STAT 151A Review Session

Henry Liev

10 December 2025

Remark 0.1 (Exam Review)

(1) Logistic Regression

- Model Interpretation
- Inference
- Computation

(2) Models for Categorical/Ordinal Data

- Multinomial
- Nested Logit
- Ordinal Logistic Regression

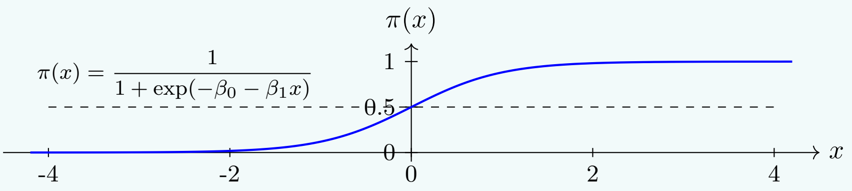
(3) Nonlinear Approaches
Basis Expansion

- Polynomial Regression
- Piecewise Constant
- Splines
- Generalized Additive Models (GAMs)

Regression Trees

Remark 0.2 (Logistic Regression)

$y_i \sim \text{Bern}(\pi_i)$
$\pi_i = \frac{1}{1+\exp(-x_i^\intercal \beta)} = \frac{\exp(x_i^\intercal \beta)}{1+\exp(x_i^\intercal \beta)}$
$x_i^\intercal \beta = \log \frac{\pi_i}{1-\pi_i}$



Estimate β using maximum likelihood
$\mathcal{L}(\beta) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
vs Linear Probability Model (i.e., fit OLS to binary Y)

- Don't get $\hat{\pi}_i's$ outside $(0,1)$
- Avoid implausible normality assumptions

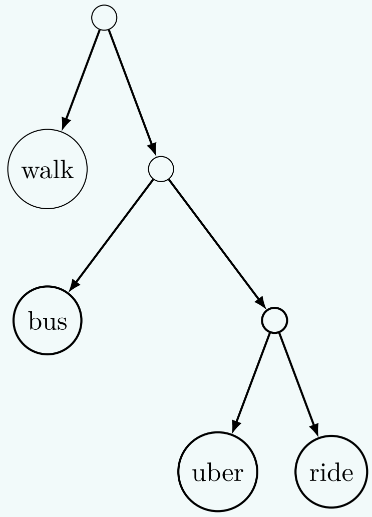
$\exp(\beta_j)$: Factor by which odds of $y_i = 1$ is different when x_j is one unit higher
Inference: Asymptotic
$\hat{\beta} \overset{n\to\infty}{\to} \beta$
$\hat{\beta} \overset{n\to\infty}{\sim} \mathcal{N}(\beta, (\mathbf{X}^\intercal \mathbf{V} \mathbf{X})^{-1}), \mathbf{V} = \text{diag}(\pi_i(1-\pi_i))$
Wald test for $H_0 : \beta_j = 0$
$\frac{\hat{\beta}_j}{\sqrt{(\mathbf{X}^\intercal \hat{\mathbf{V}} \mathbf{X})^{-1}_{jj}}} \overset{n\to\infty}{\sim} N(0, 1^2)$
Wald test for $L\beta = c$
$(L\hat{\beta} - c)(\mathbf{X}^\intercal \hat{\mathbf{V}} \mathbf{X})(L\hat{\beta} - c) \overset{n\to\infty}{\sim} \chi^2_q$
Likelihood ratio test: H_0: Small model is sufficient to describe data (i.e. large model not needed)
$-2\log\left(\frac{\mathcal{L}_{small}}{\mathcal{L}_{big}}\right) \overset{n\to\infty}{\sim} \chi^2_q$
How to actually solve for $\hat{\beta}$?
$\hat{\beta}$ is defined as maximizer of $\mathcal{L}(\beta)$ and $\log \mathcal{L}(\beta) = l(\beta)$
Trick: this means $\frac{\partial l(\beta)}{\partial \beta} = 0$ same as $\nabla_\beta l(\beta)$
Newton-Raphson Method: Guess a β, the do a Taylor expansion on $\frac{\partial l(\beta)}{\partial \beta}$ and use it to get a better guess and repeat
Key formula: $\hat{\beta}_{k+1} = \hat{\beta}_k + \left[\mathcal{I}(\hat{\beta}_k)\right]^{-1} \nabla_\beta l(\hat{\beta}_k)$
At convergence $\mathbf{X}^\intercal y - \mathbf{X}^\intercal \hat{\pi} = 0$
Convergence won't happen if data is separable (i.e. a linear function of x's can separate the 1's and the 0's, model will try to fit to infinite-valued β's)
$D_m = -2\log \mathcal{L}(\hat{\beta}_m)$
Residual deviance ↔ ErrSS, comes up in AIC and BIC
$1 - \frac{D_m}{D_0}$ is analogous to R^2, where D_0 is the intercept only model
Residuals: 2 Basic Flavors
Standardized Pearson Residuals: $\frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}\sqrt{1-h_{ii}}}, 1 - h_{ii}, h_{ii} = \mathbf{X}\hat{\mathbf{V}}^{0.5}(\mathbf{X}^\intercal \hat{\mathbf{V}} \mathbf{X})^{-1}\hat{\mathbf{V}}^{0.5}\mathbf{X}^\intercal$
Standardized Deviance Residuals: $\frac{\pm\sqrt{-2[y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]}}{\sqrt{1-h_{ii}}}, +y_i > \hat{\pi}_i, -y_i < \hat{\pi}_i$
There are also unstandardized versions of these residuals

