

Regression III

Lecture 12: Model Selection

Dave Armstrong

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

1 / 56

Goals of the lecture

- Explain the problem of collinearity
- Introduce simple diagnostics for assessing collinearity
- Consider different methods for model selection among nested and non-nested alternatives:
 - Nested model tests
 - Non-nested model tests
 - Automated variable selection
- Multi-model inference

2 / 56

Collinearity or Multicollinearity

- Multiple regression attempts to separate out the effects of each explanatory variable holding all others constant
- If two variables are perfectly related, when one is held constant, the other is as well - in these cases, the least-squares coefficients are not uniquely defined
 - A strong, but less than perfect, relationship among the X 's causes unstable OLS coefficients - standard errors are large, reflecting the imprecision in the estimation of the β 's
- If the goal is to use a regression model to predict Y , then collinearity is not problematic - as long as the model accounts for a high proportion of the variation in Y , the predictions should be quite accurate
- If we wish to understand how each of the X 's impacts Y , collinearity is potentially problematic

3 / 56

Detection of Collinearity May Not Have Practical Implications

- Even in the presence of collinearity, OLS is BLUE - therefore, if the standard errors are small, the degree of collinearity is irrelevant
 - Even if the standard errors are large, knowing that collinearity is *the* problem can only help if the study can be re-designed so that the correlation between the X 's is reduced or if more data can be collected to increase n .
- Methods that are commonly employed to compensate for collinearity (e.g., biased estimation) and variable selection - can often be worse than the collinearity itself
- Finally, collinearity is seldom a problem in the social sciences
 - Insufficient variation in explanatory variables, small samples and large error variances are more frequently the source of imprecision in estimation
 - Time-series and panel data are exceptions

4 / 56

Detecting Imperfect Collinearity

1. Start by looking at the pairwise relationships between variables (correlations above .8 (or so) are troubling)
 - Very high pairwise correlations suggest collinearity, but it is possible to have problems when the pairwise relationships aren't *that* high.
2. Look at the squared multiple correlation - the R_j^2 from a regression with X_j as the outcome variable and X_{-j} as the explanatory variables (above .6 or so is a problem here)
3. Look at the variance inflation, which is a function of R_j^2 : $VIF = (1 - R_j^2)^{-1}$
4. Large discrepancies in Eigenvalues from an Eigen-decomposition suggests high levels of collinearity.

5 / 56

Variance Inflation Factor(1)

- Collinearity need not be perfect to effect the precision of the slope estimates. The variance of the slope coefficient B_j is as follows:

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{(n - 1)S_j^2}$$

where R_j^2 is defined as above, $S_j^2 = \sum(X_{ij} - \bar{X}_j)^2 / (n - 1)$ is the variance of X_j

- The first term is the variance inflation factor (VIF) because it indicates the impact of collinearity on the precision of B_j

6 / 56

Variance Inflation Factor (2)

Important

- The R_j^2 in the VIF equation is not from pairwise correlations
 - They are *multiple correlations* from X_j regressed on all the other X 's.
- Bivariate correlations do not always give the story of collinearity, hence the term *multicollinearity* is often used
- The size of the confidence interval for β_j is proportional to the square root of the VIF
 - Consequently, it is much easier to assess the impact of collinearity if we take the square root of the VIF
- The square root of the VIF tells the factor by which the standard error and confidence interval is inflated as a function of multicollinearity

7 / 56

Variance Inflation Factor Example: Prestige Data

```
> options(useFancyQuotes = F)
> library(car)
> data(Prestige)
> mod <- lm(prestige ~ income + education, data = Prestige)
> summary(mod)
```

```
Call:
lm(formula = prestige ~ income + education, data = Prestige)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-19.4040  -5.3308   0.0154   4.9803  17.6889
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.8477787  3.2189771  -2.127  0.0359 *
income       0.0013612  0.0002242   6.071 2.36e-08 ***
education    4.1374444  0.3489120  11.858 < 2e-16 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-squared:  0.798, Adjusted R-squared:  0.7939
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

8 / 56

Variance Inflation Factor: R-Script

