Linear Regression Models

- **Part I**: Review on linear regression models
- **Part II**: Variable selection and shrinkage
  - Forward/backward/stepwise algorithms with AIC/BIC
  - Regularization methods: Ridge and Lasso
- **Part III**: Methods using derived input directions:
  - PCA and regression
- Case study: Boston Housing Data.

Part I: Review

- Setup
- LS principle and its geometric interpretation
- Understand `R` output
- Hypothesis testing in linear models
- Handle categorical features
- Collinearity
- Assumptions

Linear Regression Model

The linear regression model describes the dependence between a set of p explanatory (predictor) variables $X_1, \ldots, X_p$ and a response variable Y by

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

where $X_1 = 1$ denotes the intercept and ϵ represents the random error, the stochastic discrepancy between the linear model and Y.

Given a set of training data $(x_{i1}, \ldots, x_{ip}, y_i)_{i=1}^n$, we can express the regression model in the following matrix form

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

We often call $\mathbf{X}$ the "design matrix" (in statistical jargon). Here we assume $n > p$ and rank$(\mathbf{X}) = p$.

Matrix Representation

$$\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} &= \begin{pmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + e_1 \\ x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + e_2 \\ \cdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + e_n \end{pmatrix} \\
&= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix} \\
\mathbf{y}_{n\times 1} &= \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}
\end{aligned}$$

Estimation via Least Squares

The least squares method estimates $\boldsymbol{\beta}$ by minimizing

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \Big(y_i - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p \Big)^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T_{p\times n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})_{n\times 1} = \mathbf{0}_{p\times 1}$$

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$ [a]

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Note that the inverse of the $p \times p$ matrix $\big(\mathbf{X}^t\mathbf{X}\big)$ exists since we assume the rank of $\mathbf{X}$ is p.

[a] Known as the normal equations.

- The Fitted Values (i.e., prediction at the n observed data points $\mathbf{x}_i$'s):

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}_{n\times n}\mathbf{y}.$$

- $\mathbf{H}$: Projection (or *hat*) matrix. It is symmetric and idempotent ($\mathbf{HH} = \mathbf{H}$).

- The Residuals: $\mathbf{r}_{n\times 1} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. The residuals can be used to estimate the error variance

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{\mathsf{RSS}}{n-p}.$$

Recall that the LS estimate $\hat{\boldsymbol{\beta}}$ satisfies the normal equations

$$\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

So $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ satisifies:

- $\mathbf{X}^t\mathbf{r} = \mathbf{0}$, the cross-products between the residual vector $\mathbf{r}$ and each column of $\mathbf{X}$ are zero; especially, if the intercept is included in the model, we have $\sum_{i=1}^n r_i = 0$;
- $\hat{\mathbf{y}}^t\mathbf{r} = \hat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{r} = 0$, the cross-product between the fitted value $\hat{\mathbf{y}}$ and the residual vector $\mathbf{r}$ is zero.

That is, the residual vector $\mathbf{r}$ is orthogonal to each column of $\mathbf{X}$ and $\hat{\mathbf{y}}$.

The Geometric Interpretation

- The essence of LS: decompose the n-dim data y into two orthogonal vectors:

  $$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{r},$$

  where $\hat{\mathbf{y}}$ is a linear combination of the columns of $\mathbf{X}$, and $\mathbf{r}$ is the remaining part.

- The p column vectors of $\mathbf{X}$ spanned a p-dim linear subspace in $\mathbb{R}^n$, denoted by $C(\mathbf{X})$; every element in $C(\mathbf{X})$ can be uniquely written as a linear combination of the columns of $\mathbf{X}$, i.e., $\mathbf{X}_{n\times p}\mathbf{a}_{p\times 1}$ where $\mathbf{a} \in \mathbb{R}^p$.

- Finding $\hat{\boldsymbol{\beta}}$ that minimizes $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is equivalent to finding a vector $\hat{\mathbf{y}}$ in $C(\mathbf{X})$, which minimizes $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$. Intuitively we know how to find such a vector: $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the subspace $C(\mathbf{X})$.

- The remaining vector $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ is the projection of $\mathbf{y}$ onto $C(\mathbf{X})^{\perp}$, the space orthogonal to $C(\mathbf{X})$.

- $C(\mathbf{X})$ is the estimation space: the LS coefficient $\boldsymbol{\beta}$ is retrieved from the projection of $\mathbf{y}$ onto $C(\mathbf{X})$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\hat{\mathbf{y}} + \mathbf{r}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{y}},$$

  where the last equality is due to $\mathbf{X}^T\mathbf{r} = \mathbf{0}_{p\times 1}$.

- $C(\mathbf{X})^{\perp}$ is the error space: we estimate the error variance σ^2 from the projection of $\mathbf{y}$ onto $C(\mathbf{X})^{\perp}$:

$$\hat{\sigma}^2 = \|\mathbf{r}\|^2/(n-p).$$

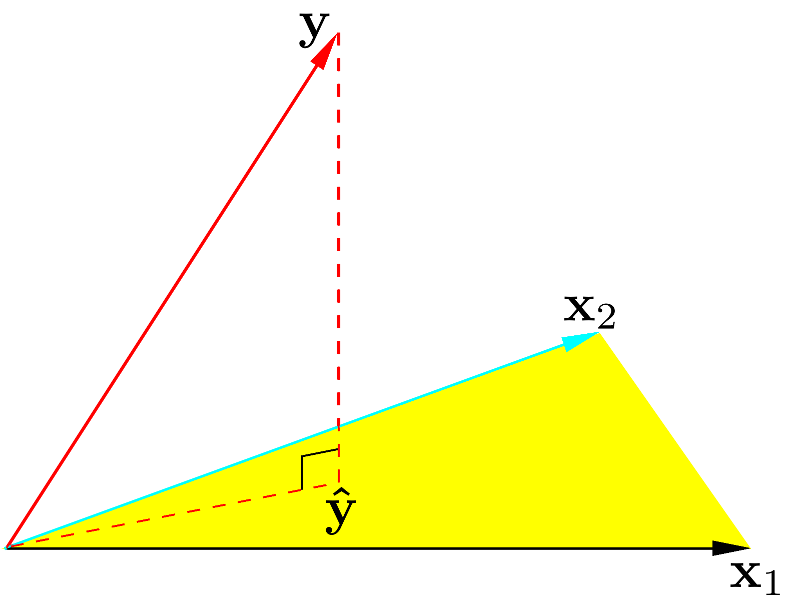

FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Goodness of Fit: R-square

We measure how well the model fits the data via R^2 (fraction of variance explained)

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2},$$

which is also equal to

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}.$$

Affine Transformation on X

- If we scale or shift a predictor, say, $\tilde{x}_{i2} = 2 \times x_{i2}$ or $(1 + x_{i2})$, how would this affect the LS fit?

- The estimated coefficients would be different since we change the predictors, but the fitted value $\hat{\mathbf{y}}$ stays the same, since any linear combination of the original predictors can be written as a linear combination of the new ones (and vice verse). Here we assume the intercept is always included.

- In general, if we apply any linear transformation on the p predictors (the new design matrix can be written as $\tilde{\mathbf{X}} = \mathbf{X}_{n \times p} A_{p \times p}$), then the fitted value $\hat{\mathbf{y}}$ and R^2 stay the same as long as the transformation does not change the rank of the design matrix.

Use R to Analyze the Boston Housing Data

- Basic command: `lm`
- How to 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.
- Note that the result from SLR might be different from the one from MLR: SLR suggests that age has a significant negative effect on housing price, while MLR suggests the opposite. Such seemingly contradictory statements are caused by correlations among predictors.
- How to handle rank deficiency?

Two-Stage LS

Partition the p predictors into two groups, and then partition the design matrix $\mathbf{X} = [\mathbf{X}_{1_{n \times p_1}}; \mathbf{X}_{2_{n \times p_1}}]$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1; \hat{\boldsymbol{\beta}}_2)$ accordingly.

We cannot obtain $\hat{\boldsymbol{\beta}}_2$ by regressing $\mathbf{y}$ onto $\mathbf{X}_2$. Instead we should run the following two-stage approach:

1. Regress $\mathbf{y}$ onto $\mathbf{X}_1$, and denote the residual vector as $\mathbf{u}_{n \times 1}$;
2. Regress each column of $\mathbf{X}_2$ onto $\mathbf{X}_1$, and concatenating the residual vectors as a new design matrix $\tilde{\mathbf{X}}_2$;
3. Finally regress the new response vector $\mathbf{u}$ onto $\tilde{\mathbf{X}}_2$.

The $p_2 \times 1$ coefficient vector returned at the last regression is $\hat{\boldsymbol{\beta}}_2$ and the residual is the same as the residual from regressing $\mathbf{y}$ onto $[\mathbf{X}_1, \mathbf{X}_2]$.

Partial Regression Coefficients

Consider a multiple linear regression model

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

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

The LS estimate $\hat{\beta}_k$ is what we could get if we (see Algorithm 3.1)

- first regress Y onto all other predictors except X_k, denote the the corresponding residuals as a new variable Y^*;
- regress X_k onto all other predictors except X_k, denote the corresponding residuals as a new variable X_k^*;
- then fit a simple linear regression model with Y^* as the response and X_k^* as the predictor.

Hypothesis Testing in Linear Regression Models

The key test is the F-test for comparing two nested models:

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

Two models are nested if the reduced model is a special case of the full model, e.g.,

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

The test statistic takes the following form

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

which follows an F distribution. (Note that $\mathsf{RSS}_a < \mathsf{RSS}_0$ and $p_a > p_0$.)

- 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, that is, the variation missed by the reduced model, when being compared with the error variance, is significantly large.

- 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 $\{S, M, L\}$.
- To include this categorical variable in a regression model, we need to generate two numerical variables. For example, for $n = 6$ obs with two taking each of the three values, we have

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

- 1st column: indicator for value "M".

- 2nd column: indicator for value "L".

- We do not need an indicator column for "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 for any categorical values with K different values, we need to generate $K-1$ binary vectors.

- We can also generate products of those indicator variables with other numerical variables to create the **interaction terms** between categorical and numerical features.

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 many of the predictors are highly correlated. In this case, the contribution of a particular predictor could be masked by other predictors, which create difficulties for statistical inference.

- 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 predictor is significant.

- What to do with collinearity? Remove some predictors or combine collinear predictions (e.g., PCA).

- Check the `seatpos` data.

Assumptions for Linear Regression Models

- We "assume" $\mathbb{E}(Y \mid X = x)$ is 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^*.

- We assume the error terms at all x_i's are uncorrelated with mean zero and constant variance[a]. 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.

- We do not need to impose any assumptions on x's. But to achieve a good performance, we would like x_i's to be uniformly sampled.

[a] A stronger version is error terms iid $\sim N(0, \sigma^2)$, but the Normal assumption can be relaxed.

Abnormal Points

- What's outlier? What's high-influential point?
- Datasets from real applications are usually large (in terms of both n and p), so do not recommend to test outliers or check high-influential points.
- 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.