BTRY 6020: Module 6
GLMs and High-dimensional Regression

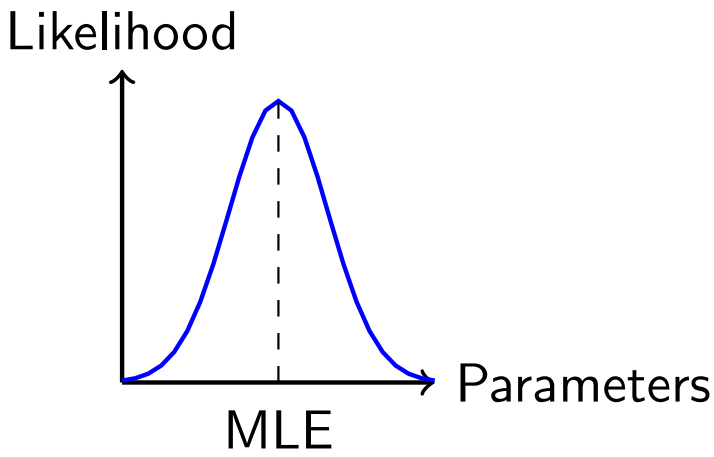
Spring 2025

Logistics

- Assessment 5 due 16th April 23:59
- Today we go over high-dimensional regression

Generalized Linear Models: Review

Key Concepts: GLM Review



- GLMs extend linear regression through:
  - Maximum likelihood estimation
  - Link functions connecting predictors to response
  - Different probability distributions for the response
- MLEs find parameter values that maximize the probability of observing our data

From Linear Regression to GLMs

Linear Regression	**Generalized Linear Models**
$E(Y_i \mid \mathbf{X_i}) = b_0 + \sum_k b_k x_{i,k}$	$g\left(E(Y_i \mid \mathbf{X_i})\right) = b_0 + \sum_k b_k x_{i,k}$
Constant variance (homoscedastic)	Variance structure depends on the mean
Gaussian errors	Various distributions (Binomial, Poisson, etc.)
Independence of observations	Independence of observations

Key Insight

GLMs extend linear models by allowing for non-normal distributions and non-constant variance through a link function $g(\cdot)$.

Model Assumptions

Three Critical GLM Assumptions

1. **Correct Mean Structure:** $g\left(E(Y_i \mid \mathbf{X_i})\right) = b_0 + \sum_k b_k x_{i,k}$
   - Link function correctly connects predictors to response

2. **Correct Variance Structure:** $\text{var}(Y_i \mid \mathbf{X_i})$ follows the specified model
   - Variance depends on mean in a specific way based on distribution

3. **Independence:** Observations are independent of each other
   - No clustering or temporal correlation

Checking Mean Structure Assumption

Key Approaches:

- Compare fitted values with actual values
- Examine Pearson residuals:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{var}}(Y_i \mid \mathbf{X_i})}}$$

- For binary data:
  - Group observations with similar predicted probabilities
  - Compare average observed outcomes in each group to predictions
  - Example: Check if observations with predicted success 65-70% actually succeed at that rate

