

Chapter 9: Building the Regression Model I: Model Selection and Validation

STAT 3240

Michael McIsaac

UPEI

Learning Objectives for Section 9.1

After Section 9.1, you should be able to

- Understand the model-building process: data collection, variable selection, model selection, model validation

9.1: Overview of Model-Building Process

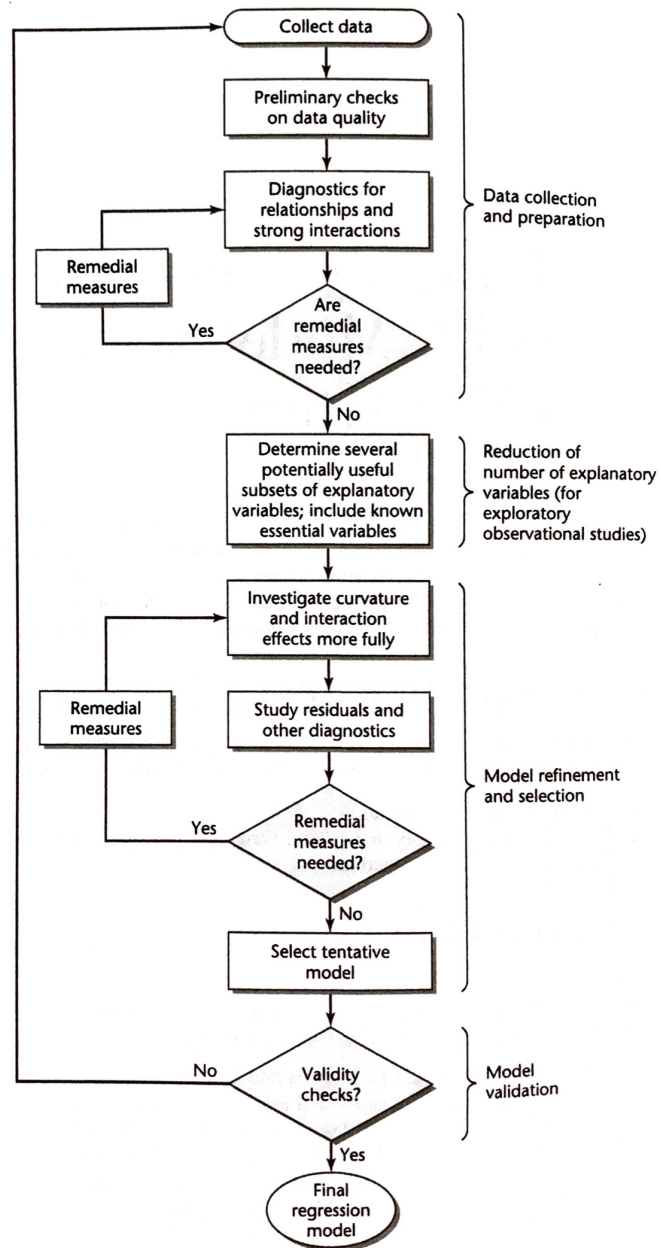
Data Collection:

- **Controlled Experiments**
- **Controlled Experiments with Covariates**
- **Confirmatory Observational Studies**
- **Exploratory Observational Studies**

CONSORT statement on randomization: www.consort-statement.org/checklists/view/32--consort-2010/87-randomisation-type

CONSORT statement on adjusted analyses: www.consort-statement.org/checklists/view/32--consort-2010/97-additional-analyses

FIGURE 9.1
Strategy for
Building a
Regression
Model.



Residential outdoor light-at-night and breast cancer risk in Canada

Experimental and epidemiologic studies suggest that light-at-night (LAN) exposure disrupt circadian rhythms, which may increase breast cancer risk. We investigated whether residential outdoor LAN is associated with breast cancer risk in Canada.

A population-based case-control study was conducted in Vancouver, British Columbia and Kingston, Ontario, Canada with incident breast cancer cases and controls frequency matched by age to cases living in the same region. This analysis was restricted to 782 cases and 833 controls who provided lifetime residential histories. Using time-weighted average duration at each home 10-20 years prior to study entry, two measures of cumulative average outdoor LAN were calculated using two satellite data sources.

Logistic regression was used to estimate the relationship between outdoor LAN and breast cancer risk, considering interactions for menopausal status and night shift work.

Gender-based inequalities in adolescent mental health in Canada

We aim to generate new knowledge about the processes by which social factors, such as gender and social positions, act in combination to systematically marginalize or privilege boys and girls leading to gender-based inequalities in mental health in Canada.