```
> vif(mod)
  income education
1.500598 1.500598

> sqrt(vif(mod))
  income education
1.224989 1.224989
```

- The VIFs for income and education are both 1.5 (they're identical because there are no other explanatory variables, the information is based only on the pairwise correlation)
- Taking the square root of the VIF makes it easier to interpret
 - The standard error for income is 1.22 times as large as it would be if income were unrelated to education. That is not a great deal, so in practical terms, these variables are not collinear in a problematic way

9 / 56

Generalized Variance Inflation Factor (1)

- The VIF is a good measure of collinearity for a single explanatory variable, but it cannot assess collinearity for sets of related regressors such as a set of dummy regressors
 - For example, correlations between dummy regressors are affected by the choice of baseline. Although, the correlations are artificial, the VIF does not take this into consideration
- The generalized variance inflation factor (GVIF) however, can accommodate sets of related regressors:

$$\text{GVIF}_1 = \frac{\det \mathbf{R}_{11} \det \mathbf{R}_{22}}{\det \mathbf{R}}$$

- \mathbf{R}_{11} is the correlation matrix of X_1 (set 1); \mathbf{R}_{22} is the correlation matrix for X_2 (set 2); and \mathbf{R} is the matrix containing the correlations among all the X variables

10 / 56

GVIF (2)

- The GVIF is independent of the coding scheme used for the factor (i.e., it doesn't matter which category is the reference category)
 - Nonetheless, choosing a large category as the reference category can give more numerically stable results
- This measure essentially shows how much the joint confidence region for the set of predictors is expanded relative to similar data for which 2 sets of predictors are unrelated to each other
 - Similar to taking the square root of the VIF, $\text{GVIF}^{\frac{1}{\text{df}}}$ gives the reduction in precision of the estimation due to collinearity (the degrees of freedom is the number of regressors for the variable)

11 / 56

GVIF Example: Prestige Data

```
> mod <- lm(prestige ~ income + education + type, data = Prestige)
> summary(mod)
```

Call:

```
lm(formula = prestige ~ income + education + type, data = Prestige)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-14.9529	-4.4486	0.1678	5.0566	18.6320

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.6229292	5.2275255	-0.119	0.905
income	0.0010132	0.0002209	4.586	1.40e-05 ***
education	3.6731661	0.6405016	5.735	1.21e-07 ***
typeprof	6.0389707	3.8668551	1.562	0.122
typewc	-2.7372307	2.5139324	-1.089	0.279

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.095 on 93 degrees of freedom
(4 observations deleted due to missingness)

Multiple R-squared: 0.8349, Adjusted R-squared: 0.8278

F-statistic: 117.5 on 4 and 93 DF, p-value: < 2.2e-16

12 / 56

GVIF Example: Prestige Data (2)

```
> vif(mod)
```

	GVIF	Df	GVIF ^{1/(2*Df)}
income	1.681325	1	1.296659
education	5.973932	1	2.444163
type	6.102131	2	1.571703

- If a factor is included in the model, the `vif` function automatically returns the GVIF
- It also returns the adjusted GVIF = $GVIF^{\frac{1}{2df}}$ which accounts for degrees of freedom
 - For quantitative variables this is simply the square root of the VIF - in other words, we no longer need to take the square root to determine the factor by which standard errors are inflated
- We see here that collinearity is not a problem - the GVIF for type is moderate, this the standard error is not really effected.

13 / 56

Coping with collinearity

- When X_1 and X_2 are strongly collinear, the data contain little information about the impact of X_1 on Y holding X_2 constant, because there is little information about X_1 when X_2 is fixed
 - It is important to note, that the estimates and standard errors for variables not characterized by collinearity are not affected by a collinearity problem elsewhere
- Although there are several strategies for dealing with collinear data, none can extract nonexistent information. Instead, the research problem is necessarily redefined, if only implicitly
 - Sometimes, the redefinition is reasonable; usually it is not
- The only *solution* to collinearity is to collect new data in a manner that avoids the problem
 - Experimental manipulation of the X 's, though this is rarely practical

14 / 56

Model Re-Specification

- Although collinearity is usually a data problem, the most common approach to handling collinearity is model re-specification
- If several variables can be conceptualized as alternative indicators of the same construct, they can either be combined in a single measure or one variable can be chosen to represent the others
 - In these cases, simple additive scales, principal components analysis or factor analysis can be used
 - High correlations among the X 's in this case are a good thing - they indicate high reliability
- If the composite measure has a substantive interpretation, this can be a good solution
- Of course, if we want to distinguish the effects of the variables that are collinear, this is not a particularly good solution

