
Lecture 14: Selecting the Best Model

Reading Assignment:

- Muller and Fetterman, Chapter 12: “Selecting the Best Model”
(Required)

Selecting the best model, while perhaps one of the most common data analysis tasks, is an **exploratory** activity that is usually accompanied by type I error rate inflation (every time we compare two models, we compromise the type I error rate of the model eventually chosen). A **confirmatory** analysis requires specifying the variables, model, and tests without knowledge of the data at hand. (This is typically the strategy used for FDA drug approval, a situation in which a great deal is known about the action of the drug through clinical trials and studies of related drugs so that such a model may be specified in advance.)

Often, investigators have many potential predictors and wish to find the best subset, and in such situations, it is important to bear in mind that p-values from hypothesis tests must be interpreted with caution in an exploratory analysis. (Alternatively, one may wish to use a correction such as the Bonferroni correction for all hypothesis tests conducted.)

Model Selection Strategy Overview

Assuming that our goal is to use purely exploratory analysis to find the best-fitting model for the data at hand, our model selection strategy consists of four steps.


1. Specify the maximum model under consideration.
2. Specify a criterion for model selection.
3. Specify a strategy for applying the criterion.
4. Conduct the analysis.

In addition, if we wish to have confidence in our model for data outside the current sample, we should evaluate the reliability of the model chosen. In a purely exploratory analysis, one often simply interprets significance of coefficients in the final model with caution, refraining from using $p \leq 0.05$ after conducting a series of tests on the data in order to choose that final model. (The practice of doing this anyway is widespread.) However, other methods may be used to assess

model reliability more formally, and we will discuss two of these, split-sample analysis and cross-validation more generally, in detail.

Specifying the Maximum Model

We will use *maximum model* or *full model* to refer to the model with the greatest number of predictors of any model. In model selection, we will consider more parsimonious models than this model. In some cases, models under consideration will be *nested* in the full model, which means that they can be created simply by deleting variables from the full model. In other cases, smaller models may not be nested in the maximum model. A common goal is to find a more parsimonious model that still describes the data (almost) as well as the full model.

When selecting the full model, one must be sure not to select a model that is too large. As a bare minimum, we need $n - p > 0$, where p is the number of columns in \mathbf{X} . In order to have fairly stable estimates, we would want $n > 5p$ or $n > 10p$. 

Consider a model with 5 covariates, x_1, x_2, x_3, x_4, x_5 , and an intercept. Several possible “full models” are listed below.


$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \quad (1)$$

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_2^2 + \beta_7 x_3^2 \quad (2)$$

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 \quad (3)$$

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5 + \beta_{10} x_1 x_2^2 + \beta_{11} x_1 x_3^2 \quad (4)$$

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 \log(x_5) \quad (5)$$

Suppose that we select Model 2 as our best model. Model 1 would be  a candidate model that is *nested* in Model 2, while Model 5 is a *non-nested* candidate model.

Specifying a Model Selection Criterion

Next, we need to choose a criterion to help us define which model is “best”. Some criteria are used to evaluate models with respect to a null or intercept-only model, while other criteria are used to evaluate models with respect to the full model or any larger model (this latter class of criteria is restricted to models *nested* in the full model).

Comparing a Candidate Model to the Null or Intercept Model

Adjusted R^2

One intuitive model selection criteria is R^2 . Recall that

$$R^2 = \frac{SSE(\beta_0) - SSE(\beta_0, \dots, \beta_{p-1})}{SSE(\beta_0)}$$

and provides us with an estimate of the percent of variability explained by the model under consideration. One problem with R^2 is that it never decreases when additional variables are added to a model, so that it will always favor selection of the largest model. An *adjusted* R^2 is given by

$$R_A^2 = 1 - \frac{n-1}{n-p}(1 - R^2),$$

where p is the number of columns of \mathbf{X} . This adjusts the usual R^2 for the number of covariates, adding a penalty for models with “too many” covariates. Typically, the model with the largest R_A^2 is said to be the best model.

F Tests

The test of corrected overall regression, given by

$$F_p = \frac{[SSE(\beta_0) - SSE(\beta_0, \dots, \beta_p)] / (p - 1)}{MSE(\beta_0, \dots, \beta_p)}, \quad \text{🗨️}$$

may be used to test whether the model under consideration offers significant improvement over the intercept-only model. For models without an intercept, the test of uncorrected overall regression may be used to test whether the model under consideration offers significant improvement over the null model.

Single Model Criteria: AIC and SBC/BIC

The Akaike Information Criterion (AIC) and the Schwarz Criterion or Bayesian Information Criterion (SBC in SAS and BIC many other places) are general metrics. In linear regression,

$$AIC = n \log \left(\frac{SSE}{n} \right) + 2p$$

$$SBC = n \log \left(\frac{SSE}{n} \right) + \log(n)p. \quad \text{🗨️}$$


For both criteria, **smaller is better.** Increasing the SSE increases the AIC and SBC, while increasing the number of predictors also increases each measure. AIC tends to favor models that are too large, while the SBC places a greater penalty on larger models. These criteria allow comparison of nested and non-nested models. NOTE: In older versions of SAS, some procedures used “bigger is better” definitions of AIC and SBC, so take care if you are using SAS versions before 8.1.

