Statistical Analysis
for the Military Decision Maker (Part II)

Professor Ron Fricker
Naval Postgraduate School
Monterey, California
Goals for this Lecture

- Linear and other regression modeling
  - What does it mean to model?
  - What are the assumptions?
  - What should I ask during a briefing?
On to Model Building!

- Up to now, we’ve discussed descriptive and inferential statistics
  - Numerical and graphical summaries of data
  - Confidence intervals
  - Hypothesis testing
- Can apply those tools and build *models* to try to explain data
  - For each “\(Y\)” in my data I also observe an “\(X\)”
  - Can I use \(X\) to say something about \(Y\)?
Why Model?

- Raw data by itself (pairs of $X$ and $Y$) often too hard to interpret
- Scatter plots informative, but can also sometimes have too much information
- Linear regression models the relationship between $X$ and $Y$ via a linear equation
- General expression for a linear equation:

$$Y = \beta_0 + \beta_1 X$$

- $\beta_1$ is the slope (change in $Y$ for a unit change in $X$)
- $\beta_0$ is the intercept (value of $Y$ when $X = 0$)
The Idea of Linear Regression

- There is some unknown, true relationship between \( X \) and the average \( Y \)
- But we only observe the individual \( Y \)s
- Hence observed data is “sprinkled around” the line
- So the model is fit to meet these assumptions
The Idea of Linear Regression

- There is some unknown, true relationship between $X$ and the average $Y$
- But we only observe the individual $Y$s
- Hence observed data is “sprinkled around” the line
- So the model is fit to meet these assumptions

The “game” is to guess the line from the data
Estimating the actual linear relationship is given by the regression of $Y$ on $X$.

$Y = \beta_0 + \beta_1 X$

Scatterplot of observed data

$Y = -0.4 + 3.08 X$
Simple Linear Regression Model

\[ Y = \beta_o + \beta_1 X + \varepsilon \quad \text{with} \quad \varepsilon \sim N(0, \sigma^2) \]
Regression Assumptions

- The $X$s and $Y$s come from a population where:
  - Mean of $Y$ is a linear function of $X$
    - Expressed by the regression line $\text{E}(y) = \beta_0 + \beta_1 x$
    - Scatter plot indicates linear model plausible
  - The errors are normally distributed: $\mathcal{N}(0, \sigma^2)$
    - In particular variance does not depend on $X$
    - So, $Y$s (equivalently errors) are independent
• Simple linear regression: One $Y$ variable and one $X$ variable ($y_i=\beta_0+ \beta_1x_i+\varepsilon_i$)

• Multiple regression: One $Y$ variable and multiple $X$ variables
  – Like simple regression, we’re trying to model how $Y$ depends on $X$
  – Only now we are building models where $Y$ may depend on many $X$s

  $y_i=\beta_0+ \beta_1x_{1i} + \cdots + \beta_mx_{mi} +\varepsilon_i$
Many Forms of Multiple Regression

- **Polynomial regression**

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots \]

- **Interaction terms**

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \]

- **More complicated models**

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 \]

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2) + \beta_3 X_1 \log(X_2) \]
• A model is a *simplified representation of reality*

• Models can be used for:
  – Compactly summarizing data
  – Understanding relationships between variables
  – Making predictions

All models are wrong, some are useful.
  – George Box
What are Other Types of (Statistical) Modeling Techniques?

• Linear models
  – Simple linear regression
  – Multiple regression

• Generalized linear models
  – Logistic regression

• Generalized additive models

• Survival analysis (models)

• Tree-based models

• Time series & forecasting models

• Econometric models
Case #4: Evaluating the Effect of IA Deployment on Navy Retention

- Do individual augmentation deployments have an effect on retention?
- Deployment generally assumed to reduce retention:

“...many factors were sources of dissatisfaction and reasons to leave the military. The majority of factors (62 percent) were associated with work circumstances such as ... the frequency of deployments ...”

GAO, NSAID-99-197BR, March 1999

- Do actions match perceptions?
What is Individual Augmentation?

• Individual sailors and officers sent to augment other (often non-Navy) units
• Differs from usual deployments
  – Individual vice unit deployment
  – Often with little notice
• Then-CNO Admiral Mullen:

  “I see this as a long-term commitment by the Navy. I’m anxious to pitch in as much as we possibly can, for the duration of this war. Not only can we do our share, but [we can] take as much stress off those who are deploying back-to-back...”

Number Starting IA Deployment by Year
(Active Component Only)

<table>
<thead>
<tr>
<th>Year</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>47</td>
</tr>
<tr>
<td>2003</td>
<td>277</td>
</tr>
<tr>
<td>2004</td>
<td>977</td>
</tr>
<tr>
<td>2005</td>
<td>2624</td>
</tr>
<tr>
<td>2006</td>
<td>4838</td>
</tr>
<tr>
<td>2007</td>
<td>8808</td>
</tr>
<tr>
<td>2008</td>
<td>1843</td>
</tr>
</tbody>
</table>

(Jan – Mar)
Deployment Locations
(Active Component Only)
Research Question:
Does IA Affect Navy Retention?

