Linear Regression

Data Analytics in Business

1 Linear regression

1.1 What is linear regression?

- Regression is a method of studying the relationship between a certain dependent variable and a set of independent variables.
- At the simplest level, we only consider a singular predictor or independent variable. This is known as simple linear regression.
- An example is predicting house prices from the number of rooms of the house
- Linear regression as its namesake suggests is the modeling of the dependent variable (response) as a linear combination of the predictor variables.
- This is shown graphically in Figure 1

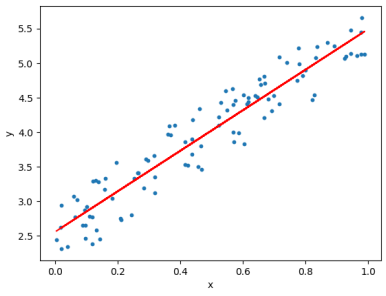

Figure 1: : Linear Regression

1.2 Mathematical Representation

So consider a response variable Y and a predictor variable X. Then our basic linear model is shown in Eq 1

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

- Y_i represents the dependent variable for observation i
- x_i represents the observed predictor for observation i
- ϵ_i is the error term
- ϵ_i are iid following $N(0, \sigma^2)$
- β_0 and β_1 are the parameters which the regression needs to solve for.

Some clarifications regarding the above syntax and jargon:

- The use of subscript i is indicative that the model is for every observation in the given dataset.
- iid means independent and identically distributed. For the case of the errors, this means that they all follow a normal distribution given by mean 0 and standard deviation σ and are also also mutually independent (errors aren't dependent on each other). Shown in Figure 2
- In case of unfamiliarity with normal distributions, refer to [1] or any of the plethora of notes available online.
- β_0 and β_1 effectively represent the intercept and slope of the line we fit to the data.

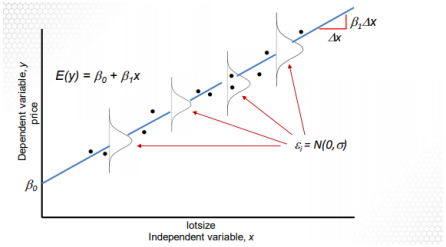

Figure 2: : Normal Errors

So the objective of linear regression is to estimate the βs from the sample data. Note as a result, naturally the β depend on the sample being used.

1.3 Ordinary Least Squares

- We need to estimate the unknown parameters β_0 and β_1.
- Ordinary Least Squares (OLS) is one method of estimating these parameters
- OLS aims to minimize the sum of the squared errors (SSE) shown in Eq 2

$$\sum_i (y_i - \hat{y})^2 \tag{2}$$

- Using squared residuals is quite intuitive: We want to calculate how 'far off' our model is from the actual data. So we compute the difference (residual) between the actual data and the predicted value.
- Instead of summing, we sum the squared residuals, as otherwise the negative and positive residuals will effectively cancel each other out, and give a poor approximation of the models accuracy
- So by squaring the residuals and then summing, we make all of them positive and so can aggregate the total residuals.
- Our regression model will test combinations of β_0 and β_1 and continuously adjust for each additional data point, aiming to minimize the squared residuals.
- Note this should remind you of calculating variation, where we square the differences between the data points and the mean and then take the sum
- Equation 3 expressed the relationship between the total variation and the residuals. This is also illustrated in Figure 3.

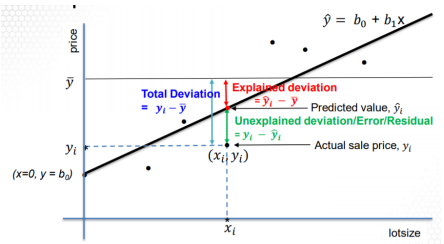

Figure 3: :Residuals

$$Total\,Variation\,(SST) = \sum (y - \bar{y})^2 \tag{3}$$

$$Explained\,Variation\,(SSR) = \sum (\hat{y} - \bar{y}) \tag{4}$$

$$SST = SSE + SSR \tag{5}$$

Note SST stands for 'Total Sum of Squares', SSE for Sum of Squared Errors and SSR for Sum of Squared Regression.