15 / 56

Options for Model Fit

- Direct tests of nested models - F (Anova), χ^2 (Analysis of Deviance, LR-Test)
- Information Criteria measures - AIC and BIC (among others)
- Distribution Free Test of Non-nested Models (Clarke)

16 / 56

Nested Model Tests

Tests like the LR test and F-test require nested models because,

- They are considering the difference between two statistics (RSS or LR)
- This difference follows an F or χ^2 distribution under the null (neither distribution permits negative values).
- So, the model with more parameters *must* provide a fit not worse than the model with fewer parameters.
 - The only way to ensure this is the case is to ensure that the models are nested

17 / 56

Likelihood Ratio Test

The LR Test uses the statistic defined by the difference in the log-likelihoods of the models.

$$LR = -2(\ell_{\text{restricted}} - \ell_{\text{unrestricted}}) \sim \chi^2_{p-q} \quad (1)$$

where there are p parameters in the unrestricted model and q parameters in the restricted model.

- The distribution is asymptotically right, but will not be exactly χ^2 in finite samples.
- Deviance is often taken as $-2\ell_{\text{model}}$, though this is not always the case (take, for example, the linear model case).

18 / 56

Information Theory

- Information theorists believe in reality, but not in the notion of “true” models.
 - Models are necessarily simplified constructions that try to approximate reality.
- There is more information in large datasets than small.
 - Information amounts to the ability to identify interesting, though substantively small effects

19 / 56

Three Principles guiding Model-based Inference

1. Parsimony
 - Encapsulates the bias-variance tradeoff.
2. Multiple Working Hypotheses
 - There is no single null hypothesis against which an alternative is to be tested.
 - rather, there is a (small-ish) set, well-specified and theoretically derived working hypotheses.
3. Strength of Evidence
 - We must be able to quantify the “strength of evidence” supporting various working hypotheses if science is to progress in the usual way.

20 / 56

K-L Information

Kullback and Leibler (1951) quantified the meaning of “information”.

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx$$

where:

- f denotes a fixed (i.e., constant) reality (reality is non-parametric [i.e., it has no parameters])
- g is a model approximating f with parameters θ .
- $I(f, g)$ is the information lost when using g to approximate f .

There is no assumption that a true model exists (much less that the true model is in our candidate set of models) nor is there an assumption that the models are nested.

21 / 56

Expected Information

We cannot use $I(f, g)$ in model selection because it requires knowledge of f and θ the parameters in g .

$$\begin{aligned} I(f, g) &= E_f [\log(f(x))] - E_f [\log(g(x|\theta))] \\ &= C - E_f [\log(g(x|\theta))] \end{aligned}$$

Estimating relative information for each model in the set results in our ability to compare across models (since C is constant for all model comparisons).

22 / 56

Akaike's Information Criterion (AIC)

The goal was to estimate: $E_y E_x [\log(g(x|\hat{\theta}(y)))]$, essentially the relative information with θ replaced with the MLE estimates $\hat{\theta}$.

- Akaike found that $\log(\mathcal{L}(\hat{\theta}|\text{data}))$ was a biased estimator of $E_y E_x [\log(g(x|\hat{\theta}(y)))]$, but that asymptotically the bias is approximately equal to K , the number of parameters in $\hat{\theta}$. Thus,

$$\log(\mathcal{L}(\hat{\theta}|\text{data})) - K = C - \hat{E}_{\hat{g}} [I(f, \hat{g})]$$

K is not arbitrary, but chosen to minimize bias in the estimated expected information.

$$\begin{aligned} AIC &= -2(\log(\mathcal{L}(\hat{\theta}|\text{data})) - K) \\ &= -2\log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K \end{aligned}$$

23 / 56

Small-sample Correction

When K is large relative to n or for any value of K for small- n , there is a correction to AIC .

$$AIC_c = -2\log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K + \frac{2K(K+1)}{n-K-1}$$

- This should be used probably always, but especially if $n/K \leq 40$ for the largest K in the model set.
- AIC_c converges to AIC as $n \rightarrow \infty$.