Our quantitative strand will use a WHO-affiliated adolescent health survey (the Health Study Approach. Behaviour in School-aged Children (HBSC) study). Analyses will involve descriptive and analytic methods that follow epidemiological traditions.

Comparative effectiveness of newer oral diabetes medications in preventing advanced diabetic retinopathy

Diabetic retinopathy (DR), the most common complication of diabetes, is the leading cause of blindness and vision impairment in working-age adults. Pivotal studies have established that pharmacotherapy decreases the risk of developing severe DR and the associated vision loss. However, these studies were conducted before the availability of the many newer diabetes medication classes. Evidence suggests that these newer medications differ in their effects on DR.

We will conduct a population-based retrospective cohort study to address the objectives. Administrative health care databases will be linked to perform the planned analyses. In our primary analysis we will use propensity score based methods to compare the risk of developing severe DR among patients treated with DPP4Is, SGLT2Is and SUs.

A Phase II Randomized, Double-Blind, Placebo-Controlled Study of the Efficacy, Safety, and Tolerability of Arbaclofen Administered for the Treatment of Social Function in Children and Adolescents with Autism Spectrum Disorders.

Warm-up Exercises

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+recre  
msummary(clothing_model)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.44e+02   3.18e+02   2.02   0.0435
## income         1.75e-02   4.13e-03   4.24   2.6e-05
## sexmale       -5.31e+02   1.42e+02  -3.75   0.0002
## food_expenditure 1.43e-01   2.61e-02   5.51   5.9e-08
## recreation_expenditure 2.19e-01   3.37e-02   6.49   2.1e-10
## miscellaneous_expenditure 1.12e-01   1.39e-01   0.81   0.4180
## marital_statusnever_married -2.46e+02   2.11e+02  -1.17   0.2425
## marital_statusother -5.10e+02   1.93e+02  -2.65   0.0083
## type_of_dwellingduplex    3.51e+02   3.41e+02   1.03   0.3028
## type_of_dwellingother   -3.65e+02   4.43e+02  -0.82   0.4111
## type_of_dwellingrow_house -1.01e+03   4.38e+02  -2.31   0.0215
## type_of_dwellingsemi_detached -1.33e+02   3.72e+02  -0.36   0.7213
## type_of_dwellingssingle_detached -2.38e+02   2.42e+02  -0.98   0.3269
##
## Residual standard error: 1470 on 487 degrees of freedom
## Multiple R-squared:  0.262,    Adjusted R-squared:  0.244
## F-statistic: 14.4 on 12 and 487 DF,  p-value: <2e-16
```

```
anova(clothing_model)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: clothing_expenditure
```

##	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## income	1	7.13e+07	7.13e+07	33.11	1.5e-08
## sex	1	1.25e+07	1.25e+07	5.82	0.016
## food_expenditure	1	1.44e+08	1.44e+08	66.90	2.5e-15
## recreation_expenditure	1	1.06e+08	1.06e+08	49.35	7.3e-12
## miscellaneous_expenditure	1	2.93e+06	2.93e+06	1.36	0.244
## marital_status	2	1.36e+07	6.79e+06	3.16	0.044
## type_of_dwelling	5	2.17e+07	4.34e+06	2.02	0.075
## Residuals	487	1.05e+09	2.15e+06		

Recap: Section 9.1

After Section 9.1, you should be able to

- Understand the model-building process: data collection, variable selection, model selection, model validation

Learning Objectives for Sections 9.3-9.5

After Sections 9.3-9.5, you should be able to

- Apply appropriate criteria to perform data-based variable selection
- Understand automatic variable selection methods
- Understand the difficulty (or, perhaps, futility) of attempting to automatically identify a "best" set of variables in exploratory model building.

9.3 Criteria for Model Selection

From any set of $p - 1$ predictors, 2^{p-1} alternative sets of included variables can be constructed.

Model selection procedures, also known as subset selection or variables selection procedures, have been developed to identify a small group of regression models that are "good" according to a specified criterion. A detailed examination can then be made of a limited number of the more promising or "candidate" models, leading to the selection of the final regression model to be employed. This limited number might consist of three to six "good" subsets according to the criteria specified, so the investigator can then carefully study these regression models for choosing the final model.

Criterion	Calculation	Evidence of a good subset of $p - 1$ predictors
R_p^2	$1 - \frac{SSE_p}{SSTO}$	Bigger is better (note: $\max\{R_p^2\}$ always increases as p increases)
$R_{a,p}^2$	$1 - \left(\frac{n-1}{n-p}\right) \frac{SSE_p}{SSTO}$ $= 1 - \frac{MSE_p}{SSTO/(n-1)}$	Bigger is better
Mallows' C_p	$\frac{SSE_p}{MSE(X_1, \dots, X_{p-1})} - (n - 2p)$	Small values; values near p
AIC_p	$n \ln SSE_p - n \ln n + 2p$	Small values (equivalent to C_p for linear regression)
SBC_p (or BIC_p)	$n \ln SSE_p - n \ln n + (\ln n)p$	Small values (note: $\ln n > 2$ for all $n \geq 8$)
$PRESS_p$	$\sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$	Small values (note: $PRESS_p$ is a leave-one-out estimate of SSE , which means it is less impacted by overfitting.)

- $p - 1$ is the number of predictors included in the model out of the available $P - 1$ variables. We will assume that $n \gg P$.

