Regression

Motivation of Linear Regression

- Linear regression looks to model the linear relationship between two or more variables
- To find if linear regression will be useful for modeling the data, the correlation coefficient "r" can be computed. The value of "r" ranges from -1 to 1, where the two extremes indicate perfect linear association, and 0 indicates there is no linear relationship between the variables
- If two variables don't have a linear correlation, they could still be related through a model with more degrees of freedom (e.g. y = cos(x), y = x^2)

> "For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data" - ESL, Stanford

Properties of Correlation

- **cor(A, B) = cor(B, A)** ← corrleation is not affected by the order of A and B
- **cor(100*A, B) = cor(A, B)** ← correlation is not affected by the units used in A and B
- **cor(A + 10, B) = cor(A, B)** ← correlation is not affected by a translation in A or B

Simple Linear Regression

Find the best fitting line to a scatterplot:

$\hat{f}(x) = \hat{B}_0 + \sum \hat{B}_j x_j$

where $\hat{B}_0$ is the intercept, $\hat{B}_j$ is one of the parameters, x_j is one of the explanatory variables, and $\hat{f}(x)$ is the dependent variable

Least Squares Estimation

In order to generate a regression line through the data, we must define a cost-function that we can minimize as much as possible. This cost-function is called the Least squares estimation, and it attempts to calculate the $\hat{B}$ parameters such that sum of squared errors is minimized.

$$SSE = \sum_{i=1}^{N}(y_i - f(x_i))^2 = \sum_{i=1}^{N}(y_i - \hat{B}_0 - \sum_{j=1}^{p}\hat{B}_j x_{ij})^2$$

This equation calculates, for every data pair (x,y) denoted as i, the difference between the actual y_i value and the estimated $\hat{y}_i$ value, and squares the difference to avoid any negative values. If there were negative values, some errors would cancel out others, and thus the end sum would not be representative of how truly "off" the predictions were from the true values.

The difference $y_i - \hat{y}_i$ is called a residual.

Types of Errors in Regression

After creating a regression on your data, you're left with a prediction $\hat{y}$ for all X. These predictions will have some error associated with them, as the line that's been generated from the regression most likely won't perfectly cross all points of data. There are a few sums of errors that help quantify the model's variation:

SSR: Sum of squares due to regression — Quantifies the variation between the predicted y values and the sample mean of y, which in the case of a regression would signify there being no relationship between the x and y-values.

$$SSR = \sum(\hat{y}_i - \bar{y}_i)^2$$

SSE: Sum of squared errors — Quantifies the variation betwen the true y values around the regression line $\hat{y}$.

$$SSE = \sum(y_i - \hat{y}_i)^2$$

SST: Sum of squares total — Quantifies the variation between the data *y* around their mean $\bar{y}$.

$$SST = SSR + SSE = \sum(\hat{y}_i - \bar{y}_i)^2 + \sum(y_i - \hat{y}_i)^2$$

$$SST = \sum(y_i - \bar{y}_i)^2$$

R-Squared

The proportion of the total variation in Y being explained by the variation in X.

$$R^2 = SSR/SST$$

If there is a large SSE or sum of squared errors, then there is a large variation between the true y values around the regression line $\hat{y}$, and thus not a lot of the variation in the data can be explained by the regression line. In this case, SSR would be small, and thus R^2 will also be small, meaning the total variation of the data cannot be explained too much by the model.

$$\downarrow SSE = \uparrow SSR = \downarrow R^2$$

If there is a low SSE or sum of squared errors, then there is low variation between the true y values around the regression line $\hat{y}$, and thus a lot of the variation in the data can be explained by the regression line. In this case, SSR would be high, and thus R^2 will also be high, meaning the total variation of the data can be explained by the model.

The value of R^2 ranges from 0:1, where

- $R^2 = 0$ means there is no relationship at all between the variation in X and the variation in Y
- $R^2 = 1$ means there is a perfect relationship between the variation in X and the variation in Y

Note: If you are comparing models but have changed the Y variable in any way (e.g. predicting log(Y) instead of Y), comparing R-squared values is no longer possible as the distributions are different.