Comparing a Candidate Model to the Full Model

Mallows C_p

Mallows C_p compares a candidate model to the full model and is given by

$$C_p = \frac{SSE(\beta_0, \dots, \beta_{p-1})}{MSE(\text{full})} - n + 2p.$$

Based on the rationale that a model not omitting useful variables should have $SSE(\beta_0, \dots, \beta_{p-1}) \approx (n - p)\sigma^2$, we seek a model with $C_p \approx p$. 

F Tests

A groupwise test of a candidate model versus the full model, given by

$$F_p = \frac{[SSE(\beta_0, \dots, \beta_{p-1}) - SSE(\text{full})] / [df(\text{candidate}) - df(\text{full})]}{MSE(\text{full})},$$

may be used to test whether the model under consideration is sufficiently close to the reduced model, provided that the model under consideration is nested in the full model. (Note that this test may be used to test the adequacy of any model nested in a larger model.)


Specifying the Selection Strategy

All Possible Regressions

Suppose our full model contains p predictors. An *all possible regressions* strategy involves fitting all possible models. In this case, each of the p predictors (including the intercept) could be in or out of the model, so that we would consider 2^p models in our selection process (including a null model). This strategy may get out of hand very quickly, since there are $2^5 = 32$ possible models with 5 columns in \mathbf{X} , $2^8 = 256$ possible models with 8 columns in \mathbf{X} , and $2^{10} = 1024$ possible models with 10 columns in \mathbf{X} . Although SAS will print the “top” models from all possible ones with using some criteria (R^2 for example), the time required for an all possible regressions strategy for other criteria (such as F tests against a maximum model) make this strategy infeasible when p is large.

Backward Elimination

Backward elimination begins with the full model and deletes variables with little value. The procedure is described below.

1. Specify the full model and set $p = p_{full}$, the number of columns of the \mathbf{X} matrix in the full model.
2. Fit all $p - 1$ variable models defined by deleting a single variable from the base model. (Typically, we do not consider eliminating the intercept.)
3. For each model, compute an added-last test for the candidate variable.
4. Find the minimum F statistic out of the set of $p - 1$ models (maximum p-value).
 - (a) If the corresponding variable is “significant” at a specified level  (typically ranging from 0.05 to 0.20), stop and select the model with p predictors as the best model.

-
- (b) If the corresponding variable is not significant, delete the variable in question and set $p = p - 1$.

5. Repeat until a model is chosen.

Forward Selection

Forward selection begins with the intercept-only (or null) model and adds variables with predictive value. The procedure is described below.

1. Fit the intercept-only (or null) model as the base model so that $p = 1$.
2. Fit all $p + 1$ variable models defined by adding a single variable to the base model.
3. For each model, compute an added-in-order test for the candidate variable.
4. Find the maximum F statistic out of the set of models (minimum p-value).
 - (a) If the corresponding variable is “significant” at a specified level

(typically ranging from 0.05 to 0.20), add the variable in question and set $p = p + 1$.

(b) If the corresponding variable is not significant, stop and choose the model with p predictors.

5. Repeat until a model is chosen.

Stepwise Selection

Strictly speaking, *stepwise selection* refers to a selection procedure with both forward and backward steps so that addition and deletion of variables may be considered. However, this term is sometimes loosely used to refer to forward selection or backward elimination as well.

SAS implementation: The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the statistic for a variable to be added must be significant at the $SLENTY=$ level. After a variable is added, however, the stepwise method looks at all



the variables already included in the model and deletes any variable that does not produce a statistic significant at the $SLSTAY =$ level. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model.

Note: It is important to note that these three strategies do not consider all possible models and might actually miss a "best" model. In addition, it is entirely possible that they may select three different models as "best" models. Famous quote from George Box: all models are wrong, but some are useful.




```
data a; do i = 1 to 500;
x1 = 10 + 5*rannor(0); * Normal(10, 25);
x2 = exp(3*rannor(0)); * lognormal;
x3 = 5+10*ranuni(0); * uniform;
x4 = 100 + 50*rannor(0); * Normal(100, 2500);
x5 = x1 + 3*rannor(0); * normal bimodal;
x6 = 2*x2 + ranexp(0);
* lognormal and exponential mixture;
x7 = 0.5*exp(4*rannor(0)); * lognormal;
x8 = 10 +8*ranuni(0); * uniform;
x9 = x2 + x8 + 2*rannor(0);
* lognormal, uniform and normal mix
x10 = 200 +90*rannor(0); * normal(200, 8100);
*x10 = 5*x2 + rannor(0);
y = 3*x2 - 4*x8 + 5*x9 + 3*rannor(0);
* true model with no intercept term;
output; end;
```



```
/*run all possible models*/  
proc reg data=a outest=est;  
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /  
selection=adjrsq sse aic ;  
output out=out p=p r=r; run; quit;
```

```
/*run all possible models without intercept*/  
proc reg data=a outest=est0;  
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /  
noint selection=adjrsq sse aic ;  
output out=out0 p=p r=r; run; quit;
```

```
/*forward selection*/   
proc reg data=a outest=est1;  
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /  
slentry=0.15 selection=forward  
ss2 sse aic;
```

```
output out=out1 p=p r=r; run; quit;
```

```
/*backward selection*/
```

```
proc reg data=a outest=est2;
```

```
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
```

```
slstay=0.15 selection=backward
```

```
ss2 sse aic;
```

```
output out=out1 p=p r=r; run; quit;
```

```
/*stepwise selection*/ 
```

```
proc reg data=a outest=est3;
```

```
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
```

```
slstay=0.15 slentry=0.15 selection=stepwise
```

```
ss2 sse aic;
```

```
output out=out3 p=p r=r; run; quit;
```

```
/*stepwise group selection*/
```

```
proc reg data=a outest=est4;  
model y={x1 x2} x3 x4 x5 x6 x7 x8 x9 x10 /  
selection=stepwise slstay=0.15 slentry=0.15  
groupnames=x1 x2 x3 x4 x5 x6  
x7 x8 x9 x10;
```