24 / 56

Δ_i values

Often, for AIC_c or AIC to be interpretable, Δ_i should be calculated such that for each model i in the model set,

$$\Delta_i = AIC_i - AIC_{\min}$$

This gives the “best” model $\Delta_i = 0$

- This captures the information loss due to using model g_i rather than the best model, g_{\min} .
- The large Δ_i , the less likely model i is the best approximation of reality f .

Conventional cut-off values for Δ_i are:

- $\Delta_i \leq 2$ indicates substantial support,
- $4 \leq \Delta_i \leq 7$ indicates less support,
- $\Delta_i \geq 10$ indicates essentially no support.

BIC

The BIC is defined as:

$$BIC = -2 \log(\mathcal{L}) + K \log(n)$$

- BIC is not technically based in “information theory” and as such is not an information criterion measure.
- The BIC is meant to approximate the Bayes Factor (or rather its log):

$$\frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(\theta_1|M_1) \Pr(D|\theta_1, M_1) d\theta_1}{\int \Pr(\theta_2|M_2) \Pr(D|\theta_2, M_2) d\theta_2}$$

- Models need not be nested and we need not appeal to the idea that there exists a “true” model, much less that the true model is in our set of candidate models.

AIC or BIC

The question of whether to use AIC or BIC is often left to how much you want to penalize additional model parameters. In actuality, the question is one of performance in picking the K-L best model.

- When there are “tapering effects”, AIC is better
- When reality is simple with a few big effects captured by the highest posterior probability models, then BIC is often better.

Non-nested Model Tests

Both AIC and BIC work for non-nested models, but neither is a *test* per se (i.e., they don't have sampling distributions which can be evaluated to produce p -values). There are a set of tests for non-nested models that do have known sampling distributions. Consider the following set of models:

$$H_1 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_1, \quad E(\mathbf{u}'_1 \mathbf{u}_1) = \sigma_1^2 \mathbf{I}$$
$$H_2 : \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}_2, \quad E(\mathbf{u}'_2 \mathbf{u}_2) = \sigma_2^2 \mathbf{I}$$

where:

- $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of length k_1 and k_2 , respectively.

These models are non-nested if, in general, it is impossible to put restrictions on $\boldsymbol{\beta}$ to arrive at $\boldsymbol{\gamma}$ and vice-versa.

Artificial Nesting

Artificial nesting is the easiest way to perform these tests:

$$H_c: \mathbf{y} = (1 - \alpha)\mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}$$

Often, if the variables in \mathbf{X} and \mathbf{Z} are overlapping, not all model parameters will be identifiable so other alternatives are required.

Encompassing Test

The encompassing test is based on the encompassing principle which suggests that the maintained model may be able to explain features of its competitors to the point where its competitors add nothing to the model. To perform the test, we estimate the model:

$$\mathbf{y} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}} + \bar{\mathbf{Z}}\bar{\boldsymbol{\gamma}} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

where:

- $\bar{\mathbf{X}}$ comprises the variables in \mathbf{X} that are not in \mathbf{Z} and $\bar{\mathbf{Z}}$ is defined similarly. \mathbf{W} is the set of variables that are in both \mathbf{X} and \mathbf{Z} .
- H_1 is rejected if $\bar{\boldsymbol{\beta}} = 0$ is not rejected with a conventional F -test.
- H_2 is rejected if $\bar{\boldsymbol{\gamma}} = 0$ is not rejected with a conventional F -test.

An equivalent test can be done with:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{Z}}\bar{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}$$

J-test

The J-test uses an analog to artificial nesting by estimating the following model:

$$H'_c: \mathbf{y} = (1 - \alpha)\mathbf{X}\boldsymbol{\beta} + \alpha\hat{\mathbf{z}} + \mathbf{u}$$

where $\hat{\mathbf{z}}$ is a consistent estimate of $\mathbf{Z}\boldsymbol{\gamma}$. Here, we use $\mathbf{Z}\hat{\boldsymbol{\gamma}}$ from an OLS estimate.

- The J-test is simply the t -test of $\alpha = 0$ both by using $\hat{\mathbf{z}}$ and $\hat{\mathbf{x}}$, similarly defined.