```
library(leaps)
```

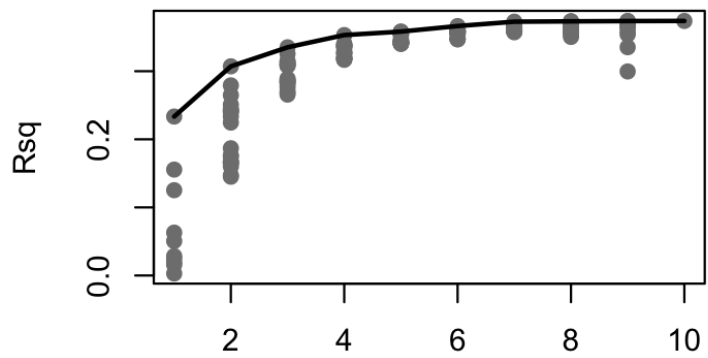
```
spending = with(spending_subset, data.frame(clothing_transformed=I(cloth
```

```
model_selection = summary(regsubsets(clothing_transformed~., data=spend-
```

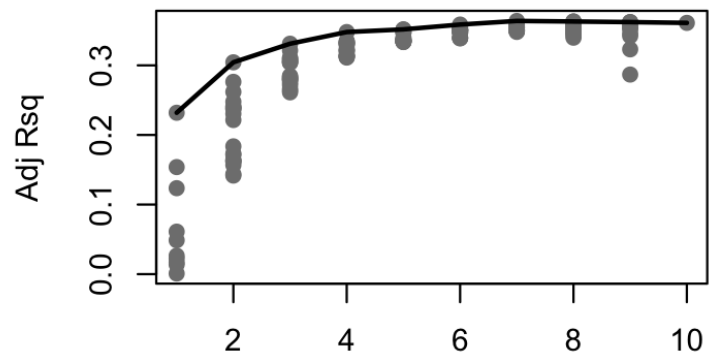
```
model_selection_10 = summary(regsubsets(clothing_transformed~., data=spe
```

Show entriesSearch:

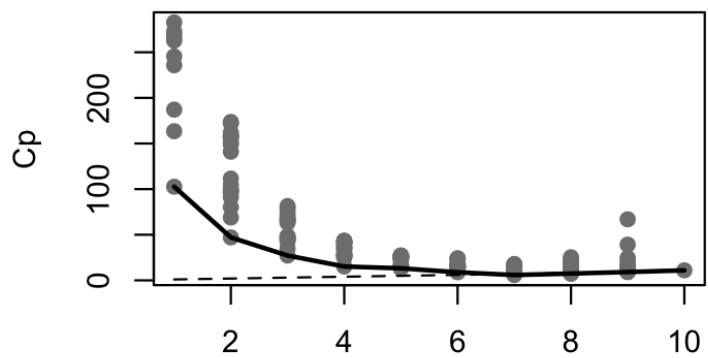
	p	income	sexmale	food	rec	misc	trans	care	alcohol
5	1	0	0	0	0	0	0	0	
7	1	0	0	0	0	0	0	0	
9	1	0	0	0	0	0	0	0	
1	1	0	0	0	0	0	0	1	
8	1	0	0	0	0	0	1	0	
6	1	0	0	0	0	1	0	0	
2	1	0	0	0	1	0	0	0	
3	1	0	0	1	0	0	0	0	
10	1	0	1	0	0	0	0	0	
4	1	1	0	0	0	0	0	0	
14	2	0	0	0	0	0	0	1	
18	2	0	0	0	0	0	0	1	
16	2	0	0	0	0	0	0	1	



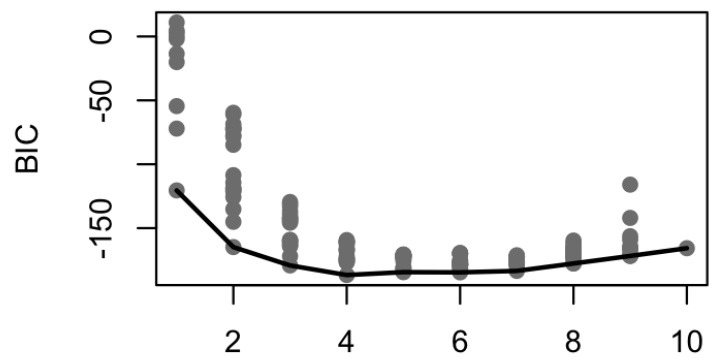
(a)



(b)



(c)



(d)

9.4 Automatic Search Procedures for Model Selection

As noted in the previous section, the number of possible models, 2^{P-1} , grows rapidly with the number of predictors. Evaluating all of the possible alternatives can be a daunting endeavor. To simplify the task, a variety of automatic computer-search procedures have been developed.

"Best" Subsets Algorithms

Time-saving algorithms have been developed in which the best subsets according to a specified criterion are identified without requiring the fitting of all of the possible subset regression models.

When the pool of potential X variables contains 30 to 40 or even more variables, use of a "best" subsets algorithm may not be feasible. An automatic search procedure that develops the "best" subset of X variables sequentially may then be helpful.

Stepwise Regression Methods

Forward stepwise regression:

1. Choose the best single predictor (the X variable with the largest t -value).
2. Out of the remaining predictors, add the most useful predictor to the existing model (the X variable with the largest t -value, given everything else that is already in the model) if it meets some prespecified threshold
3. Drop the least useful predictor from the current model if it fails to meet some prespecified threshold
4. Repeat steps 2 and 3 until no more variables can enter or leave the model.

Backward stepwise regression:

Start with the full model (include all variables), then proceed as in Forward stepwise regression, only backward...

- stepwise search procedures end with the identification of a single regression model as "best."
- "best" subsets algorithm can identify several "good" regression models for final consideration.