Comments

1. In any selection strategy, one must take steps to ensure that nonsensical models do not emerge as “winners”. For example, if x , x^2 , and x^3 are in the full model, then we do not wish to allow a backward selection strategy to delete x while letting x^2 and x^3 remain in the model. In such cases, we will define variable “groups” to be tested as units. In this case, a logical strategy would be to treat (x, x^2, x^3) as a group. In backward elimination, we would first test the group. If it was significant, we might then move to test x^3 , and if significant, we would retain all three variables in the model.
2. The p-values obtained using a series of F tests should not be viewed strictly but merely as guides due to inflated type I error rates due to multiple testing.
3. Ideally, the best error term is the one from the full model because it should have the least bias. In forward selection in particular, the

estimate of the error term may be biased upwards, leading the process to select a final model that is too small.

4. A group-wise, backward selection strategy based on F tests is superior in a wide variety of settings.



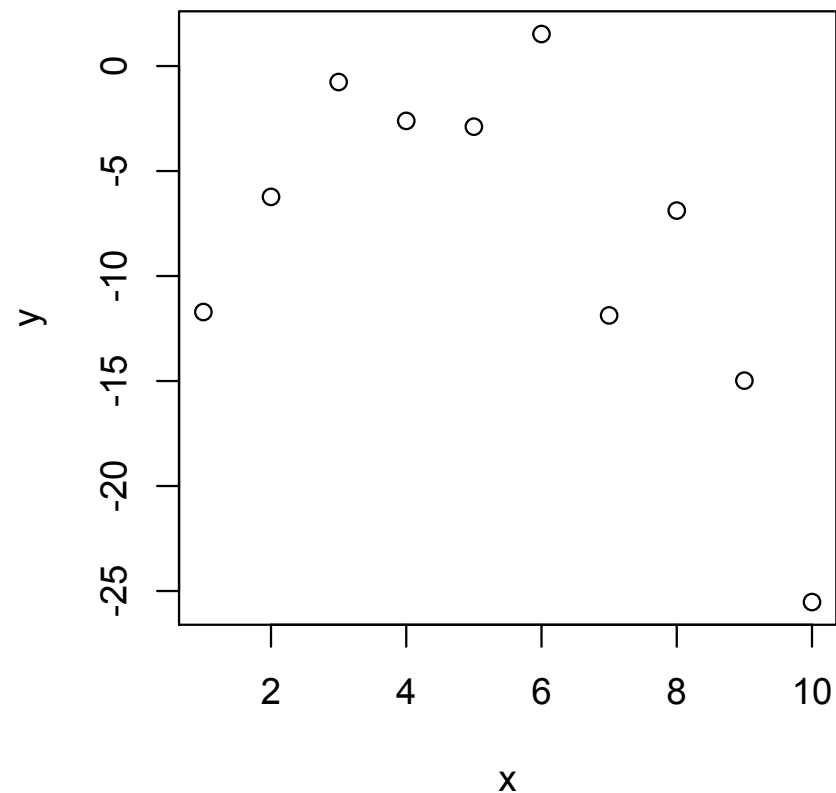
Fixed Tests Model Selection Strategy

If one wishes to avoid “data dredging” and to better preserve the type I error rate, a fixed tests implementation should be considered. This strategy is outlined below.