Remark 0.3 (Multinomial Logistic Regression with m Categories)

$y_i \sim \mathrm{Mult}(1, [\pi_{i1}, \dots, \pi_{im}])$
$\mathbb{P}(y_i = j) = \pi_{ij} = \frac{\exp(\gamma_{0j}+\gamma_{1j}x_{1i}+\cdots+\gamma_{pj}x_{pi})}{1+\sum_{l=1}^{m-1}\exp(\gamma_{0l}+\gamma_{1l}x_{1i}+\cdots+\gamma_{pl}x_{pi})}, j \leq m-1$
$\pi_{im} = 1 - \sum_{l=1}^{m-1} \pi_{il}$
Number of parameters: $(p+1)(m-1)$
Interpretation
$\log\left(\frac{\pi_{ij}}{\pi_{im}}\right) = \gamma_{0j} + \gamma_{1j}x_{1i} + \cdots + \gamma_{pj}x_{pi}$
$\exp(\gamma_{lj})$ is factor by which conditional odds of category j (vs. category m) differs when x_l is 1 amount higher
$\log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = (\gamma_{0j} - \gamma_{0k}) + (\gamma_{1j} - \gamma_{1k})x_{1i} + \cdots + (\gamma_{pj} - \gamma_{pk})x_{pi}$
Inference: Use maximum likelihood, mostly use likelihood ratio tests

Remark 0.4 (Nested Dichotomies with m Categories)

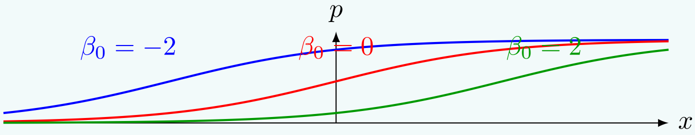
Organize m categories into a binary tree



Fit a logistic regression for each binary split on the data remaining at that split
Number of parameters: $(m-1)(p+1)$
To get predicted probabilities $\hat{\pi}$: Multiply component $\hat{\pi}'s$ across the tree
To get Inference, adding (independent) LRT statistics across trees
Nested Logit: Better if your tree is a really good description of decision process
Multinomial Logit: Better if you're not sure about decision tree (?)

Remark 0.5 (Ordinal Logistic Regression)

Suppose categories are $1 \leq 2 \leq \cdots \leq m$
$\log\left(\frac{\mathbb{P}(y_i \leq j)}{\mathbb{P}(y > j)}\right) = \alpha_j + \underbrace{\beta_1 x_{1i} + \cdots + \beta_p x_{pi}}_{\text{Same } \forall j}$
Benefit: $(m-1)+p$ parameters
Cost: Stronger assumption, proportional odds



$\exp(\beta_j)$: Factor by which odds that y_i is no greater than (same level) differs when x_j is 1 unit higher
$\exp(\alpha_j - \alpha_k) = \frac{odd(y_i=j|x_i)}{odds(y_i=k|x_i)}$: Factor by which odds of j differs from odds of k with same covariate data

Remark 0.6 (Nonlinear Regression)

[label = -]
What if $\mathbb{E}(y_i|x_i)$ is not linear in x_i even after a nonlinear transformation
Basis expansion: Replace individual column of $\mathbf{X}$ with many nonlinear transformations of that column

- Polynomial Regression:
$$(\vec{1}, \vec{x}) \to (\vec{1}, \vec{x}, \vec{x}^2, \vec{x}^3, \dots)$$
- Piecewise Constant:
$$(\vec{1}, \vec{x}) \to I\{x_i \leq c_1\}, I\{x_i \in (c_1, c_2]\}, \dots, I\{x_i > c_k\}$$
- Regression Spline (Focus on cubic spline):
$$(\vec{1}, \vec{x}, \vec{x}^2, \vec{x}^3, (\vec{x}_1, c_1)_+^3, \dots, (\vec{x}_1, c_k)_+^3)$$
Cubic polynomial at every x, but not always the same one
Continuous and 1st and 2nd order differentiable Natural Splines → Same but may force function to be linear outside smallest and largest knot (makes it extrapolate less poorly)
Cubic regression splines: $k+4$
Natural Splines: $k+2$
- Generalized Additive Models (GAMs): Many x_j's, add a separate (spline) basis for each continuous variable
Categorical variables as usual
Plot the "partial fit" (basis terms from a given x_j) against x_j itself to see how y is being modeled as a function of x_j

For all basis expansion methods: Changing X only after this, it's business as usual:
Asses fit, check NLM assumptions, inference, plug into logistic regression
Biggest difference: don't care about individual β_j's anymore

Remark 0.7 (Regression Trees)

List of binary decision rules, not linear combinations
Each leaf is a subset of n data points
Average (or majority votes) of each leaf is prediction
How to fit?
Repeatedly, choose a split (out of all possible splits) that minimizes error criterion
Regresson: Minimizes ErrSS
Classification: Minimize Gini Index or Entropy
Grow the tree really big (small leaves)
Prune the tree by using a penalized error criterion: ErrSS + $\alpha\left|T\right|, \left|T\right|$ is the number of leaves
Solve for a range of α and find the best subtree of each size
Compare those using cross validation
Trees vs Linear Models
Trees: Interpretable, good at interactions, (sometimes) easier to deplay
Linear Models: Inference, more stable, good at trends, larger set of possible predictions