```
summary(regsubsets(clothing_transformed~., method="exhaustive", data=spe
```

```
##           income sexmale food rec misc trans care alcohol weeks married
## 1  ( 1 )  " "      " "      " "  " "  " "  " "  "*"  " "      " "
## 2  ( 1 )  " "      " "      " "  "*"  " "  " "  " "  "*"  " "      " "
## 3  ( 1 )  " "      " "      "*"  "*"  " "  " "  "*"  " "      " "
## 4  ( 1 )  "*"      " "      "*"  "*"  " "  " "  "*"  " "      " "
## 5  ( 1 )  "*"      " "      "*"  "*"  " "  " "  "*"  "*"      " "
## 6  ( 1 )  "*"      "*"      "*"  "*"  " "  " "  "*"  " "      "*"
## 7  ( 1 )  "*"      "*"      "*"  "*"  " "  " "  "*"  "*"      "*"
## 8  ( 1 )  "*"      "*"      "*"  "*"  " "  "*"  "*"  "*"      "*"
## 9  ( 1 )  "*"      "*"      "*"  "*"  "*"  "*"  "*"  "*"      "*"
## 10 ( 1 )  "*"      "*"      "*"  "*"  "*"  "*"  "*"  "*"      "*"

```

Since the algorithm returns a best model of each size, the results do not depend on a penalty model for model size: it doesn't make any difference whether you want to use AIC, BIC, CIC, DIC, ...

```
summary(regsubsets(clothing_transformed~., method="forward", data=spend-
```

##		income	sexmale	food	rec	misc	trans	care	alcohol	weeks	married
## 1	(1)	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "
## 2	(1)	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	" "
## 3	(1)	" "	" "	"*"	"*"	" "	" "	"*"	" "	" "	" "
## 4	(1)	"*"	" "	"*"	"*"	" "	" "	"*"	" "	" "	" "
## 5	(1)	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	" "	" "
## 6	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	"*"	" "	" "
## 7	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	"*"	" "	"*"
## 8	(1)	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	" "	"*"
## 9	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
## 10	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