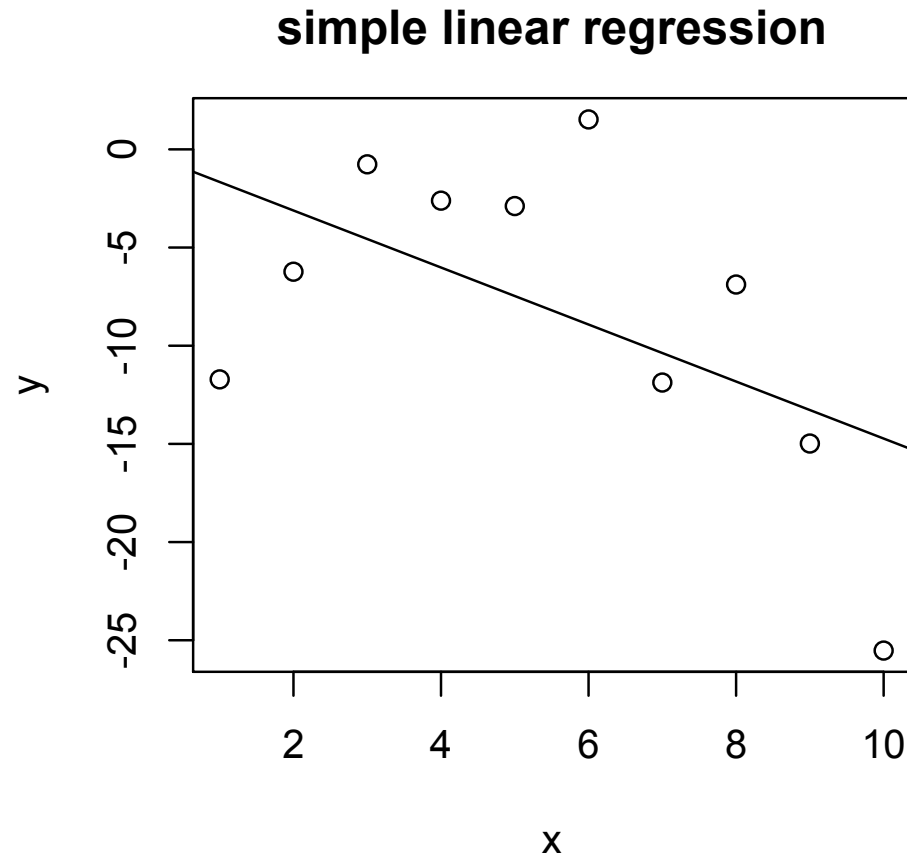
1. Very thoughtfully specify the maximum model.
2. Select groups and use a groupwise strategy, selecting a sequence for testing models and adjusting α for tests.
3. Assess diagnostics.
 - (a) Fit the full model and check for collinearity.
 - (b) Parsimoniously change or reduce the full model to eliminate collinearity.
 - (c) Conduct assumption diagnostics.
4. Conduct the *planned* tests.
5. Conduct assumption diagnostics on the resulting reduced model.

Assessing Reliability

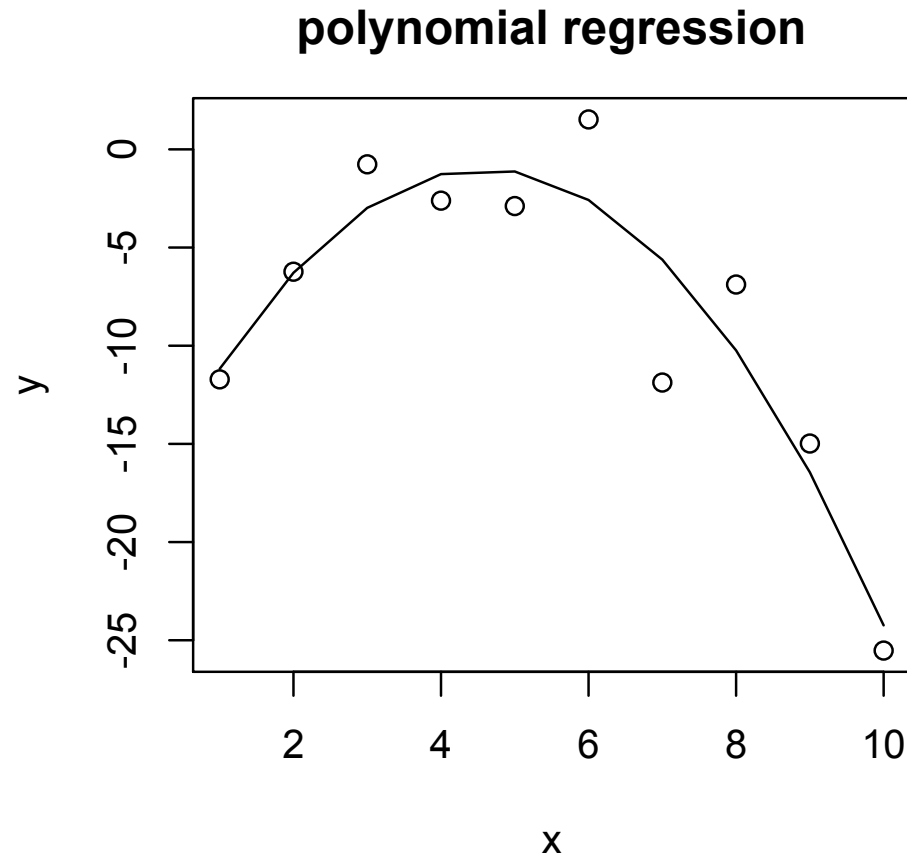
Why is assessing reliability an important step in model fitting?
Consider the following data.



