Multiple Linear Regression
Metrics and Variable Selection

Laura E. Barnes & Julianne Quinn

University of Virginia
Charlottesville, VA 22904

Agenda

1. Review of Multiple Regression
2. Measurement of Model Performance
3. Variable Selection for Multiple Regression
4. Model Comparison

Review of Multiple Regression

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.
- Linear regression uses stochastic models with two components:
  - Deterministic: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and
  - Stochastic: ϵ
- Matrix notation: $y = f(X) + \epsilon = X\beta + \epsilon$
- Regression is the solution to an optimization problem. We find a linear fit to the data that minimizes the sum of square errors.

Assumptions for Multiple Regression

- Assumptions for Optimization
  - The data for the input variables or predictors, $x_1, \cdots, x_p$ are known.
  - The predictors are linearly independent.
  - The response variable, y, is quantitative.
- Assumptions for Inference
  - For a sample size of n, the distribution of the $\epsilon_i, i = 1, \cdots, n$ are independent, identical and Gaussian with
  - $E(\epsilon_i) = 0$, and
  - $Var(\epsilon_i) = \sigma^2$

Least Squares Estimates

- We find optimal estimates for the coefficients β where the criterion is least squares. The optimization problem is:
$$\text{minimize}(X\beta - y)^T(X\beta - y)$$
- The least square estimate:
$$\hat{\beta} = (X^TX)^{-1}X^Ty$$
- The estimate for σ^2 where k is the number of predictors or input variables:
$$\hat{\sigma}^2 = \frac{(X\hat{\beta} - y)^T(X\hat{\beta} - y)}{n-k-1} = \frac{y^T(I-H)y}{n-k-1}$$
- $H = X(X^TX)^{-1}X^T$ is the hat matrix: $\hat{y} = X\hat{\beta} = Hy$

Sum of Squares Decomposition, F-Test, t-tests

- The sum of squares has a convenient decomposition:

$$\text{Total S.S.} = \text{Residual S.S} + \text{Model S.S}$$

- Model Utility Test:

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

$$H_A : \beta_i \neq 0, i \in \{1, \cdots, p\}$$

- t-tests:

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

Interpret Coefficients

- Predict *EQPDMG* using TEMP + TRNSPD + TONS:

Variable	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	2788.700	3248.789	0.858	0.39069
TEMP	-142.800	49.524	-2.883	0.00394
TRNSPD	2972.50	66.853	44.463	< 2e-16
TONS	9.380	0.241	38.918	< 2e-16

- F test results: 1422 on 3 and 39667 d.f, p-value: $< 2.2e-16$
- Write down the estimated model and interpret it.

Steps to Build Multiple Regression Models

- Suppose we have an interesting problem like reducing severity of rail accidents. Also, we have data related to this problem.
- What will you do to solve this problem?

Evidence-Informed Problem Solving

1. Background, goals, sources of evidence
2. Hypothesis(es)
3. Visualization and Graphical Analysis
4. Build multiple regression models:
    - How well do the models perform?
    - What explanatory variables should we use?
    - ...

Sum of squares decomposition

- Sum of squares decomposition:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2$$

$$\text{Total S.S.} = \text{Residual S.S} + \text{Model S.S}$$

Multiple Coefficient of Determination: R^2

- Sum of squares decomposition:

$$\text{Total S.S.} = \text{Residual S.S} + \text{Model S.S}$$

- Coefficient of determination:

$$R^2 = 1 - \frac{\text{Residual S.S}}{\text{Total S.S}}$$

  - What's the range of R^2? $[0, 1]$
  - What does $R^2 = 1$ mean? A perfect fit to data; but not necessarily a perfect model.
  - Will R^2 decreases if we add variables to a regression model? No.

Multiple Coefficient of Determination: R^2

Properties of R^2

- The range of R^2 is from 0 to 1. Near 0 value indicate little linear association the independent variables and the dependent variable. A value near 1 means a strong association.
- R^2 cannot go down when another predictor is added to the model.
- R^2 can almost always be made close to 1 by using a model with k predictors where k is very close to n

Criterion Based Assessments

- What criteria can we use to compare performance? Our goal is to predict accurately on new data.
- Why don't we use R^2?
- Most criteria reward good fits and penalize size (complexity).

Common Criterion-Based Methods

- Adjusted-R^2:

$$
\begin{aligned}
R^2_{adj} &= 1 - \frac{\text{Residual M.S}}{\text{Total M.S}} \\
&= 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}
\end{aligned}
$$

Properties

- Adjusted R^2 penalizes the addition of extraneous predictors to the model.
- Adjusted R^2 is smaller than R^2

Common Criterion-Based Methods

- Akaike Information Criterion (AIC):

$$AIC = nlog(R.S.S/n) + 2p$$

- Bayesian Information Criterion (BIC):

$$BIC = nlog(R.S.S/n) + plog(n)$$

- Mallows C_p:

$$C_p = nlog(R.S.S/n) + p - n$$

Common Criterion-Based Methods

- The preferred model gives the smaller value of AIC and BIC.
- In most cases, BIC penalizes the complexity more strongly AIC does.
- You may see different values for AIC or BIC in different functions. This results from the use of different definitions of the terms and different logs. All that matters is order and consistency.

