Introduction to Statistical Learning with Visualization

Dr YINGHUI WEI

Email: yinghui.wei@plymouth.ac.uk

Overview

Real-life data visualization

Click on each item below:

- ONS: infographic for TSA
- ONS: Survival to age 100
- Guardian website: The best of infographics
- BBC website: taking data visualization eye candy to efficiency
- Bank of England: data visualization competition
- NASA: Global Climate Change

Introduction to Visualization

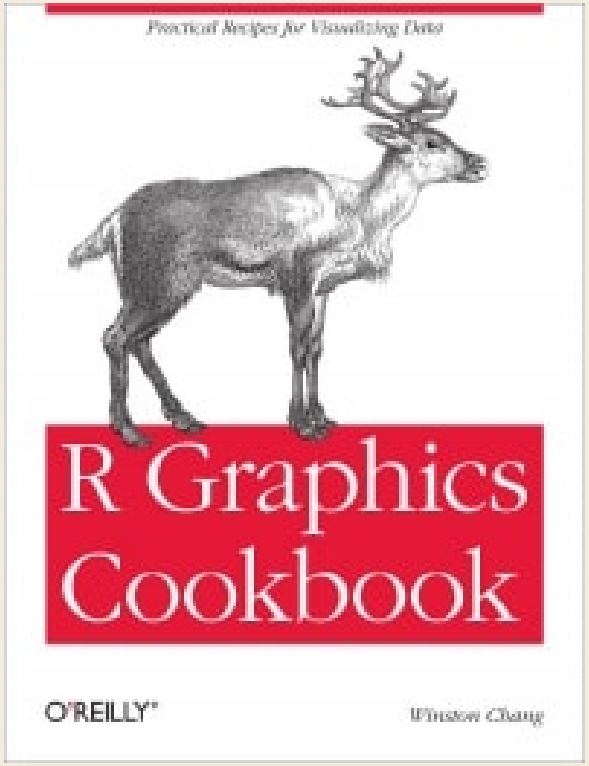
- Visualizations help people see things that were not obvious to them by looking at the raw data.
- Even when data volumes are very large, patterns can be spotted quickly and easily.
- Visualizations convey information in a universal manner and make it simple to share ideas with others.

Graphs

Some widely used graphs:

- Bar
- Boxplot
- Histogram
- Pie
- Scatter plot
- Contour plot
- Line graphs
- Heat map