```
summary(regsubsets(clothing_transformed~., method="backward", data=spend
```

##		income	sexmale	food	rec	misc	trans	care	alcohol	weeks	married
## 1	(1)	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "
## 2	(1)	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	" "
## 3	(1)	" "	" "	"*"	"*"	" "	" "	"*"	" "	" "	" "
## 4	(1)	"*"	" "	"*"	"*"	" "	" "	"*"	" "	" "	" "
## 5	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	" "	" "	" "
## 6	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	" "	" "	"*"
## 7	(1)	"*"	"*"	"*"	"*"	" "	" "	"*"	"*"	" "	"*"
## 8	(1)	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	" "	"*"
## 9	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
## 10	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

9.5: Some Final Comments on Automatic Model Selection Procedures

- no automatic search procedure will always find the "best" model
 - indeed, there may exist several "good" regression models whose appropriateness for the purpose at hand needs to be investigated.
- Subject-specific expertise and judgment needs to play an important role in model building for exploratory studies.
- When a qualitative predictor variable is represented in the pool of potential X variables by a number of indicator variables (e.g., province is represented by several indicator variables), it is often appropriate to keep these indicator variables together as a group to represent the qualitative variable, even if a subset containing only some of the indicator variables is "better" according to the criterion employed.
- Similarly, if second-order terms X^2 or interaction terms need to be present in a regression model, one would ordinarily wish to have the first-order terms in the model as representing the main effects.

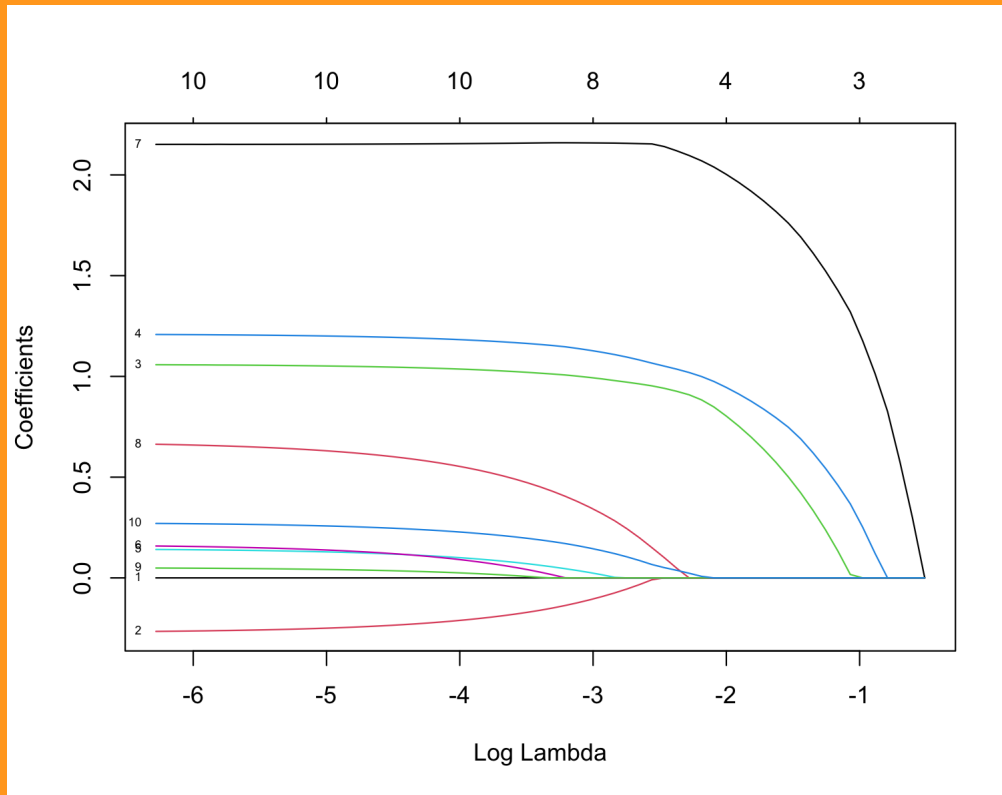
Penalized Regression

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

$$\beta_{lasso} = \operatorname{argmin}[SSE(\beta) + \lambda \cdot \|\beta\|_1],$$

where the L_1 -norm $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ measures the *complexity* of the model.

So, instead of only seeking to minimize SSE (as the least squares estimator does), the *lasso* penalizes models that are too complex. I.e., the lasso errs toward models with parameters set to zero (so models that exclude lots of not-overly-important predictor variables); this can help avoid overfitting (and can act as a variable selection tool).



```
coef(fit.lasso, s=exp(c(-.5, -1, -2, -3, -Inf))) %>% round(5)
```

```
## 11 x 5 sparse Matrix of class "dgCMatrix"
##           s1      s2      s3      s4      s5
## (Intercept) 6.722 6.29851 5.38710 4.92781 4.60921
## income      .      .      0.00000 0.00001 0.00001
## sexmale     .      .      .      -0.10465 -0.26551
## food        .      0.00000 0.00005 0.00006 0.00006
## rec         .      0.00003 0.00010 0.00012 0.00013
## misc        .      .      .      0.00001 0.00005
## trans       .      .      .      .      0.00000
## care        .      0.00030 0.00050 0.00054 0.00054
## alcohol     .      .      .      0.00002 0.00005
## weeks       .      .      .      .      0.00136
## married     .      .      .      0.14461 0.27041
```

