Basic Regression

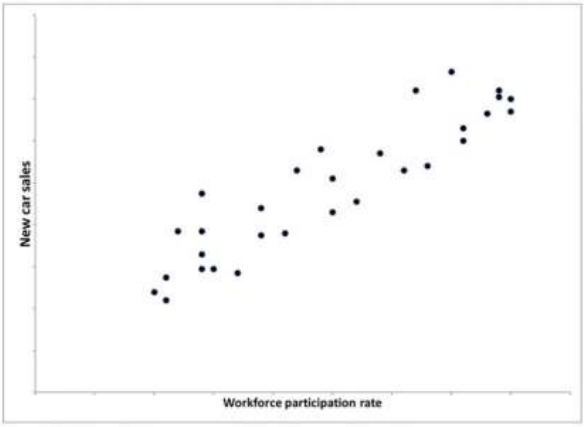
Monday, September 23, 2024 1:38 PM

Regression Introduction

Regression: a statistical technique that relates a dependent/response variable to one or more independent/predictor variables. It can answer two different types of questions:

1. How do systems work?
    a. What is the value of a home run?
    b. What is the effect of economic factors on the presidential election?
    c. What impact does education have on income?
    d. What are the key factors in purchasing a car?
2. What will happen in the future? / making predictions about future events
    a. How tall will a child be when they grow up?
    b. What will be the price of oil in 1.5 years?
    c. What will be the housing demand in the next six months?
    d. How long will this insurance applicant live?

There are several different types of regression, with the most basic being Simple Linear Regression (SLR): linear regression with one predictor. For example, there appears to be a relationship between the workforce participation rate and how many new cars are sold. This makes logical sense; the more work people have, the more money they have to spend on things like new cars.



With SLR, we can look for a linear relationship between the variables. Suppose the following:

- $y = response\ variable$ (new car sales)
- $x_1 = predictor\ variable$ (workforce participation)

Then the regression equation is $y = a_0 + a_1 x_1$

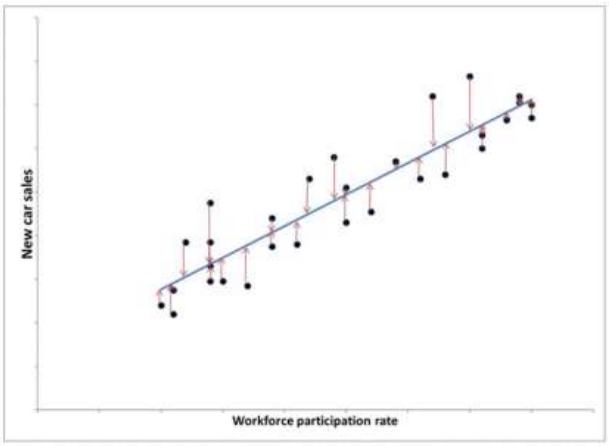
Note: with m predictors, this line becomes: $y = a_0 + \sum_{j=1}^{m} a_j x_j$

The quality of the line's fit can be described by the metric the Sum of Squared Errors (SSE): for each data point, the error is defined as the distance between true value and the estimated value (true number of cars sold vs estimated cars sold).

Defining those values below provides the metric of the error:

- $y_i = true\ value\ for\ data\ point\ i$
- $\widehat{y_i} = model's\ predicted\ value\ for\ data\ point\ i$

Then the equation for the i*th* data point's prediction error is $y_i - \widehat{y_i} = y_i - (a_0 + a_1 x_{i1})$ and the Sum of Squared Errors is $\sum_{i=1}^{n} (y_i - (a_0 + a_1 x_1))^2$



The best-fit regression line minimizes the SSE and is defined by the constants a_0 and a_1.

Recall that the three basic types of questions that analytics can answer are descriptive, predictive, and prescriptive. Regression can answer both descriptive and predictive questions, so it is a commonly used technique.

Maximum Likelihood and Information Criteria

Likelihood: a measurement of how well a model explains observed data by calculating the probability of seeing that data under different parameter values of the model. The parameters that give the highest probability are known as the Maximum Likelihood and they are the best-fit parameters. Good statistical software will be able to calculate this metric, but it does become much more computationally complex.

Information Criteria are likelihood-based measures of model fit that include a penalty for complexity. Some different Information Criteria are described below.