Recommended Text for Visualization



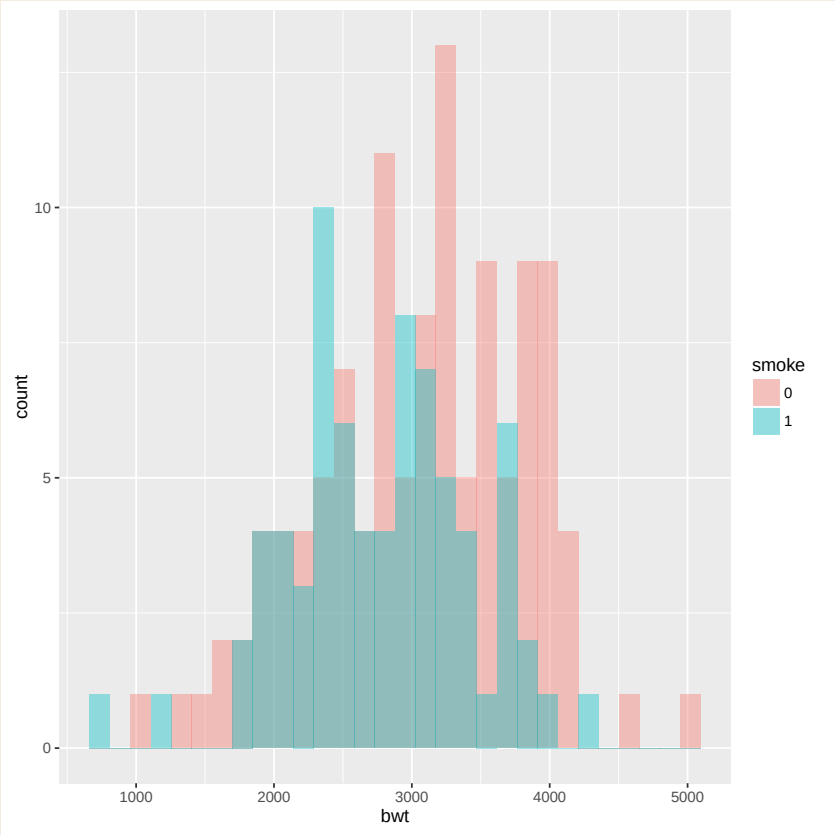
`http://www.cookbook-r.com/`

Visualization Example - Birth Weight

In this application, we investigate risk factors associated with low birth weight.

Visualization Example - Birth Weight

In this application, we investigate risk factors associated with low birth weight.