```
msummary(lm(clothing_transformed~., data=spending))
```

```
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.60e+00   2.10e-01  21.90  < 2e-16
## income       1.15e-05   2.91e-06   3.94  9.4e-05
## sexmale     -2.72e-01   9.88e-02  -2.75  0.00613
## food        6.39e-05   1.79e-05   3.57  0.00039
## rec         1.28e-04   2.33e-05   5.49  6.4e-08
## misc        5.44e-05   9.45e-05   0.58  0.56521
## trans       4.20e-06   6.49e-06   0.65  0.51765
## care        5.39e-04   7.08e-05   7.61  1.4e-13
```

Recap Sections 9.3-9.5

After Sections 9.3-9.5, you should be able to

- Apply appropriate criteria to perform data-based variable selection
- Understand automatic variable selection methods
- Understand the difficulty (or, perhaps, futility) of attempting to automatically identify a "best" set of variables in exploratory model building.

Learning Objectives for Section 9.6

After Section 9.6, you should be able to

- Understand the need for model validation.
- Understand how to validate models using replicate studies and data splitting
- Assess model validity given appropriate output.

9.6: Model Validation

The final step in the model-building process is the validation of the selected regression models.

Model validation usually involves checking a candidate model against independent data.

Three basic ways of validating a regression model are:

1. Collection of new data to check the model and its predictive ability.
2. Comparison of results with theoretical expectations, earlier empirical results, and simulation results.
3. Use of a holdout sample to check the model and its predictive ability.

With recent computational advances, **bootstrapping** has also become an appealing modification to using a holdout sample.

Collection of New Data to Check Model

One validation method is to re-estimate the model form chosen earlier using the new data.

The estimated regression coefficients and various characteristics of the fitted model are then compared for consistency to those of the regression model based on the earlier data.

If the results are consistent, they provide strong support that the chosen regression model is applicable under broader circumstances than those related to the original data.

I do not understand the difference between the validation step with new data for an exploratory observational study versus conducting a confirmatory observational study instead.

A second validation method is designed to calibrate the predictive capability of the selected regression model.

When a regression model is developed from given data, it is inevitable that the selected model is chosen, at least in large part, because it fits well the data at hand (i.e., due to sampling noise). For a different set of random outcomes, one may likely have arrived at a different model.

A result of this model development process is that the error mean square MSE will tend to understate the inherent variability in making future predictions from the selected model.

A means of measuring the actual predictive capability of the selected regression model is to use this model to predict each case in the new data set and then to calculate the mean of the squared prediction errors, to be denoted by MSPR, which stands for **mean squared prediction error**:

$$MSPR = \frac{\sum_{i^*=1}^{n^*} (Y_{i^*} - \hat{Y}_{i^*})^2}{n^*}$$

where

- Y_{i^*} is the value of the response variable in the i^* th validation case
- \hat{Y}_{i^*} is the predicted value for the i^* th validation case based on the model-building dataset
- n^* is the number of cases in the validation data set

Recall that

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - p}$$

- MSE is measured on the data set used to develop the prediction model
- MSPR measures how well the model works for predictions in a *different* data set

Comparison with Theory, Empirical Evidence, or Simulation Results

Comparisons of regression coefficients and predictions with theoretical expectations, previous empirical results, or simulation results should be made.

Unfortunately, there is often little theory that can be used to validate regression models.

Data Splitting

By far the preferred method to validate a regression model is through the collection of new data. Often, however, this is neither practical nor feasible.

An alternative when the data set is large enough is to split the data into two sets.

The first set, called the model-building set or the **training sample**, is used to develop the model.

The second data set, called the **validation** or prediction set, is used to evaluate the reasonableness and predictive ability of the selected model. Data splitting in effect is an attempt to simulate replication of the study.

There are difficult decisions to be made when splitting the data:

- How big should the training sample be? Alternately, how big should the validation sample be?
- Should we split the data completely randomly, or should we try to somehow balance the distribution of predictors across the two samples?
- Should we even have a validation set or should we use all of our data in developing the best model possible?

Using a single validation sample is usually insufficient; we still have sample noise in the validation sample.

In addition, splitting the data causes us to perform variable selection on a smaller data set, which increases the problems associated with overfitting (choosing variables based on sample noise, rather than any true underlying trends).

Often, the validation process will be extended to *K-fold cross-validation*:

- the data are split into K parts
- the first $K-1$ parts are treated as the training sample, with the K th part being the validation sample
- this process is repeated with each of the K parts being treated as the validation sample once

This gives us a sense of how *optimistic* MSE tends to be for that particular model (i.e., the average difference between MSE and MSPR). We don't want to base *model selection* on overly-optimistic MSEs.

This idea can be (and probably *should be*) taken further using *bootstrapping* (sampling with replacement from within our sample).

In the end, we might want to fit an overall model (using all of the data), and report MSE along with some measure of how optimistic it may be.

Split the data into training and test samples

