Explanatory Statistical Modeling

From Linear Models to GLMs, Ordinal Models, and Survival

Diogo Ribeiro

Senior Data Scientist & Mathematician

October 28, 2025

Full Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Welcome

- What you'll get today:
  - Practical modeling workflow for explanatory analysis.
  - How to pick models for different outcome scales.
  - Diagnostics, validation, and communication.
- Materials: code snippets in R and Python.

About Me

- **Diogo Ribeiro** — Senior Data Scientist & Mathematician.
- Focus: time series, streaming analytics, and interpretable models.
- Motto: lean models, clean code, reproducible pipelines.

Explanatory vs Predictive

Explanatory

Emphasizes understanding relationships and effects. Inference-oriented.

Predictive

Emphasizes generalization error on unseen data. Forecast-oriented.

Key Point

Pick metrics, validation, and modeling choices aligned with your goal.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Data Types & Structures

- Numeric (int/double)
- Categorical (nominal/ordinal)
- Logical/Boolean
- Dates/Times

- Vectors, matrices
- Data frames/tibbles
- Missingness & encoding

R Refresher

```r
x <- 42.0            # numeric (double)
y <- 42L             # integer
f <- factor(c("A","B","A"))

df <- data.frame(x = 1:5, y = c(3, NA, 5, 8, 13))
mean(df$y, na.rm = TRUE)
```

Python Refresher (pandas)

```python
import pandas as pd
import numpy as np

df = pd.DataFrame({"x": range(5), "y": [3, np.nan, 5, 8, 13]})
df["y"].mean(skipna=True)
```

EDA: First Things First

- Distributions, outliers, missing data patterns.
- Group differences and relationships.
- Data leakage risks (especially in longitudinal/sensor data).

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Samples, Populations, & SEs

- Sample → estimate parameters.
- Standard errors quantify sampling variability.
- Confidence intervals communicate uncertainty.

Hypothesis Testing Basics

- Null vs alternative hypotheses.
- p-values, effect sizes, and power.
- Multiple testing: control FWER/FDR.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. **Linear Models**
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Simple Linear Regression

$y = \beta_0 + \beta_1 x + \varepsilon$

- Interpret β_1: expected change in y per unit x.
- Assumptions: linearity, independence, homoscedasticity, normality.

Multiple Linear Regression

$y = \alpha + \sum_k \beta_k x_k + \varepsilon$

- Interpret coefficients ceteris paribus.
- Address collinearity (VIF), feature selection, interactions.

R Example: LM

```r
url <- "https://peopleanalytics-regression-book.org/data/ugtests.csv"
ug <- read.csv(url)

m <- lm(Final ~ Yr1 + Yr2 + Yr3, data = ug)
summary(m)
confint(m)
```

Diagnostics: Linear Models

- Residual plots and Q-Q plots.
- Leverage and Cook's distance.
- Transformations or robust regression when assumptions fail.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. **Logistic Regression (Binary)**
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Logit Model

$$\log \frac{p}{1-p} = \beta_0 + \beta^\top x \quad \Rightarrow \quad \exp(\beta) = \text{odds ratios}$$

- Report OR with CIs.
- Calibrate probabilities (reliability diagrams).

R Example: Binary Logistic

```r
url <- "https://peopleanalytics-regression-book.org/data/speed_dating.csv"
sp <- read.csv(url)
sp_m <- subset(sp, gender == 1)
fit <- glm(dec ~ attr + intel + prob, data = sp_m, family = binomial())
est <- summary(fit)$coefficients
cbind(est, OR = exp(est[, "Estimate"]))
```

Model Assessment

- ROC, PR, and calibration.
- Pseudo-R^2 (McFadden, Cox–Snell, Nagelkerke).
- Train/test splits or cross-validation.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Proportional Odds Model

- Ordered categories via cumulative logits.
- Same slopes across thresholds; different intercepts.
- Check proportional odds assumption.

R Example: Ordinal

```r
library(MASS)
url <- "https://peopleanalytics-regression-book.org/data/managers.csv"
man <- read.csv(url)
man$performance_group <- ordered(man$performance_group,
  levels = c("Bottom","Middle","Top"))
mod <- polr(performance_group ~ test_score + group_size +
                yrs_employed + concern_flag, data = man)
summary(mod)
```

Interpreting Ordinal Models

- Effects on odds of being at/above each threshold.
- Marginal effects to translate into probabilities.
- Violations: partial proportional odds models.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Poisson and Negative Binomial

- Poisson: $\mathbb{E}[y] = \mathbb{V}[y] = \mu$.
- Overdispersion $\Rightarrow$ consider Negative Binomial.
- Offsets for exposure/time at risk.