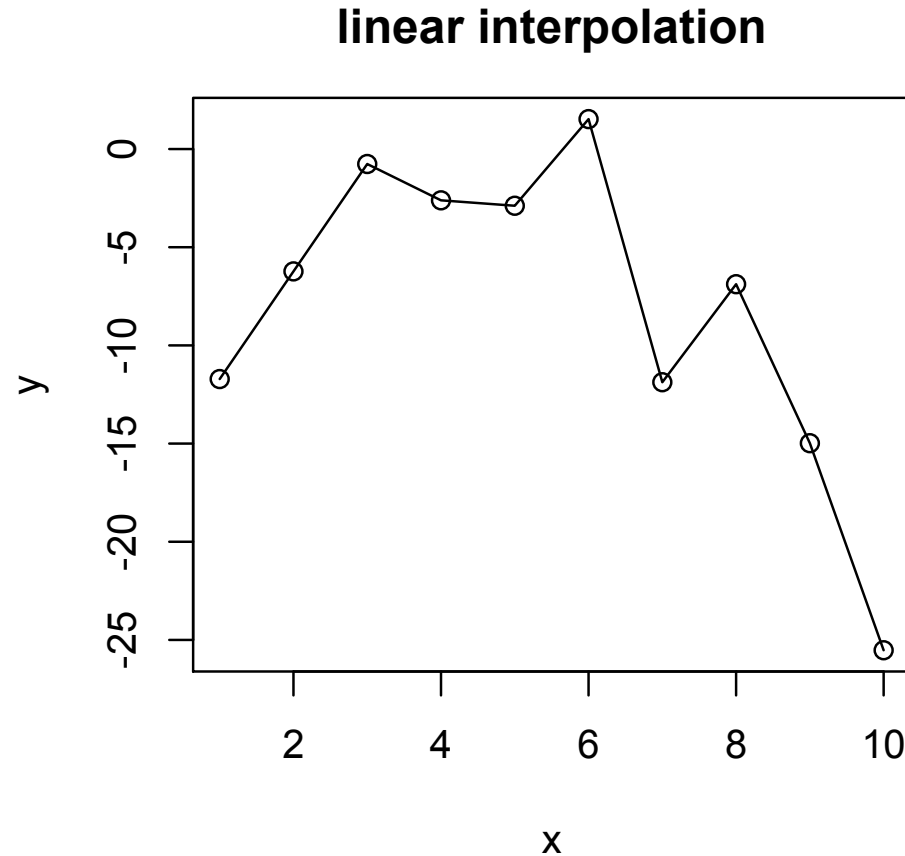
One possible model for the data is a simple linear regression, shown below.



Another approach is a regression model with linear and quadratic terms in x .



A third approach is a linear nonparametric approach (“connect the dots”).




Why not choose the model with the best fit to the data?

How well will the model predict future data from the same distribution?

Split-Sample Analysis

The simplest model validation method is one-time *data-splitting*. In this method, the data are split into *training* (model development) and *test* (model validation) samples by a random process. This splitting must occur before any analysis begins. Then, regression is performed on the training set, with future performance evaluated in the test set.

The split-sample procedure is outlined below.

1. Before looking at the data, split the sample using a random process (this process may be stratified by subgroups defined by values of \mathbf{X} such as smoking status or race). The split fraction  could be 50-50 (so that half the data fall into the training sample and half fall into the test sample). For large data sets, often a fraction as small as 10% could be placed in the training sample while still providing a sufficient sample size for exploratory data analysis. If the sample size is too small to allow a reasonable fraction in the training sample, the value of a split-sample

approach will be greatly diminished.

2. Conduct any and all desired exploratory analyses on the training data, including diagnostics.
3. Based on the exploratory analyses, select a “best” model based on selected criteria. The parameter estimates for this model will be called $\hat{\beta}_1$.
4. Compute the predicted values \hat{y} for this model and call them \hat{y}_1 .
5. Compute the squared multiple correlation, $R_1^2 = r^2(\mathbf{y}_1, \hat{\mathbf{y}}_1)$ for this model.
6. Compute the cross-validation correlation.
 - (a) Compute the predicted values in the test sample as $\hat{\mathbf{y}}_2 = \mathbf{X}_2 \hat{\beta}_1$. (That is, use the $\hat{\beta}$ from the training data with the covariates from the test data to compute predicted values.)
 - (b) Compute the squared cross-validation correlation as $R_{*2}^2 = r^2(\mathbf{y}_2, \hat{\mathbf{y}}_2)$.

-
- (c) Compute the estimated shrinkage on cross-validation as $R_1^2 - R_{*2}^2$. Smaller shrinkages are better. (One rule of thumb is that shrinkage < 0.05 indicates results are generalizable and that shrinkage > 0.10 is cause for concern.) Often, one computes the proportion relative shrinkage by dividing by R_1^2 .
- (d) The MSE may also be used to evaluate future performance.



