MTH 4230 Spring 2021
Module 9 Notes and Exercises (Project 8)

Model Selection

In the Module 7 notes, the matrix form of the multiple linear regression data model for a set of response variable values $\mathbf{Y}$ as a function of a set of vector-valued predictor variable $\mathbf{x} = (x_1, \ldots, x_{p-1})$ is given by

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$ and $\mathbf{Y}$ are as indicated in the Module 6 notes and $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$.

In the Module 8, we emphasized the potential complexity of this model and importance of transformations and interactions. In these notes, we survey model selection techniques. Hypothesis testing methods for comparing models and automated model searching procedures are defined. Also, we exemplify transformation selection methods by way of the Box-Cox procedure for identifying power function transformations.

Hypothesis Tests for Model Selection. The following hypothesis testing methods can be used for model selection.

Testing for Lack of Fit. Assuming the residual error terms satisfy the model requirements of independence, normality, and constant variance, one can test for the lack of fit to a deterministic model. **This test requires repeated observations of each treatment considered**, referred to as the *number of trials* of each selected predictor level. Let m denote the number of distinct groups of (different) observations so that $\frac{n}{m}$ is the number of observations of each combination of factor levels considered in the repeated observations.

Letting $\overline{Y}_j$ denote the average response for the observation associated with the jth level, $j = 1, \ldots, m$, we define the Pure Error Sum of Squares (SSPE) by,

$$\sum_j \sum_i (Y_{ij} - \overline{Y}_j)^2$$

We then define the Lack of Fit Sum of Squares (SSLF) by $SSE - SSPE$ where,

$$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

The ratio $\left(\frac{SSE-SSPE}{m-p}\right) / \left(\frac{SSPE}{n-m}\right) =: \frac{MSLF}{MSPE}$ gives a test statistic for the lack of fit hypothesis test, outlined below. Here, MSLF abbreviates Mean Squares Lack of Fit and Mean Squares Pure Error, respectively.

F-test for lack of fit

- **Null Hypothesis** $H_0 : E[Y] = \mathbf{X}\,\boldsymbol{\beta}$
- **Alternative Hypothesis** $H_A : E[Y] \neq \mathbf{X}\,\boldsymbol{\beta}$
- **Test Statistic**

$$F^* = \frac{MSLF}{MSPE} = \frac{\left(\frac{SSE-SSPE}{m-p}\right)}{\left(\frac{SSPE}{n-m}\right)}$$

- **Rejection Region**

$$[F_{\alpha,m-p,n-m}\ ,\ \infty)$$

- P**-value**

$$P[F > F^*]$$

where F is a the F-distribution with $m-p$ numerator degrees of freedom and $n-m$ denominator degrees of freedom.

Tests Based on Nested Models. When predictors are removed from a full model (reduction of predictiors) or when predictors are added to a base model (additional predictors), we can compare the full model to the reduced model using a hypothesis test.

Let $SSR = SST - SSE$ for any multivariate regression model; we will use subscripts to indicate factors in a particular regression model SSR and SSE values. Note that SST does not depend on the particular predictors used. In the case of two predictors X_1 and X_2, for example, we can decompose the total sum of squares in terms of X_1 as follows.

$$SST = SSR_{X_1} + SSE_{X_1}$$

The Sum of Squares Error associated with the model that includes both predictors, X_1 and X_2, is denoted SSE_{X_1,X_2}, we can define the marginal increase in the regression sum of squares by $SSE_{X_1|X_2} := SSE_{X_1} - SSE_{X_1,X_2}$ to get,

$$SST = SSR_{X_1} + SSR_{X_2|X_1} + SSE_{X_1,X_2}$$

We now substitute the previous identity into the decomposition of SST in terms of X_1 *and* X_2, given by

$$SST = SSR_{X_1,X_2} + SSE_{X_1,X_2}$$

After the substitution, we have

$$SSR_{X_1,X_2} = SSR_{X_1} + SSR_{X_2|X_1}$$

Similar derivations, with more predictors, can gives endless formula for these type of decompositions, for example,