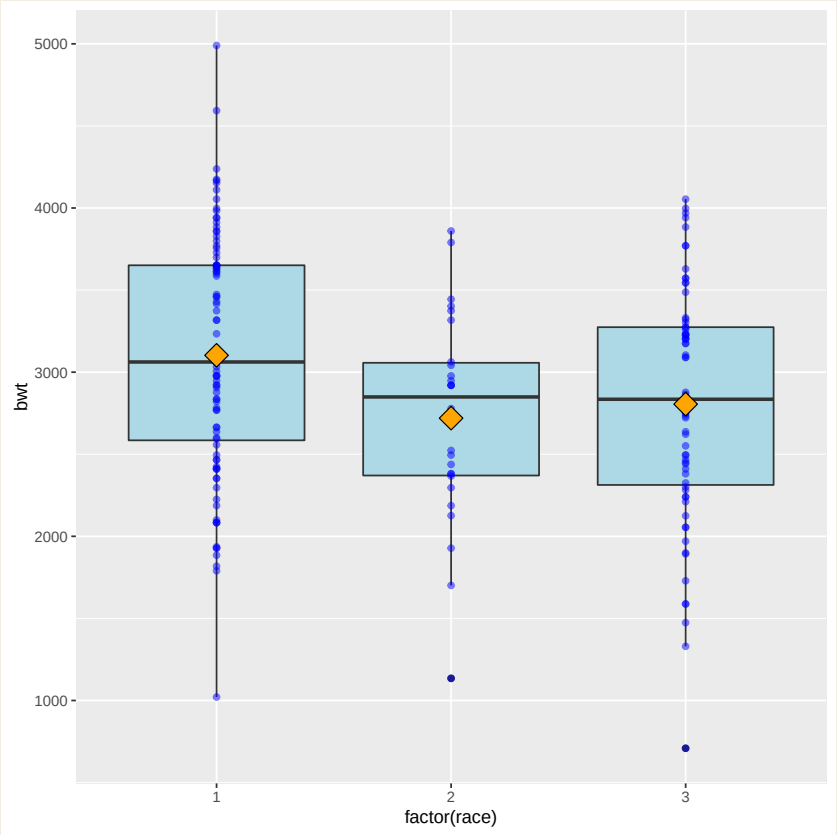


Visualization Example - Birth Weight

Box plot

Visualization Example - Birth Weight

Box plot

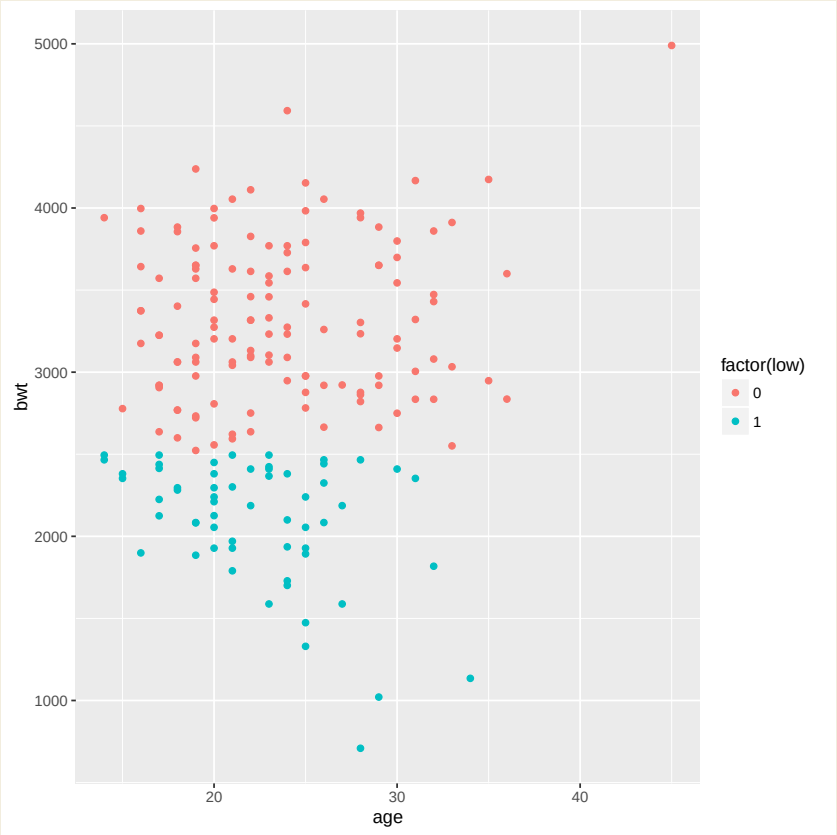


Visualization Example - Birth Weight

Scatter plot

Visualization Example - Birth Weight

Scatter plot

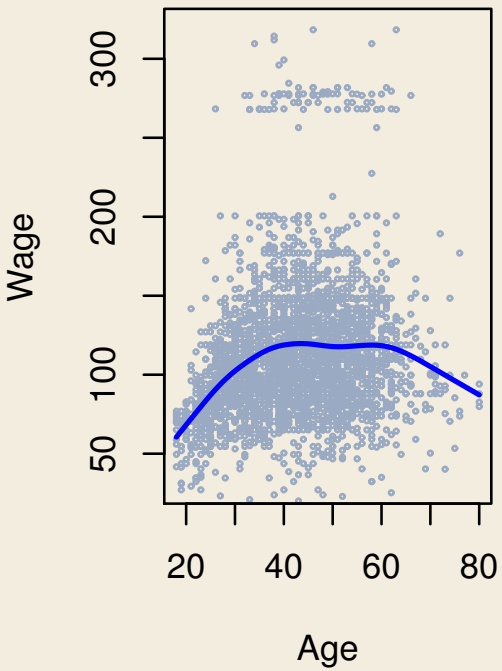


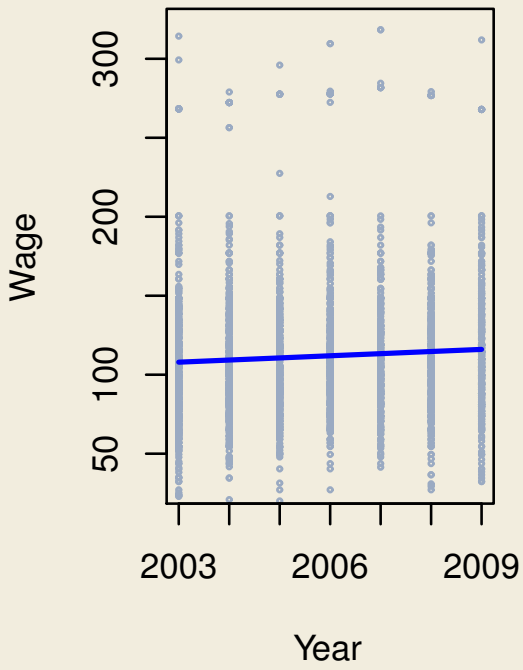
Visualization Example - Wage

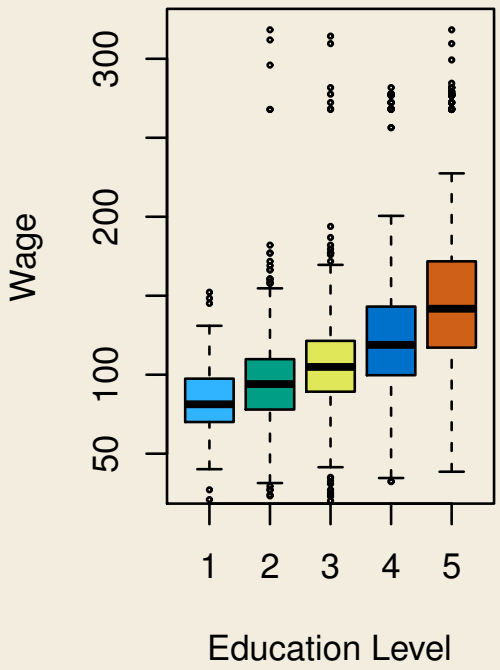
In this application, we examine a number of factors that relate to wages for a group of males from the Atlantic region of the United States. In particular, we wish to understand the association between an employee's age and education, as well as the calendar year, on his wage.

