NYC DATA SCIENCE ACADEMY

Multiple Linear Regression

Data Science Bootcamp

Outline

- Part 1: Multiple Linear Regression
- Part 2: Assumptions & Diagnostics
- Part 3: Research Questions of Interest
- Part 4: Extending Model Flexibility
- Part 5: Review

PART 1

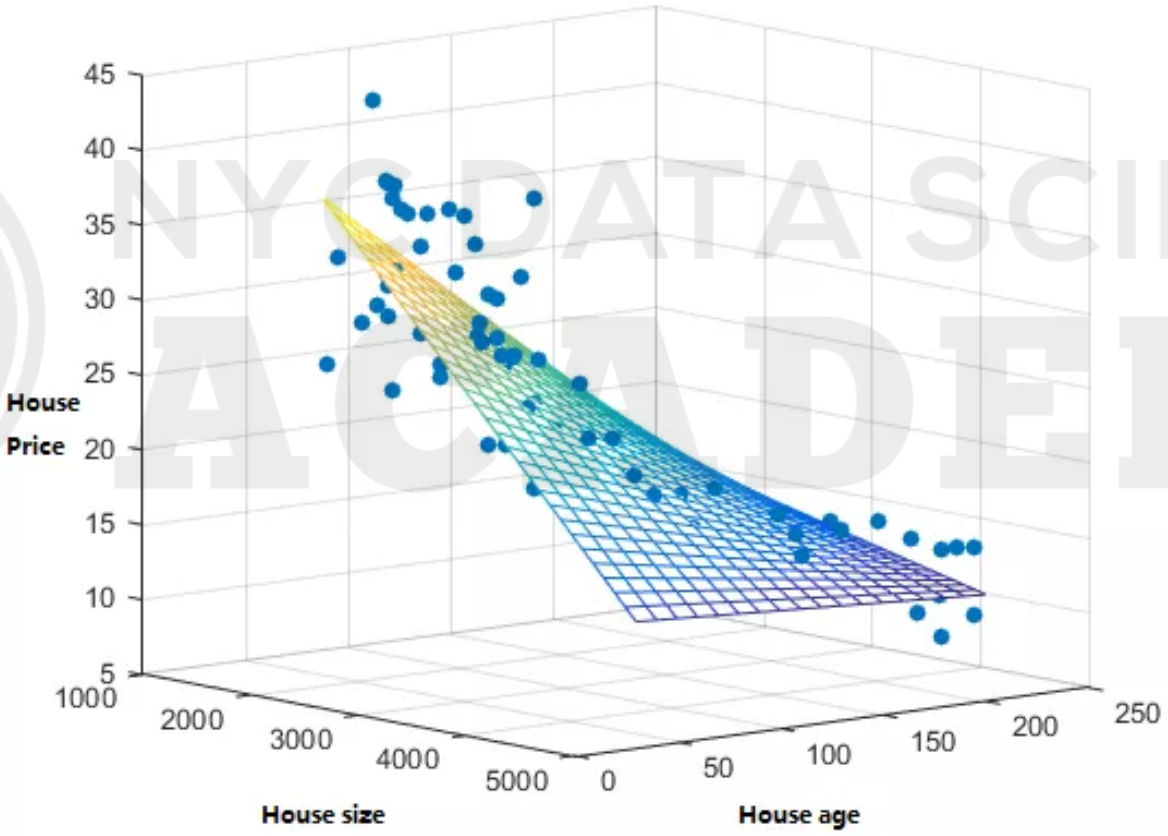
Multiple Linear Regression

What is Multiple Linear Regression?

- Multiple linear regression is a supervised machine learning method that aims to uncover a linear relationship between a set of variables X_i and a single outcome variable Y:
  - The explanatory/independent/input variables X_1, X_2, X_3, …, X_p.
  - The response/dependent/output variable Y.

- The ultimate goal is to use this relationship to make predictions about observations not within our dataset. We answer the question:
  - If I have the values of X_1, X_2, X_3, …, X_p, what should my best guess for Y be?

