Sociology 740 John Fox

Lecture Notes

Review for the Second Exam

1. Diagnostics: Unusual and Influential Data

- Outliers, leverage and influence.
- Assessing leverage: Hat-values.
- Detecting outliers: Studentized residuals.
- Measuring influence: Influence on coefficients and Cook's D.
- Joint influence: Added-variable plots.

2. Diagnostics: Collinearity and Model Selection

- Nature of the problem.
- Variance-inflation factors (VIF) and generalized variance-inflation factors (GVIF).
- Putative solutions:
  - model respecification.
  - variable selection.
  - biased estimation.
  - prior information.
- Model selection criteria: Mallows's C_p, cross-validation, generalized cross-validation, AIC, BIC.
- Model validation.

3. Diagnostics: Non-Normality, Non-Constant Error Variance, and Nonlinearity

- Non-normality: Quantile-comparison plot, histogram or density estimate; boxplot; transformations.
- Non-constant error variance: Plotting residuals against fitted values; spread-level plot; transformations, WLS regression, and "corrected" standard errors.
- Nonlinearity: Component+residual plots; transformations, polynomial regression, and regresson splines.
- Discrete data: testing for nonlinearity ("lack of fit") and non-constant error variance (Levene's test).

- Maximum-likelihood methods (treat as optional):
  - Box-Cox transformation of Y.
  - Box-Tidwell transformation of the X's.
  - constructed variables and score tests.
  - score test for non-constant error variance.

4. Logit and Probit Models for Dichotomous Data

- Linear probability, logit, and probit models for dichotomous data.
- Interpretation of coefficients in the logit model:
  - $B_j/4$ is the effect on the estimated probability of "success" $\widehat{\pi}$ of increasing X_j by 1 (or, for a dummy variable, in comparison to the baseline category), holding other Xs constant, when $\widehat{\pi}$ remains near .5.
  - $\exp(B_j) = e^{B_j}$ is the *multiplicative* effect on the estimated odds of "success" $\widehat{\pi}/(1 - \widehat{\pi})$ of increasing X_j by 1 holding other Xs constant.
- Wald and likelihood-ratio tests; analysis of deviance.

5. Logit and Probit Models for Polytomous Data

- Polytomous (multinomial) logit model.
- Nested dichotomies.
- Proportional-odds model (ordered logit model).

6. Generalized Linear Models

- Format of GLMs:
  - conditional distribution of Y:
    - exponential families: Gaussian, binomial, Poisson, gamma, inverse-Gaussian — fit by ML.
    - others: quasi-binomial, quasi-Poisson (for overdispersed binomial and Poisson data) — fit by quasi-likelihood.
    - dispersion parameter ϕ and conditional variance of Y.

Family	*Canonical Link (see below)*	*Range of* Y_i	$V(Y_i \mid \eta_i)$
Gaussian	identity	$(-\infty, +\infty)$	ϕ
binomial	logit	$\dfrac{0, 1, ..., n_i}{n_i}$	$\mu_i(1 - \mu_i)$
Poisson	log	$0, 1, 2, ...$	μ_i
gamma	inverse	$(0, \infty)$	$\phi\mu_i^2$
inverse gaussian	inverse-square	$(0, \infty)$	$\phi\mu_i^3$

- linear predictor: $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$.
- link function, $g(\mu_i) = \eta_i$; and inverse-link (mean) function, $g^{-1}(\eta_i) = \mu_i$.

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \dfrac{\mu_i}{1-\mu_i}$	$\dfrac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

▶ Poisson and quasi-Poisson models for count data

- Interpretation of coefficients: $e(B_j) = \exp(B_j)$ is the multiplicative effect on expected response count of increasing X_j by 1 (or, for dummy variable, in comparison to baseline category), holding other Xs constant.
- Same estimated coefficients for Poisson and quasi-Poisson models, but SEs for quasi-Poisson model multiply by $\sqrt{\widehat{\phi}}$ (and thus are typically larger).

▶ Analysis of deviance.

▶ Diagnostics: studentized residuals, hat-values, Cook's D, dfbeta and dfbetas, added-variable plots, component+residual plots.

7. Overview of Linear and Generalized-Linear Models

Explanatory Variables	*Response Variable*	*Type of Model*
Quantitative (e.g., education, years)	Quantitative (e.g., income, dollars)	Regression
Categorical (e.g., region, gender)	Quantitative	Analysis of Variance
Mixed	Quantitative	Dummy Regression/ Analysis of Covariance/ General Linear Model

Explanatory Variables	*Response Variable*	*Type of Model*
Any combination	Dichotomous (e.g., yes/no)	Binary/Binomial Logit Model
Any combination	Polytomous, Unordered (e.g., vote)	Multinomial Logit Model, Nested Dichotomies?
Any combination	Polytomous, Ordered (e.g., education categories)	Proportional-Odds Model?, Continuation Dichotomies?
Any combination	Count	Poisson/Quasi-Poisson Generalized Linear Model