```
library(tidyverse)
library(caret)

set.seed(123)
training.samples <- spending$clothing_transformed %>% createDataPartiti

train.data <- spending[training.samples, ]
test.data <- spending[-training.samples, ]
```

Build the model using the training data

```
model <- lm(clothing_transformed ~ income+sex+food+rec+care+married, dat
msummary(model)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.70e+00   1.78e-01  26.48  < 2e-16
## income      1.17e-05   2.99e-06   3.91  0.00011
## sexmale     -2.59e-01   1.09e-01  -2.39  0.01749
## food        7.27e-05   1.93e-05   3.76  0.00020
## rec         1.47e-04   2.61e-05   5.61  3.7e-08
## care        5.15e-04   7.89e-05   6.53  2.0e-10
## married     3.15e-01   1.19e-01   2.65  0.00839
##
## Residual standard error: 0.989 on 394 degrees of freedom
## Multiple R-squared:  0.374,    Adjusted R-squared:  0.365
```

Make predictions and compute the R^2 and MSPR on the TRAINING data

```
predictions <- model %>% predict(train.data)
data.frame(
  SSE    = anova(model)["Residuals", "Sum Sq"],
  PRESS  = PRESS(model),
  MSE    = anova(model)["Residuals", "Mean Sq"],
  MSPR   = RMSE(predictions, train.data$clothing_transformed)^2,
  R2     = R2(predictions, train.data$clothing_transformed))
```

```
##      SSE PRESS    MSE    MSPR    R2
## 1 385.7 400.7 0.9789 0.9618 0.3742
```

Make predictions and compute the R^2 and MSPR on the TEST data

```
predictions <- model %>% predict(test.data)
data.frame(
  MSPR   = RMSE(predictions, test.data$clothing_transformed)^2,
  R2     = R2(predictions, test.data$clothing_transformed))
```

```
##      MSPR    R2
## 1 1.045 0.3439
```

Now consider the building of a very complex model (one that is more likely to be prone to overfitting)

```
model <- lm(clothing_transformed ~ .^2, data = train.data)
msummary(model)
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	2.98e+00	6.78e-01	4.39	1.5e-05
## income	4.97e-05	1.87e-05	2.65	0.0083
## sexmale	-5.03e-01	5.37e-01	-0.94	0.3493
## food	1.71e-04	9.76e-05	1.75	0.0811
## rec	4.51e-05	1.70e-04	0.27	0.7911
## misc	7.70e-04	5.46e-04	1.41	0.1589
## trans	1.13e-05	4.41e-05	0.26	0.7979
## care	7.46e-04	4.83e-04	1.54	0.1236
## alcohol	-1.12e-04	1.39e-04	-0.81	0.4206
## weeks	3.37e-02	1.50e-02	2.25	0.0250
## married	-3.56e-02	5.90e-01	-0.06	0.9519
## income:sexmale	-1.24e-05	8.32e-06	-1.49	0.1379
## income:food	5.36e-10	1.61e-09	0.33	0.7394
## income:rec	2.82e-09	2.02e-09	1.40	0.1636
## income:misc	-4.68e-09	7.14e-09	-0.66	0.5128
## income:trans	3.10e-10	6.95e-10	0.45	0.6556
## income:care	-9.43e-09	5.99e-09	-1.57	0.1162
## income:alcohol	1.10e-09	1.71e-09	0.64	0.5216
## income:weeks	-6.89e-07	3.18e-07	-2.17	0.0308
## income:married	-5.95e-06	9.07e-06	-0.66	0.5123

Make predictions and compute the R^2 and MSPR on the TRAINING data

```
predictions <- model %>% predict(train.data)
data.frame(
  SSE    = anova(model)["Residuals", "Sum Sq"],
  PRESS  = PRESS(model),
  MSE    = anova(model)["Residuals", "Mean Sq"],
  MSPR   = RMSE(predictions, train.data$clothing_transformed)^2,
  R2     = R2(predictions, train.data$clothing_transformed))
```

```
##      SSE PRESS      MSE      MSPR      R2
## 1 321 467.3 0.9303 0.8004 0.4792
```

Make predictions and compute the R^2 and MSPR on the TEST data

```
predictions <- model %>% predict(test.data)
data.frame(
  MSPR = RMSE(predictions, test.data$clothing_transformed)^2,
  R2    = R2(predictions, test.data$clothing_transformed))
```

```
##      MSPR      R2
## 1 1.422 0.1904
```

- Which model is preferable? The complex model or the simpler model?