Vuong Test

The Vuong test is a likelihood ratio test specified as follows:

$$\tilde{L}R_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n) = \log(\mathcal{L}_1) - \log(\mathcal{L}_2) - \frac{k_1 - k_2}{2} \log n$$

This statistic has a standard normal distribution under the null hypothesis that the two models are not different from each other.

Cox Test

The Cox test is also a likelihood based test.

1. First, estimate both models and calculate $\hat{l}_{fg} = \log(\mathcal{L}_f) - \log(\mathcal{L}_g)$
2. Calculate T as the difference between \hat{l}_{fg} and $E(\hat{l}_{fg})$ where the second term is the model estimated with predicted values from the null model as the dependent variable.
3. The statistic has a standard normal distribution under the null hypothesis.

33 / 56

Distribution Free Test

Clarke (2003) puts forth a distribution-free test that is really a “paired sign test”. The statistic is calculated as:

$$d_i = \log(\mathcal{L}_{\beta, x_i}) - \log(\mathcal{L}_{\gamma, z_i})$$
$$B = \sum_{i=1}^n I_{0,+\infty}(d_i)$$

- The d_i are the difference in individual log-likelihoods for the two models
- The second equation above counts up the number of positive d_i values.
- We are testing to see whether B is significantly bigger than a random binomial variable that has a $p = .5$ and n the same as the number of rows in X and Z .

34 / 56

Examples in R

```
> library(car)
> data(Prestige)
> mod1 <- lm(prestige ~ income + women, data = na.omit(Prestige),
+           y = T)
> mod2 <- lm(prestige ~ education + type + women, data = na.omit(Prestige),
+           y = T)
> smod1 <- summary(mod1)
> smod2 <- summary(mod2)
> library(lmtest)
> jtest(mod1, mod2, Prestige)

J test

Model 1: prestige ~ income + women
Model 2: prestige ~ education + type + women
              Estimate Std. Error t value Pr(>|t|)
M1 + fitted(M2)  0.82365    0.064343 12.8009 < 2.2e-16 ***
M2 + fitted(M1)  0.31531    0.079309   3.9758 0.0001395 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> encomptest(mod1, mod2, Prestige)

Encompassing test

Model 1: prestige ~ income + women
Model 2: prestige ~ education + type + women
Model E: prestige ~ income + women + education + type
              Res.Df Df    F    Pr(>F)
M1 vs. ME          92  -3 53.481 < 2.2e-16 ***
M2 vs. ME          92  -1 15.807 0.0001395 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

35 / 56

Examples in R

```
> coxtest(mod1, mod2, Prestige)

Cox test

Model 1: prestige ~ income + women
Model 2: prestige ~ education + type + women
              Estimate Std. Error z value Pr(>|z|)
fitted(M1) ~ M2 -65.625      3.2933 -19.9268 < 2.2e-16 ***
fitted(M2) ~ M1 -15.203      3.1316  -4.8546 1.206e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> vuong <- (logLik(mod2) - logLik(mod1)) - ((mod2$rank - mod1$rank)/2) *
+ log(nrow(na.omit(Prestige)))
> vuong

'log Lik.' 37.10624 (df=6)

> il1 <- log(dnorm(mod1$y, mod1$fitted, smod1$sigma))
> il2 <- log(dnorm(mod2$y, mod2$fitted, smod2$sigma))
> cuts <- qbinom(c(0.05, 0.95), size = length(mod1$y), prob = 0.5)
> d <- sum(il1 > il2)
> pbinom(d, length(il1), 0.5)

[1] 1.948207e-08

> AIC(mod1)

[1] 763.8879

> BIC(mod1)

