Applied Statistical Methods II

A review of STAT 2132

Non-linear regression: the model

- Assumed $Y_i = f(X_i, \gamma) + \epsilon_i$, where the function f was known and $\epsilon_i \sim N(0, \sigma^2)$.
- Goal: Estimate and perform inference on the unknown γ
- Examples
  - General exponential: $f(X_i, \gamma) = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i)$. In most problems, $\gamma_2 \leq 0$. What happens as $X_i = 0$ and $X_i \to \infty$?
  - Logistic regression non-linear regression model (this is a POOR name): $f(X_i, \gamma) = \frac{\gamma_0}{1+\gamma_1 \exp(\gamma_2 X_i)}$, $\gamma_2 \leq 0$. What is the behavior of f when $X_i = 0$, $X_i \to \infty$?
- Note that $\epsilon_i \sim N(0, \sigma^2)$. This is **critical**.
- These are NOT GLMs, since $g\{E(Y_i)\} \neq x_i^T \beta$ for some link function g.

Non-linear regression: the model

- Assumed $Y_i = f(X_i, \gamma) + \epsilon_i$, where the function f was known and $\epsilon_i \sim N(0, \sigma^2)$.
- Goal: Estimate and perform inference on the unknown γ
- Examples
  - General exponential: $f(X_i, \gamma) = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i)$. In most problems, $\gamma_2 \leq 0$. What happens as $X_i = 0$ and $X_i \to \infty$?
  - Logistic regression non-linear regression model (this is a POOR name): $f(X_i, \gamma) = \frac{\gamma_0}{1+\gamma_1 \exp(\gamma_2 X_i)}$, $\gamma_2 \leq 0$. What is the behavior of f when $X_i = 0$, $X_i \to \infty$?
- Note that $\epsilon_i \sim N(0, \sigma^2)$. This is **critical**.
- These are NOT GLMs, since $g\{E(Y_i)\} \neq x_i^T \beta$ for some link function g.

Non-linear regression: the model

- Assumed $Y_i = f(X_i, \gamma) + \epsilon_i$, where the function f was known and $\epsilon_i \sim N(0, \sigma^2)$.
- Goal: Estimate and perform inference on the unknown γ
- Examples
  - General exponential: $f(X_i, \gamma) = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i)$. In most problems, $\gamma_2 \leq 0$. What happens as $X_i = 0$ and $X_i \to \infty$?
  - Logistic regression non-linear regression model (this is a POOR name): $f(X_i, \gamma) = \frac{\gamma_0}{1+\gamma_1 \exp(\gamma_2 X_i)}$, $\gamma_2 \leq 0$. What is the behavior of f when $X_i = 0$, $X_i \to \infty$?
- Note that $\epsilon_i \sim N(0, \sigma^2)$. This is **critical**.
- These are NOT GLMs, since $g\{E(Y_i)\} \neq x_i^T \beta$ for some link function g.

Non-linear regression: fitting

$Y_i = f(X_i, \gamma) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$

- We fit using least squares: $Q(\gamma) = \sum_{i=1}^{n} \{Y_i - f(X_i, \gamma)\}^2$
- This is equivalent to fitting with ML.
- We fit these using Gauss-Newton (GN)
  - Idea: Approximate Q using a first order Taylor expansion
  - Each update was equivalent to solving an ordinary least squares problem
- Problems: first order approximation depends heavily on the curvature of the objective function
- GN is **very** sensitive to starting points (you saw this on HW)

Non-linear regression: fitting

$Y_i = f(X_i, \gamma) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$

- We fit using least squares: $Q(\gamma) = \sum_{i=1}^{n} \{Y_i - f(X_i, \gamma)\}^2$
- This is equivalent to fitting with ML.
- We fit these using Gauss-Newton (GN)
  - Idea: Approximate Q using a first order Taylor expansion
  - Each update was equivalent to solving an ordinary least squares problem
- Problems: first order approximation depends heavily on the curvature of the objective function
- GN is **very** sensitive to starting points (you saw this on HW)

Non-linear regression: fitting

$Y_i = f(X_i, \gamma) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$

- We fit using least squares: $Q(\gamma) = \sum_{i=1}^{n} \{Y_i - f(X_i, \gamma)\}^2$
- This is equivalent to fitting with ML.
- We fit these using Gauss-Newton (GN)
  - Idea: Approximate Q using a first order Taylor expansion
  - Each update was equivalent to solving an ordinary least squares problem
- Problems: first order approximation depends heavily on the curvature of the objective function
- GN is **very** sensitive to starting points (you saw this on HW)