• With almost 20,000 AC sailors and Navy officers IA deployed in the past 6 years, Navy leadership interested in whether it’s hurting retention

• RADM Masso, Deputy Chief of Naval Personnel:

  “Since 2002, 82 percent of our IA’s have come from the Reserve component, yet I see letters of resignation from officers listing a fear of IA duty as being the reason they are getting out. IA duty affects two percent of the surface warfare officer (SWO) community, yet if you speak to a junior officer on the waterfront, you would think that half of their wardroom are IA’s.”

Almost 20,000 AC Navy Personnel IA Deployed Since March 2002

Officer vs. Enlisted

Warrant Officer Ranks

Officer Ranks

Enlisted Pay Grades
Deployed Sailors Largely in Security, Medical, IT, Admin, & Supply Ratings

Enlisted Ratings

Count

Rating
Previous Work on Deployment Effects

• From prior studies of effects of Perstempo:
  – Some deployment positively related to retention, too much can be negative
  – Hostile deployments generally positively related to retention

• See:
  – Hosek and Totten (1998, 2002) for enlisted personnel studies
  – Fricker (2001) for study of military officers
Modeling Effects of IA

• Approach: Model individuals at their reenlistment decision point or end of initial service obligation
  – Compare between those that had an IA deployment prior to their decision versus those that did not
• Relevant cohort: those “at risk” of (1) an IA and (2) leaving the Navy
  – Also subset to only those with deployment experience
• “IAer:” An individual who made a stay-in/get-out decision after an IA deployment
  – If stay-in/get-out decision observed prior to IA, then individual was a “non-IAer” at that time
The Data

• **IA data (OPNAV Pers-4)**
  - Information on Navy personnel deployed as IAs
    • 21,340 records (Mar 02 – Mar 08 + future IAs)
  - Relevant fields
    • Identifiers: Name, rank, SSN
    • IA scheduling: Date deployed, est. BOG, est. return date
    • Other IA information: Location, billet title, UIC

• **USN data (DMDC)**
  - Information on all Navy personnel for past decade
    • 893,461 records (Oct 97 – Sept 07)
  - Relevant fields
    • Identifiers: Name, rank, SSN
    • Demographics: rate/designator, gender, race, family status
    • Deployment experience
### Navy (DMDC) Data

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>893,461</td>
<td>Total active duty Navy personnel (10/97-9/07)</td>
</tr>
<tr>
<td>-174,049</td>
<td>Officers and records with duplicate SSNs</td>
</tr>
<tr>
<td>-448,949</td>
<td>No decision after 3/02, all data missing, or invol. sep.</td>
</tr>
<tr>
<td>-36,637</td>
<td>No deployment experience (prior to decision)</td>
</tr>
<tr>
<td>-382</td>
<td>No data year prior to decision</td>
</tr>
<tr>
<td>233,444</td>
<td></td>
</tr>
</tbody>
</table>

### IA Data

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,469</td>
<td>Total Navy IA personnel (3/02-9/07)</td>
</tr>
<tr>
<td>-4,534</td>
<td>Officers and warrant officers</td>
</tr>
<tr>
<td>-8,972</td>
<td>No decision after IA deployment</td>
</tr>
<tr>
<td>1,963</td>
<td></td>
</tr>
</tbody>
</table>
Comparing the Populations by Gender (Enlisted Only)

- Whole Enlisted (n=719,412)
- All Enlisted IAers (n=10,888)
- Enlisted Deployers (n=233,444)
- Enlisted IAers w/ Decisions (n=1,963)
Comparing the Populations by Race/Ethnicity (Enlisted Only)

![Bar chart showing race/ethnicity distribution]

- Whole Enlisted Navy
- All Enlisted IAers (that match)
- Enlisted Deployers
- Enlisted IAers w/ Decisions
Comparing the Populations by Family Status (Enlisted Only)

- **Whole Enlisted Navy**
- **All Enlisted IAers (that match)**

- **Enlisted Deployers**
- **Enlisted IAers w/ Decisions**
Comparing the Populations by Pay Grade (Enlisted Only)
Modeling the Decision Point: Stay In or Get Out of the Navy

- Model a binary decision point
  - Function of fixed (e.g., gender) and variable (e.g., family status) characteristics

All must have at least one deployment pre-decision
IAers must have IA pre-decision
Variable data values
Stay-go decision point