Visual Diagnostics

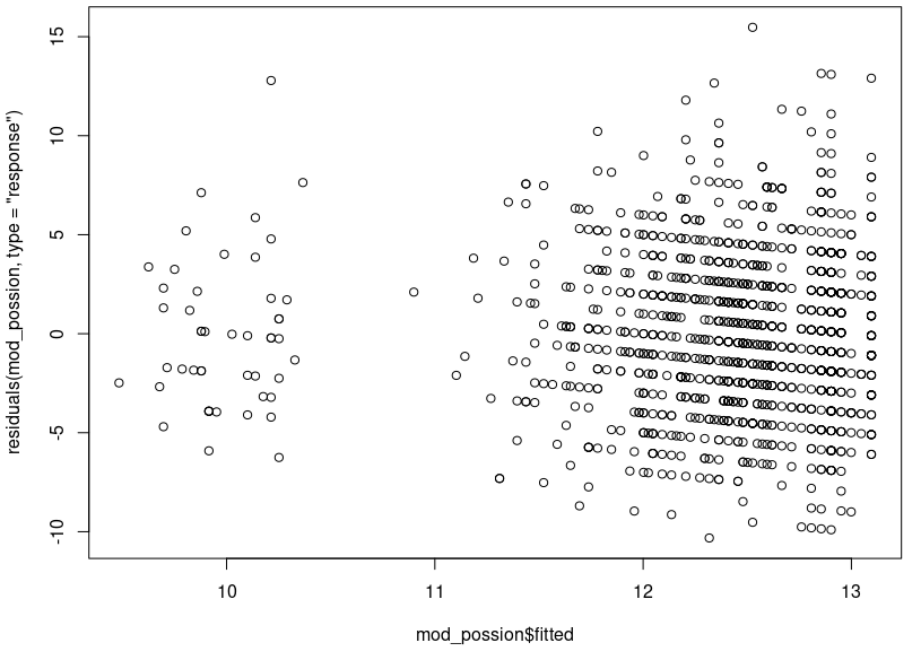
Plots of residuals vs. fitted values should show no systematic patterns if the mean structure is correct.

Diagnostics for Model Assumptions in GLMs

Key Diagnostic Approaches

- Residual plots help us evaluate if our model assumptions are met
- We'll examine two types of models with NFL data:
  - Poisson regression for count data (penalty counts)
  - Logistic regression for binary data (field goal success)

Poisson Regression Diagnostics: Response Residuals



What we observe:

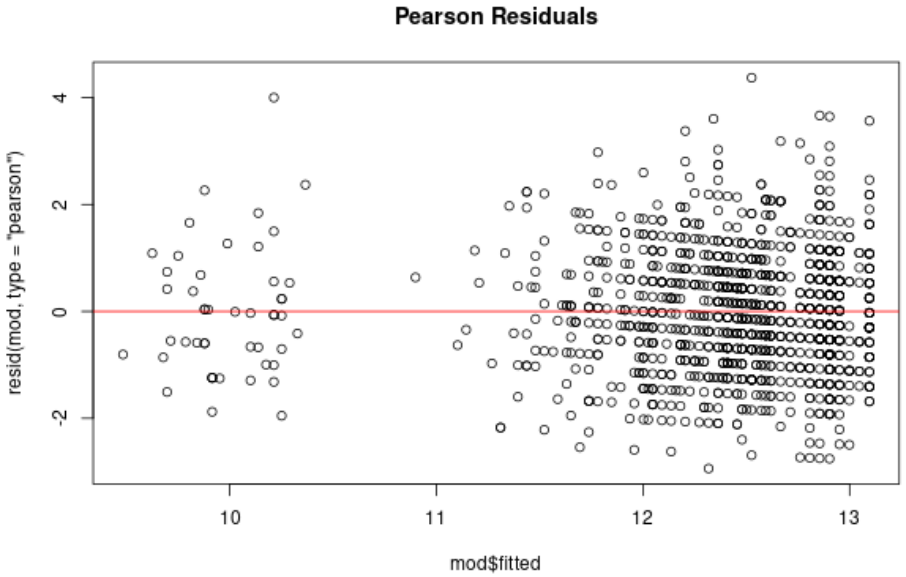
- Funnel shape pattern
- Wider spread as fitted values increase
- Asymmetric spread (more positive residuals)

What this suggests:

- Possible overdispersion issue
- Expected for count data

```
plot(mod$fitted, resid(mod, type = "response"))
```

Poisson Regression Diagnostics: Pearson Residuals



What we observe:

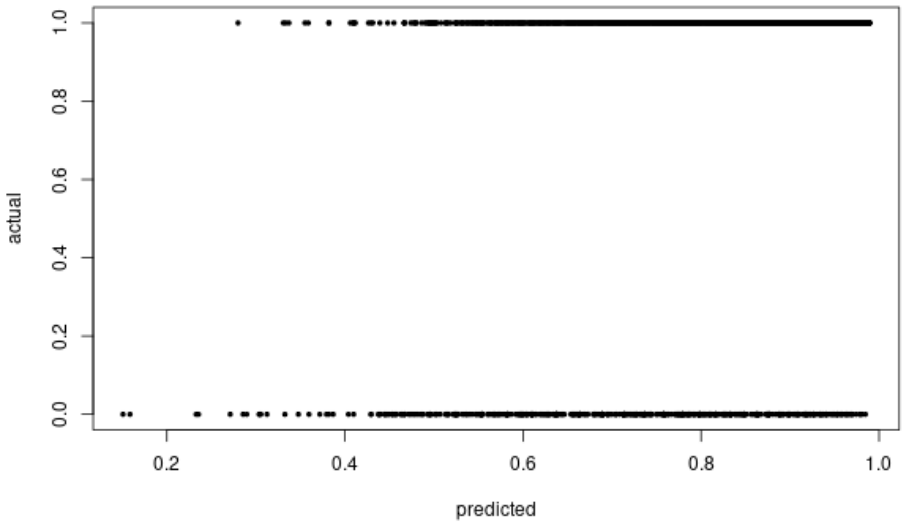
- Residuals standardized by estimated standard deviation
- Still shows spreading pattern
- Mean of squared Pearson residuals 1.3

What this indicates:

- Mild overdispersion
- Mean structure may be reasonable

```
plot(mod$fitted, resid(mod, type = "pearson"))
```

Logistic Regression Diagnostics: Raw Data



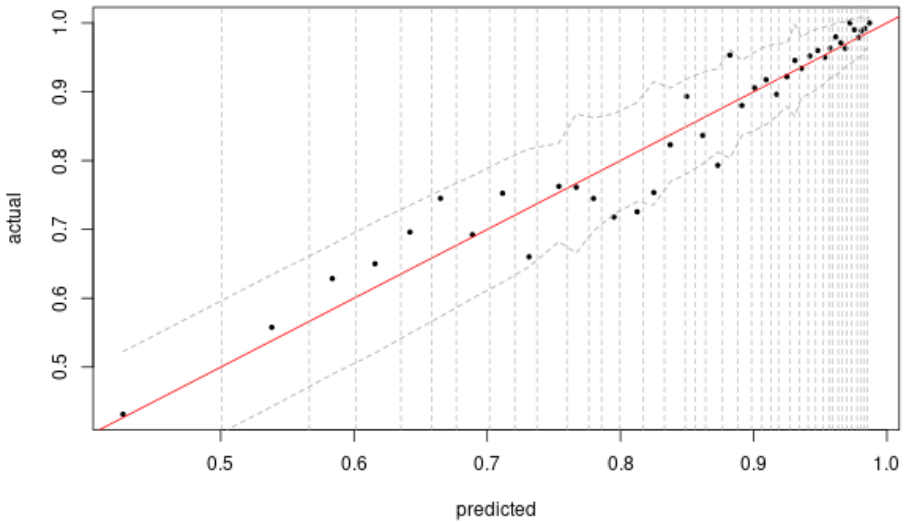
What we observe:

- Binary outcome (0 or 1)
- Predicted probabilities mostly between 0.4-1.0
- More 1s at higher predicted probabilities

Limitations:

- Hard to assess fit directly from this plot
- Binary data will always appear in this pattern

Logistic Regression Diagnostics: Calibration Plot



How to read this plot:

- Points represent grouped observations with similar predicted values
- Red line shows perfect calibration
- Dotted lines show confidence bands

What this shows:

- Good calibration - points near the line
- Mean structure assumption appears satisfied
- Model predicts probabilities accurately