1.4 Coefficient of Determination

- The coefficient of determination or more commonly known as R^2 expresses the proportion of variance explained by the independent variables in a regression model.
- Effectively this is a measure of the overall strength of the relationship between the dependent variable (Y) and independent variables (X).
- So we can use R^2 to measure the ability of a regression model with a higher R^2 desired/
- However, there are limitations to using R^2 as a goodness measure:
  - Every time you add a predictor to model, the R^2 will increase, even if this is due to simply chance.
  - If there are too many predictors in a model, the model starts to overfit. This simply means the model starts to model the noise in the data. This is undesired as this means the model is not capturing the actual relationship in the data.
  - Although the model will be very successful for the data is is modeled on, it will perform poorly when new data (out of sample) data is used.
- R^2 is calculated from Eq 5:

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \\ &= \frac{Explained\,Deviation}{Total\,Deviation} \end{aligned}$$

- One way to bypass R^2's promotion of over-fitting is to use the adjusted R^2, which adds a penalty on the number of fitting variables.
- The adjusted R^2 is shown below in Eq 6

$$Adjusted\,R^2 = 1 - \frac{SSE}{n-p-1} / \frac{SST}{n-1} \tag{6}$$

n is the number of observations in the data and p is the number of independent variables.

1.5 Null hypothesis and t tests

- After fitting a regression model, it is important to see if the model is good or not.
- For this we consider a null hypothesis, which claims that there is no relationship between the independent variables and the response variable.
- This effectively means the null hypothesis claims that our βs are 0.
- The t statistic is the coefficient divided by the standard error. The standard error is the standard deviation of the coefficient. We can think of it as the precision at which the regression coefficient is observed.
- What is most significant is the P value (or probability value). This compares the t value calculated for the coefficient against a known library of t value distributions (Student's t distribution).
- The p value indicates the probability of finding a t value of this size given the null hypothesis is true.
- So if the p value is high it implies the null hypothesis is valid
- If the null hypothesis is low, we can reject the null hypothesis.
- Typically p values 5% are sufficient to reject the null hypothesis.
- For a more thorough explanation refer to [2]

1.6 F test

- Without going into any significant depth of statistics, the F test is another measure of rejecting or accepting the null hypothesis.
- Please refer to the supplementary notes on Anova testing, which leads to F statistics.
- The F statistic is given by Equation 7 where m is the number of groups in the overall dataset (number of variables) and n is the number of data points in each group (observations)

$$\frac{\frac{SSB}{m-1}}{\frac{SSW}{m(n-1)}} \tag{7}$$

- As a general point, the F statistic value from the given data can be then compared to fixed tables of F values (based on degrees of freedom of both the numerator and denominator) to give you a probability which with we can reject or accept the hypothesis.

2 Problems of Linear Regression

2.1 Assumptions

1. Linearity Assumption: $E[Y] = \beta 0 + \beta_1 x$. We are assuming the expected value of Y for any given X is a linear function of X (approximates to a straight line)

2. Assumption about errors: We assume the error terms are independently and identically distributed (iid) normal random variables with 0 mean and variance σ^2

3. Assumptions about predictors: For multiple regression (more than 1 predictor) we assume the predictors are linearly independent.

2.2 Common Problems

1. Non-linearity of the response-predictor relationships (Failure of linearity assumptions)

2. Correlation of error terms

3. Non-constant variance of error terms

4. Outliers (as consider squared residuals, outliers can have signficant impact)

5. High-leverage points (Observations made at extreme or outlying values of the "independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation")

6. Collinearity (failure of predictor assumption)

2.3 Dealing with problems

2.3.1 Non-linearity of response-predictor

- A large downfall in modern data science is the lack of preliminary analysis when dealing with data.

- In reality the hardest part of data analytics is not implementing the model, but choosing which model and which parameters.