Non-linear regression: inference

$Y_i = f(X_i, \gamma) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $\gamma \in \mathbb{R}^p$

- Used the first order Taylor approximation to derive asymptotic distribution
  - Under regularity conditions, $\hat{\gamma} - \gamma \approx N\left(0, \sigma^2 \left(J^T J\right)^{-1}\right)$
  - $J \in \mathbb{R}^{n \times p}$ is the matrix of first derivatives.
  - What is J is linear regression? Does asymptotic dist'n match usual dist'n of OLS estimator when f is linear?
- Could also perform bootstrap
  - Set $r_i = Y_i - f(X_i, \hat{\gamma})$
  - Create synthetic datasets as $Y_i^{(b)} = f(X_i, \hat{\gamma}) + r_i^{(b)}$.
  - Re-estimate γ and compute statistics.
  - Approximate dist'n of statistics using bootstrap dist'n $\Rightarrow$ confidence intervals.
  - Why is the assumption that ϵ_i's are i.i.d so important here?

Non-linear regression: inference

$Y_i = f(X_i, \gamma) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $\gamma \in \mathbb{R}^p$

- Used the first order Taylor approximation to derive asymptotic distribution
  - Under regularity conditions, $\hat{\gamma} - \gamma \approx N\left(0, \sigma^2 \left(J^T J\right)^{-1}\right)$
  - $J \in \mathbb{R}^{n \times p}$ is the matrix of first derivatives.
  - What is J is linear regression? Does asymptotic dist'n match usual dist'n of OLS estimator when f is linear?
- Could also perform bootstrap
  - Set $r_i = Y_i - f(X_i, \hat{\gamma})$
  - Create synthetic datasets as $Y_i^{(b)} = f(X_i, \hat{\gamma}) + r_i^{(b)}$.
  - Re-estimate γ and compute statistics.
  - Approximate dist'n of statistics using bootstrap dist'n $\Rightarrow$ confidence intervals.
  - Why is the assumption that ϵ_i's are i.i.d so important here?

Generalized linear models (GLMs)

- Started out simple: logistic regression
  - $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
  - $\text{logit}\{\pi(X_i)\} = \log\left\{\frac{\pi(X_i)}{1-\pi(X_i)}\right\} = X_i^T\beta$
  - $\frac{\pi(X_i)}{1-\pi(X_i)}$ is the odds evaluated at X_i
  - Parameters could be interpreted as odds ratios
- $g = \text{logit}$ was called the **link function**. h transforms the **mean** of Y_i.
- In this case, g ensures that $E(Y_i \mid X_i) \in [0, 1]$.
- In general, required g to be continuously differentiable and strictly increasing. Other examples:
  - $g = \text{probit} = \Phi^{-1}$.
  - $g =$ Complementary log-log
  - In general, g can be the inverse of the CDF of a random variable that also has a density.
- NO TRANSFORMATION OF Y_i OCCURS IN GLM.

Generalized linear models (GLMs)

- Started out simple: logistic regression
  - $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
  - $\text{logit}\{\pi(X_i)\} = \log\left\{\frac{\pi(X_i)}{1-\pi(X_i)}\right\} = X_i^T\beta$
  - $\frac{\pi(X_i)}{1-\pi(X_i)}$ is the odds evaluated at X_i
  - Parameters could be interpreted as odds ratios
- $g = \text{logit}$ was called the **link function**. h transforms the **mean** of Y_i.
- In this case, g ensures that $E(Y_i \mid X_i) \in [0, 1]$.
- In general, required g to be continuously differentiable and strictly increasing. Other examples:
  - $g = \text{probit} = \Phi^{-1}$.
  - $g =$ Complementary log-log
  - In general, g can be the inverse of the CDF of a random variable that also has a density.
- NO TRANSFORMATION OF Y_i OCCURS IN GLM.