Visualization Example - Wage

In this application, we examine a number of factors that relate to wages for a group of males from the Atlantic region of the United States. In particular, we wish to understand the association between an employee's age and education, as well as the calendar year, on his wage.



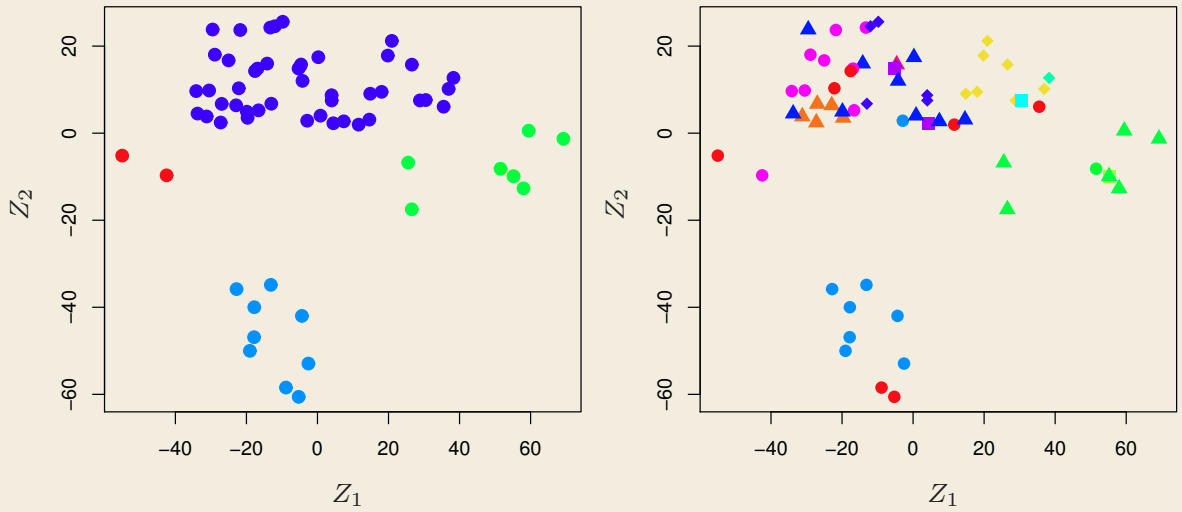




Visualization Example - Gene Expression Data

In this application, we consider the a data set consisting of 6,830 gene expression measurements for each of 64 cancer cell lines. Instead of predicting a particular output variable, we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements.

