Multiple Linear Regression – Workflow and Considerations

Introduction

* **OLS**: In statistics, ordinary least squares (OLS) is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed responses (values of the variable being predicted) in the given dataset and those predicted by a linear function of a set of explanatory variables.
  + **RSS minimization**: In determining ordinary least squares estimates, the objective is to estimate coefficients that minimize the residual sum of squares.
  + **Gradient descent**: Gradient descent is one method to find the minimum of a function through trial and error. We use it in most supervised learning to minimize an objective function (assume this objective function looks like a canyon where the bottom of the canyon has the lowest error). For example, we use gradient descent in linear regression to minimize the RSS. To find a linear regression model, we can first choose random coefficients for our features. This corresponds to a point on the objective function. To decrease the error, we can take the derivative of the cost function at that point and change our coefficients in the direction of that gradient. The learning rate and the magnitude of the derivative determine how big of a step we take in that direction. Once the derivative becomes ≈0, we are at the local minimum and have found the coefficients for our linear regression that produce the least error.
* **Pros/cons of using MLR**: Good interpretability but low accuracy if there is no linear trend, model can be invalid if assumptions are violated.
* **Bias-variance tradeoff**: Compared to other machine learning models, a multiple linear regression is considered a simpler model, and thus have higher bias and lower variance. However, when the number of predictors are high, variance increases and the issue of overfitting can occur. In this scenario, Ridge or Lasso regularization can be used to reduce variance at the cost of increasing bias.

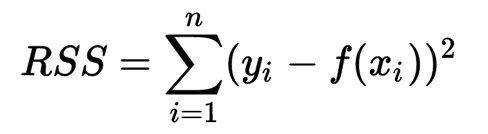
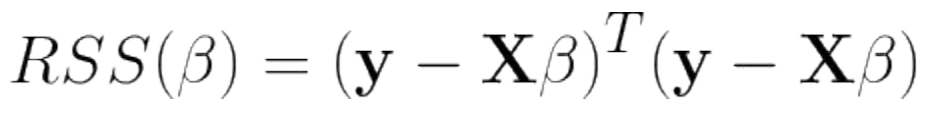
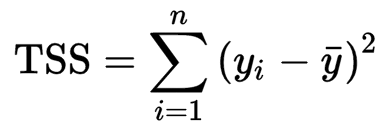
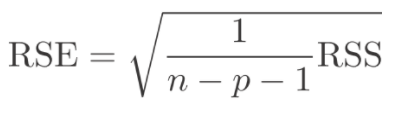
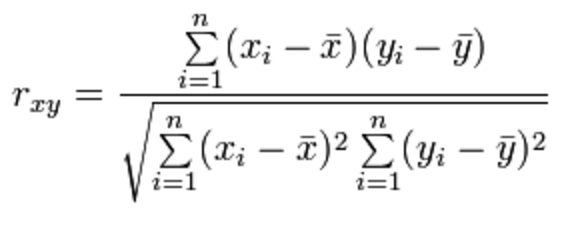
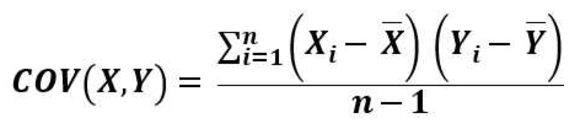
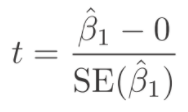
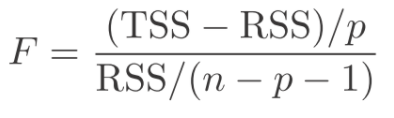
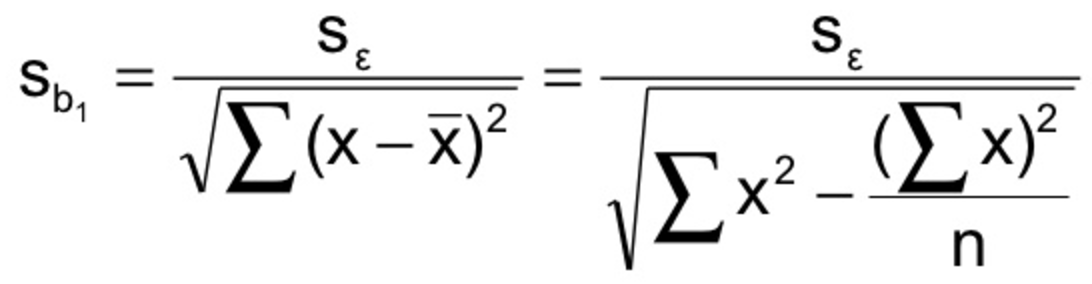
Data Preparation

* **Tidy data**: Each variable forms a column. Each observation forms a row. Each type of observational unit forms a table.
* **No missing values:** A multiple linear regression does not accept missing values. They must be treated prior to fitting the model.
* **Variable scaling**: While scaling allows you to compare all variables relatively, you lose interpretability in the process.