$$\begin{aligned} SSR_{X_2,X_1} &= SSR_{X_2} + SSR_{X_1|X_2} \\ SSR_{X_1,X_2,X_3} &= SSR_{X_1} + SSR_{X_2,X_3|X_1} \\ SSR_{X_1,X_2,X_3} &= SSR_{X_1,X_2} + SSR_{X_3|X_1,X_2} \\ SSR_{X_1,X_2,X_3,X_4} &= SSR_{X_1,X_2} + SSR_{X_4,X_3|X_1,X_2} \end{aligned}$$

With the so-called Extra Sums of Squares termed, as defined above, we can now define several special cases of a class of hypothesis tests for comparing models based on this sum of squares decomposition.

F-test for $\beta_j = 0$

Let $j \in \{1, \ldots, p-1\}$. We can test whether one coefficient in a full model is equal to zero as follows. Here we use a three-predictor model ($p = 4$) with the single coefficient of interest given by β_3.

- **Null Hypothesis** $H_0 : \beta_3 = 0$
- **Alternative Hypothesis** $H_A : \beta_3 \neq 0$

- **Test Statistic**

$$F^* = \frac{MSR(X_3|X_2, X_1)}{MSE(X_1, X_2, X_3)} = \frac{\left(\frac{SSR_{X_1,X_2,X_3} - SSR_{X_1,X_2}}{4-3}\right)}{\left(\frac{SSE_{X_1,X_2,X_3}}{n-4}\right)}$$

- **Rejection Region**

$$[F_{\alpha,1,n-4}\ ,\ \infty)$$

- ***P*-value**

$$P[F > F^*]$$

where F is a the F-distribution with 1 numerator degrees of freedom and $n-4$ denominator degrees of freedom.

F-test for if several $\beta_j = 0$

Let $J \subset \{1, \ldots, p-1\}$ be a subset of indices within a full model. We can test whether several coefficients in a full model are equal to zero as follows. Here we use a four-predictor model ($p = 5$) with the coefficients of interest given by β_3 and β_4.

- **Null Hypothesis** $H_0 : \beta_3 = \beta_4 = 0$
- **Alternative Hypothesis** $H_A : \beta_3 \neq 0$ or $\beta_4 \neq 0$
- **Test Statistic**

$$F^* = \frac{MSR(X_3|X_2, X_1)}{MSE(X_1, X_2, X_3)} = \frac{\left(\frac{SSR_{X_1,X_2,X_3} - SSR_{X_1,X_2}}{4-3}\right)}{\left(\frac{SSE_{X_1,X_2,X_3}}{n-4}\right)}$$

- **Rejection Region**

$$[F_{\alpha,1,n-4}\ ,\ \infty)$$

- ***P*-value**

$$P[F > F^*]$$

where F is a the F-distribution with 1 numerator degrees of freedom and $n-4$ denominator degrees of freedom.

The hypothesis tests exemplified above represent only a glimpse of many model comparison tests that can be formulated to test specific hypotheses associated with model.

Model Selection Criteria. The hypothesis test based model selection methods above are employed when researchers can reduce the set of candidate models down to a just a couple of nested models that can be compared in pairwise iterations. When large groups of possible models are considered, more efficient, computational model selection procedures using various model criteria are employed. Many methods exist; some of the popular approaches are indicated below.

In cases where variables in a linear model can be reduced to a subset of size k where 2^k is small enough to reasonably list all of the numbers from 1 to 2^k, automated selection procedures are employed according to various model selection criterion. Care has to be taken when considering large numbers of candidate models. In particular, we need to verify that candidate models satisfy the model assumptions.