Population Regression Function and ε

When performing a regression on data, the objective is always to estimate the "true" relationship between a dependent variable *y* and explanatory variables *X = (x_1, x_2, ... , x_p)*. Because we are performing a regression on a sample of data, the regression line $\hat{y}$ is a linear model for the sample, and not for the population, so it is safe to assume there will be a difference between the prediction we've made $\hat{y}$ and the true relationship line between y and X.

This theoretical "true" relationship line between y and X is called the Population Regression Function, denoted as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where the βs are the true coefficients of the relationship between Y and X, and ϵ is the error term denoting the difference between the data and the population regression function Y. This error term is only theoretical, since we can never truly know the population regression function and therefore can never calculate the values for ϵ.

Degrees of Freedom

When fitting a regression to data, a particular question might arise: If I have k explanatory variables, and I'm trying to predict Y, what is the minimum amount of observations I need to perform a regression?

Case 1: One explanatory variable, regression: $y = \beta_0 + \beta_1 x$

- **One observation:** With one observation, no line can be plotted since there are infinitely many lines that cross a single point. Therefore, one observation is too little to fit a regression ❌
- **Two observations:** With two observations, a line can be plotted between them. However, this regression line will always have an $R^2 = 1$, since the line crosses through both of the points. Therefore, there are still too few observations to fit a regression ❌
- **Three observations:** With three observations, a line can now be regressed that will most likely have an $R^2 < 1$, meaning there is now some significance to the line of best fit. In this case, three observations is sufficient to perform a regression ✓

Case 2: Two explanatory variables, regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- **One - Three observations:** Much like case 1, a regression cannot be performed with one or two points, because there are not sufficient points to draw a plane through them. At three observations, a plane can be drawn through them, but this plane will have an $R^2 = 1$, since the plane fits perfectly through the three points. Therefore, up to three observations cannot fit a regression ❌
- **Four observations:** With four observations, a plane can now be regressed that will most likely have an $R^2 < 1$, meaning there is now some significance to the plane of best fit. In this case, four observations is sufficient to perform a regression ✓

This brings us to degrees of freedom, which quantify the relationship between the number of observations, the number of explanatory variables, and the statistical power a regression line has to model a given data. The equation for degrees of freedom is:

$$df = n - k - 1$$

where *df* = degrees of freedom, *n* = number of observations, and *k* = number of explanatory variables (X).

So in Case 1 above, with n = 3 observations and k = 1 explanatory variable, *df* comes out to be $df = 3 - 1 - 1 = 1$, meaning that with three observations the regression had one degree of freedom. With n = 2 observations, the model had 0 degrees of freedom, and thus a regression could not be performed. With n = 1 observation, the value of degrees of freedom goes negative, so there is definitely not a regression line that can be plotted in that case.

In Case 2 above, the same logic can be applied to figure out how many observations are necessary for k = 2 explanatory variables. We need at least k = 1 degree of freedom to perform a regression, thus the number of observations required is $n = df + k + 1 = 1 + 2 + 1 = 4$ observations, which aligns with the explanation from above.

Degrees of freedom are closely related to R^2:

> 💡 As more explanatory (X) variables are added to a model, *df* decreases, and as a result R^2 will ONLY increase.

This can make R^2 deceiving, because you could be adding lots of junk X variables that aren't adding any explanatory power to your model but your R^2 still increases, giving the impression that your model is being better trained by the variables added to it, when the reality is that you are just removing degrees of freedom which increase R^2.

Adjusted R-Squared

$$\bar{R}^2 = 1 - (\frac{SSE}{SST})\frac{n-1}{n-k-1}$$

The adjusted R^2 takes into account the degrees of freedom of the model and adjust the score provided by R^2 to reflect a closer estimate of how much variation in Y is really explained by the variation in X.