Addressing Mean Structure Issues

When to take action

Reconsider your mean structure when:

- Systematic patterns exist in residuals
- Calibration plots show poor fit
- Model consistently over/under-predicts

Potential Solutions

- Try a different distribution family
  - Negative binomial for overdispersed counts
  - Beta-binomial for overdispersed proportions
- Change the link function
  - Probit instead of logit
  - Log vs. identity
- Transform or add predictors

Variance Assumption in GLMs

Key Principle

- In GLMs, variance depends on the mean
- Each distribution family implies a specific variance structure:

$$\text{var}(Y_i \mid \mathbf{X_i}) = f(\mu_i)$$

- If this relationship is misspecified, inference suffers

Evidence of Variance Issues

- Funnel-shaped residual plots
- Mean of squared Pearson residuals far from 1.0
- Confidence intervals too narrow/wide
- Our example: Pearson residuals 1.3 suggests mild overdispersion

Two Approaches to Address Variance Issues

- **Change the distribution family:** Alters both mean and variance structure
- **Account for over/underdispersion:** Adjust standard errors while keeping coefficient estimates

$$\text{True Variance} = \text{Model Based Variance} \times \text{Dispersion factor } \phi$$

Example: Addressing Overdispersion in Count Models

Poisson Regression

$$\text{var}(Y_i \mid \mathbf{X_i}) = E(Y_i \mid \mathbf{X_i}) = \theta(\mathbf{X_i})$$

- Variance = Mean (equality)
- Restrictive assumption
- Often violated in real data

Solutions for Overdispersion

Negative Binomial:

$$\text{var}(Y_i \mid \mathbf{X_i}) = \theta(\mathbf{X_i}) + \frac{1}{r}\theta(\mathbf{X_i})^2$$

Quasi-Poisson:

$$\text{var}(Y_i \mid \mathbf{X_i}) = \phi \cdot \theta(\mathbf{X_i})$$

When to use each approach

- **Negative Binomial:** When variance increases quadratically with the mean
- **Quasi-Poisson:** When variance is proportional to the mean
- For our NFL penalties example (dispersion factor 1.3), either approach would work

Independence Assumption

The requirement

Each observation must be independent of other observations:

- No clustering effects
- No temporal correlation
- No spatial correlation

Detecting violations

- Plot residuals against time/space
- Check residuals by group
- Examine autocorrelation

Addressing dependence

- **Random effects models:**
  - Account for clustering
  - Allow for correlation within groups
- **GEE (Generalized Estimating Equations):**
  - Focus on population-average effects
  - Flexible correlation structures
- **Note:** These methods are more complex and interpretation changes

Summary: Checking and Addressing GLM Assumptions

Mean Structure

- **Check:**
  - Residual plots
  - Calibration plots
- **Fix:**
  - Different distribution
  - Different link function
  - Transform/add predictors

Variance Structure

- **Check:**
  - Squared Pearson residuals
  - Funnel shapes in plots
- **Fix:**
  - Different distribution
  - Quasi-likelihood
  - Robust standard errors

Independence

- **Check:**
  - Temporal patterns
  - Group-based patterns
- **Fix:**
  - Random effects
  - GEE approaches
  - Time series methods

Our NFL Examples

- **Poisson model:** Shows mild overdispersion (1.3), might benefit from quasi-Poisson
- **Logistic model:** Shows good calibration, mean structure appears appropriate
- **Both models:** Independence assumption requires additional diagnostics

High-dimensional Regression

Motivation

"Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat" (Tsai et. al., Scientific Report 2020)

- Authors are interested in identifying genetic markers associated with yield, protein content, disease resistance in spring barley and winter wheat

 [1]

[1] Photo from: https://fieldcropnews.com/2020/09/winter-barley-is-making-a-comeback-in-ontario/