What is Multiple Linear Regression?



Image

What is Multiple Linear Regression?

- Why can't we just fit p simple linear regression models, one for each of our X_i variables, instead of using the multiple linear regression procedure?
  - While theoretically this approach can be successful, it is highly inefficient.
  - How would you make a single prediction from a slew of different models, all which yield different predictions for the outcome?
  - All of these separate models would inherently ignore the other predictor variables during their construction of the coefficient estimates, which can lead to misleading results.

- Multiple linear regression is an extension of simple linear regression; it accommodates the flexibility of considering more than one explanatory variable at a time.

Multiple Linear Regression: Mathematically

- Ultimately, we wish to quantify the relationship as follows:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- Once again, note that this equation is linear in form.

- In order to regress Y onto X_i, we need to estimate $(p + 1)$ coefficients/parameters:
  - β_0: The intercept of the surface; the expected value of Y when all X_i are 0.
  - β_i: The slope of the i^{th} predictor; the expected change in Y when X_i shifts by one unit and all other variables are held constant.

- Once we have estimates for β_0 and all β_i, we can use the equation above to estimate the value of Y given the values of X_i.

Multiple Linear Regression: Mathematically

- The prediction for Y based on the i^{th} values of X is as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

- Recall the difference between the i^{th} observed response (the actual value) and the i^{th} response prediction (the estimated value) is called the residual or error e_i:

$$e_i = y_i - \hat{y}_i$$

- We would like the residual to be as small as possible for all observations in our dataset.
  - In other words, this is essentially an optimization problem where we would like to minimize error as much as possible.
  - How do we do this now that we have (p + 1) coefficients to estimate?

Multiple Linear Regression: Mathematically

- Consider the sum of the squared error terms for each observation in our dataset. We call this the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2$$

- For multiple linear regression, this is equivalent to:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Multiple Linear Regression: Mathematically

- Task: find the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ that reduce the sum of the squared vertical distances from the observations to the regression surface (i.e., the RSS) as much as possible.

- Procedure: derive formulas for these estimates using basic linear algebra.

Multiple Linear Regression: Mathematically

- Denote $\mathbf{X}$ as the n by $(p + 1)$ matrix with each row as an observation in our dataset; note that the first column will, by default, be a vector of 1's as a placeholder for the β_0 intercept term.
- Denote $\mathbf{y}$ as the n by 1 vector with each entry representing the response values in our dataset.

- We can rewrite the residual sum of squares in matrix notation as follows:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- This is a quadratic function with $(p + 1)$ parameters.

Multiple Linear Regression: Mathematically

- Differentiating with respect to β, we obtain the following:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}$$

- Because the second derivative is necessarily positive, we know that this estimate is a minimum.

Multiple Linear Regression: Mathematically

- Assume that $\mathbf{X}$ is of full column rank, and thus that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is positive definite. We can then set the first derivative equal to 0 as follows:

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \stackrel{!}{=} 0$$

$$\Rightarrow \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\beta = 0$$

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta$$

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

Multiple Linear Regression: Mathematically

- Thus, the least squares coefficient estimates for multiple linear regression are given by:

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

- Given our data, these are the best estimates for $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ as they ensure the sum of the squared vertical distances from the observations to the regression line (i.e., the RSS) is at a minimum.

- To obtain the fitted values of the regression, we simply pass our data matrix $\mathbf{X}$ through the matrix of coefficient estimates as follows:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

PART 2

Assumptions & Diagnostics

Assumptions of Multiple Linear Regression

- Recall that we wish to quantify the relationship between X_1, X_2, X_3, …, X_p and Y; however, even though we now have multiple predictors incorporated into our model, it is likely that in reality there are still other factors that tend to influence the behavior of Y, and thus other sources of variability.

- In multiple linear regression, we still have to account for these unmeasured discrepancies by including the error term ϵ:

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

- Recall that the error term accounts for the fact that the statistical model does not yield an exact fit to the data.

