Chapter 3

Multiple Linear Regression

STAT303-2

Northwestern | WEINBERG COLLEGE OF ARTS & SCIENCES

For Simple Linear Regression: (SLR)

- We had one predictor: X
- The assumption for the underlying relationship was: $Y = \beta_0 + \beta_1 X + \epsilon$

Multiple Linear Regression: (MLR)

- A generalization of the same idea to multiple predictors: $X_1, X_2, \ldots, X_p$
- The linear assumption still there – now between $X_1, \ldots, X_p$ and Y.
- The (assumed) underlying relationship becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$car\ price \approx \beta_0 + \beta_1 * mileage + \beta_2 * mpg + \cdots + \beta_p * engineSize$$

Multiple Linear Regression - Training

The Multiple Linear Regression (MLR) model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Each predictor has its own coefficient for MLR.
- SLR had a line with the slope as the coefficient of the only predictor.
- What we fit for MLR is called a **hyperplane.**
- Training the model is very similar to SLR – same method, more derivatives.

- We need to find the optimum parameters - $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ - using the formula:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

And the data:

Careful with the notation! $\begin{bmatrix} X_1^1 \\ X_2^1 \\ \ldots \\ X_p^1 \end{bmatrix} \leftarrow$ $X^1, X^2, \ldots X^N$

$Y^1, Y^2, \ldots Y^N$

Multiple Linear Regression - Training

RSS is set up:

$$RSS = e_1^2 + e_2^2 + \cdots + e_N^2$$

where N is the total number of observations.

We can write it as:

$$RSS = \left(Y^1 - \hat{Y}^1\right)^2 + \left(Y^2 - \hat{Y}^2\right)^2 + \cdots + \left(Y^N - \hat{Y}^N\right)^2$$

which, for MLR, becomes:

$$RSS = \left(Y^1 - \hat{\beta}_0 - \hat{\beta}_1 X_1^1 - \cdots - \hat{\beta}_p X_p^1\right)^2 + \left(Y^2 - \hat{\beta}_0 - \hat{\beta}_1 X_1^2 - \cdots - \hat{\beta}_p X_p^2\right)^2 + \cdots + \left(Y^N - \hat{\beta}_0 - \hat{\beta}_1 X_1^N - \cdots - \hat{\beta}_p X_p^N\right)^2$$

- Find the parameters that minimize RSS – optimum parameters
    - Take the partial derivative with respect to each parameter.
    - Set all to zero.
    - Solve for the optimum coefficients.

Multiple Linear Regression - Prediction

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- We have a trained/optimized MLR model after the optimum parameters are found.
- To predict the response for a new/test observation(s**) with the same predictors**, just plug in each predictor value to the corresponding X.
- Find RMSE/MAE and/or some visualization.

Multiple Linear Regression – Inference

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$car\ price \approx \beta_0 + \beta_1 * mileage + \beta_2 * mpg + \cdots + \beta_p * engineSize$$

Same insights as SLR:

- The parameters themselves →
- Uncertainty
  - Confidence intervals
  - Prediction intervals
  - Statistical significance
- How much $\hat{f}$ explains the variation in the data
  - Coefficient of Determination (R^2)

- For $0 \leq j \leq p$,
- β_j is the average effect on Y of a unit amount of increase in X_j holding all the other predictors fixed.
- Note that while i is the common index for the observations, j is the common index for the features/variables/dimensions.

Multiple Linear Regression – Uncertainty

Same idea as SLR:

- Find the $SE(\beta_j)$, $0 \leq j \ \leq p$.
  - Calculate 95% confidence intervals (CIs) for each parameter.
  - Find statistical significance.

$$[\hat{\beta}_0 - 2 * SE(\hat{\beta}_0), \hat{\beta}_0 + \ 2 * SE(\hat{\beta}_0)]$$

$$[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + \ 2 * SE(\hat{\beta}_1)]$$

$$\ldots$$

$$[\hat{\beta}_1 - 2 * SE(\hat{\beta}_p), \hat{\beta}_1 + \ 2 * SE(\hat{\beta}_p)]$$

Note: Standard errors of regression coefficients in MLR:

- Similar derivation, assuming constant variance of random error (ϵ), just as in SLR
- Not getting into the formulas this time – introduced the idea in SLR and we will use Python for the rest
- Have the same relationship with the number of observations in the data and the error variance as the standard errors in SLR
  - Higher error variance, (hence, RSS) higher SE
  - Higher number of observations with constant RSS, lower SE

Note that predictions intervals for each coefficient are still valid!