-
1. Conduct regression diagnostics on the pooled data.
 - (a) If shrinkage is “small enough,” pool the data to provide best available estimates of β .
 - (b) If shrinkage is “poor,” pool data to conduct a second-round *exploratory* analysis.
 2. Report results honestly.

One big advantage of data-splitting is that hypothesis tests are confirmed in the test sample. In addition, this method is simple to implement. However, it has several disadvantages, including the following.

- Data-splitting greatly reduces the sample size for both model development and model testing.
- Different splits may lead to different results, and with small sample sizes, our test set may just be lucky or unlucky.
- Data-splitting does not validate the final model but rather a model

developed on only a subset of the data. The training and test sets are recombined for fitting the final model, which is not validated.

Cross-Validation

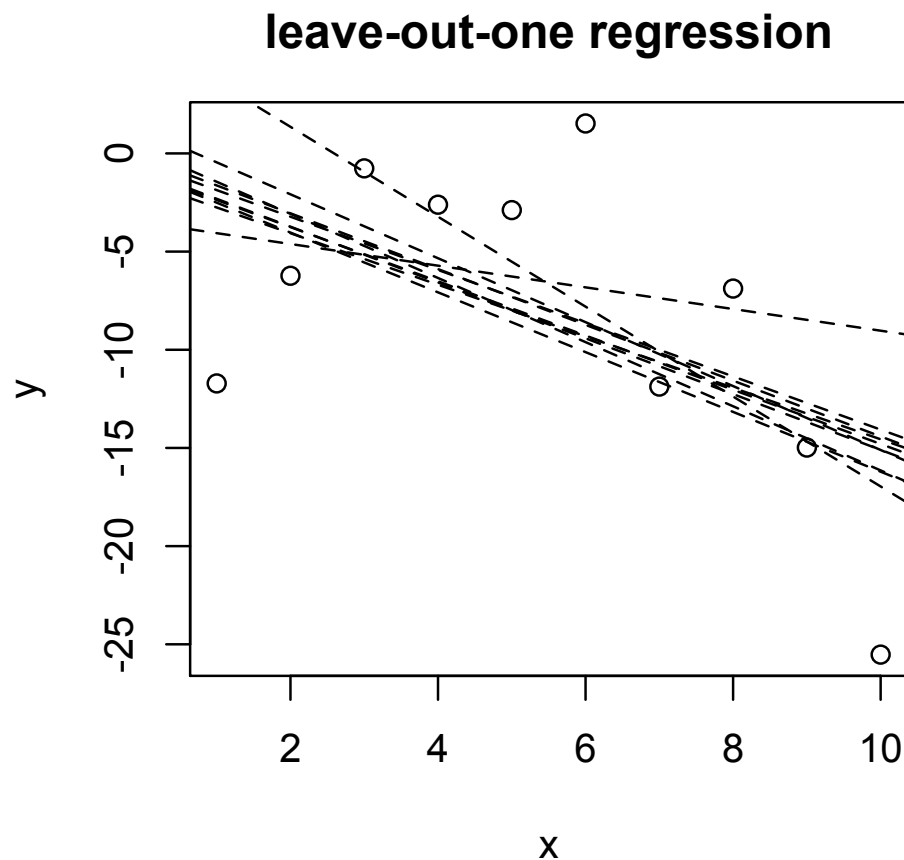
Cross-validation is a generalization of data-splitting that overcomes some of its disadvantages.

Leave-out-one cross-validation is similar to jackknifing. In leave-out-one cross-validation, one observation is omitted from the analysis, and the response for that one subject is predicted using a model obtained from the remaining $n - 1$ subjects. Note the error of that prediction. When you've done this n times (once for each subject), report the mean squared error over all the predictions. This method does not “waste” data, but is computationally expensive.





Consider leave-out-one cross-validation for the previous data. Omitting the j^{th} observation, $j = 1, \dots, 10$, we have the lines below for the linear models.



The errors are calculated as the distance from the fitted line omitting each subject to each subject's observed outcome. Summing these squared distances, we obtain SSE . The estimated error from the linear model can be compared to that of the quadratic and “connect-the-dots” models to see which model is preferred.

K-fold cross validation randomly breaks the dataset into k partitions. For each partition, train on the points not in the partition, and find the test-set errors for the partition of interest. Repeat for all k partitions, and then report the mean squared error over all the predictions. This type of cross-validation (often done for $k = 10$) is a compromise between the leave-out-one and the split sample approaches.

Final Comments

Model selection and validation tend to be personal topics with different investigators preferring a variety of different methods. The major consensus is that exploratory analyses must be treated with caution, and test results must be presented honestly (and not as “confirmatory” when extensive model exploration has been carried out).

Frank Harrells comments:

Here are some of the problems with stepwise variable selection.

- It yields R-squared values that are badly biased to be high.
- The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
- The method yields confidence intervals for effects and predicted values that are falsely narrow.
- It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.

-
- It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large).
 - It has severe problems in the presence of collinearity.
 - It is based on methods (e.g., F tests for nested models) that were intended to be used to test prespecified hypotheses.
 - Increasing the sample size doesn't help very much.