Example: Rail Accidents

- Two models:

$$EQPDMG \sim TEMP + TRNSPD + TONS \tag{1}$$
$$EQPDMG \sim TEMP + TRNSPD + TONS + CARS \tag{2}$$

- Results for multiple criteria:

	Model (1)	Model (2)
R^2	0.09712	0.09718
R^2_{adj}	0.09705	0.09709
AIC	1086803	1086802
BIC	1086846	1086854

- Interpret the models based on each metric.

Importance of Variable Selection

- Okham’s (Occam’s) Razor
- Extra terms can add noise to the predictions . More explanatory variables is not necessarily better.
- Correlation among the variables creates instabilities in the model. Small changes have big effects that are not wanted. This is called multicollinearity.
- Leaving out variables can also cause inaccurate understanding and predictions.
- Getting the data to add predictors can be costly.

Complexity of Variable Selection

- If the number of available explanatory variables is very small, such as 2-3, we might build models with all combinations of explanatory variables.
- What if we have k explanatory variables? We will have $2^k - 1$ main effects models.
- Actually, we need consider more than the main effects models as we will see later. In addition, we will add qualitative variables and consider transformation of some or all of the variables.
- What is the next step?

Automated Selection

- Automated procedures have been developed to choose variables.
- Since the number of models to consider is so large, none of the procedures are optimal on any particular criteria. However, they can provide a convenient approach to narrowing your search and providing some insight.
- Three simple automated selection techniques:
  - Forward selection
  - Backward selection
  - Stepwise regression
- Regression models with automated variable selection is a hot topic of recent research.

Automated Selection Techniques

- All the automated techniques require some criterion:
  - We used to use F tests and the significance level on F tests.
  - Most modern approaches use AIC or BIC or some other version of them
- Forward selection: starts with the mean model and adds terms one at a time. It picks the term that does the best on the criterion. It stops when the threshold is reached.
- Backward selection: starts with the complete model. It then eliminates terms one at a time and stops when a threshold is met.

Stepwise Regression

- Stepwise regression:
  - A sequential process that adds or drops one new variable to the model at each step.
  - It always selects the variable that provides the greatest improvement to the selection criterion given the variables already in the model.
  - Stepwise stops when no variable can be added or dropped according to a threshold.
- Stepwise works best when started by backward selection (i.e., start with a large model).
- Stepwise is computationally expensive. For large variable sets expect a wait.
- In R stepwise defaults to using AIC.

Partial F Tests

- How can we tell whether the reduction in sum of squares by a inclusive model (one that includes all of the variables of the smaller model and more) is statistically significant? Partial F tests
- Suppose the smaller model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \epsilon$$

and the larger model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \beta_{i+1} x_{i+1} + \cdots + \beta_k x_k + \epsilon$$

- We use the F statistic to test:

$$H_0 : \beta_{i+1} = \cdots = \beta_k = 0$$

$$H_A : \beta_j \neq 0, j \in \{i+1, \cdots, k\}$$

Partial F Tests

- Partial F test example: full model vs. stepwise model:

	Res.Df	RSS	Df	Sum of Sq	F	$Pr(> F)$
1	2615	2.7961e+14				
2	2617	2.7965e+14	-2	-4.066e+10	0.1901	0.8269

- If we accept the null hypothesis, which model is preferred? smaller one
- Partial F tests provide a convenient, parametric method to judge between models.
- Also allows us to assess all of the variables.

Test Sets

- Partial F test is a good approach, but what if we want to compare non-nested models?
- Recall our goal for model selection. In using test sets we sample a subset (without replacement) of the data and do not use it to build the model. We use this subset for testing.
- A common choice is to sample 1/3 of the data for testing (called the test set or out-of-sample set) and 2/3 for model building (called the training set or the sample).
- Check that your test set is representative.
- What criterion do we use to choose among the models? Predicted Mean Squared Error (PMSE)

Example: Rail Accident Test Sets Results

- Test set $n = 13,224$; models in comparison:

$$EQPDMG \sim TEMP + TRNSPD + TONS + CARS$$
$$EQPDMG \sim TEMP + TRNSPD$$

- PMSE Result:

	PMSE
Model 1	32386889683
Model 2	33913904861

Cross-Validation

- Test sets method requires enough data.
- If we don't have enough data, we can use cross validation.
- In cross validation we start by sampling without replacement to get k parts, called folds.
- We build a model on $k - 1$ of the folds and test on the remaining fold.
- We repeat this using each fold as a test set for the other $k - 1$ folds.
- If we have n data points how much of it do we use for training? n
- How much do we use for testing? n
- Models are compared based on the averages of PMSE.

Rail Accident Cross-Validation Example

- Folds $k = 10$; models in comparison:

$$\begin{aligned} &EQPDMG \sim TEMP + TRNSPD + TONS + CARS \\ &EQPDMG \sim TEMP + TRNSPD \end{aligned}$$

- CV Results:

	MSE
Model 1	547889555 66
Model 2	558409204733