Statistical Computing

Michael Mayer

Spring 2025

Statistical Computing: What will we do?

Chapters

1. R in Action
2. Statistical Inference
3. Linear Models
4. Model Selection and Validation
5. Trees
6. Neural Nets

Remarks

- Chapters 3 to 6:
  Statistical ML in Action
- Two weeks per chapter
- Exercises at end of chapter notes

Linear Models

Outline

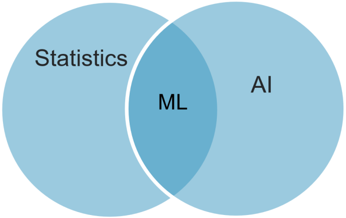
- Start of “Statistical ML in Action”
- Linear Regression
- Generalized Linear Models (GLM)
- Modeling Large Data

Statistical ML in Action

What is ML?

Collection of statistical algorithms used to

1. predict things (supervised ML) or to
2. investigate data structure (unsupervised)



Focus on supervised ML

- Regression
- Classification

Chapters

3. Linear Models
4. Model Selection and Validation
5. Trees
6. Neural Nets

Model Setup

$$T(Y \mid \boldsymbol{X} = \boldsymbol{x}) \approx f(\boldsymbol{x})$$

This means: Approximate property T of response Y (often $T = \mathbb{E}$) by function f of p-dim covariate vector $\boldsymbol{X} = (X^{(1)}, \dots, X^{(p)})$ with value $\boldsymbol{x} = (x^{(1)}, \dots, x^{(p)})$

- Estimate f by $\hat{f}$ from data by minimizing objective

$$Q(f) = \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda\Omega(f)$$

- L: loss function in line with T, e.g. squared error $L(y, z) = (y - z)^2$ for $T = \mathbb{E}$
- $\lambda\Omega(f)$: optional penalty
- $\boldsymbol{y} = (y_1, \dots, y_n)^T$: observed values of Y
- $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$: n feature vectors; $x_i^{(j)}$: i-th value of $X^{(j)}$; $\boldsymbol{x}^{(j)}$: n values of feature $X^{(j)}$

Linear Regression

- Postulate model equation

$$\mathbb{E}(Y \mid \boldsymbol{x}) = f(\boldsymbol{x}) = \beta_o + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$$

- Interpretation of parameters β_j? Ceteris Paribus!
- Optimal $\hat{\beta}_j$? Minimize as objective the sum of squared errors/residuals

$$\sum_{i=1}^{n} (\underbrace{y_i - \hat{y}_i}_{\text{Residual}})^2$$

- Predicted/fitted values $\hat{y}_i = \hat{f}(\boldsymbol{x}_i)$
- This means: we work with the squared error loss and no penalty

Example

Simple linear regression: $\mathbb{E}(Y \mid x) = \alpha + \beta x$

Aspects of Model Quality

Predictive performance

- $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Root-MSE (RMSE)
- Relative performance:
  $R^2 = 1 - \text{MSE}/\text{MSE}_0$
- $\text{MSE}_0 \rightarrow$ intercept-only model

Validity of assumptions

- Model equation is correct
- Normal linear model

$$Y = f(\boldsymbol{x}) + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2)$$

Example

Typical Problems

Missing values

Outliers

Overfitting

Collinearity

Categorical Covariates

- One-Hot-Encoding
- Dummy coding
- Interpretation?

Example

Example of One-Hot-Encoding

color	D	E	F	G	H	I	J
E	0	1	0	0	0	0	0
E	0	1	0	0	0	0	0
E	0	1	0	0	0	0	0
I	0	0	0	0	0	1	0
J	0	0	0	0	0	0	1
J	0	0	0	0	0	0	1
I	0	0	0	0	0	1	0
H	0	0	0	0	1	0	0
E	0	1	0	0	0	0	0
H	0	0	0	0	1	0	0

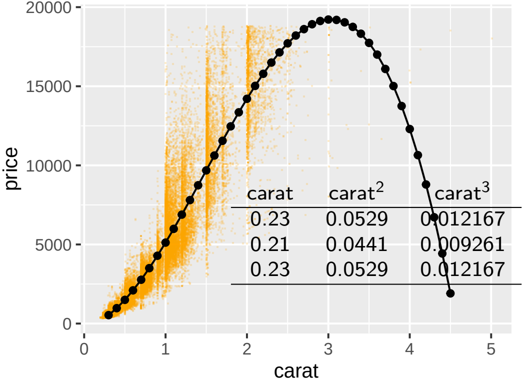
Linear Regression is Flexible

1. Non-linear terms
2. Interactions
3. Transformations like logarithms

These elements are essential but tricky!

Non-Linear Terms

Cubic terms for carat



carat	carat^2	carat^3
0.23	0.0529	0.012167
0.21	0.0441	0.009261
0.23	0.0529	0.012167

Use systematic predictions

Deal with non-linear associations to Y?

