Use R to Analyze the Boston Housing Data

- Basic command: `lm`
- Rank deficiency
- RSS *vs.* prediction error (training error *vs.* test error)

Interpret the LS coefficients

- $\hat{\beta}_j$ measures the average change of Y per unit change of X_j, with all other predictors held fixed.
- Seemingly contradictory results from SLR and MLR: SLR suggests that "age" has a significant negative effect on housing price, while MLR suggests the opposite.

Partial Regression Coefficients

Consider a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_p X_p + \text{err}.$$

The LS estimate $\hat{\beta}_k$ describes the partial correlation between Y and X_k **adjusted for the other predictors**.

The LS estimate $\hat{\beta}_k$ can be obtained as follows (see Algorithm 3.1 from ESL):

1. Y^*: residual from regressing Y onto all other predictors except X_k
2. X_k^*: residual from regressing X_k onto all other predictors except X_k
3. Regress Y^* onto X_k^*

Hypothesis Testing in Linear Regression Models

The key test is the F-**test**. Compare two nested models

- H_0: reduced model with p_0 coefficients;
- H_a: full model with p_a coefficients.

Nested: if the reduced model is a special case of the full model, e.g.,

$$H_0 : Y \sim X_1 + X_2, \quad H_a : Y \sim X_1 + X_2 + X_3.$$

Note that $\text{RSS}_a < \text{RSS}_0$ and $p_a > p_0$.

F-test

Test statistic:

$$F = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(p_a - p_0)}{\mathsf{RSS}_a/(n - p_a)},$$

which $\sim F_{p_a - p_0, n - p_a}$ under the null.

- Numerator: variation (per dim) in the data not explained by the reduced model, but explained by the full model, i.e., evidence supporting H_a.
- Denominator: variation (per dim) in the data not explained by either model, which is used to estimate the error variance.

Reject H_0, if F-stat is large, i.e., the variation missed by the reduced model, when being compared with the error variance, is significantly large.

Special Cases of the F-test

- The so-called t-test for each regression parameter (see the `R` output) is a special case of F-test. For example, the test for the j-th coef β_j compares
  - $H_0 : Y \sim 1 + X_1 + \cdots + X_{j-1} + \qquad X_{j+1} + \cdots + X_p$
  - $H_a : Y \sim 1 + X_1 + \cdots + X_{j+1} + X_j + X_{j+1} + \cdots + X_p$
- The overall F-test (at the bottom of the `R` output) compares
  - $H_0 : Y \sim 1$
  - $H_a : Y \sim 1 + X_1 + \cdots + X_{j+1} + X_j + X_{j+1} + \cdots + X_p$

Handle Categorical Variables

Consider a categorical predictor, `Size`, taking values from $\{S, M, L\}$, which needs to be coded as two numerical predictors.

$$\begin{pmatrix} S \\ S \\ M \\ M \\ L \\ L \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}_{6\times 2}$$

- 1st column: indicator for value "M".
- 2nd column: indicator for value "L".
- No need to code "S", which is chosen as the **reference level** and its effect is absorbed into the intercept. (You can choose any value as the reference group.)
- In general, code a categorical predictor with K values as $(K-1)$ binary vectors.

Categorical Variables and Interactions

We can also generate products of those indicator variables with other variables to create the **interaction terms**. Suppose there is another numerical predictor, `Price`, denoted by $\{x_i\}_{i=1}^6$, and we fit a linear regression model including `Size`, `Price`, and their interaction. The design matrix looks like follows

$$\begin{pmatrix} S \\ S \\ M \\ M \\ L \\ L \end{pmatrix} \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 & 0 & 0 & x_1 & 0 & 0 \\ 1 & 0 & 0 & x_2 & 0 & 0 \\ 1 & 1 & 0 & x_3 & x_3 & 0 \\ 1 & 1 & 0 & x_4 & x_4 & 0 \\ 1 & 0 & 1 & x_5 & 0 & x_5 \\ 1 & 0 & 1 & x_6 & 0 & x_6 \end{pmatrix}$$

How to interpret the LS coefficients?

Collinearity

- We often encounter problems in which some predictors are highly correlated, e.g., the `seatpos` data. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference on $\boldsymbol{\beta}$.
- Typical symptoms of collinearity: high pair-wise (sample) correlation between predictors; R^2 is relatively large, overall F test is significant, but none of the predictors is significant.

Collinearity

- We often encounter problems in which some predictors are highly correlated, e.g., the `seatpos` data. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference on β.
- Typical symptoms of collinearity: high pair-wise (sample) correlation between predictors; R^2 is relatively large, overall F test is significant, but none of the predictors is significant.
- What to do with collinearity? Remove some predictors or combine collinear predictions (e.g., PCA).

Collinearity

- We often encounter problems in which some predictors are highly correlated, e.g., the `seatpos` data. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference on $\boldsymbol{\beta}$.
- Typical symptoms of collinearity: high pair-wise (sample) correlation between predictors; R^2 is relatively large, overall F test is significant, but none of the predictors is significant.
- What to do with collinearity? Remove some predictors or combine collinear predictions (e.g., PCA).
- How would collinearity affect prediction of Y?

LINE: Assumptions for Linear Regression

- L: $f^*(x) = \mathbb{E}(Y \mid X = x)$ is "assumed" to be a linear function of x. This is not really an assumption, but a restriction. If the truth f^* is not a linear function, then regression just returns us the best linear approximation of f^*.
- INE: error terms at all x_i's are iid $\mathcal{N}(0, \sigma^2)$ (can be relaxed to be uncorrelated with mean zero and constant variance). This assumption is related to the objective function, an unweighted sum of the squared errors at all x_i's. If the errors have unequal variances (heteroscedasticity) or correlated, then we should use a different objective function.
- No assumptions on X's. But to achieve a good performance, we would like $\mathbf{x}_i$'s to be uniformly sampled.

Outliers

- Outlier test based on leave-one-out prediction error. Let $\hat{\boldsymbol{\beta}}_{(-i)}$ be the LS estimate of $\boldsymbol{\beta}$ based on $(n-1)$ samples excluding the i-th sample $(\mathbf{x}_i, y_i)$, then

$$\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{(-i)}}{\text{some normalizing term}} \sim \mathcal{N}(0,1), \text{ if } i\text{th sample is NOT an outlier.}$$

- Datasets from real applications are usually large (in terms of both n and p). Do not recommend to test outliers. Why?
  - Need to adjust for multiple comparison; cannot detect a cluster of outliers.
- But do recommend to do some of the following:
  - Run the `summary` command in `R` to know the range of each variable;
  - Apply log, square-root or other transformations on right-skewed predictors and Y.
  - Apply winsorization to remove the effect of extreme values.