Here we indicate some common model selection criterion, but do not motivate these criteria thoroughly. The theoretical motivations for these criterion can be involved, but it should be noted that a key feature of several of the criteria is *the extent to which the criteria penalizes the model for including more variables.* It is important to penalize for additional variables in the model, because additional variables tend to increase the standard error of parameter estimates.

Coefficients of Determination. Two coefficients, the classical coefficient of multiple determination, R^2, and the **adjusted coefficient of multiple determination**, R_A^2, which penalizes for more variables, are defined below.

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SST} \\ R_A^2 &= 1 - \frac{\left(\frac{SSE}{n-p}\right)}{\left(\frac{SST}{n-1}\right)} = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SSE}{SST}\right) \end{aligned}$$

Maximizing either statistic corresponds to optimizing the fit of the model. It should be noted that, for a fixed sample size, maximizing R^2 and R_A^2 with respect to variable combinations is equivalent to minimizing SSE and MSE, respectively.

Mallow's C_p Criteria. Mallow's C_p criteria compares the total mean squared error of n fitted values with respect to different subsets of $p-1$ predictor variables from a full model with population mean response μ_i associated with the ith observation. This criteria relies on the theoretical observation that, if $\hat{Y}_i$ denotes the fitted values associated with the sub-model for the ith observation, then

$$E\left[\hat{Y}_i - \mu_i\right] = \left(E\left[\hat{Y}_i\right] - \mu_i\right)^2 + \sigma^2_{\hat{Y}_i}$$

The statistic C_p estimates the total mean square error with respect to fitted values with $p-1$ predictors in the submodel, standardized by σ^2, and is calculated by the following formula:

$$C_p = \frac{SSE_p}{MSE_T} - (n - 2p)$$

Where SSE_p is sum squares error associated with fitted values for the sub-model with $p-1$ predictors and MSE_T denotes mean squares error associated with sum full model with $P - 1 > p - 1$ predictors.

Minimizing this statistic over sub-models corresponds to optimizing the fit of the sub-model. Notably, values of C_p near p (the lowest possible value of C_p, as it turns out) indicate an unbiased model.

Akaike's Information Criteria. Akaike's Information Criteria (AIC) is an alternative measure for which minimal AIC values are desired. Minimizing similar to minimizing SSE (hence maximizing R^2) but with a more sophisticated penalty for additional variables, balanced with sample size, that is motivated by theoretic statistics considerations.

Precisely, AIC is defined by,

$$n\ln(SSE) - n\ln(n) + 2p$$

Minimizing this statistic over sub-models corresponds to optimizing the fit of the sub-model.

Scharwz' Bayesian Information. Scharwz' Bayesian Information or the Bayesian Information Criteria (BIC) is an alternative measure for which minimal BIC values are desired. Minimizing similar to minimizing SSE (hence maximizing R^2) but with a more sophisticated penalty for additional variables, balanced with sample size, that is motivated by theoretic statistics considerations.

Precisely, BIC is defined by,

$$n\ln(SSE) - n\ln(n) + (\ln(n))p$$

Minimizing this statistic over sub-models corresponds to optimizing the fit of the sub-model.

PRESS Criteria. The PRESS Criteria is a novel, computational approach that essentially predicts the response for each of the n observations using a model built from all of the other observations fo the data set, and then takes the squared difference of the actual response and the predicted response using the omitted observation model for each i. Letting $\hat{Y}_{i(i)}$ denote the predicted response when the ith observation is omitted, we define the PRESS measure of the model to be,

$$PRESS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)}\right)^2$$

Minimizing this statistic over sub-models corresponds to optimizing the fit of the sub-model.

