Regression:
Multiple Regression

CS 418. Introduction to Data Science

Linear Regression
Multiple Linear Regression

- **Multiple linear regression** models a **linear relationship** between one quantitative **response variable** and more than one **predictor variable**.
- The **response variable** is the variable being modeled or predicted and the **predictor variables** are the variables used to predict the response.
- The **multiple linear regression** model is

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p + \varepsilon}$$

  where:
  - Y is the **response variable**.
  - $X_1, X_2, \ldots, X_p$ are the **predictor variables**.
  - $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the **regression parameters**.
  - ε is the **regression error term**.

Multiple Linear Regression
Assumptions

- A **multiple linear regression** model makes the following assumptions:
  - **Linearity**.
    - There is a **linear relationship** between the response variable and the predictor variables.
  - **Normality**.
    - Errors are **normally distributed** with a **mean of 0**.
  - **Homoscedasticity**.
    - Errors have **constant variance**.
  - **Independence**.
    - Errors are **independent** of each other.
  - **No collinearity**.
    - Predictor variables are **linearly independent**.

Multiple Linear Regression
Adjusted R^2

- R^2 measures the **proportion of variance** of the response variable that is **explained** by the model. The value of R^2 is **between 0 and 1**. The higher the value, the better the fit of the model.
- The value of R^2 never decreases when adding more predictor variables.
- The **adjusted** R^2 is a modified version of R^2 that has been adjusted for the number of predictor variables in the model.

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

  where n is the sample size and p is the number of predictor variables.
- The value of the **adjusted** R^2 increases only if the new predictor variables improve the fit of the model more than would be expected by chance. Otherwise, it decreases.
- The value of the **adjusted** R^2 can be **negative**.

Multiple Linear Regression
AIC and BIC

- The **Akaike information criterion** (**AIC**) and the **Bayesian information criterion** (**BIC**) are two commonly used criteria for **model comparison** and **selection**.

$$AIC = 2p - 2\ln L$$

$$BIC = ln(n)\,p - 2\,ln\,L$$

  where L is the likelihood function, p is the number of predictor variables, and n is the sample size.

- The model with the **lowest AIC** or **BIC** is preferred.
- Both **AIC** and **BIC** penalize models with **more predictor variables**, but **BIC** penalizes model **complexity** more heavily.
- In general, it might be best to use **AIC** and **BIC** together, although, in some cases, **AIC** may choose a more complex model than **BIC**.

Multiple Linear Regression
Categorical Predictors

- A **categorical predictor variable** has qualitative values representing one of a finite number of categories.
- A **multiple linear regression** model can incorporate categorical predictors if these are recoded into one or more **dummy variables**.
  - A **dummy variable** is a binary variable that is 1 for a specified category of a categorical predictor and 0 for all other categories.
  - ***Example*****:** for a "gender" predictor variable, a "male" dummy variable is 1 for males or 0 for females, while a "female" dummy variable is 1 for females or 0 for males.
  - To avoid **collinearity**, one dummy variable must be omitted from the model. The category represented by the omitted dummy variable is known as the **reference level** of the categorical predictor.

Multiple Linear Regression
Nonlinear Relationships

- **Nonlinear relationships** can be incorporated to a **multiple linear regression model** by using **transformations** of the response and/or the predictor variables.
  - **Polynomial regression:**
    $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$
    Used when there is a **polynomial** (e.g., quadratic) relationship between the response and the predictor variable.
  - **Interaction model.**
    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$
    Used when the effect of a predictor variable on the response depends on the values of another predictor variable.
  - **Logarithmic variable transformation.**
    $$\log Y = \beta_0 + \beta_1 X + \varepsilon$$
    Used when the distribution of a variable is **skewed** (that is, higher values are more spread out than lower values).

Multiple Linear Regression
Regularization (I)

- **Regularization techniques** add an additional constraint (or **penalty**) to the regression model to **discourage large coefficients**.
  - **Ridge regression** or regression with $\boldsymbol{L_2}$ **regularization** minimizes the sum of squared errors plus the **sum of the squares of the coefficients**. That is,

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{m}\beta_j X_{ij}\right)^2 + \alpha\sum_{j=0}^{m}(\beta_j)^2$$

    where α is a user-defined parameter.
  - **LASSO (least absolute shrinkage and selection operator) regression** or regression with $\boldsymbol{L_1}$ **regularization** minimizes the sum of squared errors plus the **sum of the absolute values of the coefficients**. That is,

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{m}\beta_j X_{ij}\right)^2 + \alpha\sum_{j=0}^{m}|\beta_j|$$

    where α is a user-defined parameter.

Multiple Linear Regression
Regularization (II)

- **Regularization techniques** add an additional constraint (or **penalty**) to the regression model to **discourage large coefficients**.
  - In **LASSO regression**, some coefficients will become 0 and can be dropped.
  - In **ridge regression**, the coefficients will approach 0, but will not be dropped.
  - **Elastic net** linearly combines the L_1 and L_2 penalties of **ridge regression** and **LASSO regression**. That is,

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{m} \beta_j X_{ij}\right)^2 + \alpha\left(\lambda \sum_{j=0}^{m}\left(\beta_j\right)^2 + (1-\lambda)\sum_{j=0}^{m}\left|\beta_j\right|\right)$$

where α and λ are user-defined parameters.

Multiple Linear Regression
References

- Daniel Chen. *Pandas for Everyone* (2018).
- Joel Grus. *Data Science from Scratch* (2015).
- Cathy O’Neil and Rachel Schutt. *Doing Data Science* (2013).