“I keep saying that the sexy job in the next 10 years will be statisticians”

— Hal Varian, chief economist at Google

... “And I’m not kidding”

Outline

BMa&CompSc and Honours

What jobs can a statistician get?

What jobs have I got?

Doing a PhD in statistics
I chose this path after high school, not with a particular endpoint in mind, maths was just my strong suit in high school

Combination of

- pure maths
- applied maths
- statistics
- computer science

In my experience these are not distinct areas
Honours

I majored in statistics because I enjoyed analysing data and solving problems

Honours = subject work and a project

My project was “Using Linear Models to Estimate Home Ground Advantage of AFL Teams”

HGA: constant, linear, team only, ground only, interaction, away team disadvantage
<table>
<thead>
<tr>
<th>Team Effect</th>
<th>Season</th>
<th>1997</th>
<th>1999</th>
<th>2001</th>
<th>2003</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelaide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bris Lions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlton</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collingwood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essendon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fremantle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geelong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hawthorn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kangaroos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melbourne</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Port Adel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richmond</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St Kilda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sydney</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W Bulldogs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advantage Parameter</td>
<td>Estimate</td>
<td>Standard Error</td>
<td>P-value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>----------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{Adel}$</td>
<td>19.0</td>
<td>4.1</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{Bris}$</td>
<td>9.5</td>
<td>4.9</td>
<td>0.055</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{Geelong}$</td>
<td>8.7</td>
<td>5.6</td>
<td>0.121</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{Melb}$</td>
<td>5.7</td>
<td>2.3</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{Perth}$</td>
<td>24.0</td>
<td>4.1</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_{Sydney}$</td>
<td>10.6</td>
<td>4.8</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Statistician?

Business/public sector will snap you up - they need quantitative people!

Stats = data management
There is a snowballing culture of collecting data without the knowledge of how to use it

“I’ve got this data, what can I do with it?”
Stats gives you the tools to deal with this
Jobs

Statistics Center Abu Dhabi

- Advertised in 2010 to Australians
- ABS for UAE
- Tax free salary packages: $US170,000 - $US215,000 + benefits
- Senior Statistical Researcher × 6, Data Dissemination Specialist, Data Quality Specialist, Econometrician, Head of Methodology & Analysis, Head of Social Statistics
Total Pay: AU$46,787 - AU$147,376
### Recent job advertisements

<table>
<thead>
<tr>
<th>Company</th>
<th>Description</th>
<th>Position</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche</td>
<td>Pharmaceutical</td>
<td>Statistical Analysis and Programmer</td>
<td></td>
</tr>
<tr>
<td>Queensland Health</td>
<td>Health Stats</td>
<td>Analyst Intern</td>
<td>$50,000-$55,000</td>
</tr>
<tr>
<td>Mango solutions</td>
<td>Consulting</td>
<td>R Developer</td>
<td></td>
</tr>
<tr>
<td>The University of Bristol</td>
<td>Clinical Sciences</td>
<td>Medical Statistician/Data Manager</td>
<td>£30,000-£34,000</td>
</tr>
<tr>
<td>Insureware</td>
<td>Insurance</td>
<td>Actuary</td>
<td>$55,000-$80,000</td>
</tr>
<tr>
<td>Smart Odds</td>
<td>Statistical Consultants</td>
<td>Quantitative Analyst</td>
<td></td>
</tr>
</tbody>
</table>
Position: SSABSA (Now SACE Board)

Description: Project Officer (Scaling)

- Perform the end-of-year scaling
- Creating and manipulating database queries
- Programming
- Write reports
Second Job

**Position:** Data Management and Analysis Centre (DMAC) (Discipline of Public Health, University of Adelaide)

**Description:** Statistician

- Survival Analysis for the AOANJRR
- Consulting (sample size, analysis)
- Contribute to research papers
- Data preparation and cleaning
Motivating scenario for PhD - Prostate cancer

Current method for screening: prostate-specific antigen (PSA)
- Poor sensitivity and specificity
- Baseline PSA is variable over patients
- PSA is influenced by other benign prostate conditions
- PSA: ineffective yes/no threshold based on one marker
Proteomic mass spectrometry

... Expression of proteins/peptides from a blood sample
- Use these profiles to differentiate cancer from healthy patients
- Use maths/stats to create effective screening test
- Primary dataset of prostate cancer/healthy patients
The data

- **Cancer**
- **Control**
Our two data points from each class from spectra/peptide expression
We need more observations to create a rule or model.
Regression:

\[ \begin{align*}
\min & \sum r_i^2 \\
\text{or} \\
\min & \sum f(r_i)
\end{align*} \]
Optimal separation:

\[
\max \sum d_i \\
or \\
\max \sum f(d_i)
\]
3 Dimensions:
Separate by a plane
What about 4D+?
Challenges

Unfortunately not as simple as all that

- The data is not as clean as shown - preprocessing is required
- Which proteins/peptides do we use - how many?
- With this data - there is no separating line/plane/hyperplane (SVM, RF, LDA)
- More than two classes
- How can we best visualise/summarise the data?
- Training and test data to prove effectiveness of model
Thank you

tyman.stanford@adelaide.edu.au