Coefficients of Partial Determination. Using the same sum of squares error definition that precedes the F-test for $\beta_j = 0$ from the previous subsection, we can define descriptive measures called *coefficients of partial determination* that measure the marginal contribution of the factors in the larger model not already included in the sub-model. If the response variable, Y, is clearly indicated, we can use the following notation for various coefficients of partial determination exemplified below

$$
\begin{aligned}
R^2_{2|1} &= \frac{SSR_{X_2|X_1}}{SSE_{X_1}} \\
R^2_{1|2} &= \frac{SSR_{X_1|X_2}}{SSE_{X_2}} \\
R^2_{1|23} &= \frac{SSR_{X_1|X_2,X_3}}{SSE_{X_2,X_3}} \\
R^2_{4|123} &= \frac{SSR_{X_4|X_1,X_2,X_3}}{SSE_{X_1,X_2,X_3}} \\
R^2_{43|12} &= \frac{SSR_{X_4,X_3|X_1,X_2}}{SSE_{X_1,X_2}}
\end{aligned}
$$

Transformation Selection. Methods of data transformation can be used to transform response variables of interest that may satisfy the model assumptions. Here, we consider the Box-Cox Procedure

Box-Cox Procedure. The Box-Cox procedure is a method used when a power transformation of the response variable is considered appropriate. The Box-Cox procedure adds an additional parameter, λ, to the General Linear Model in the following form:

$$\mathbf{Y}^{\lambda} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Computational statistical commands in *R* and other statistical software can compute maximum likelihood estimators of this parameter based on statistical theory.

Project 8

Exercise (5 points): Use the your group project results for Project for this exercise. Consider two of the models you considered in this project that have a nested (sub-model) structure. Compare the two models using an F-test.

Exercise (5 points): Use the `sat` data frame for this exercise. This exercise is based on Exercises 4, Chapter 3, in *Linear Models with R.*

(1) Fit a model with total SAT score as the response and `expend`, `ratio` and `salary` as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?
(2) Now add takers to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent

Exercise (5 points): Using transformations to obtain a good fit, fit a linear model to **one of** the following response variables and predictor variable(s) combination. Use a goodness-of-fit test to check your model, if appropriate; if a test for lack of fit is not possible for your data, explain why. Decide if it is reasonable to leave the response untransformed. This exercise is based on Exercises 2-7, Chapter 9, in *Linear Models with R*.

(1) Use `yield` as a response and `nitrogen` as a predictor from the `cornit` data set.
(2) Use `O3` as a response and `temp`,`humidity`, and `ibh` as predictors from the `ozone` data set.
(3) Use `pressure` as a response and `temperature` as a predictor from the `pressure` data set.
(4) Use `volume` as a response and `girth`, `humidity`, and `height` as predictors from the `trees` data set.
(5) Use `cheddar` as a response and three other variables as predictors from the `cheddar` data set.

Exercise (5 points): This exercise uses a method called *Ridge Regression* that penalizes estimates of coefficients for being large. Ridge regression, covered in Section 11.3, makes the assumption that the regression coefficients should not be very large. This exercise is based on Exercises 3-4, Chapter 11, in *Linear Models with R*.

For **one of** the following response variables and predictor variables combinations, (a) build a linear regression model with all predictors, (b) build a linear regression with variables selected according to the AIC procedure, and (c) build a linear regression with variables selected using a ridge regression procedure.

(1) Fit models using the `seatpos` data with `hipcenter` as the response and all other variables as possible predictors.
(2) Fit models using the `fat` data with `siri` as the response and all other variables except `brozek` and `density` as possible predictors.

References

[1] Cornillon, Pierre-Andre. *R for Statistics, 1rst ed.*. Chapman and Hall, (2012).
[2] Draper, N.R., Smith, H. *Applied regression analysis 3rd ed.* Wiley (1998)
[3] Faraway, J. *Linear Models with R, 2nd ed.*. Chapman and Hall, (2014).
[4] Faraway, J. (April 27, 2018) *Linear Models with R* translated to Python. Blog post. https://julianfaraway.github.io/post/linear-models-with-r-translated-to-python/
[5] Fahrmeir, Kneib, Lang, Marx , *Regression.* Springer-Verlag Berlin Heidelberg (2013).
[6] Kutner, Nachtsheim, Neter, Li *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin (2005).