Visualization Example - Gene Expression Data

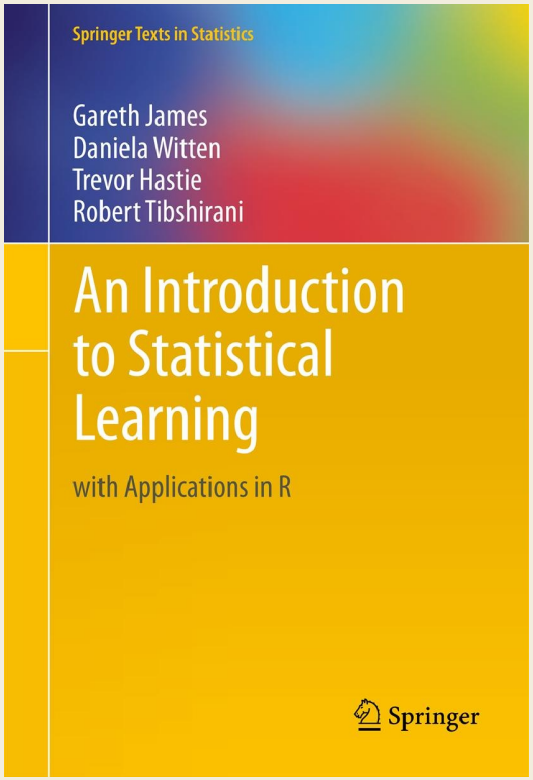


Left: Representation of the NCI60 gene expression data set in a two-dimensional space, Z_1 and Z_2 . Each point corresponds to one of the 64 cell lines. There appears to be four groups of cell lines, which we have represented using different colors. Right: Same as left panel except that we have represented each of the 14 different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

Introduction to Statistical Learning

- Statistical learning refers to a set of tools for modelling and understanding complex data sets.
- The field encompasses many methods such as regression models, classification, regression trees, bagging, random forest and boosting methods.
- With the explosion of "Big Data" problems, statistical learning has become a very hot field in medicine, engineering as well as marketing, finance, and other business disciplines.

Recommended Text for Statistical Learning



Ch 1. Introduction
Ch 2. Statistical Learning
Ch 3. Linear Regression
Ch 4. Classification
Ch 5. Resampling Methods
Ch 6. Linear Model Selection and Regularization
Ch 7. Moving Beyond Linearity
Ch 8. Tree-Based Methods
Ch 9. Support Vectors Machines
Ch 10. Unsupervised Learning

`http://www-bcf.usc.edu/~gareth/ISL/index.html`

Variables Type

Quantitative variables take on numerical values:

- Examples include a person's age, height, or income, the value of a house, categorical and the price of a stock.

Qualitative variables take on values in one of K different classes, or categories:

- Examples include a person's gender (male or female), the brand of product purchased (brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia).

Classification Questions

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user IP address, past transaction history, and so forth.