R Example: Poisson GLM

```r
df <- data.frame(y = rpois(100, lambda = 2),
                 x = rnorm(100), log_exp = log(1))
po <- glm(y ~ x + offset(log_exp), data = df, family = poisson())
summary(po)
```

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. **Survival Analysis**
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Survival Basics

- Time-to-event outcomes; censoring is common.
- Kaplan–Meier for $S(t)$ curves.
- Cox PH: $\lambda(t|x) = \lambda_0(t)\exp(\beta^\top x)$.

R Example: Survival

```r
library(survival)
time <- c(5, 8, 12, 18, 20)
cens <- c(1, 0, 1, 1, 0)  # 1=event, 0=censored
x <- c(0, 1, 0, 1, 1)
cox <- coxph(Surv(time, cens) ~ x)
summary(cox)
```

Assumptions & Checks

- Proportional hazards diagnostics (Schoenfeld residuals).
- Ties handling (Efron/Breslow).
- Alternative AFT models when PH fails.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Experiment Design

- Randomization, control/treatment, stratification.
- Power, minimum detectable effect, test duration.
- Guard against novelty and interference.

Analysis & Pitfalls

- Intention-to-treat vs per-protocol.
- Sequential peeking and alpha spending.
- Multiple comparisons corrections.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. **Diagnostics & Robustness**
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Robust Regression

- M-estimators (Huber, Tukey).
- Resistant to outliers in y.
- Compare with OLS: coefficient stability.

Influence Diagnostics

- Leverage, Cook’s distance, DFFITS, DFBETAS.
- Actionable thresholds and domain knowledge.

Model Uncertainty

- Competing models & information criteria (AIC/BIC).
- Sensitivity analysis.
- Pre-registration for confirmatory studies.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. **Communicating Results**
12. Hands-on Exercises
13. Case Studies
14. Appendix

Explain with Clarity

- Report estimates with SE/CI.
- Visualize partial effects/marginal means.
- State assumptions, limitations, and scope.

Tables & Figures

- Coefficient tables: tidy, labeled, units included.
- Visuals: avoid chartjunk; annotate key effects.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. **Hands-on Exercises**
13. Case Studies
14. Appendix

Exercise 1: Linear Model

1. Fit $y \sim x_1 + x_2$.
2. Check residuals and leverage.
3. Explain β in plain language.

Exercise 2: Logistic Model

1. Fit $y \in \{0, 1\}$ with 3 predictors.
2. Report OR with 95% CI.
3. Plot calibration.

Exercise 3: Ordinal Model

1. Fit proportional odds model.
2. Check proportional odds assumption.
3. Translate to category probabilities.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. **Case Studies**
14. Appendix

Case: Employee Performance

- Outcome: ordered performance groups.
- Predictors: test scores, tenure, team size.
- Model: ordinal logistic with checks.

Case: Click-through Rate

- Outcome: binary click/no click.
- Predictors: layout, position, user segment.
- Model: logistic with interactions.

Case: Time to Churn

- Outcome: time-to-churn with censoring.
- Predictors: usage frequency, support tickets.
- Model: Cox PH with diagnostics.

Agenda

1. Welcome & Context
2. Data Foundations
3. Inference Basics
4. Linear Models
5. Logistic Regression (Binary)
6. Ordinal Logistic Regression
7. Count Models & Overdispersion
8. Survival Analysis
9. Designing A/B Tests (Bonus)
10. Diagnostics & Robustness
11. Communicating Results
12. Hands-on Exercises
13. Case Studies
14. Appendix

Useful R Packages

- `stats`, `MASS`, `survival`, `brant`
- `car` (VIF), `sandwich`, `lmtest`
- `ordinal`, `MASS::polr`

Useful Python Packages

- `statsmodels`, `scikit-learn`
- `lifelines` (survival), `pandas`, `numpy`

Glossary

- **OR**: odds ratio.
- **CI**: confidence interval.
- **PH**: proportional hazards.

References

- Gelman & Hill (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*.
- Harrell (2015), *Regression Modeling Strategies*.
- Fox (2016), *Applied Regression Analysis and GLMs*.

Key Takeaways

- Match model to outcome and study goal.
- Diagnose and validate before concluding.
- Communicate effects with uncertainty and clarity.

Thank you!

Slides: **Diogo Ribeiro**