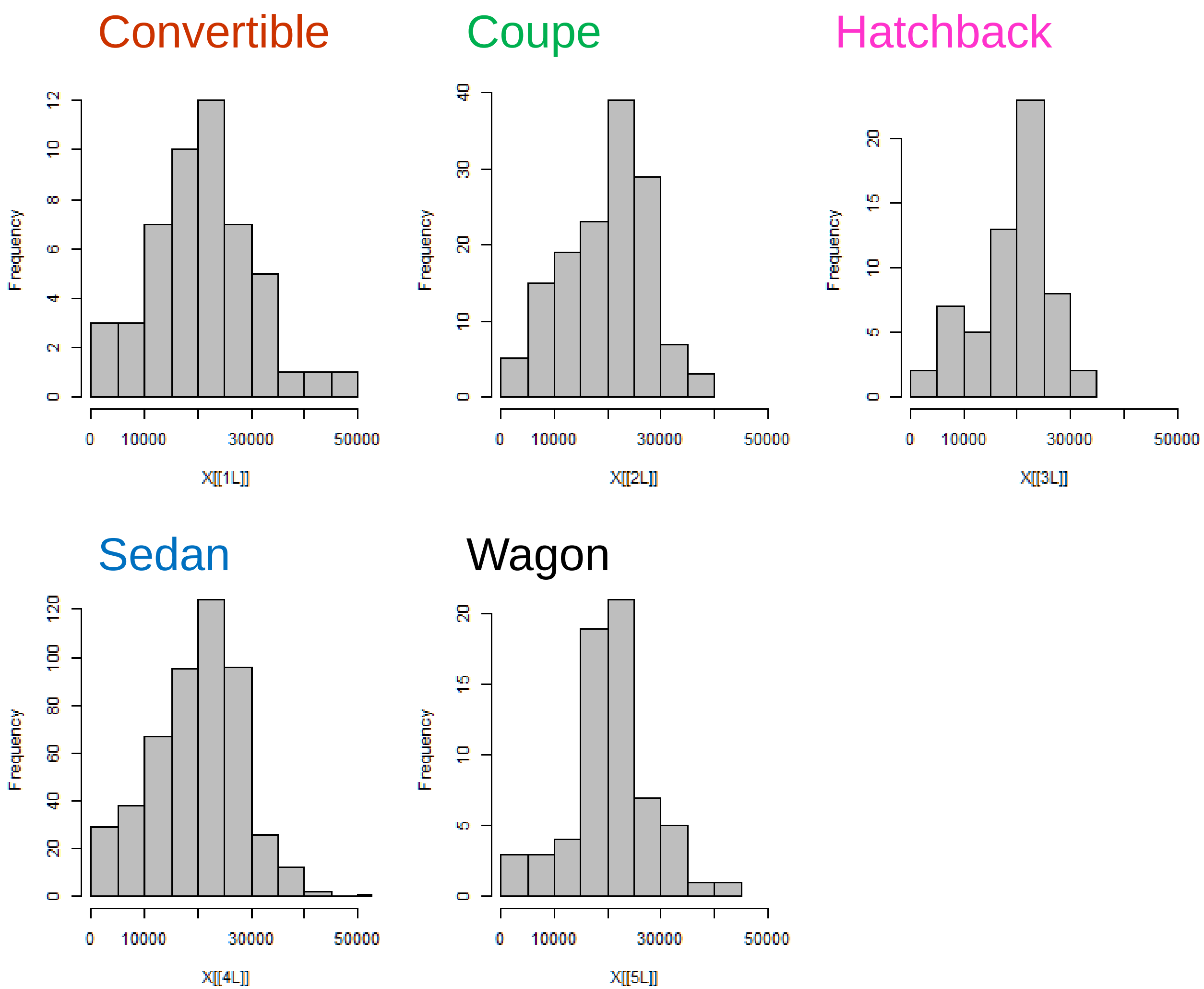
Modeling Car Price: Ecological Fallacy

```
> tapply(CarData$Mileage, CarData$Type, mean)
Convertible          Coupe    Hatchback          Sedan          Wagon
      20705          19859        19091          19728          20584
```