Multiple Linear Regression – Uncertainty

Same idea as SLR:

- Find the $SE(\beta_j)$, $0 \leq j \leq p$.
  - Calculate 95% confidence intervals (CIs) for each parameter.
  - Find statistical significance.

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Mostly similar to SLR:

- Find the t-statistic for each coefficient. (β_j , $0 \leq j \leq p$)
- Find the p-value for each coefficient.
- Determine statistical significance **for each coefficient.**

An important addition to this for MLR: F-test

Is there an underlying linear relationship between the response and **all** the predictors?

Multiple Linear Regression – F-test

$\mathbf{H_0}: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$\mathbf{H_A}$: At least one β_j , $1 \leq j \leq p$ is non-zero.

$$\underbrace{\sum_{i=1}^{N}(Y^i - \bar{Y})^2}_{\text{Total sum of squares}} \qquad \underbrace{\sum_{i=1}^{N}(Y^i - \hat{Y}^i)^2}_{\text{Residual sum of squares}}$$

This hypothesis test is performed by computing the F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(N - p - 1)}$$

- F, just like t, is a Random Variable that belongs to a distribution.
- The distribution of F is called the F-distribution.
- Calculate the F-statistic plugging in the necessary values after training.
- Check where the F-statistic falls in the distribution and find the **probability** of that value. **(using Python)**
  - **This probability is the probability of the predicted non-zero coefficients being due to random chance.**

Question: Given that we already have individual p-values for each variable, why do we need the overall F-statistic?

- Consider number of predictors, *p = 100*, and assume that H_0: $\beta_0 = \beta_1 = \cdots = \beta_p = 0$ is true. Can you answer now?
- About 5% of the p-values associated with each variable will be below 0.05 by chance. In other words, we expect to see approximately five *small* p-values even in the absence of any true association between the predictors and the response.
- However, F-statistic adjusts for the number of predictors, and thus does not suffer from this problem.
- There is only 5% chance that the F-statistic will result in a p-value below 0.05, regardless of the number of predictors.

Multiple Linear Regression – F-test

An important extension of the F-test with **all the predictors**:

Is there an underlying linear relationship between the response and a **subset of *q*** predictors?

The residual sum of squares for the model that has all the predictors outside the subset. **(Train another model for this.)**

$\mathbf{H_0}: \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$

$\mathbf{H_A}$: At least one β_j , $p - q + 1 \leq j \leq p$ is non-zero.

Again, compute the F-statistic, with a slightly different formulation this time: $F = \dfrac{(RSS_0 - RSS)/q}{RSS/(N - p - 1)}$

- This test may help in variable selection
- Suppose we have a large number of predictors, *p*.
- However, a few predictors are highly statistically significant *(p-value<<0.05).*
- In such a case, we may test if the model with only the highly significant predictors is sufficient to explain the response
- This may eliminate a large number of unnecessary predictors from the model

Multiple Linear Regression – R^2

- The same idea and formula – how much $\hat{f}$ explains the variation in the data.

$\hat{Y}^i = \hat{\beta}_0 + \hat{\beta}_1 X_1^i + \hat{\beta}_2 X_2^i + \cdots + \hat{\beta}_p X_p^i$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}[Y^i - \hat{Y}^i]^2}{\sum_{i=1}^{N}[Y^i - \bar{Y}]^2} = 1 - \frac{RSS}{TSS}$$

- For SLR, R^2 was the square of Pearson correlation coefficient – correlation between X and Y.

$$R^2 = Cor(Y, X)^2$$

For MLR, R^2 is the square of **the correlation between the real response and the predicted response.**

$$R^2 = Cor(Y, \hat{Y})^2$$

Reference

Source for slides: https://www.statlearning.com/resources-second-edition