Assumptions of Multiple Linear Regression

- Recall that for simple linear regression, the validity of our model depends on the following assumptions:
  - Linearity
  - Constant Variance
  - Normality
  - Independent Errors

- For multiple linear regression, these assumptions should be checked and confirmed for each predictor variable to help ensure a valid model.

- Furthermore, a valid multiple linear regression model depends on an additional assumption of little to no multicollinearity between the predictor variables.

Assumptions of Multiple Linear Regression: Multicollinearity

- What does it mean?
  - The assumption of multicollinearity describes the relationships among the independent variables in our model. To satisfy this assumption, we need to verify that our independent variables are uncorrelated with each other. If they are not, our coefficient estimates could be unstable.

- How do we check?
  - We can inspect scatterplots and correlations between pairs of independent variables to see if there appears to be a relationship among them.
  - Once the multiple regression is fit, we can also inspect the variance inflation factors of each predictor. (We will learn about this shortly.)

Assumptions of Multiple Linear Regression: Multicollinearity

- Why care about multicollinearity? Why is it unwanted?
  - If two or more variables are highly correlated, they essentially contain the same information and introduce redundancies within our predictors.
  - If variables are multicollinear, then it is more difficult to make inferences about the relationships between our predictors and response variable (i.e., it is hard to separate the effect of each individual variable).

- What can happen if multicollinearity exists?
  - The standard errors of our estimates will be inflated.
  - The power and reliability of our regression coefficients will decrease.
  - It can create a need for a larger sample size.

- Ultimately, with a violation in multicollinearity the interpretations of the model coefficients would be faulty and misleading.

Accuracy of the Coefficient Estimates

- Consider a multiple regression model with only two predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Let ϱ_{12} denote the correlation between X_1 and X_2, and S_{xj} denote the standard deviation of X_j. It can be shown that:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)S_{x_j}^2} \times \frac{1}{1-\rho_{12}^2}, \quad j = 1, 2$$

- Consider the first fractional component:
  - σ^2 represents the overall scatter around the regression surface. A higher σ^2 implies a higher variance in the coefficient estimate.
  - n represents the sample size. A greater sample size implies a lower variance in the coefficient estimate.
  - S_{xj} represents the variability of a particular covariate. Greater variability in a single predictor implies a lower variance in the coefficient estimate.

The Variance Inflation Factor

- What about the remaining piece? This second part of the parameter estimate variance is called the variance inflation factor. For our two-predictor example:

$$VIF_{x_1 x_2} = \frac{1}{1-\rho_{12}^2}$$

- The VIF is a measure of how the model factors influence the uncertainty of the coefficient estimates.

- Notice that this value increases as the absolute value of the correlation increases:
  - Thus, correlation amongst the predictors implies a greater variance in the estimated regression coefficients. Multicollinearity is bad!

The Variance Inflation Factor

- Extending this idea to a general multiple regression model with p predictors, we have the following:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)S_{x_j}^2} \times \frac{1}{1-R_j^2}, \quad j = 1, 2, \ldots, p$$

- Here, R_j^2 denotes the R^2 value obtained from the regression of X_j onto the other (p - 1) predictors (i.e., the amount of variability in X_j explained by all the other predictor variables).

- A higher R_j^2 value implies greater uncertainty in our coefficient estimate, and ultimately is an indicator of multicollinearity with the variable in question.

The Variance Inflation Factor

- The VIF tells us the factor by which the estimated variance of a predictor variable is larger in comparison to if the predictor variable were to be completely uncorrelated with the other variables in the model.

- In general, predictors that have VIF values that exceed 5 need to be dealt with; these variables' VIFs indicate the corresponding regression coefficients are poorly estimated due to multicollinearity.
  - The information of predictors with high VIF values is contained within the remaining predictor variables in the model; having a high VIF makes a predictor a candidate to be removed from the model entirely.