As k increases, Adjusted R^2 will tend to decrease, reflecting the reduced power in the model. In the case where you've added new explanatory variables that help explain the data very well, then the adjusted R^2 will also increase, but in the case where the extra

variables don't do much you'll see a decrease in the adjusted score.

Regression Outputs

F-Statistic

The F-Statistic is used to reject the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$, which would imply that none of the explanatory variables have any relationship with the dependent variable Y.

If the p-value of the F-Statistic is less than the level of significance, then we can reject the null hypothesis and conclude that the beta values are in fact significant in the regression.

R-Squared

The proportion of the total variation in Y being explained by the variation in X.

Adjusted R-Squared

The proportion of the total variation in Y being explained by the variation in X, adjusted for degrees of freedom in the model

Variables Section

- **Coef**. = Coefficients of all explanatory variables
  - **Interpretation of a coefficient β_i —** For every increase of 1 unit of x_i, the output y increases by β_i on average, all other variables held constant.
- **Std. Err**. = Standard Error
  - The typical variation of the coefficient
- **t-statistic**
  - Divide Coefficient by the standard Error
  - A greater t-statistic indicates a more statistically significant variable, while a t-statistic close to 1 indicates a less significant variable
- **p-value (two-tailed)**
  - describes how extreme a given coefficient is under the null hypothesis that the coefficient = 0
  - This p-value is found on a t-distribution, looking for the probability of getting a t-statistic equal to the calculated value assuming the null hyptoehsis that the t-statistic = 0
- **confidence interval**
  - A 95% confidence interval creates bounds for the coefficient's value, estimating with 95% confidence that the coefficient is within the bounds
  - If the confidence interval for a variable includes 0, then you need to be wary that the given variable is not statistically significant for your model

Functional Form & Transformations

Non-Linear Relationships and Logarithms

When performing a regression, we are attempting to linearly fit each x variable to the y variable, which may not always be easy to do depending on the distribution of the data. When looking at the numerical x variables you'll be using to calculate your regression, look at the scatter plots of y against each x.

Parabolic relationships

In the case that an x variable is correlated on a seemingly parabolic relationship with the y variable, try squaring the input variable x and inputting that into the regression model with the original x variable.

Hyperbolic / Logarithmic relationships

When performing a regression, it is beneficial for the explanatory (x) variables to follow some sort of normal distribution. Sometimes however they do not follow a normal distribution, as they may be skewed to the right or to the left.

A solution to a skewed variable x_s is to take the natural logarithm of it $ln(x_s)$ and input that as a factor in the regression. This way, the variable data will have been scaled so as to present a more normal distribution.

> 💡 Note: The overall regression model is still linear in nature, even if you square a variable or take the natural log of another. Since you input the adjusted variables like all other variables, the model has no way of knowing what relationship you've presented. It will only see the new values that it wants to fit linearly, and thus the overal model will remain a linear regression model.

Interpreting logarithmic coefficients

In the case where you've input $ln(x_i)$ as an explanatory variable for your model, you'll have a coefficient βi. In the case that your output variable is a continuous, non-logarithmic variable, the interpretation for β_i would be:

- For every 1% increase in x_i, the value of y changes by $\beta_i/100$ on average, holding all other variables constant.

If the output variable is a logarithmic variable as well, the interpretation becomes:

- For every 1% increase in x_i, the value of y changes by β_i (as a %) on average, holding all other variables constant.

Categorical X variables and Interaction Terms

Multi-Level Categorical Variables

- **Dummy Variable Trap**
  - When separating categorical variables into their own columns, you can't leave all categories up
  - This is because one variable will be explained by the values of the other variables, and thus this variable adds no information to the regression
  - To avoid this, you must remove one categorical variable and leave the rest. The removed categorical variable will act as a "baseline" for the rest, and the dummy variable trap is avoided

Interaction Terms

- In the case where one X variable affects the relationship of Y with another X variable, then the model would benefit from an interaction term
-