[1] 774.2278
```

36 / 56

Variable Selection

- Another common approach to collinearity is the use of some variable selection technique
- An attempt is made to reduce the number of regressors to a less highly correlated set
- Often, stepwise methods are used toward this end
 - Forward stepwise methods add explanatory variables to the model one at a time. At each step, the variable that yields the largest increment in R^2 is selected. The procedure stops with the increment is smaller than some preset condition
 - Backward stepwise methods are similar except that the procedure starts with the full model and deletes variables one at a time
 - Forward/backward methods combine the two approaches

37 / 56

Stepwise Regression using AIC and BIC

- Applicable to any model (including the linear model) for which a log-likelihood can be found.
- The Akaike Information Criterion (AIC) is a measure of fit that penalizes the fit according to the number of variables
 $AIC = -2ll + 2p$ where ll is the log-likelihood and p is the number of parameters in the model
- The Bayesian Information Criterion (BIC) is similar to the AIC, except that it applies a larger penalty that also includes sample size (prefers more parsimonious models): $BIC = -2ll + p \log n$ where n is the number of observations
- For both measures our goal is to make them as small as possible. That is, we want the best balance of fit and size

38 / 56

Stepwise Regression using AIC and BIC in R

- We can obtain the best fitting models according to BIC and AIC several different ways in R
- The easiest way is to use the step function
 - The formula $-2ll + kp$ is used (by default $k = 2$ (AIC), $K = \log n$ gives the BIC)
 - By default a backward stepwise method is used

```
> data(Ericksen)
> mod <- lm(undercount ~ minority + crime + poverty + language +
+   highschool + housing, data = Ericksen)
> mod.step.aic <- step(mod)
```

Start: AIC=68.94

```
undercount ~ minority + crime + poverty + language + highschool +
  housing
```

	Df	Sum of Sq	RSS	AIC
- housing	1	0.150	151.87	67.003
- poverty	1	0.159	151.88	67.007
- highschool	1	2.379	154.10	67.965
<none>			151.72	68.938
- crime	1	6.335	158.06	69.638
- language	1	8.019	159.74	70.337
- minority	1	34.331	186.05	80.401

39 / 56

Stepwise Regression: Some Cautions

- Stepwise methods are often abused by researchers who wrongly interpret the order of entry of the X 's as an indication of their "importance"
 - Suppose that there are two highly correlated X 's that have nearly identical large correlations with Y ; only the one that enters first will be statistically significant and thus enter the regression equation
 - A small modification to the data, or a new sample could easily reverse the result
- Ultimately, stepwise methods can fail to turn up the optimal subset of regressors of a given size
 - Recall the difference between type I and type II F -tests
- Ideally we will enter variables in a stepwise fashion only if we have a theoretical explanation for their position in the causal order

40 / 56

Subset Methods

- The goal of subset methods is to examine *which subsets give the best fit to the data with the smallest number of predictors*
- Even when the number of variables is large, it is feasible to examine all subsets
 - If there are p potential predictors, then there are 2^p possible models
- Subset techniques have the advantage over stepwise regression of revealing alternative *nearly equivalent* models and thus avoid the appearance of a uniquely “correct” result
- Several measures can be used to determine the best model subset
 - R^2
 - AIC
 - BIC
 - Mallow's C_p -statistic

41 / 56

Subset Methods: Mallow's C_p -statistic

- Mallow's C_p -statistic is defined as:

$$\begin{aligned}C_p &= \frac{\sum E_i^2}{S_E^2} + 2p - n \\ &= (K + 1 - p)(F_p - 1) + p\end{aligned}$$

- S_E^2 is for the full model containing k explanatory variables; RSS ($\sum E_i^2$) is from the subset model with p explanatory variables
- F_p is the incremental F -test for the hypothesis that the regressors omitted from the subset have slope 0. If the hypothesis is true, $E(F_p) \approx 1$, and thus $C_p \approx p$
- C_p minimizes the sum of squared residuals and thus maximizes the R^2
- A good model, then, has C_p as close to p as possible
- A plot of C_p against p allows us to choose the model

42 / 56

Model Selection Example: Ericksen Data (1)

```
> library(leaps)
> X <- Ericksen[, c("minority", "crime", "poverty", "language",
+ "highschool", "housing")]
> y <- Ericksen$undercount
> rmods <- regsubsets(x = X, y = y, method = "exhaustive", all.best = TRUE
+ nbest = 10)
```

43 / 56

Model Selection Example: Ericksen Data (2)

