STATS 330

Handout 14
Regression models: Prediction or explanation?

University of Auckland

Prediction or Explanation

Regression models are use to relate a response variable to a set of explanatory variables.

The motivation for doing this can either be prediction or explanation.

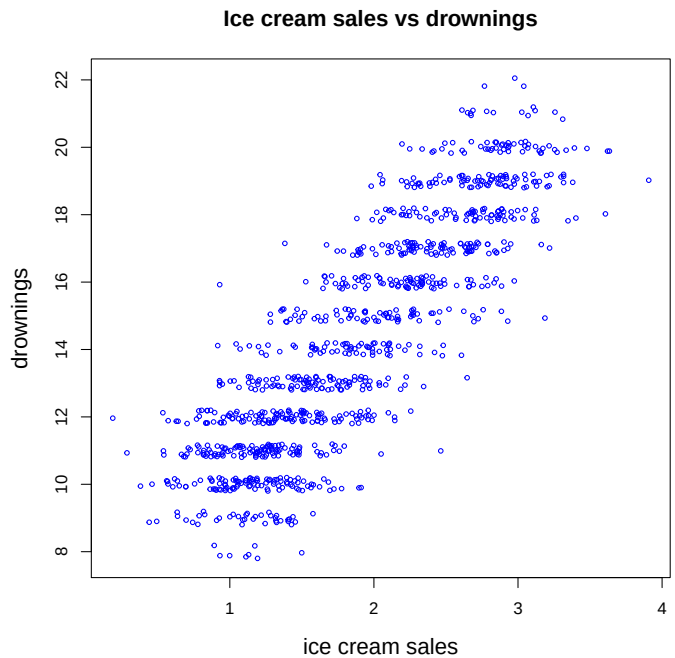
Prediction The goal is to develop a model to predict the response based on observed values of the explanatory variables.

Explanation The goal is to understand how the response is related to the explanatory variables.

This distinction is crucial as a model that works well for prediction may not be useful for explanation and vice versa.

Example: To explain or to predict? I

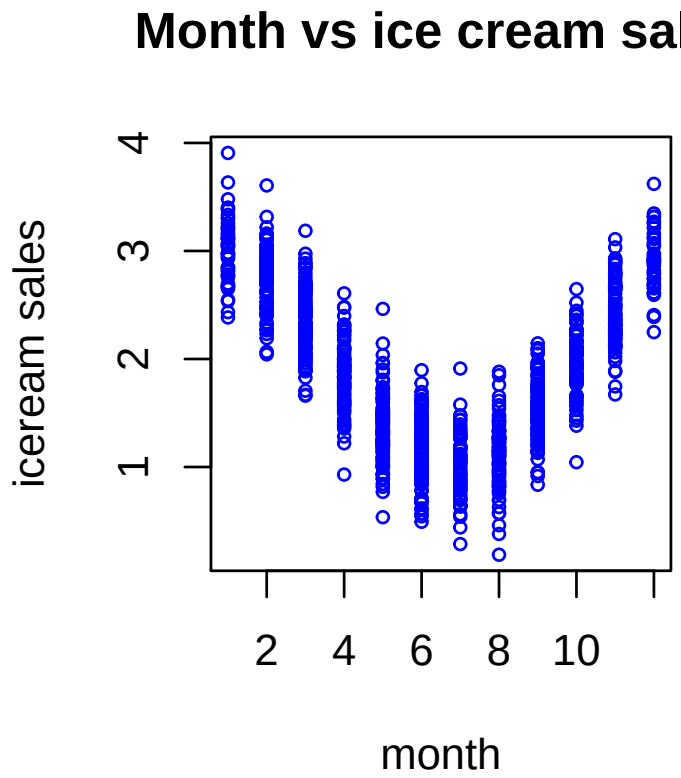
- Suppose we have data on ice cream sales and the number of drownings per month at different beaches.
- This is how the (jittered) scatter plot looks:

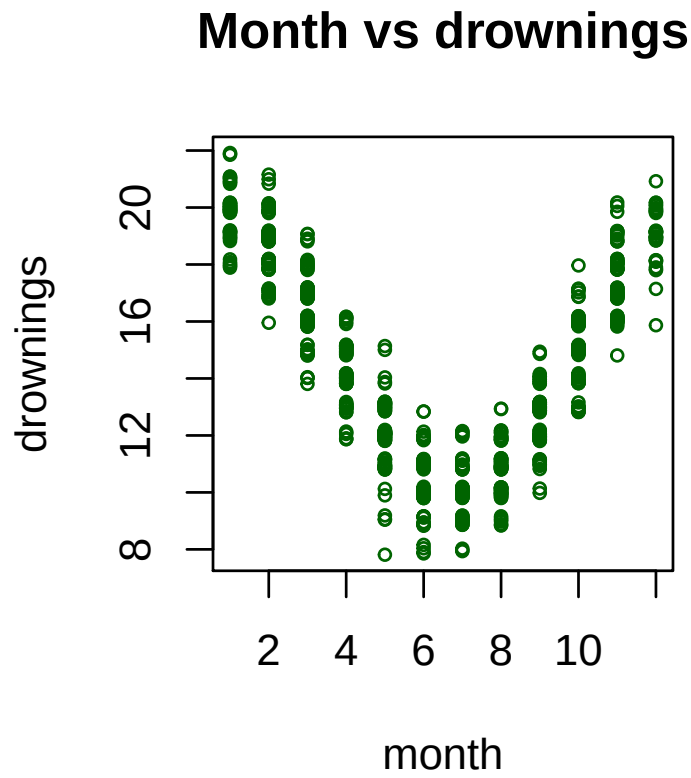


- Conclusion: Ice cream consumption causes drowning!

Example: To explain or to predict? II

- It is clear that we can use ice cream sales to predict the number of drownings.
- However, to explain the situation we need further investigation (or expert advice).
- This is what can explain the situation (time of the year)





Explanatory and predictive models I

We use regression models as a tool to either predict or explain.

- All of the different types of models that we have studied (Normal, logistic, Poisson, negative binomial, . . .) can be used for either purpose.
- This set of slides will clarify the distinction and between predictive and explanatory applications and give an overview of the important issues for each type of application.
- In general,
  - For a **predictive model**, the goal is to predict the response, based on the observed values of the explanatory variables.
  - For an **explanatory model**, the goal is to explore the relationships between the explanatory variables and the response.

Two Types of Explanatory Modelling I

In this course we will divide explanatory modeling into the following sub-categories:

- Descriptive modeling: regression models are used to describe the associations (e.g., correlations) between the response and the explanatory variables.
- Causal modeling: regression models are used to describe the causal effects that explanatory variables have on the response.

These Things Stay the Same

Certain aspects of regression modeling are the same for all types of applications.

1. It is important to think about how the data was collected as this establishes the context for which it applies.
   - Consider the data as being a sample from a population—any models we build only apply to that population.
2. Data cleaning is an essential step—clearly we should not use observations that are errors.
3. Model diagnostics should be applied to a fitted model—if necessary, suitable modifications should be made to the model.

What is Different? I

Important differences occur with respect to model evaluation, model selection and model interpretation.

- Predictive models
  - models are evaluated in terms of their ability to predict the response for new observations.
  - a key feature of predictive modeling is obtaining a realistic estimate of prediction error.
  - model selection usually involves minimizing some measure of prediction error.
  - achieving low prediction error requires balancing how well the model fits the data and model complexity.
  - the estimated regression coefficients are of little interest.

What is Different? II

- Explanatory models
  - models are chosen to answer specific questions about the way the response is related to the explanatory variables.
  - the estimated regression coefficients are crucial in answering these questions.
  - how well the model fits the data (i.e., R^2) is of little concern.
  - often a number of different models are needed to answer different questions about how the response is related to the explanatory variables.

What is Coming? I

In the next set of slides we will look at predictive modeling in more depth.

- We will look at a number of examples using different types of regression models: normal, Poisson, logistic.
- Methods of selecting models and evaluating their predictive ability will be discussed.

What is Coming? II

The set of slides after that will look at the two types of explanatory modeling: first descriptive modeling and then causal modeling. Important topics include:

- Identifying models to explore specific aspects of the relationship between the response and the explanatory variables.
- Causal effects and the circumstances under which causality can be inferred.
- Using causal diagrams to identify what model should be used to estimate a particular causal effect.