PART 3

Research Questions of Interest

Typical Research Questions

- Following a successfully fitted multiple regression, the typical questions to which we would like answers are:
  - Are the predictor variables (X_1, X_2, X_3, …, X_p) useful in helping to predict the response Y?
    - Are any of the variables useful?
    - Is just a subset of the variables useful?
  - How well does our model fit the data?
    - How can we quantify and assess the extent of this relationship?
  - If I have the values of X_1, X_2, X_3, …, X_p for a specific observation, what should my best guess for Y be?
    - How accurate is this prediction?

The Overall F-Test: Is There a General Relationship?

- For multiple linear regression, the principal hypothesis test is as follows:
  - Null Hypothesis (H_0): $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
  - Alternative Hypothesis (H_A): At least one of the β's does not equal 0

- What would it mean if the null hypothesis were true?
  - We would expect that the population mean of Y would be β_0 no matter what the values of X_1, X_2, X_3, …, X_p.
  - In other words, this would mean that all of our predictors have no effect on Y!

- What would it mean if the null hypothesis were false?
  - We would expect that Y would vary with different values of X_1, X_2, X_3, …, X_p.
  - In other words, this would mean that at least some of our predictors do have an effect on Y.

The Overall F-Test: Is There a General Relationship?

- We perform this hypothesis test by calculating an F-statistic as follows:

$$F^* = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

- Recall that:
  - TSS tells us how much variation there is in the dependent variable overall (i.e., the total).
  - RSS tells us how much variation there is in the dependent variable that our model did not explain (i.e., the residual).

The Overall F-Test: Is There a General Relationship?

- For the numerator of this F-statistic:

$$E[(TSS - RSS)/p] \begin{cases} = \sigma^2 & H_0 \; true \\ > \sigma^2 & H_a \; true \end{cases}$$

- For the denominator of this F-statistic:

$$E[RSS/(n-p-1)] = \sigma^2, \; H_0 \; true$$

- If there is no relationship between the response and the predictors (H_0), we should expect to see F* values close to 1.
- If there is a relationship between the response and the predictors (H_a), we should expect to see F* values greater than 1.

The Partial F-Test: Is a Subset of Predictors Useful?

- Sometimes it is desirable to test whether a specific subset of variables is useful in predicting our response; assume, for simplicity, that the final q predictors are in question:
  - Null Hypothesis (H_0): $\beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$
  - Alternative Hypothesis (H_A): At least one of these q β's does not equal 0

- What would it mean if the null hypothesis were true?
  - We would expect that the population mean of Y is unaffected by the values of X_{p-q+1}, X_{p-q+2}, …, X_p; these variables have no effect on Y.

- What would it mean if the null hypothesis were false?
  - We would expect that the population mean of Y is affected by the values of X_{p-q+1}, X_{p-q+2}, …, X_p; these variables have an effect on Y.

The Partial F-Test: Is a Subset of Predictors Useful?

- We perform this hypothesis test by calculating an F-statistic as follows:

$$F^* = \frac{(RSS_{Reduced} - RSS_{Full})/q}{RSS_{Full}/(n-p-1)} \sim F_{q,n-p-1}$$

- Where:
  - RSS_{Full} is the residual sum of squares for the full model with all p predictors.
  - $RSS_{Reduced}$ is the residual sum of squares for the reduced model that excludes the q variables in question.

- This test is essentially assessing the difference between $RSS_{Reduced}$ and RSS_{Full}:
  - If this difference is small, then the full model doesn't add much information to the model (H_0).
  - If this difference is large, then the full model does add information to the model (H_a).

Variable/Model Selection

- Suppose that we have a significant overall F-test and now have a set of variables that we have determined are useful for predicting the response. How do we know which variable(s) specifically contribute to the model?

- Simply compute the partial F-test for each individual variable, and select the significant variables? Caution!
  - Consider the case in which we have 100 variables that truly have no association with the response variable.
  - By virtue of statistical theory, with a threshold of $\alpha = 0.05$, 5 of these variables will incorrectly surface as significant (type I error).

- While this might be a good starting point, we need a better way to compare the efficacy across models.

