UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Statistics

Logistic Regression

Sara Geremia

May 9th, 2024

What is a model?

A model is a simplification of the reality:

- **Predictive Relationships**: In statistics, models often aim to understand how one variable (y) depends on another (x). For instance, in a linear regression model, this relationship is explored to predict future outcomes based on new data.

- **Predictive Equations**: Equations derived from models can predict data that hasn't been observed yet. For example, in physics, the equations of motion can predict the future position of a moving object given its current state.

All models are wrong but some are useful

Any model, by its nature, is an oversimplification of reality:

- **Imperfect Representations**: Models rely on simplifications and assumptions to make complex systems understandable and manageable. For example, economic models might assume rational behavior among agents, which isn't always true in real life.

- **Limitations**: Every model has limitations in scope and precision. For example, weather models might predict general trends accurately but fail to capture micro-climates or unexpected weather phenomena precisely.

Simple vs Multivariate Regression

- One dependent variable Y predicted from one independent variable X
- One regression coefficient
- R^2: proportion of variation in dependent variable Y predictable from X

- One dependent variable Y predicted from a set of independent variables $(X_1, X_2 \ldots . X_p)$
- One regression coefficient for each independent variable
- R^2: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

Significance Tests

Testing R^2

- Test R^2 through an F test
- Test of competing models (difference between R^2) through an F test of difference of R^2s

Testing β

- Test of each partial regression coefficient (β) by t-tests

Different Ways of Building Regression Models

- Simultaneous: All independent variables entered together
- Stepwise: Independent variables are removed or included in order to find the subset of variables in the data set resulting in the best-performing model
- Hierarchical: Variables are entered into the model in predefined stages or blocks. Each stage represents a group of variables that are theoretically related or logically connected.

Stepwise Regression

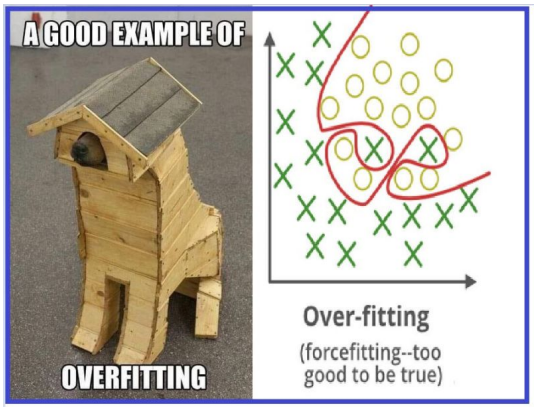
The stepwise regression consists of iteratively adding and removing predictors, in the predictive model, to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

- **Forward selection**: starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
- **Backward selection**: starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
- **Stepwise selection**: is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

All models are wrong but some are useful

The more complex, the less understandable:

- **Complexity vs. Understandability**: Trade-off between the complexity of a model and its understandability. Simple models are easy to understand and use but may not be very accurate. Complex models might be more accurate but can become so intricate that they are difficult to interpret and apply.
- **Overfitting**: In data science, a highly complex model might fit the training data perfectly (overfitting), but it performs poorly on new, unseen data. This is because the model captures noise instead of the underlying pattern.

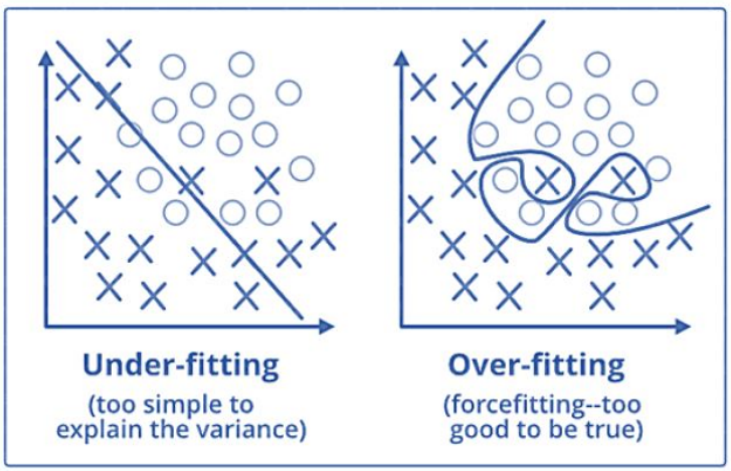


Under-fitting and Over-fitting

We want models that are as accurate and simple as possible:

We need to find the trade-off between accuracy and complexity, where:

$$F = accuracy - complexity$$



Throwback: Bayes' Theorem

Given two events E e H, such that $P(E) \neq 0$ e $P(H) \neq 0$,

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

or for a generic set of mutually exclusive and collectively exhaustive events $H_i, i = 1, \ldots, k$, and E, with $P(E) \neq 0$ and $P(H_i) \neq 0$

$$P(H_i|E) = \frac{P(H_i)P(E|H_i)}{P(E)} = \frac{P(H_i)P(E|H_i)}{\sum_{i=1}^{k} P(H_i)P(E|H_i)}$$

Throwback: Bayes' Theorem

Rewrite Bayes rule to include a particular model:
Given a model m, a response variable y and a set of parameters θ,

$$P(\theta|y,m) = \frac{P(y|\theta,m)P(\theta|m)}{P(y|m)}$$

- $(\theta|y,m)$ is the posterior, it combines prior beliefs about a parameter with the information provided by the observed data.
- $P(y|\theta,m)$ is the likelihood, representing the probability of the response y given the parameters and the model m
- $P(y|m)$ is the model evidence, also called marginal likelihood. It is the probability of the observed data y under the model m.

Model evidence

$$P(\theta|y,m) = \frac{P(y|\theta,m)P(\theta|m)}{P(y|m)}$$

When the model evidence $P(y|m)$ is high, the model is considered good. Conversely, when it is low, the model is considered poor.

A good model is a model that optimizes the tradeoff between accuracy and complexity. This balance is essential for generalization, which refers to the model's ability to perform well on new, unseen data.
We can always decompose model evidence to:

$$P(y|m) = accuracy - complexity$$

Approximations of model evidence

$$P(y|m) = accuracy - complexity$$

Accuracy measures how well a model's predictions match the actual data. Various metrics can be used to quantify this accuracy. We met R^2, SSE, $\mathcal{X}^2$.

Complexity measures the capacity of a model to capture data patterns and often relates to the number of parameters, the model's structure, and its flexibility. Higher complexity increases the risk of overfitting.

Approximations of model evidence

- **Akaike Information Criterion (AIC):**

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

- **Bayesian Information Criterion (BIC):**

$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

These criteria are similar in terms of accuracy but differ in how they account for complexity.

- Both penalize for the number of parameters k to address overfitting.
- Lower AIC and BIC values indicate lower penalty terms hence a better model.
- BIC has a greater penalty for complexity than AIC.