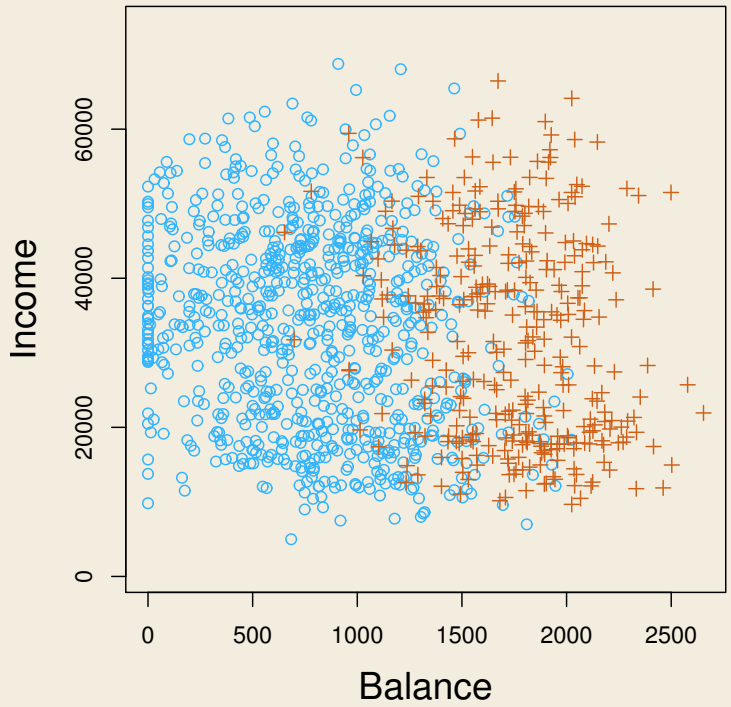
Classification Techniques

There are many possible classification techniques, or classifiers, that one might use to predict a qualitative response. Three most widely-used classifiers:

- Logistic Regression
- Linear Discriminant Analysis
- K-Nearest Neighbours

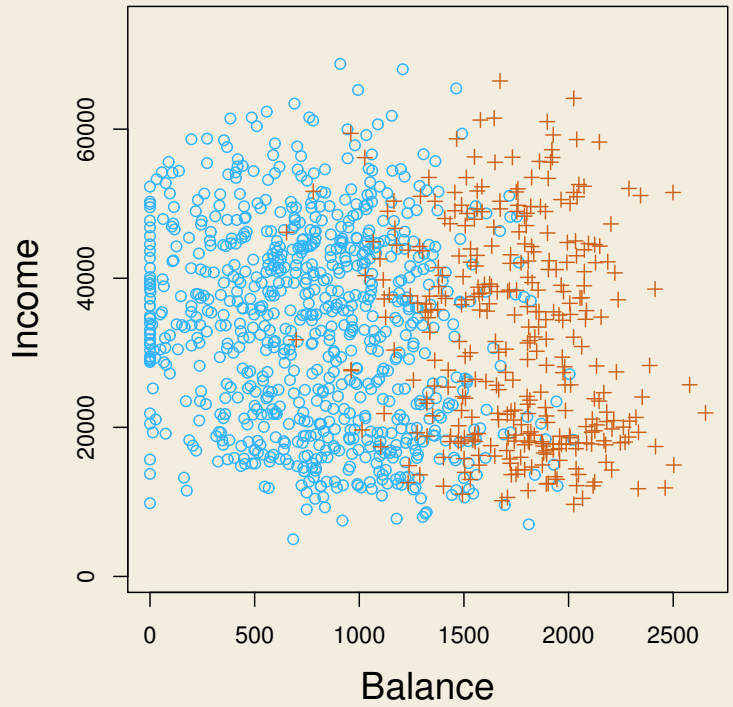
Default Data

We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.