Variable/Model Selection: Akaike Information Criterion (AIC)

- The AIC of a model is defined as:

$$AIC = -2\ln(L) + 2p$$

- Where:
  - L is the maximized value of the likelihood function for the model.
  - p is the number of estimated parameters in the model.

- The better fitted models are identified by smaller AIC values; having a goal of a small AIC results in:
  - A reward for the goodness of fit of the model.
  - A penalty for the complexity of the model.

Variable/Model Selection: Bayesian Information Criterion (BIC)

- The BIC of a model is defined as:

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

- Where:
  - L is the maximized value of the likelihood function for the model.
  - p is the number of estimated parameters in the model.
  - n is the number of observations.

- The better fitted models are identified by smaller BIC values; having a goal of a small BIC results in:
  - A reward for the goodness of fit of the model.
  - A penalty for the complexity of the model.

Variable/Model Selection: AIC or BIC?

- Both AIC and BIC appear mathematically very similar. Which should we choose?

- Favor AIC when the primary goal of multiple regression is prediction (i.e., when you build a model in order to accurately predict new outcomes).
  - As n increases, predictive accuracy increases because more subtle effects are acknowledged by the model. AIC will favor this, whereas BIC will not.

- Favor BIC when the primary goal of multiple regression is descriptive (i.e., when you build a model in order to find the most meaningful relative factors that influence the outcome).
  - The penalty term of BIC is a more stringent than AIC because it not only considers model size, but also observations. Thus, BIC favors a more parsimonious model.

Variable/Model Selection: Stepwise Procedures

- Now that we have a few ways to compare across models, how can we generate the models to compare? How about we try all of them? No!
  - If we have p parameters, then there are a total of 2^p models that we could theoretically create.

- Instead, we need to come up with another procedure to help us hone in on a model that we deem is appropriate using:
  - Forward selection
  - Backward selection
  - Both selection

- **NB:** These procedures may not find the optimal subset of variables to help build our model, but help balance the computational aspect of the search.

Variable/Model Selection: Stepwise Procedures

- The forward selection procedure:
  - Begin with a model that only has an intercept term, then sequentially add the predictor that most improves the fit (based on AIC, BIC, etc.).

- The backward selection procedure:
  - Begin with a model that includes all parameter terms, then sequentially remove the predictor that has the least impact on the fit.

- The both selection procedure:
  - Begin with either a model that only has an intercept term or a model that includes all parameter terms, then sequentially add or remove the predictor that that has the most/least impact on the model, respectively.

Assessing Model Fit: Residual Standard Error

- In simple linear regression, we motivated the residual standard error as an estimate of σ. The general form of this equation is as follows:

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-p-1}}$$

- The residual standard error is the standard deviation of the residuals about the regression surface.
  - A smaller RSE implies that the errors are more tightly bound around the regression surface.

- Again, this measure is in units associated with the dependent variable, so a "small" or "large" RSE needs to be assessed in reference to the problem at hand.

Assessing Model Fit: The Coefficient of Determination R^2

- Recall that for simple linear regression, we used the coefficient of determination R^2 to quantify the proportion of total sample variability in the dependent variable that is explained by the regression model:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Unfortunately, for multiple linear regression, there are some limitations:
  - Each time a predictor is added into the model, R^2 increases, even if the predictor is just latching onto noise; a model with a high R^2 could appear to have a better fit simply because it has more variables.
  - R^2 does not take into account model complexity, and thus does not have any preventative measures against overfitting.

Assessing Model Fit: The Adj. Coefficient of Determination R^2_{Adj}

- To compensate for the possibility of adding irrelevant predictor variables into the model, we can measure model fit by assessing the adjusted coefficient of determination R^2_{Adj}:

$$R^2_{Adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

- It can be shown that the addition of predictor variables to a model only leads to an increase in R^2_{Adj} if the corresponding F-test exceeds 1 (and is therefore likely statistically significant).

Prediction

- Once a valid model is chosen, prediction becomes very easy and straightforward. Recall that the general form of an estimated multiple regression relationship is as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}$$

- At this point, all β values have already been numerically estimated; furthermore, we know all X_i values because they come from our observations!