Akaike Information Criterion (AIC): $AIC = 2K - 2\ln(L^*)$

- $L^* = maximum\ likelihood\ value$
- $k = number\ of\ parameters\ being\ estimated$
  - Penalty term $2k$ balances the likelihood with simplicity, helping to avoid overfitting.
- Smaller k and higher L^* → Small AIC → better model

AIC works best with infinitely many data points, which is unrealistic. Thus a correction term can be added to the AIC formula when working with smaller data sets:

$$AIC_c = 2K - 2\ln(L^*) + \frac{2k(k+1)}{n-k-1}$$

AIC values from different models can also be compared against one another, using a metric known as Relative Likelihood. This metric can tell you the relative probability that one of the models is better than the other. Specifically it tells you if the model with the lower AIC value is better than the model with higher AIC.

$$Relative\ Likelihood = e^{(AIC_1 - AIC_2)/2}$$

For example, when comparing Model 1 with AIC 75 and Model 2 with AIC 80, the relative likelihood is $e^{(75-80)/2} \approx .082$. Implying that model 2 is 8.2% as likely as model 1 to be better, so the first model is probably better than the second model.

Bayesian Information Criterion (BIC): $BIC = k\ln(n) - 2\ln(L^*)$

- $L^* = maximum\ likelihood\ value$
- $k = number\ of\ parameters\ being\ estimated$
- $n = number\ of\ data\ points$

The BIC is very similar to the AIC, with the only difference being the way it deals with the number of parameters and data points. In general, the BIC's penalty term is larger than the AIC's penalty terms, so the BIC encourages models with fewer parameters than the AIC does. You should only use BIC when then are more data points than parameters.

When comparing models based on their BIC, there is a rule of thumb:

- $|BIC_1 - BIC_2| > 10$
  - The smaller-BIC model is "very likely" better
- $6 < |BIC_1 - BIC_2| < 10$
  - The smaller-BIC model is "likely" better
- $2 < |BIC_1 - BIC_2| < 6$
  - The smaller-BIC model is "somewhat likely" better
- $0 < |BIC_1 - BIC_2| < 2$
  - The smaller-BIC model is "slightly likely" better

In summary: there is no hard and fast rule for using AIC, BIC, or Maximum Likelihood. All three can give valuable information and looking at all three can help you decide which is the best to use in a given situation.

Using Regression

When using Regression to answer questions about how systems work, the answer is in the coefficients.

For example, say we want to determine how many average of runs a Home Run is worth. In this case, the Response variable is how many runs a team scored during a season and the Predictors could include variables like the number of home runs, triples, doubles, singles, outs, double plays, stolen bases, etc. We can fit a regression model where the regression equation looks something like the below:

$$runs\ scored = a_0 + a_1[num\ HR] + a_2[num\ triples] \ldots + a_7[num\ stolen\ bases]$$

Taking team data from the last seasons, a basic linear regression gives HR a coefficient of approximately 1.4, meaning every home run adds about 1.4 runs to the team's total number of home runs.

$$runs\ scored = a_0 + 1.4[num\ HR] + a_2[num\ triples] \ldots + a_7[num\ stolen\ bases]$$

When using Regression to make predictions about the future, then then the answer is in the predicted responses.

For example, say we want to predict how tall a 2 year old American child will be when they grow into an adult, we can compile information on many Americans. For each person, the Response variable is their height as an adult and their Predictors could be things like father's height, mother's height, age 2 height, gender, etc. After fitting the regression model, the equation could look something like the below:

$$adult\ height = a_0 + a_1[father's\ height] + a_2[mother's\ height] + a_3[age\ 2\ height] + a_4[gender]$$

Then if we have a two year old child, we can use this equation to predict their adult height.

Causation VS Correlation

Causation: when one thing causes another thing

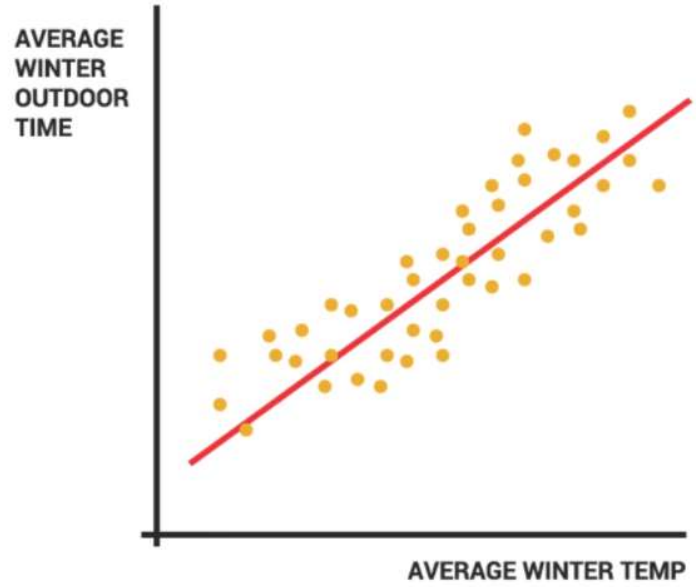
Correlation: two things tend to happen or not happen together; neither of them might cause the other.