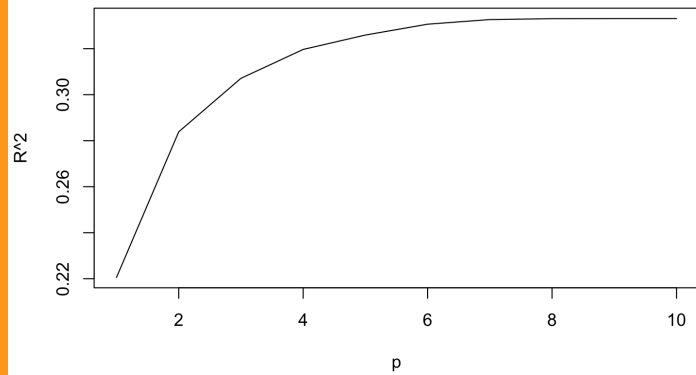
Warm-up Exercises

```
library(leaps)
```

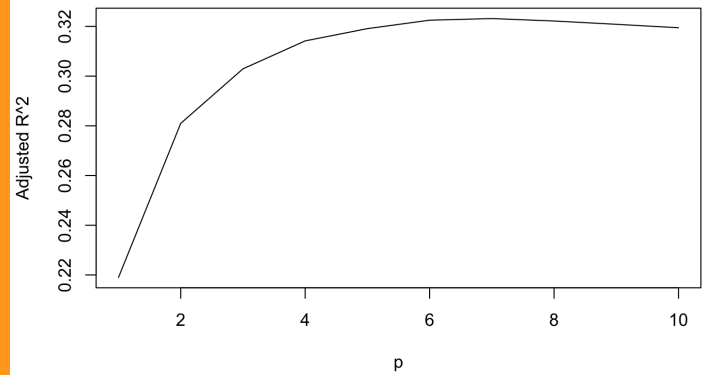
```
spending = with(spending_subset, data.frame(clothing=clothing_expenditure,
```

```
model_selection = summary(regsubsets(clothing~., data=spending, nvmax=100))
```

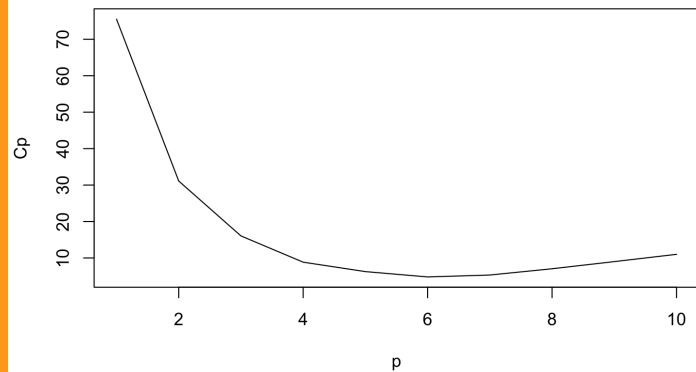
```
plot(model_selection$rsq, type='l')
```



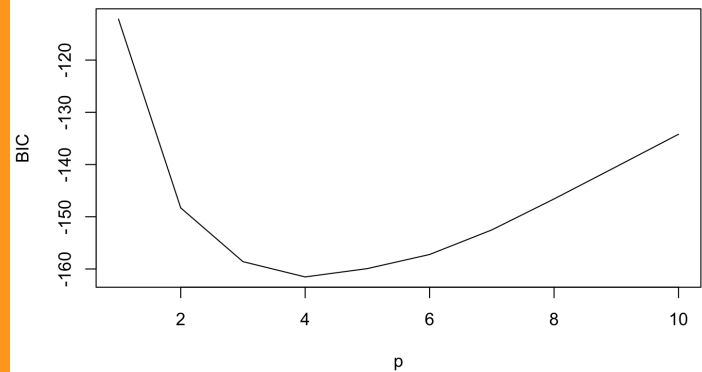
```
plot(model_selection$adjr2, type='l')
```



```
plot(model_selection$cp, type='l')
```



```
plot(model_selection$bic, type='l')
```



```
model_selection$outmat
```

```
##           income sexmale food rec misc trans care alcohol weeks married
## 1  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 2  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 3  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 4  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 5  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 6  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 7  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 8  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 9  ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
## 10 ( 1 )  " "      " "      " "  " "  " "  " "  " "      " "      " "
```

Recap Section 9.6

After Section 9.6, you should be able to

- Understand the need for model validation.
- Understand how to validate models using replicate studies and data splitting
- Assess model validity given appropriate output.