- Examples:
  - IAer:
  - Non-IAer:
  - Non-IAer:
Analytical Issues

• Analysis based on observational information from administrative datasets
• Can’t identify volunteers versus non-volunteers
• Must (imperfectly) infer some critical data on decision points
  – Expiration of enlistment contract or end of initial service obligation period
  – Deployment experience
Junior Officer Results: Comparing Raw Rates

- Odds IAer retained = 1.94
- Odds non-IAer retained = 0.76
- Odds ratio = 2.56
- “Statistically significant” result ($p<0.0001$)
Junior Officer Logistic Regression Model Results

<table>
<thead>
<tr>
<th></th>
<th>Log odds ($\beta$)</th>
<th>Std. error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.235</td>
<td>0.146</td>
<td>-1.61</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.356</td>
<td>0.042</td>
<td>-8.47</td>
</tr>
<tr>
<td>White</td>
<td>0.286</td>
<td>0.119</td>
<td>2.39</td>
</tr>
<tr>
<td>Black</td>
<td>0.585</td>
<td>0.132</td>
<td>4.41</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.392</td>
<td>0.132</td>
<td>2.96</td>
</tr>
<tr>
<td>Indian</td>
<td>0.441</td>
<td>0.197</td>
<td>2.23</td>
</tr>
<tr>
<td>Asian</td>
<td>0.326</td>
<td>0.134</td>
<td>2.43</td>
</tr>
<tr>
<td>Other</td>
<td>0.549</td>
<td>0.208</td>
<td>2.64</td>
</tr>
<tr>
<td>Married</td>
<td>-0.176</td>
<td>0.077</td>
<td>-2.28</td>
</tr>
<tr>
<td>Single w/dep</td>
<td>-1.243</td>
<td>0.096</td>
<td>-12.98</td>
</tr>
<tr>
<td>Single w/o dep</td>
<td>-1.154</td>
<td>0.080</td>
<td>-14.39</td>
</tr>
<tr>
<td>DesigOther</td>
<td>0.235</td>
<td>0.046</td>
<td>5.14</td>
</tr>
<tr>
<td>DesigSub</td>
<td>0.171</td>
<td>0.072</td>
<td>2.36</td>
</tr>
<tr>
<td>DesigSupply</td>
<td>0.573</td>
<td>0.077</td>
<td>7.44</td>
</tr>
<tr>
<td>DesigSurface</td>
<td>0.231</td>
<td>0.052</td>
<td>4.47</td>
</tr>
<tr>
<td>IA</td>
<td>0.944</td>
<td>0.074</td>
<td>12.74</td>
</tr>
</tbody>
</table>

- Model for junior officers:
  - Coefficient for IA = 0.944, so adj. O.R. = 2.57
  - Virtually equivalent to raw O.R. = 2.56
Enlisted Personnel Results: Comparing Raw Rates

- Odds IAer retained = 2.01
- Odds non-IAer retained = 1.55
- Odds ratio = 1.30
- “Statistically significant” result ($p<0.0001$)
• Model controlled for pay grade, gender, race/ethnicity, family status, AFQT, education, and year of decision

• Model for all IAers:
  – Coefficient for IA_Deployer_Ind = 0.427, so adjusted O.R. = 1.53

• Model just Iraq and Afghanistan IAers:
  – Coefficient for IA_Deployer_Ind = 0.660, so adjusted O.R. = 1.93

• Remember raw O.R. = 1.30
Comparing Retention Rates by Gender

PCT Retained by Gender and IA Status

<table>
<thead>
<tr>
<th>Gender</th>
<th>Non-IAer</th>
<th>IAer</th>
<th>Non-IAer</th>
<th>IAer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>61.18</td>
<td>66.17</td>
<td>57.52</td>
<td>70.94</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparing Retention Rates by Family Status

PCT Retained by Family and IA Status

<table>
<thead>
<tr>
<th>Family Status</th>
<th>Non-IAer</th>
<th>IAer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>53.89</td>
<td>59.23</td>
</tr>
<tr>
<td>Single w/ Children</td>
<td>62.11</td>
<td>68.67</td>
</tr>
<tr>
<td>Married</td>
<td>65.36</td>
<td>69.88</td>
</tr>
<tr>
<td>Joint Marriage</td>
<td>63.26</td>
<td>73.53</td>
</tr>
</tbody>
</table>

36
Comparing Retention Rates by Race/Ethnicity

PCT Retained by Race/Ethnicity and IA Status

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Non-Iaer</th>
<th>Iaer</th>
<th>Non-Iaer</th>
<th>Iaer</th>
<th>Non-Iaer</th>
<th>Iaer</th>
<th>Non-Iaer</th>
<th>Iaer</th>
<th>Non-Iaer</th>
<th>Iaer</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>58.3</td>
<td>65.0</td>
<td>66.7</td>
<td>68.8</td>
<td>59.7</td>
<td>69.5</td>
<td>67.6</td>
<td>70.0</td>
<td>54.5</td>
<td>77.5</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