- As it's namesake suggest, linear regression assumes a linear relationship between the independent variables and the response variable.

- From merely looking at data it isn't obvious what actual relationships exist in the data.

- So, it is extremely useful to visualize the data before fitting any models. Typically plotting the response against each predictor is advised
- Consider Figure 4. If we blindly attempt to fit a linear model to this data, it will prove extremely unsuccessful. It is obvious by eye that a polynomial (2nd order) model will be much more successful.

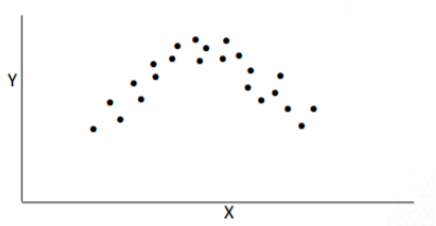

Figure 4: :Residuals

2.3.2 Correlation of Error Terms

- Recall in performing linear regression we assumed the error terms are uncorrelated. If they aren't we get autocorrelation.
- This basically means knowing the value of any singular error term should not influence the value of any other error term
- If correlation does exist then:
  1. Estimated standard errors will be underestimate the true standard errors
  2. Confidence and prediction intervals will be narrower than they should be as well as p values being lower.
- Autocorrelation can be detected by the Durbin-Watson test, although we will not look at this in detail (can be trivially done by lmtest package in R).

2.3.3 Heteroskedasticity

- "The assumption is that the spread of the responses around the straight line is the same at all levels of the explanatory variable (i.e., we have constant variance or homoskedasticity)"
- There may be non constant errors in the data (increasing errors with increasing independent variable). These can be easily detected with the residuals vs fitted values plot.

- If there is non-context error, our hypotheses tests and confidence intervals can be misleading
- If there is heteroskedasticity, we may require a transformation of variable. For example $ln(Y)$.

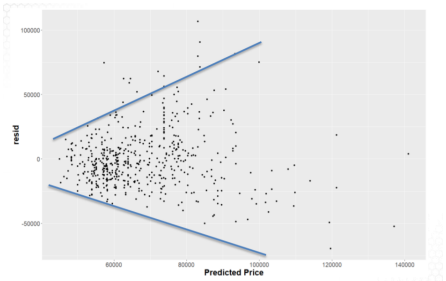

Figure 5: : Heteroskedasticity of a data sample. The increasing projection of errors is clearly observed.

2.3.4 Outliers

- Outliers are points that has a y_i value that is far from its predicted value. $\hat{y}_i$
- Again you can find outliers by plotting residuals against predicted values of Y
- These outliers could be due to incorrect data or simply due to non linear relationship
- Generally don't assume that an outlier should be removed as it may be due to a model deficiency such as a missing predictor.

2.3.5 High leverage Points

- Leverage points are those which have predictor values outside the normal range of observations
- A point with high leverage causes significant changes to the model upon its deletion.
- Points of high influence can be identified using Cook's distance which measures the difference between the regression coefficients obtained
  - From the full data
  - By deleting the point i
- Typically points with Cook's distance ¿ 1 are considered highly influential

2.3.6 Multicollinearity

- If 2 or more variables are linearly related we say there is collinearity in the data.
- This is problematic as it makes it difficult to identify the effect of each predictor on the response and consequently which predictors to include in the model.
- We can detect (multi)collinearity using Variance Inflation Factors (VIF)
- So to calculate the Variance Inflation Factors, we regress predictor variable X_j against all other X variables, and label the corresponding R^2 value as R_j^2
- $VIF_j = \frac{1}{1-R_j^2}$
- If X_j has a strong linear relationship to other X variables, then R_j is close to 1, and VIF_j will be large.
- Values of VIF greater than 5 signify multicollinearity (generally)
- We can mitigate the impact of multicollinearity simply by selecting 1 variable
- Can also use Principal Component Analysis, but we will not discuss this to extent

[1]https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/normal-distributions-library/a/normal-distributions-review
[2]https://www.youtube.com/watch?v=-FtlH4svqx4