Variable Selection

* **F-test and variable selection:** If the F-statistic indicates significance in at least one of the predictors, one can examine the individual coefficient p-values to determine which predictors are significant. However, when the number of predictors is high, we are expected to falsely identify some of the predictors as significant. Therefore, other statistics such as R2, AIC, or BIC can be examined instead.
* **AIC/BIC**: AIC and BIC are both penalized-likelihood criteria. They can be used for choosing best predictor subsets in regression. The AIC or BIC for a model is usually written in the form [-2logL + kp], where L is the likelihood function, p is the number of parameters in the model, and k is 2 for AIC and log(n) for BIC. Computationally, BIC penalizes model complexity more heavily. The only way they should disagree is when AIC chooses a larger model than BIC.
* **Forward selection**: When using forward selection, we begin with a null model and then iterative add single predictors which minimizes the chosen measure, such as RSS, AIC, or BIC.
* **Backward selection**: When using backward selection, we begin with a saturated model and then iterative remove single predictors which minimizes the chosen measure, such as RSS, AIC, or BIC.
* **Mixed selection**: When using mixed selection, we begin with a null model, and then iterative add single predictors which minimizes the chosen measure, such as RSS, AIC, or BIC. In each iteration, if any of the predictors’ p-value increases beyond a certain threshold, then it will be removed.

Model Assessment

* **R2**: Also known as the coefficient of determination, the R2 value is a measure of how well your model explains the variability in the data. The closer to 1.0 your R2 value is, the better your linear regression models explains the trend. It is calculated in the following way: R2= (TSS – RSS)/TSS = 1 – (RSS/TSS). TSS measures the variability inherent in the response variable before the regression is performed. RSS measures the amount of variability that is left unexplained after performing the regression. If the regression does a perfect job in explaining all of the variability inherent in the response variable, then RSS approaches 0, and R2 approaches 1, indicating that 100% of the variability in the response variable *Y* can be explained using the regression *X*.
* **R2adj**: Adding more features will always spuriously increase unadjusted R2, so the adjusted R2 considers the number of features and the size of the dataset with the equation: R2adj = 1 - [RSS/(n-p-1)] / [TSS/(n-1)]. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.
* **RSS**: Residual sum of squares, also called SSE (sum of squared errors). For a simple linear regression, it is calculated by: . For a multiple linear regression, it is calculated by : . This is essentially the sum of squared distances between the predicted response values and the actual response values.
* **TSS**: Total sum of squares. This is the sum of the squared differences between the response variable and its mean. It is calculated by: 
* **RSE**: Residual standard error. Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is computed using the formula: , where n = number of observations, and p = number of predictors. While adding predictors can lower the RSS, the RSE may increase, making estimations/predictions less stable.
* **MSE**: Mean squared errors. It is calculated as such: RSS/n, where n = number of observations.
* **Correlation**: Measure of the linear relationship between two or more variables. For two variables, it can be calculated as such:.
  + **Correlation Matrix**: Tool used to examine pairwise correlations among two or more variables.
  + **Variance Inflation Factor (VIF)**: The VIF is the ratio of the variance of *B*j when fitting the full model divided by the variance of *B*j if fit on its own. Oftentimes, it is used with a correlation matrix to determine which predictors to remove in the presence of predictors with high VIFs.
* **Covariance**: Measure of positive or negative linear relationship between two variables. It can be calculated as .
* **Evaluation of Coefficients Estimates**
  + **T-test**: A t-test is an analysis of two population means using statistical examination. In a multiple linear regression, the t-test is used to determine if a predictor’s coefficient is significantly different from 0. The t-statistic can be calculated using the formula: . This formula measures the number of standard deviations that *B1* is away from 0. Once the t-statistic is evaluated, it can be used along with the degrees of freedom to determine a p-value. Generally speaking, a p-value of less than 0.05 indicates that *B1*is significantly different from 0, or that there is a significant relationship between the predictor and the response variable.
  + **F-test**: An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. In a multiple linear regression, the F-test of Overall Significance tests whether any of the independent variables in a multiple linear regression model are significant. It is calculated by: , where p = number of predictors. Should the F-test of Overall Significance indicate significance in at least one of the predictors, t-tests should be used to identify which predictors are significant predictors to the response variable.
  + **P-value**: P-value is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true.
  + **Standard error of coefficients**: A measure of the statistical accuracy of an estimate, equal to the standard deviation of the theoretical distribution of a large population of such estimates. In a simple linear regression with only one predictor, the standard error of the slope can be estimate using the formula: .

Model Assumptions

* Homoscedastictiy (or constant variance in residuals)
  + Examine residual plot
  + If heteroscedastic, perform transformation (e.g. Box-Cox, log, power)
* Normal distribution of residuals
  + Examine normal probability plot
  + If non-normal, try to subset data or remove outliers
* Linearity
  + Examine correlation matrix
  + If non-linear, try to add non-linear terms
* Independent errors
  + If errors are dependent, try to remove predictors
* No or little multicollinearity (additivity)
  + If multicollinearity is present, remove predictors with VIF > 5
* No outliers
* No high leverage points

Model Validation

* Cross validation

Predictions

* Confidence interval
  + Coefficients
  + Prediction estimates
* Prediction interval
  + Prediction estimates