For example, say we do a regression analysis of the Response variable hours spent outdoors per day in winter and the Predictor variable of the average daily winter temperature. The model may show a correlation between the Response and Predictor variables, with a low p-value of the Predictor's coefficient.

$$y = a_0 + a_1 x_1$$

So, do higher temperatures cause people to go outdoors more?

- Probably; it is logical that cities with higher winter temperatures will have more people spending time outside than colder places.



If we look at the Regression equation in reverse, however, does it make sense?

$$x_1 = b_0 + b_1 y$$

Does people spending more time outdoors in winter cause higher winter temperatures?

- No, this is silly

These two approaches have the same correlation between the Predictor and Response variables, but no causation. Both approaches can be made to make predictions, but there isn't a causal relationship between the two.

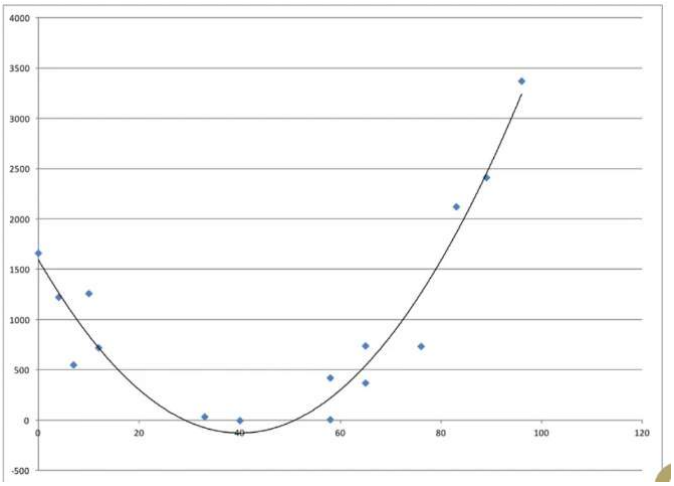
So when is there causation? Here are some rules of thumb:

- When the cause is before effect
- The idea of the causation makes sense

- There are no outside factors causing the relationship

Transformations and Interactions

Say we want to perform a simple linear regression on some data that doesn't look linear. How do we do it?



We can transform the data so the fit is linear:

- Quadratic Regression: $y = a_0 + a_1x_1 + a_2x_2^2$

We can transform the response variable so the fit is linear:

- Response Transform: $\log(y) = a_0 + a_1x_1 + a_2x_2$

We could also transform both the data and the response.

We could feature engineer an interaction variable, such as the product of both parent's heights when predicting their child's adult height.

Software, like Box-Cox transformation, can be used to automate finding the best-fit coefficients.

Output

We use helpful indicators from the Regression model to interpret its output. Some commonly used indicators are listed below.

p-Values: for each coefficient, its p-value estimates the probability that the coefficient equals zero. This is a kind of hypothesis test.

- Rule of thumb: p-value > 0.05, then the coefficient likely is zero and the corresponding attribute can be removed from the model
- Higher alpha thresholds imply that more factors can be included, but that runs the risk of including an irrelevant factor
- Lower alpha thresholds imply that fewer factors can be included, but that runs the risk of excluding a relevant factor

Note: p-values come with two warnings

1. With large amounts of data, p-values get small even when the attributes are not at all related to the Response variable.
2. p-values are just probabilities, so even when they are meaningful it's not absolute certainty. For example, with 100 attributes with a p-value of 0.02 each, each attribute has a 2% chance of not being significant - therefore we can expect that 2 out of the 100 attributes are actually irrelevant.

Confidence Interval (CI): an interval which is expected to contain the true coefficient value, so we can see a range of where the coefficient likely lies and how close it is to zero. Related to the p-value.

T-Statistic: The coefficient divided by its standard error. Related to the p-value.

Another helpful indicator is the coefficient itself. When multiplied by its attribute value, it still may not make much of a difference to the overall equation even if its p-value is very low and we then know that the attribute isn't very important.

R-Squared Value: Known as the coefficient of determination, this is the estimate of how much variability the model accounts for.

- If a model has an r-squared value of 59%, that means the model accounts for about 59% of the variability in the data and the remaining 41% could be randomness or other factors not included in the model.

The Adjusted R-Squared value adjusts the metric for the number of attributes used in the model.

Note: when using R-Squared values, some things are just not easily modeled. Especially things that affect real-life systems where humans are involved. In many cases, an R-squared value of 0.4 or 0.3 is quite good.