- To determine the estimate of Y_i for a particular observation X_i, all we need to do is plug in the observed values for X_i.

PART 4

Extending Model Flexibility

Adding & Interpreting Qualitative Variables

- So far, we have only considered adding variables in our model that are quantitative in nature; what about qualitative variables?

- What if we wanted to include variables that are not numerical, but instead categorical, such as:
  - Gender?
  - Ethnicity?
  - College Major?
  - City?

- This is possible, but we need to do some mathematical manipulation. After all, it doesn't make sense to multiply by "male" or subtract "New York."

Adding & Interpreting Qualitative Variables

- In order to code qualitative variables in a useful way for use in multiple regression, we have to create indicator/dummy variables.
  - Indicator/dummy variables are binary representations of whether or not a condition is satisfied; 1 represents yes, 0 represents no.

- Suppose we have a variable `State` that takes values of `NY`, `FL`, or `CA`. To create dummy variables, we can define:

$$X_{NY} \begin{cases} 0 & i^{th}\ obs. \neq NY \\ 1 & i^{th}\ obs. = NY \end{cases} \quad X_{FL} \begin{cases} 0 & i^{th}\ obs. \neq FL \\ 1 & i^{th}\ obs. = FL \end{cases} \quad X_{CA} \begin{cases} 0 & i^{th}\ obs. \neq CA \\ 1 & i^{th}\ obs. = CA \end{cases}$$

- For a qualitative variable that has k categories, how many dummy variables do we need? Only $(k - 1)$! Why?

Adding & Interpreting Qualitative Variables

- We only need (k - 1) dummy variables because the information contained in the k^{th} dummy variable can be inferred.
  - Same concept as determining a single unknown observation's value when the mean of all observations is known.

- The category of the dummy variable we choose to omit from the model is called the baseline; it is the category to which we compare all other categories.

Adding & Interpreting Qualitative Variables

- Suppose we want to predict the `Price` of gas using a simple model with just the `State` dummy variables; we will choose `NY` to be our baseline:

$$Price = \beta_0 + \beta_{FL}X_{FL} + \beta_{CA}X_{CA}$$

- In this model:
  - β_0 is the average price of gas in New York.
  - β_{FL} is the average difference in gas price between Florida and New York.
  - β_{CA} is the average difference in gas price between California and New York.

- Notice how the indicator variables switch on and off depending on the state in consideration. Essentially, indicator variables represent shifts between categories in our data.

Polynomial Regression

- While linear regression assumes that there is a linear relationship between the dependent and independent variables, the true relationship may not be linear.
  - How can we model non-linear behavior using a linear regression model?

- We can add higher-order terms into our model by stacking simple degree transformations:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots$$

- Notice that this equation is still linear; we are simply manipulating the values of X to create new predictor variables and then passing them into the model.

- Be careful about adding too many higher-order terms as they can unnecessarily overfit your data and produce a model that is not generalizable.

PART 5

Review

Review

- Part 1: Multiple Regression
  - What is Multiple Linear Regression?
  - Multiple Linear Regression: Mathematically
- Part 2: Assumptions & Diagnostics
  - Assumptions of Multiple Linear Regression
    - Multicollinearity
  - Accuracy of the Coefficient Estimates
  - The Variance Inflation Factor
- Part 3: Research Questions of Interest
  - Typical Research Questions
  - The Overall F-Test
  - The Partial F-Test
  - Variable/Model Selection
    - Akaike Info. Criterion (AIC)
    - Bayesian Info. Criterion (BIC)
    - Stepwise Procedures
  - Assessing Model Fit
    - Residual Standard Error
    - The Coef. of Determination R^2
    - The Adjusted Coef. of Determination $R^2_{Adj.}$
  - Prediction
- Part 4: Extending Model Flexibility
  - Adding & Interpreting Qualitative Variables
  - Polynomial Regression
- Part 5: Review