1 Linear Regression

1.1 Concept

Linear regression is a parametric model that is additive and linear in the provided features. It is a classic technique used in many fields, and its widespread popularity greatly pre-dates the popularity of machine learning. Its general form is

\[ \hat{y} = \hat{\beta}X \]  

Where \( \hat{y} \) is a vector of predicted \( y \) values and \( X \) is a matrix whose rows correspond to observations and whose columns correspond to features.

When there is only one feature on the right hand side, the model is called a “simple linear regression.” When there are multiple features on the right hand side, the model is called “multiple linear regression.”

When used for inference, we are interested in \( \hat{\beta} \). However, when used for prediction, we are only interested in \( \hat{y} \), and we cannot say that the \( \hat{\beta} \)s reflect any sort of causal relationship between the features and the outcome. For more information on how to test the significance of regression coefficients, please see Chapter 3 of ISLR for a reference on \( t \)-tests (in the simple model) and \( F \)-tests (in the multivariate model).

1.2 Method

To find the coefficients \( \hat{\beta} \) in a linear regression, we find the value of \( \hat{\beta} \) that minimizes the residual sum of squares (RSS) in the training data. The classic formula for \( \hat{\beta} \) uses matrix algebra and is

\[ \hat{\beta} = (X'X)^{-1}X'y \]  

We will estimate \( \hat{\beta} \) using statistical software.

It is worth noting that the traditional measure of fit for linear regression is \( R^2 \), but \( R^2 \) mechanically increases with the inclusion of additional features. Therefore, in the prediction setting, the \( R^2 \) on the training data is less important than the mean squared error (MSE) on the test data.

1.3 Implementation and Considerations

There are a few things to watch out for as far as the features that you feed into a linear regression.

- There must be fewer features than observations. Later in the semester, we will cover penalized regression methods that do variable selection to yield estimable linear models, even when the number of available features exceeds the number of observations. Common penalized regression methods are lasso and ridge regression.

- You can use quantitative or qualitative features for the \( X \)s. When using qualitative features, generate indicator variables for all but one category. The omitted category will serve as the “baseline,” meaning that the coefficients on the included categories can be thought of as the differential effect of being in that category compared to the baseline (omitted) one.
The reason you omit one category when making indicator variables is to avoid linear dependence. If all categories were represented, the indicator columns would all sum to 1, which would mean they were linearly dependent. More generally, you cannot have collinearity or multi-collinearity, which means you cannot have features that are (close to) perfectly correlated.

You can interact two features (e.g. create a feature that is the product of two other features), and such interactions are valid on categorical and continuous features. However, when you include an interaction, you should also include each of the features on their own as well. Interactions have intuitive appeal if you think there are synergies between two features in terms of their effect on $y$.

You can exponentiate features and include the exponentiated features in your model. The resulting model is sometimes called polynomial regression and is appropriate when there appears to be a non-linear relationship between a feature and the outcome.

Check for influential points – those that are both outliers (they have an unusual or extreme $y$ value) and high leverage (they have an unusual or extreme $x$), as these points can greatly influence the model fit. You may want to exclude them or at least check your model’s sensitivity to including them versus excluding them.

If you are interested in inference (e.g. looking at the $\hat{\beta}$s to understand a causal relationship), it is important to be aware of Omitted Variables Bias (OVB). OVB occurs when you have two correlated features that each have an effect on $y$ but only one is included in the regression. In that case, the coefficient on the included feature is biased, because it is partially picking up the true effect of the feature on the outcome and is also partially picking up the effect of the omitted feature on the outcome (since the omitted feature is correlated with the included feature).

As an example of OVB, suppose $X_1$ and $X_2$ are positively (but not perfectly) correlated. If they are also both positively correlated with $y$, then when $X_2$ is omitted from the regression, the coefficient on $X_1$ will be higher than when both $X_1$ and $X_2$ are included. This is because the coefficient on $X_1$ will now pick up both the effect of $X_1$ on $y$ and part of the effect of $X_2$ on $y$ (since $X_1$ is a proxy for $X_2$ because the two features are positively correlated).

When implementing linear regression, you should also take a look at your residuals and make sure there are no red flags:

- When you plot residuals, they should appear randomly scattered. Any non-linearity or patterns in the residuals suggest your model is not appropriate.
- Linear regression assumes residuals are uncorrelated. Evidence of correlated residuals indicates a problem with your model or your data that should be investigated.
- Residuals should have constant variance. If you plot your residuals and their variance seems to be a function of $x$, then the errors are heteroskedastic (a fancy word for “a function of $x$”). In this case, traditional statistical measures of significance are invalid, but other valid methods are available.

1.4 Comparison to KNN

The main difference of note between linear regression and KNN is that linear regression is a parametric model whereas KNN is a non-parametric model. There are a few general differences between parametric and non-parametric models that are worth noting:

- Non-parametric models are more flexible whereas parametric models impose stronger assumptions
- When there is a small number of observations per feature, parametric models tend to outperform non-parametric models

In addition to the general differences between parametric and non-parametric models, a key difference between KNN and linear regression is that linear regression is quite simple to fit. In fact, it only requires estimation of a few $\beta$s, whereas KNN is much more computationally intensive. Because of this simplicity, linear regression is also more interpretable than KNN.