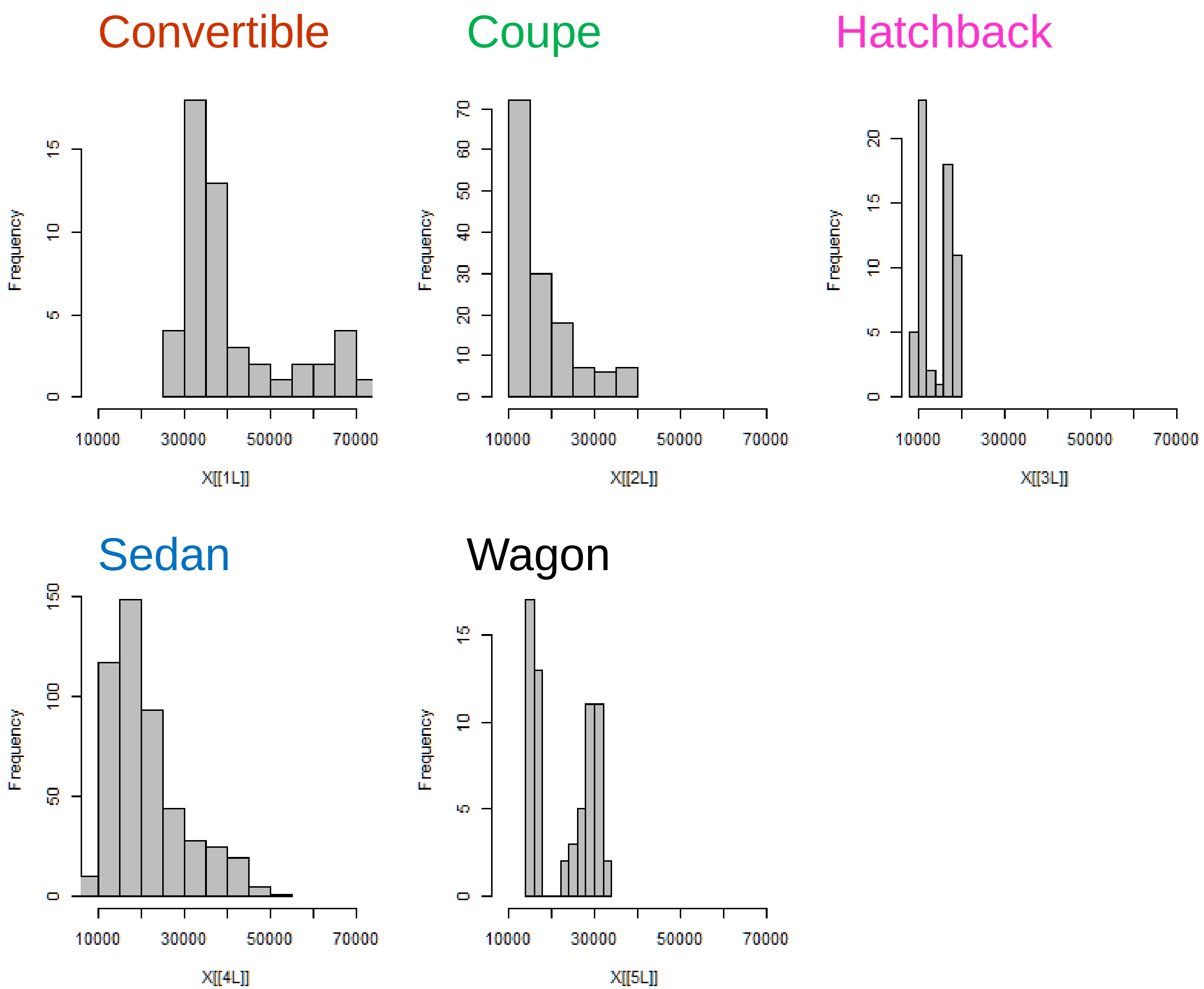




Mileage


Convertible
Coupe
Hatchback
Sedan
Wagon
Frequency
X[[1L]]
X[[2L]]
X[[3L]]
X[[4L]]
X[[5L]]
0
10000
30000
50000
0
2
4
6
8
10
12
0
10
20
30
40
0
5
10
15
20
0
20
40
60
80
100
120

Price


Convertible
Coupe
Hatchback
Sedan
Wagon
Frequency
10000
30000
50000
70000
X[[1L]]
X[[2L]]
X[[3L]]
X[[4L]]
X[[5L]]