$\rightarrow$ invest more parameters

1. Polynomial terms
   - E.g., cubic regression

     $$\mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

   - Don't extrapolate!
2. Regression splines

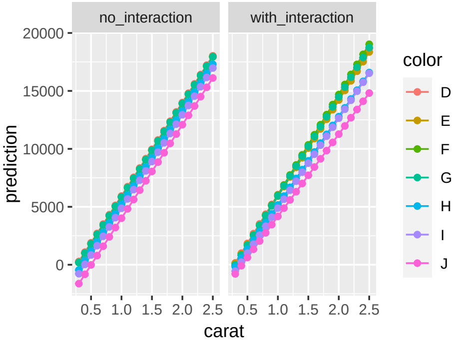
Interactions

- Additivity of effects not always realistic

$$\mathbb{E}(Y \mid \boldsymbol{x}) = \beta_o + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$$

- Adding interaction terms brings necessary flexibility $\rightarrow$ more parameters
- Interaction between features X and Z
  - Multiplication (for categoricals?)
  - For categorical Z, effects of X are calculated by level of Z
  - Like separate models per level of Z

Carat and color



Transformations of Covariates

Examples

- Dummy variables for categoricals
- Decorrelation
- Logarithms against outliers

Effects are interpreted for transformed covariates

Logarithmic Covariates

- $\mathbb{E}(Y \mid x) = \alpha + \beta \log(x)$
- Properties of logarithm allow interpretation for original covariate
- A 1% increase in X is associated with an increase in $\mathbb{E}(Y)$ of about $\beta/100$
- Why?

$$\begin{aligned}
\mathbb{E}(Y \mid 1.01x) - \mathbb{E}(Y \mid x) &= \alpha + \beta \log(1.01x) - \alpha - \beta \log(x) \\
&= \beta \log\left(\frac{1.01x}{x}\right) \\
&= \beta \log(1.01) \approx \beta/100
\end{aligned}$$

Example

Logarithmic Responses

We see: log-transforming X allows to talk about relative effects in X

Idea: log-transformed Y allows to talk about relative effects on Y

Assume for a moment that

$$\mathbb{E}(\log(Y) \mid x) = \alpha + \beta x \implies \log(\mathbb{E}(Y \mid x)) = \alpha + \beta x$$

- Multiplicative model $\mathbb{E}(Y \mid x) = e^{\alpha + \beta x}$
- Relative interpretation: "A one-point increase in X is associated with a relative increase in $\mathbb{E}(Y)$ of $100\%(e^{\beta} - 1) \approx 100\%\beta$"
- If also $\log(X)$?

But assumption is wrong $\rightarrow$ biased predictions for Y $\rightarrow$ GLMs

Examples

Example: Realistic Model for Diamond Prices

- Response: log(price)
- Covariates: log(carat), color, cut and clarity

Generalized Linear Model (GLM)

(One) extension of linear regression

Model equation

Two equivalent formulations

$$g(\mathbb{E}(Y \mid \boldsymbol{x})) = \eta(\boldsymbol{x}) = \beta_o + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$$

$$\mathbb{E}(Y \mid \boldsymbol{x}) = g^{-1}(\eta(\boldsymbol{x})) = g^{-1}(\beta_o + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)})$$

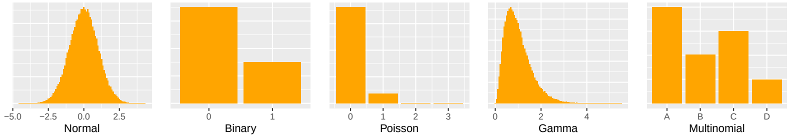
Components

- Linear function/predictor η
- Link function g to map $\mathbb{E}(Y \mid \boldsymbol{x})$ to linear scale
- Distribution of Y conditional on covariates $\rightarrow$ loss function (unit deviance)

Typical GLMs

Regression	Distribution	Range of Y	Natural link	Unit deviance
Linear	Normal	$(-\infty, \infty)$	Identity	$(y-\hat{y})^2$
Logistic	Binary	$\{0, 1\}$	logit	$-2(y\log(\hat{y}) + (1-y)\log(1-\hat{y}))$
Poisson	Poisson	$[0, \infty)$	log	$2(y\log(y/\hat{y}) - (y-\hat{y}))$
Gamma	Gamma	$(0, \infty)$	$1/x$ (typical: log)	$2((y-\hat{y})/\hat{y} - \log(y/\hat{y}))$
Multinomial	Multinomial	$\{C_1, \ldots, C_m\}$	mlogit	$-2\sum_{j=1}^{m} 1(y = C_j)\log(\hat{y}_j)$

Empirical distributions of 100'000 values



- Predictions?
- Log-Link?
- For binary Y: $\mathbb{E}(Y) = P(Y = 1) = p$
- MSE $\to$ Deviance
- Losses in ML?

Why GLM, not Linear Regression?

Linearity assumption not always realistic

1. Binary Y:
   Jump from 0.5 to 0.6 success probability less impressive than from 0.89 to 0.99
2. Count Y: Jump from $\mathbb{E}(Y)$ of 2 to 3 less impressive than from 0.1 to 1.1.
3. Right-skewed Y:
   Jump from 1 Mio to 1.1 Mio deemed larger than from 2 Mio to 2.1 Mio.

Logarithmic Y not possible in the first two cases

GLM solves problem by suitable link g

Further advantages?

Interpretation of Effects guided by Link

Identity link

Like linear regression

Log link

Like linear regression with log response

- Multiplicative model for response
- Now in mathematically sound way

Logit link

- Additive model for $\text{logit}(p)$
- $\text{logit}(p) = \log(\text{odds}(p)) = \log\left(\frac{p}{1-p}\right)$
- Remember: $p = P(Y = 1) = \mathbb{E}(Y)$
- Multiplicative model for $\text{odds}(p)$
- Coefficients e^{β} interpreted as odds ratios

Examples with Insurance Claim Data

1. Poisson regression for claim counts
2. Binary logistic regression for claim (yes/no)

Modeling Large Data

As per 2025

- On normal laptops, we can model datasets up to 8 GB in size (1 Mio iris data)
- Cloud computing allows 1000 times more
- We focus on in-memory situations
  $\rightarrow$ data fits in RAM

Aspect and example technology

1. Data storage $\rightarrow$ Apache Parquet
2. Data loading $\rightarrow$ Apache Arrow
3. Preprocessing $\rightarrow$ data.table
4. Modeling $\rightarrow$ H2O

Example