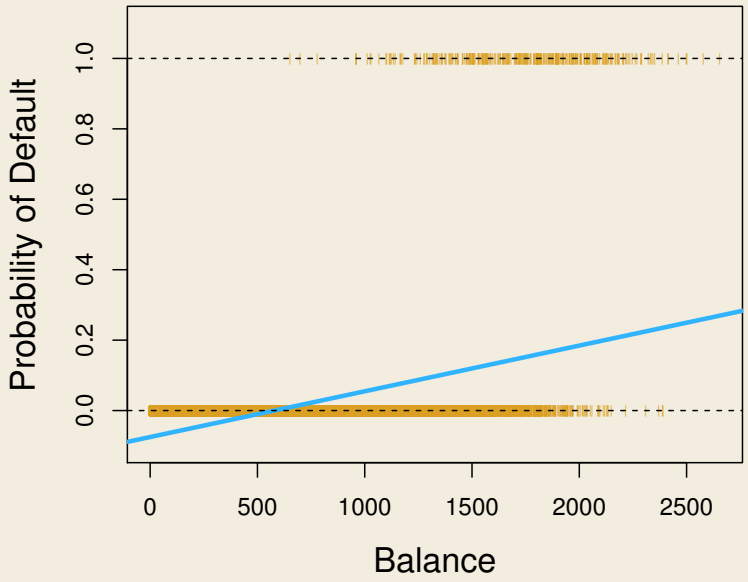
Default Data

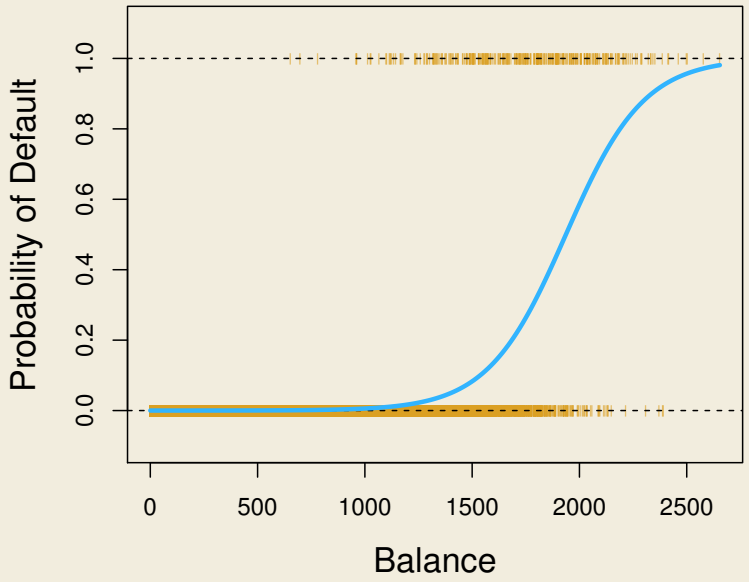
We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.



Which variable would you choose as a predictor of *default* (orange).

Default Data





Model

Logistic Model

$$\begin{aligned} \text{logit}\left(p(X)\right) &= \beta_0 + \beta_1 X \\ \log\left(\frac{p(X)}{1-p(X)}\right) &= \beta_0 + \beta_1 X \end{aligned}$$

Model

Logistic Model

$$\begin{aligned} \text{logit}\left(p(X)\right) &= \beta_0 + \beta_1 X \\ \log\left(\frac{p(X)}{1-p(X)}\right) &= \beta_0 + \beta_1 X \end{aligned}$$

- $p(X)$ represents the probability of success, $p(Y = 1|X)$.
- The fraction $\dfrac{p(Y = 1|X)}{1 - p(Y = 1|X)}$ represents the odds of the event.
- Values of the odds close to 0 and $+\infty$ indicate very low and very high probabilities of default, respectively.

Maximum Likelihood

We use maximum likelihood to estimate the model parameters. The likelihood function is:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)) .$$

We find the estimated values of β_0 and β_1 which maximise this likelihood. The estimates can be obtained by using standard packages in R.

Results

For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance are given in the following table:

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
Balance	0.0055	0.0002	24.9	< 0.0001

Table: Estimated coefficients in univariate logistic model for the Default data

Making Predictions

Once the coefficients have been estimated, it is a simple matter to compute the probability of default for any given credit card balance. For example, using the coefficient estimates given in Table 1, we predict that the default probability for an individual with a balance of $ 1, 000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

Multiple Logistic Regression

We now consider the problem of predicting a binary response using multiple predictors. The multivariable logistic model is:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p,$$

and so

$$p(X) = \frac{exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1 + exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}$$

Parameters will be estimated using maximum likelihood.