```
> sqrt(vif(mod))
minority    crime    poverty    language    highschool    housing
2.041873    1.713406    1.908846    1.257935    1.776573    1.267794
```

- The model summary indicates the possibility of high multicollinearity - despite a high R^2 and low residual variance
- The VIF's confirm this speculation
 - The square root of the VIF is approximately 2 for minority and poverty, indicating that they are highly related to the other X 's in the model
- We now try to select variables that limit collinearity using subset method

44 / 56

Model Selection Example: Ericksen Data (3)

- Subset selection is implemented in **R** using two packages: `leaps` and `car`
 - Using the `regsubsets` function, you specify the full model and how many subsets you want
- The `subsets` function in `car` graphs the models with the subset size on the horizontal axis and the statistic used for fit on the vertical axis
 - The `subsets` function allows you to specify the following statistics: Mallows C_p "cp", R^2 "rsq", adjusted R^2 "adjrs2", RSS "rss" or BIC "bic"
 - You can also specify the number of predictors you want in the model (below specifies 3 to 5 predictors)

45 / 56

Subsets plot for the Ericksen Data

```
> library(car)
> pdf("subsetfig.pdf", height = 6, width = 6)
> subsets(rmods, statistic = "cp", legend = F)
```

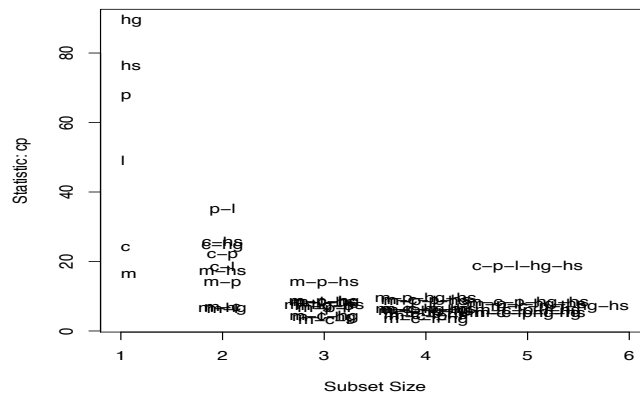
	Abbreviation
minority	m
crime	c
poverty	p
language	l
highschool	hg
housing	hs

```
> dev.off()

null device
      1
```

46 / 56

Subsets plot for the Ericksen Data (2)



47 / 56

Variable Selection Methods: Cautions (1)

- Variable selection has applications to statistical modelling even when collinearity is not an issue
 - If we have a very large number of predictors and we simply want a parsimonious predictive model, subset methods are very valuable
 - Subset methods are also useful when our analysis is intentionally exploratory - i.e., we have a vague idea of many possible predictors, but not a concrete hypothesis about how they are related to the outcome variable
- When tackling collinearity, however, variable selection results in a re-specified model that usually does not address the original research question
 - If the original model is correctly specified, then coefficient estimates following variable selection are *biased*. However, the bias may not be overwhelming if you started off with a severe collinearity problem

48 / 56

Variable Selection Methods: Cautions (2)

- If our goal is to assess the individual predictors (or their relative impacts), variable selection models have serious implications
 - Standard errors calculated following variable selection overstate the precision of results - they do not control for relevant predictors
 - A new sample may give different results, leading to inconsistent interpretation of "effects"
- Finally, when regressors occur in sets (e.g., dummy regressors representing categorical variables), these sets should be kept together during selection. Likewise, when there are hierarchical relations among regressors (interaction effects), these relations should be respected

49 / 56

Model Selection Uncertainty

Generally we present what we think to be the single best model after a more or less exhaustive model search.

- Our estimates of sampling variability of parameters is often too small because we "forget" to include model selection uncertainty.
- This uncertainty captures the extent to which we are unsure about this model and have considered other alternatives.
- Others have proposed solutions to the problem (e.g., Leamer, with the idea of "Leamer Bound"), but we will consider two alternatives here - AIC/BIC weights and Bayesian Model Averaging.

50 / 56

Akaike Weights

We can construct Akaike weights in the following way:

$$w_i = \frac{\exp\left(\frac{-\Delta_i}{2}\right)}{\sum_i \exp\left(\frac{-\Delta_i}{2}\right)}$$

- $\exp\left(\frac{-\Delta_i}{2}\right)$ is the likelihood of the model given the data.
- w_i gives (essentially) the probability that model i is the K-L best model approximation of f .

These estimates can be used to give us measures of sampling variance that are *unconditional* on the model selected.

$$\widehat{\text{var}}(\hat{\theta}) = \left[\sum_i w_i \left[\widehat{\text{var}}(\hat{\theta}_i | g_i) + (\hat{\theta}_i - \hat{\theta})^2 \right]^{\frac{1}{2}} \right]^2$$

where $\hat{\theta} = \sum_i w_i \hat{\theta}_i$

51 / 56

Example

Let's think about the Duncan Dataset and estimating different models.