54.4 68.8
Comparing Retention Rates by Pay Grade

PCT Retained by Pay Grade and IA Status

- **n=1**
- **n=9**
- **n=13**
Conclusions

• Thus far, IA deployment generally associated with higher retention rates
  – Consistent effects for both junior officers and enlisted personnel
    • Perhaps a paygrade effect for enlisted?
  – Self-selection and other effects present
    • Paygrade correlated with volunteer status?
• Thus far, hypothesis seemingly untrue: IA deployment causes significant decrease in propensity to stay in the Navy
Take-Aways

• Good modeling must appropriately account for structure(s) in the data and/or the underlying phenomenon
  – E.g., applying linear regression to censored data would produce incorrect results

• Empirically-based methods and models can move the discussion beyond opinion

• The concept of statistical significance may be irrelevant if the model is of the whole population
Case #5: Modeling IEDs

- **Goal:** Provide operational and tactical level staffs with an analytically-based *daily* assessment of probable IED locations

- **Potentially useful for:**
  - Deciding where to employ a limited number of neutralizing assets
  - Determining high threat areas for convoys
  - Assessing the effect of counter-measures

- **Output:** an easily-interpretable map overlay of future IED event likelihood
General Approach

• Employ data mining techniques to flexibly model high dimensional data
  – IED problem is inherently spatio-temporal
  – Model factors that vary in:
    • Space: Proximity to key infrastructure; etc.
    • Time: Religious events; political events; etc.
    • Space-Time: Number of IED events; Coalition force activity (we are the first to consider this factor)
• Does not lend itself to existing methods
  – It is very difficult to capture the effects of spatio-temporal factors
Make Predictions In and Around Small Road Segments

For each small region around a road segment…

...provide an assessment of how likely an IED will be there tomorrow.
Model Incorporates All Available Relevant Data
Time is a Critical Dimension in the Data

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Block ID (space)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of events w/in “X” meters in past “Y” days</td>
<td></td>
</tr>
<tr>
<td>Coalition force activity on route “A”</td>
<td></td>
</tr>
<tr>
<td>Distance to nearest “infrastructure”</td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing predictors and block ID in a 3D space-time diagram](image.png)
Output is Visual Depiction of Potential IED “Hot Spots”

Low probability of IED tomorrow

High probability of IED tomorrow
Data Processing Requires Many Steps

Data Sources

- Coalition Force Activity
- IED Event Data
- Geographic Information Data
- Route Data (names and check points)

Data Manipulation

- Clean data/Calculate
- Clean data/Calculate
- Visualize/calculate
- Compile/calculate
- Build Models

Modeling

Tools Used

- Excel/VBA
- ArcGIS
- S-Plus

Format

Probability map
Short timeframe binary response for each block

Conditional probability value for each block

Baseline event related predictors

Non-event related time varying predictors

Medium timeframe binary response for each block

Most important time invariant predictors

All time invariant predictors

Long timeframe binary response for each block

C&RT

Logistic Regression

Logistic Regression

Coefficients for the additive model that will produce tomorrow’s likelihood map
• The tool shows promise for assessing probable IED locations
  – Useful as a supplement to the tools already in use for operational and tactical level staffs
• The process captures a changing time-space relationship
  – These factors are allowed to change as the nature of the problem changes
• Model coefficients are interpretable
  – What factors play a positive or negative role in IED occurrence
Operators Think the Output is Useful

“This is exactly the type of tool that operators at the operational and tactical level need – a tool that will help them prioritize and allocate the scarce resources that they have.”

– Quote from an OR analyst in the Counter-IED cell currently operating in Western Iraq after viewing preliminary results
Take-Aways

- Modeling complex phenomenon may require more complicated and/or non-traditional methods
- Good data is critical to building good models
- Judging model “fit,” particularly for prediction models, may have little to do with conventional metrics ($p$-values, etc)
  - Statistical significance may have little meaning or use – prediction accuracy is what’s relevant
Some Briefing Questions

• Why did you choose this particular model (modeling approach)?

• What are the underlying assumptions of your model?
  – Is what you observe in your data consistent with these assumptions?
  – How robust are your results to violations in these assumptions?

• Did you do a sensitivity analysis?
  – At what point does the solution change and how far?

• In what ways does this model not reflect reality?
  – What simplifying assumptions did you have to make in order to fit the model?
Case #1: www.amstat.org/publications/jse/v5n2/datasets.starr.html.


Case #4: Fricker, R.D., Jr., and S.E. Buttrey, Assessing the Effects of Individual Augmentation (IA) on Active Component Navy Enlisted and Officer Retention, Naval Postgraduate School Technical Report, NPS-OR-08-003, 2008. (http://faculty.nps.edu/rdfricke/docs/NPS-OR-08-003.pdf)