Fitting logistic regressions

- $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
- $g\{\pi(X_i)\} = X_i^T\beta$, $g =$ logit

- Generally fit with maximum likelihood.
- In terms of mathematical convenience, we generally prefer $g =$ logit when optimizing. Why?
- Maximizing the likelihood is akin to finding the root of the score function. What is the expectation of score, evaluated at the true parameter?
- Used a second order Taylor approximation to derive asymptotic distribution of $\hat{\beta}$.
- We could also use the likelihood ratio test to perform inference.
- The deviance was analogous to the residual sum of squares

Fitting logistic regressions

- $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
- $g\{\pi(X_i)\} = X_i^T\beta$, $g = \text{logit}$

- Generally fit with maximum likelihood.
- In terms of mathematical convenience, we generally prefer $g = \text{logit}$ when optimizing. Why?
- Maximizing the likelihood is akin to finding the root of the score function. What is the expectation of score, evaluated at the true parameter?
- Used a second order Taylor approximation to derive asymptotic distribution of $\hat{\beta}$.
- We could also use the likelihood ratio test to perform inference.
- The deviance was analogous to the residual sum of squares

Additional inference with logistic regression

- $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
- $g\{\pi(X_i)\} = X_i^T\beta$, $g = \text{logit}$

- Pearson's chi-squared
  - Was applicable when there were degenerate covariate patterns.
  - It tests H_0 : linea model is correct.
  - Has nice asymptotic properties when the number of samples in each covariate patter is large.
- Deviance test.
  - Most applicable when there were degenerate covariate patterns.
  - Allows you to form a deviance table, which is analogous to an ANOVA table
- Primarily use Wald and likelihood to perform inference and compute confidence intervals for β.

Additional inference with logistic regression

- $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
- $g\{\pi(X_i)\} = X_i^T\beta$, $g = \text{logit}$

- Pearson's chi-squared
  - Was applicable when there were degenerate covariate patterns.
  - It tests H_0 : linea model is correct.
  - Has nice asymptotic properties when the number of samples in each covariate patter is large.
- Deviance test.
  - Most applicable when there were degenerate covariate patterns.
  - Allows you to form a deviance table, which is analogous to an ANOVA table
- Primarily use Wald and likelihood to perform inference and compute confidence intervals for β.

Additional inference with logistic regression

- $Y_i \mid X_i \sim Ber\{\pi(X_i)\}$
- $g\{\pi(X_i)\} = X_i^T\beta$, $g =$ logit

- Pearson's chi-squared
  - Was applicable when there were degenerate covariate patterns.
  - It tests H_0 : linea model is correct.
  - Has nice asymptotic properties when the number of samples in each covariate patter is large.
- Deviance test.
  - Most applicable when there were degenerate covariate patterns.
  - Allows you to form a deviance table, which is analogous to an ANOVA table
- Primarily use Wald and likelihood to perform inference and compute confidence intervals for β.

Other important GLMs

- GLM is an umbrella term.
- There are many different type so GLMs (Poisson, Gamma, Negative-binomial)
- We focused on Poisson to model count data (i.e. when $Y_i \in \{0, 1, 2, \ldots\}$)
- We typically used log-link. Why?
- You had a little exposure to the Negative-binomial.

Other important GLMs

- GLM is an umbrella term.
- There are many different type so GLMs (Poisson, Gamma, Negative-binomial)
- We focused on Poisson to model count data (i.e. when $Y_i \in \{0, 1, 2, \ldots\}$)
- We typically used log-link. Why?
- You had a little exposure to the Negative-binomial.

GLMs: a general framework

- In general, a GLM uses exponential families to model data.
- Exponential families are a large family of distributions.
  - Include discrete distributions: Binomial, Poisson, Negative-binomial
  - Also continuous distributions: Normal, Gamma
- $\ell(\theta_i; Y_i) = Y_i\theta_i - b(\theta_i) + c(Y_i)$
  - θ_i is called the **canonical parameter**
  - The derivatives of $b(\theta_i)$ give the **cumulants** (mean, variance, skew, etc.)
  - $E_{\theta_i}(Y_i) = b'(\theta_i)$, $\text{Var}_{\theta_i}(Y_i) = b''(\theta_i)$
  - c does not depend on θ_i.
- In GLM, we assume that $g\{E_{\theta_i}(Y_i)\} = g\{b'(\theta_i)\} = X_i^T\beta$
  - Therefore, GLM can also be thought of as a way of modeling θ_i.
  - What is the canonical link function g, i.e. when is $g\{E_{\theta_i}(Y_i)\} = \theta_i$? What is special about this link?
  - When $\theta_i = X_i^T\beta$, $\ell(\theta_i; Y_i) = \ell(\beta; Y_i)$ is concave in β!