Problem

- There were $p = 4,056$ genetic markers in spring barley and $p = 11,154$ genetic markers in winter wheat
- Genetic markers take values $\{0, 1, 2\}$
- Tested roughly $n = 1300$ breeding lines for barley and wheat
- Number of covariates (genetic markers) is greater than sample size (breeding lines)
- Model selection problem where p is very very large

Problem

- When the number of covariates, p, is large relative to the number of samples, n, the estimates of each coefficient can be imprecise
- When $p > n$, linear regression isn't just imprecise, it's ill-defined
- The RSS can be made 0
- Different sets of coefficients can be used to make the RSS 0

Problem

When $p > n$, there is not a unique set of coefficients which can make RSS 0

- Consider the case where we have a dummy variable for each observation and one continuous covariate
- For any slope for the continuous covariate $\hat{b}_1$
- Set the coefficient of the dummy variable equal to the observation for the sample so that $\hat{b}_{\text{obs i}} = y_i - b_1 x_{i,1}$
- Then, $\hat{y}_i = y_i$

$$\hat{y}_i = b_1 x_{i,1} + \hat{b}_{\text{obs i}} = b_1 x_{i,1} + y_i - b_1 x_{i,1} = y_i$$

Solution

A solution to pick a specific value for $\hat{\mathbf{b}}$ is to minimize not just the RSS, but

$$\min_{\hat{\mathbf{b}}} \sum_i (y_i - \hat{y}_i)^2 + \text{Penalty}(\hat{\mathbf{b}})$$

- The Penalty($\hat{\mathbf{b}}$) term can be various quantities
- If Penalty($\hat{\mathbf{b}}$) term is the number of non-zero entries in $\hat{\mathbf{b}}$, then we get something that almost is like AIC/BIC
- Other choices for Penalty($\hat{\mathbf{b}}$)

$$L_1 \text{ Penalty}: \ \lambda \sum_k |\hat{b}_k| = \lambda \|\hat{\mathbf{b}}\|_1$$

$$L_2 \text{ Penalty}: \ \lambda \sum_k \hat{b}_k^2 = \lambda \|\hat{\mathbf{b}}\|_2^2$$

Solution

The LASSO (Least absolute shrinkage and selection operator) estimator solves the following:

$$\min_{\hat{\mathbf{b}}} \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_k |\hat{b}_k|$$

The Ride Regression estimator solves the following:

$$\min_{\hat{\mathbf{b}}} \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_k b_k^2$$

Penalized Regression

Why is this good?

- Allows us to select a specific value for $\hat{\mathbf{b}}$
- Computationally, much easier than fitting every sub-set of variables and comparing the RSS (or AIC or BIC)
- Fit one model with all covariates included

Lasso

The LASSO (Least absolute shrinkage and selection operator) estimator solves the following:

$$\min_{\hat{\mathbf{b}}} \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_k |\hat{b}_k|$$

- Usually, the solution $\hat{\mathbf{b}}$ is "sparse" where many coefficients are set to 0
- Also "encourages" solutions where $|\hat{b}_k|$ is closer to 0
- Similar to a model selection procedure
- User sets λ to a specific value in advance
- Larger values of λ typically mean more estimated coefficients are 0

Choosing λ

The Ride Regression estimator solves the following:

$$\min_{\hat{\mathbf{b}}} \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_k b_k^2$$

- Usually, the solution $\hat{\mathbf{b}}$ is not "sparse" where all coefficients non-zero
- Can improve predictions (similar to model selection), but still includes all covariates
- Also "encourages" solutions where $|\hat{b}_k|$ is closer to 0
- User sets λ to a specific value in advance
- Larger values of λ typically mean estimated coefficients are closer to 0

Summary

- Penalized regression can be a useful way to navigate the complexity trade-off
- Plays a similar role as model selection
- Lasso regression "picks" covariates and also affects the estimated coefficients
- Ridge regression includes all covariates and affects the estimated coefficients
- As seen in lab, both can outperform linear regression when p is large relative to n