```
> c1 <- t(combn(4, 1))
> c2 <- t(combn(4, 2))
> c3 <- t(combn(4, 3))
> c4 <- t(combn(4, 4))
> l <- list()
> k <- 1
> for (i in 1:nrow(c1)) {
+   l[[k]] <- c1[i, ]
+   k <- k + 1
+ }
> for (i in 1:nrow(c2)) {
+   l[[k]] <- c2[i, ]
+   k <- k + 1
+ }
> for (i in 1:nrow(c3)) {
+   l[[k]] <- c3[i, ]
+   k <- k + 1
+ }
> for (i in 1:nrow(c4)) {
+   l[[k]] <- c4[i, ]
+   k <- k + 1
+ }
> P <- Prestige[, c("income", "education", "women", "type", "prestige")]
> data <- lapply(l, function(x) P[, c(x, 5)])
> mods <- lapply(data, function(x) lm(as.formula(paste("prestige ~ ",
+   ifelse(length(names(x)) > 2, paste(names(x)[1:(ncol(x) -
+   1)]), collapse = " + "), colnames(x)[1]), sep = "")),
+   data = x)
> AICs <- sapply(mods, AIC)
> delta.aic <- AICs - min(AICs)
> w <- exp(-delta.aic/2)/sum(exp(-delta.aic/2))
```

52 / 56

Multi-model Sampling Variance in R

```
> coefs <- vars <- matrix(0, ncol = 6, nrow = length(mods))
> colnames(coefs) <- names(coef(mods[[15]]))
> for (i in 1:length(mods)) {
+   coefs[i, match(names(coef(mods[[i]])), colnames(coefs))
+   vars[i, match(names(coef(mods[[i]])), colnames(coefs))]
+ }
> theta.bar.hat <- w %*% coefs
> part.2 <- sweep(coefs, 2, theta.bar.hat)^2
> var.theta.bar.hat <- (w %*% sqrt(vars + part.2))^2
> theta.bar.hat/var.theta.bar.hat
```

```
(Intercept) income education women typeprof typ
[1,] -0.02443621 18819.7 8.901594 18.55387 0.3973234 -0.4262
```

53 / 56

More Multi-model Inference

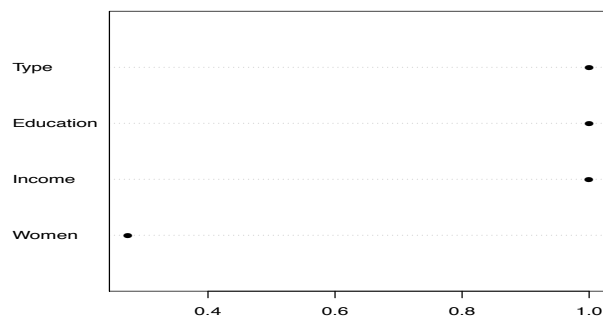
We could also think about two other uses of Akaike weights:

- We could actually use $\hat{\theta}$ as our point estimate that will be less biased due to model selection.
- Summing w_i across all of the models including variable j can give a sense of how “important” variables are.

54 / 56

Important Variables in R

```
> impvars <- (w %*% (coefs != 0))[, 2:5]
> names(impvars) <- c("Income", "Education", "Women", "Type")
> impvars <- impvars[order(impvars)]
> pdf("impdot.pdf", height = 6, width = 6)
> dotchart(impvars, pch = 16)
> invisible(dev.off())
```



55 / 56

BIC and Posterior Model Probabilities

The BIC can be used to calculate posterior model probabilities (under the assumption that for R candidate models, they all have prior probability $\frac{1}{R}$).

$$\Delta BIC_i = BIC_i - BIC_{min}$$

then,

$$p_i = \Pr\{g_i | \text{data}\} = \frac{\exp\left(-\frac{1}{2}\Delta BIC_i\right)}{\sum_i \exp\left(-\frac{1}{2}\Delta BIC_i\right)}$$

This gives the probability that model i is the “quasi-true” model (i.e., the best in the set of candidate models).

56 / 56