GLMs: a general framework

- In general, a GLM uses exponential families to model data.
- Exponential families are a large family of distributions.
  - Include discrete distributions: Binomial, Poisson, Negative-binomial
  - Also continuous distributions: Normal, Gamma
- $\ell(\theta_i; Y_i) = Y_i\theta_i - b(\theta_i) + c(Y_i)$
  - θ_i is called the **canonical parameter**
  - The derivatives of $b(\theta_i)$ give the **cumulants** (mean, variance, skew, etc.)
  - $E_{\theta_i}(Y_i) = b'(\theta_i)$, $\text{Var}_{\theta_i}(Y_i) = b''(\theta_i)$
  - c does not depend on θ_i.
- In GLM, we assume that $g\{E_{\theta_i}(Y_i)\} = g\{b'(\theta_i)\} = X_i^T\beta$
  - Therefore, GLM can also be thought of as a way of modeling θ_i.
  - What is the canonical link function g, i.e. when is $g\{E_{\theta_i}(Y_i)\} = \theta_i$? What is special about this link?
  - When $\theta_i = X_i^T\beta$, $\ell(\theta_i; Y_i) = \ell(\beta; Y_i)$ is concave in β!

GLMs: a general framework

- In general, a GLM uses exponential families to model data.
- Exponential families are a large family of distributions.
  - Include discrete distributions: Binomial, Poisson, Negative-binomial
  - Also continuous distributions: Normal, Gamma
- $\ell(\theta_i; Y_i) = Y_i\theta_i - b(\theta_i) + c(Y_i)$
  - θ_i is called the **canonical parameter**
  - The derivatives of $b(\theta_i)$ give the **cumulants** (mean, variance, skew, etc.)
  - $E_{\theta_i}(Y_i) = b'(\theta_i)$, $\text{Var}_{\theta_i}(Y_i) = b''(\theta_i)$
  - c does not depend on θ_i.
- In GLM, we assume that $g\{E_{\theta_i}(Y_i)\} = g\{b'(\theta_i)\} = X_i^T\beta$
  - Therefore, GLM can also be thought of as a way of modeling θ_i.
  - What is the canonical link function g, i.e. when is $g\{E_{\theta_i}(Y_i)\} = \theta_i$? What is special about this link?
  - When $\theta_i = X_i^T\beta$, $\ell(\theta_i; Y_i) = \ell(\beta; Y_i)$ is concave in β!

GLMs: a general framework

- In general, a GLM uses exponential families to model data.
- Exponential families are a large family of distributions.
  - Include discrete distributions: Binomial, Poisson, Negative-binomial
  - Also continuous distributions: Normal, Gamma
- $\ell(\theta_i; Y_i) = Y_i\theta_i - b(\theta_i) + c(Y_i)$
  - θ_i is called the **canonical parameter**
  - The derivatives of $b(\theta_i)$ give the **cumulants** (mean, variance, skew, etc.)
  - $E_{\theta_i}(Y_i) = b'(\theta_i)$, $\text{Var}_{\theta_i}(Y_i) = b''(\theta_i)$
  - c does not depend on θ_i.
- In GLM, we assume that $g\{E_{\theta_i}(Y_i)\} = g\{b'(\theta_i)\} = X_i^T\beta$
  - Therefore, GLM can also be thought of as a way of modeling θ_i.
  - What is the canonical link function g, i.e. when is $g\{E_{\theta_i}(Y_i)\} = \theta_i$? What is special about this link?
  - When $\theta_i = X_i^T\beta$, $\ell(\theta_i; Y_i) = \ell(\beta; Y_i)$ is concave in β!

Inference with GLMs

$\ell(\theta_i; Y_i) = Y_i\theta_i - b(\theta_i) + c(Y_i)$, $g\{b'(\theta_i)\} = X_i^T\beta$

- Could derive asymptotic dist'n for β with second order Taylor expansions
  - Gives us Wald statistics.
- Could also use likelihood ratio
- Score test
- However, we had to be careful of **over-dispersion**
  - In general, over-dispersion occurs when $\text{Var}(Y_i) > V\{E_{\theta_i}(Y_i)\}$.
  - In above model, $V\{E_{\theta_i}(Y_i)\} = b''(\theta_i)$
  - i.e. the variance is larger than we think it is.
  - What will happen to inference/confidence intervals? Will confidence intervals be too wide or too narrow?