Multiple Logistic Regression

For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income (thousands of dollars) and student status (Yes or No) are given in the following table:

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.09	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[yes]	-0.6468	0.2362	-2.74	0.0062

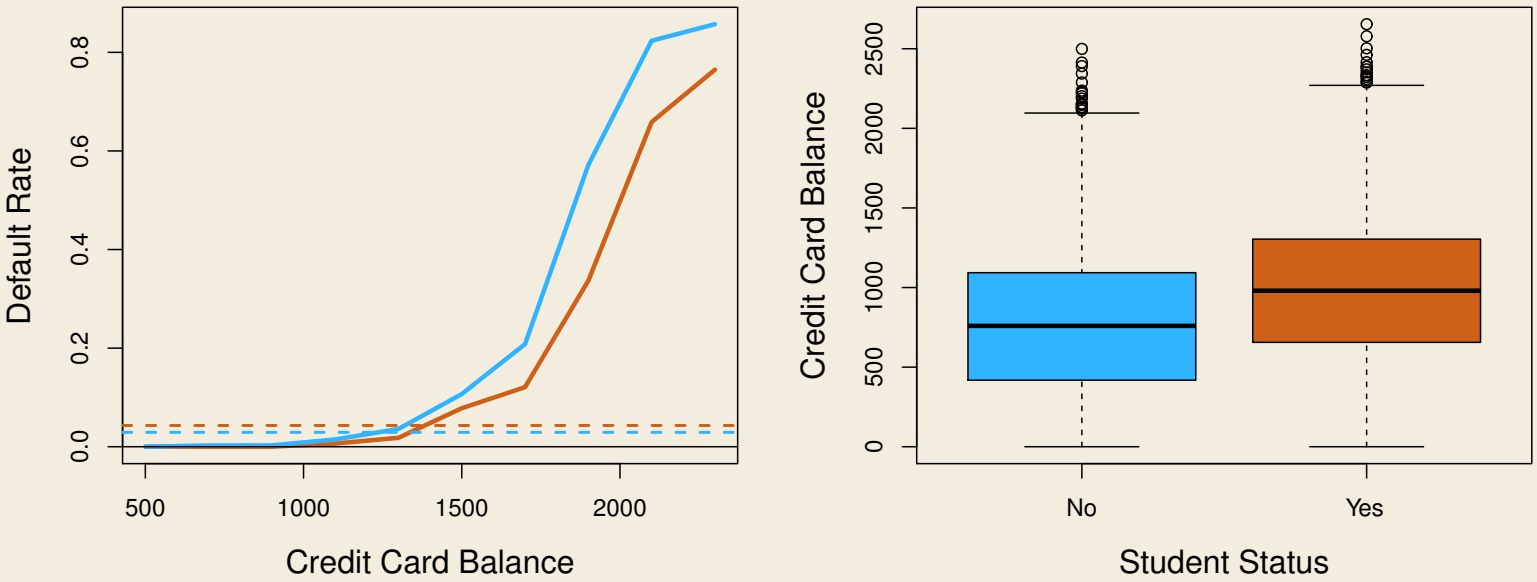
Table: Estimated coefficients in multivariable logistic model for the Default data

Making Predictions

For example, a student with a credit card balance of \$1, 500 and an income of \$40, 000 has an estimated probability of default of

$$\hat{p}(x) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058$$

Multiple Logistic Regression



This illustrates the dangers associated with performing regressions involving only a single predictor when other predictors may also be relevant. As in the linear regression setting, the results obtained using one predictor may be quite different from those obtained using multiple predictors, especially when there is correlation among the predictors.

References

- James G., Witten D., Hastie T., Tibshirani R. (2013), *An introduction to Statistical Learning**, Springer.
- Wei Y. (2016) Visualization Notes, Plymouth University.
- GGplot2 Cheat Sheet, `https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf`

* Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Progress Checklist

- Start now!
- Study your project
- Read relevant chapters in the recommended books
- Using R, produce graphs and apply statistical learning methods to your project
- Understand the underlying methods
- Be able to interpret your results
- Write up your presentation slides...practice your presentation... and finally write up your scientific report