For Post-Selection Statistical Inference, you may check Robert Tibshirani's talk.

Coding Schemes for Regression

Reading Assignment:

- Muller and Fetterman, Chapter 12: "Coding Schemes for Regression" (Required)

Goals

1. Understand various coding schemes for ANOVA and relationships

between ANOVA and multiple regression models.

2. Understand basic strategy and techniques of multiple comparisons.
3. Recognize the impact of various amounts of missing data.

The study of ANOVA is motivated by desire to model and test hypotheses about two or more group means.

Regression or ANOVA?

In theory and computation, ANOVA is a special case of regression analysis with dummy variables as predictors.

An *indicator* or *dummy* variable is used to represent group membership. Suppose we are interested in the effect of two weight loss regimens, diet and exercise, over a period of two months, compared to subjects on neither regimen (control subjects). Then we can create three dummy variables to denote regimen membership as follows.



$$\begin{aligned}x_1 &= \begin{cases} 1 & \text{diet} \\ 0 & \text{exercise or neither} \end{cases} \\x_2 &= \begin{cases} 1 & \text{exercise} \\ 0 & \text{diet or neither} \end{cases} \\x_3 &= \begin{cases} 1 & \text{neither} \\ 0 & \text{diet or exercise} \end{cases}\end{aligned}$$

For analysis, we will use (x_1, x_2, x_3) (or some combination of them) in our \mathbf{X} matrix to represent the diet regimens.

The name ANOVA reflects the expression of tests in terms of variances and the nature of computational strategies predating computers.

An ANOVA *coding scheme* defines a set of rules for representing all of the information in one or more categorical variables as one or more interval variables. We will discuss the following five coding schemes:

1. reference cell,
2. cell mean,
3. classical ANOVA,
4. effect, and
5. polynomial.

Each coding scheme merits consideration as well as a description of its relationship to other schemes. This chapter only considers estimation, and all consideration of testing will be left to subsequent chapters.

Classical ANOVA Coding

Classical ANOVA coding uses a less than full rank \mathbf{X} matrix ($\text{rank}(\mathbf{X}_{n \times (p+1)}) = p$) intentionally. This \mathbf{X} matrix is equal to the cell mean coding \mathbf{X} matrix with the addition of a column of 1's for the



intercept. The model is given by



$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} has rank p (less than full rank). 

Often, we represent $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ as $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_p)$.

Expected Values for Classical ANOVA Coding

Group	$E[y_i]$	Group Mean
1	$1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$	μ_1
2	$1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_2$	μ_2
\vdots	\vdots	\vdots
p	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_p = \beta_0 + \beta_p$	μ_p

Group 1 is the reference group, and $\beta_0 + \beta_1 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_2) - (\beta_0 + \beta_1)) = (\beta_2 - \beta_1)$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_2) - (\beta_0 + \beta_p)) = (\beta_2 - \beta_p)$.

Because \mathbf{X} is not full rank, our estimates of β are not unique because β is not estimable. However, we may obtain a unique solution by imposing an additional constraint. A common constraint is to require $\sum_{j=1}^p \beta_j = 0$. This ensures that β_0 is the grand mean because

$$\begin{aligned} \frac{(\beta_0 + \beta_1) + (\beta_0 + \beta_2) + \cdots + (\beta_0 + \beta_p)}{p} &= \frac{p\beta_0 + \sum_{j=1}^p \beta_j}{p} \\ &= \beta_0. \end{aligned}$$

Many authors describe $\beta_0 = \mu$ as the grand mean.

Classical ANOVA coding leads to a less than full rank model that may be numerically unstable. Because all parameters are not estimable, we must use the theory for the less than full rank model. However, it is easier just to use a full rank coding scheme.

SAS uses classical ANOVA coding but imposes the constraint that $\beta_j = 0$ for one j (a *reference cell* constraint).

Note that parameter estimates under different constraints may have different meanings. For example, $\hat{\beta}_0$ under the first constraint is the grand mean, while $\hat{\beta}_0$ under the second constraint is the mean for the reference group.

Reference Cell Coding

One Group

Suppose that a geneticist wishes to address the impact of a potentially toxic environmental exposure on mice. The intercept-only model assumes that all responses differ randomly from a common mean response, called the *grand mean*. Thus we have the model

$$\begin{aligned} \mathbf{y}_{n \times 1} &= \mathbf{X}_{n \times 1} \boldsymbol{\beta}_{1 \times 1} + \boldsymbol{\varepsilon}_{n \times 1} \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [\beta_0] + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \end{aligned}$$

This ANOVA model is often called the *grand mean model*. In this model, \mathbf{X} is full rank, and $E[y_i] = \beta_0$. Often, we replace β_0 with μ so that $E[y_i] = \mu$.