Inference with GLMs

$$\ell(\theta_i; Y_i) = Y_i\theta_i - b(\theta_i) + c(Y_i),\ g\{b'(\theta_i)\} = X_i^T\beta$$

- Could derive asymptotic dist'n for β with second order Taylor expansions
  - Gives us Wald statistics.
- Could also use likelihood ratio
- Score test
- However, we had to be careful of **over-dispersion**
  - In general, over-dispersion occurs when $\text{Var}(Y_i) > V\{E_{\theta_i}(Y_i)\}$.
  - In above model, $V\{E_{\theta_i}(Y_i)\} = b''(\theta_i)$
  - i.e. the variance is larger than we think it is.
  - What will happen to inference/confidence intervals? Will confidence intervals be too wide or too narrow?

Over dispersion with GLM

- We introduced a "fudge factor" ϕ, called dispersion parameter.
  - Assumed $\text{Var}_{\theta_i}(Y_i) = \phi V\{E_{\theta_i}(Y_i)\}$
  - We estimated ϕ using Pearson's χ^2 statistic: $\hat{\phi} = \frac{1}{n-p}\sum_i \frac{(\text{Observed}_i - \text{expected}_i)^2}{\text{Expected variance}_i}$
  - Recall expression for Poisson...
- This model for the variance is akin to assuming

$$\ell(\theta_i; \phi, Y_i) = \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i; \phi)$$

- This led us to quasi-likelihood, generalized estimating equations, etc.

Over dispersion with GLM

- We introduced a "fudge factor" ϕ, called dispersion parameter.
  - Assumed $\text{Var}_{\theta_i}(Y_i) = \phi V\{E_{\theta_i}(Y_i)\}$
  - We estimated ϕ using Pearson's χ^2 statistic: $\hat{\phi} = \frac{1}{n-p}\sum_i \frac{(\text{Observed}_i - \text{expected}_i)^2}{\text{Expected variance}_i}$
  - Recall expression for Poisson...
- This model for the variance is akin to assuming

$$\ell(\theta_i; \phi, Y_i) = \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i; \phi)$$

- This led us to quasi-likelihood, generalized estimating equations, etc.

Study designs

- We spent some time on study designs
  - Randomized experiment vs. observational data
  - Complete block design vs. incomplete block design
- Most important take-away: remember the adage “crap in, crap out”.
  - Drawing meaningful conclusions is impossible is impossible in poorly designed experiments.
  - The gold standard: randomized experiment, since we can randomize to control for confounding variables
  - Observational data: need to be wary of confounders
  - There is a whole field of statistics and economics devoted to being able to make causal statements from observational data (i.e. instrumental variables, Mendelian randomization, propensity weighting)

Factor level means

- We spent quite some time discussing models where variables were treated as factors.
- Some important take-aways:
  - Additive effects vs. interactions
  - ANOVA in balanced and unbalanced designs.
  - Simultaneous inference (Tukey, Scheffe, Bonferroni)

ANCOVA

- Analysis of Covariance (ANCOVA)
- Addresses the question: how do we perform inference when some aspects of experiment are controlled, and others are purely observational.
- Include observed, non-randomized factors in model can help reduce residual variance.
- Generally assume the treatment and observational covariates do not interact
- Ideas here are similar to what you learned in 2131

Mixed effect models 1

- Started off simple: a balanced, repeated measures design.
- Same number of observations on each individual.
- We could use nice mathematical properties to perform inference on intra- and inter-individual variance parameters
- We could fit these models using simple method of moments.
- Data are usually not this simple

Mixed effect models 2

- Moved to the more general case $Y = X\beta + \epsilon$, $\epsilon \sim (0, \sum_{r=1}^{b} \theta_r B_r)$
- Could fit this with maximum quasi-likelihood using normal likelihood
- It's "quasi" because ϵ does not have to be normally distributed to get consistent estimates for θ_r
- Used full model MLE and REML to estimate $\theta_1, \ldots, \theta_b$
- Generally, REML is preferred b/c it accounts for the degrees of freedom lost when estimating β
- OLS is equivalent to REML when $\sum_{r=1}^{b} \theta_r B_r = \sigma^2 I_n$
- Estimate β with generalized least squares.
- Mixed effects will be important for the exam...

Additional topics

- Generalized method of moments
- Factor analysis in high dimensional data
- Fundamental problem: loadings $\boldsymbol{L} \in \mathbb{R}^{p \times K}$ and factors $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ are not identifiable.
- Formulated PCA as a problem in factor analysis that parametrizes $\boldsymbol{L}$ and $\boldsymbol{C}$ to make them identifiable.
- Discussed how to estimate K.