In this model,

$$\begin{aligned}\widehat{\beta}_0 &= (\mathbf{J}'_n \mathbf{J}_n)^{-1} \mathbf{J}'_n \mathbf{y} \\ &= n^{-1} \sum_{i=1}^n y_i \\ &= \bar{y}.\end{aligned}$$

Two Groups

Now suppose that the mice are from different mouse strains so that n_1 of the mice are black six mice, and $n - n_1 = n_2$ of the mice are Swiss

albino mice. Then we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times 2} \boldsymbol{\beta}_{2 \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{J}_{n_1 \times 1} & \mathbf{0}_{n_1 \times 1} \\ \mathbf{J}_{n_2 \times 1} & \mathbf{J}_{n_2 \times 1} \end{bmatrix}_{n \times 2}.$$

A dummy (indicator) variable indicates the species for each



observation; that is,

$$x_1 = \begin{cases} 0, & \text{black six} \\ 1, & \text{Swiss albino} \end{cases} .$$

Predictor values assigned by the scientist are often called treatment levels.

Three or More Groups

For p mouse species, we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & & & & \vdots \\ 1 & 0 & \dots & \dots & \dots & \vdots \\ 1 & 1 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank p .

Expected Values for Reference Cell Coding		
Group	$E[y_i]$	Group Mean
1	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0$	μ_1
2	$1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$	μ_2
\vdots	\vdots	\vdots
p	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_0 + \beta_{p-1}$	μ_p

Group 1 is the reference group, and $\beta_0 = \mu_1$ is the mean for group 1.

The difference in mean response between group 2 and group 1 is

$(\mu_2 - \mu_1) = ((\beta_0 + \beta_1) - \beta_0) = \beta_1$, and the difference in mean response between group 2 and group p is

$(\mu_2 - \mu_p) = ((\beta_0 + \beta_1) - (\beta_0 + \beta_{p-1})) = (\beta_1 - \beta_{p-1})$.

Cell Mean Coding

With *cell mean coding*, all of the β 's equal group means. Cell mean coding is the most natural coding scheme and the easiest to understand (and the easiest to explain to investigators!). For p groups, we create p indicator variables and do not include an intercept in the

model, which is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \dots & \dots & \dots & \vdots \\ 0 & 1 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank p .

Expected Values for Cell Mean Coding

Group	$E[y_i]$	Group Mean
1	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0$	μ_1
2	$0 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_1$	μ_2
\vdots	\vdots	\vdots
p	$0 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_{p-1}$	μ_p

The vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})$ is often written $\boldsymbol{\beta} = (\mu_1, \dots, \mu_p)$ or $\boldsymbol{\beta} = (\alpha_1, \dots, \alpha_p)$.

Group 1 is the reference group, and $\beta_0 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = (\beta_1 - \beta_0)$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = (\beta_1 - \beta_{p-1})$. Does this model span an intercept?

Effect Coding

Effect coding provides a useful coding scheme. This full rank scheme has design matrix \mathbf{X} with a column of 1's and $p - 1$ other columns, like the reference cell coding scheme. However, the *effect coding scheme* reassigns the value of 0 to -1 for dummy variable covariates for

subjects in the reference group. The model is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & \dots & -1 \\ \vdots & \vdots & \text{🗨️} & & & \vdots \\ 1 & -1 & -1 & -1 & \dots & -1 \\ 1 & 1 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank p .

The effect coding indicator variable may be created in two steps.

1. Create a dummy variable indicating group membership, with the value 1 if the observation belongs to the group and 0 otherwise.
2. If the observation represents the reference group, then reassign the value of 0 to -1 .

Expected Values for Effect Coding

Group	$E[y_i]$	Group Mean
1	$1 \cdot \beta_0 - 1 \cdot \beta_1 - \cdots - 1 \cdot \beta_{p-1} = \beta_0 - \sum_{j=1}^{p-1} \beta_j$	μ_1
2	$1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$	μ_2
\vdots	\vdots	\vdots
p	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_0 + \beta_{p-1}$	μ_p

Group 1 is the reference group, and $\beta_0 - \sum_{j=1}^{p-1} \beta_j = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_1) - (\beta_0 - \sum_{j=1}^{p-1} \beta_j)) = 2\beta_1 + \sum_{j=2}^{p-1} \beta_j$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_1) - (\beta_0 + \beta_{p-1})) = (\beta_1 - \beta_{p-1})$.

Note that the mean of all group means is

$$\frac{(\beta_0 - \sum_{j=1}^{p-1} \beta_j) + (\beta_0 + \beta_1) + \cdots + (\beta_0 + \beta_p)}{p} = \frac{p\beta_0}{p} = \beta_0.$$

This parameter is the mean of the particular cells in the current design. This is different from the grand mean parameter in classical ANOVA coding (when there is no restriction), which represents the hypothetical mean of all observations in the population.

Relationships Among Coding Schemes

Any full rank \mathbf{X} may be expressed as a full rank linear transform of any other (with both based on the same categorical predictors). Any parameter estimable or testable in one coding scheme is also estimable or testable in any other.

With p groups, only p parameters are estimable.

Parameters for any coding scheme may be expressed as linear functions of cell means.

Next: One-way ANOVA

Reading Assignment:

- Muller and Fetterman, Chapter 13: "One-Way ANOVA"