Some Useful Formulas From LM's, GLM's, and MLM's

Do not bring this to the exam. A copy will be provided for you.

Basic Statistics

	Traditional	Simulation
Confidence Intervals	If parameter estimate (mean, MLE, etc.) follows normal distribution, use "estimate ± 2×SE"	Estimate parameter(s) from real data, simulate 1000 data sets from estimated parameters, compute 1000 estimates of what you want, compute 2.5%-ile and 97.5%-ile.
Hypothesis Tests	Mathematically derive (or look up) distribution of T(data) under H_0. If T(real data) is in tail of distribution, reject H_0.	Estimate parameters from real data, simulate 1000 data sets to get distribution of T(data). If T(real data) is in tail of distribution, reject H_0.

Linear Regression

- *Lazy Way:* $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$
- *Long Way:* $y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \epsilon_i = X_i\beta + \epsilon_i,\ i = 1, \ldots n$
  - y_i are responses, X_{ik} are intercept (1's) and predictors, β_k are coef's, ϵ_i are iid $N(0, \sigma^2)$ "errors".
- *Matrix Way:* $Y = X\beta + \epsilon$
  - $Y_{n\times 1}$ are responses, $X_{n\times K}$ are intercept (1's) and predictors, $\beta_{K\times 1}$ are coef's, $\epsilon_{n\times 1}$ are iid $N(0, \sigma^2)$ "errors".

Some other formulae

- $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2 = \frac{1}{n-k}(y - X\hat{\beta})^T(y - X\hat{\beta})$
- $Y \sim N(X\beta, \sigma^2 I) \Rightarrow \hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$, $\hat{y} \sim N(Hy, H\sigma^2)$, where $H = X(X^TX)^{-1}X^T$
- $R^2 = \frac{\text{(raw variance)–(residual variance)}}{\text{(raw variance)}} = 1 - \frac{\hat{\sigma}^2}{\text{Var}(Y)} = [Cor(Y, \hat{Y})]^2$

Generalized Linear Models

GLM's

- $y_i \sim f(y|\mu_i, \ldots),\ \mu_i = E[y_i], i = 1 \ldots n$
- $g(\mu_i) = \theta_i = X_i\beta$

Logistic Regression

- $y_i \in \{0, 1\};\ y_i \sim$ Bern(p_i), $p_i = P[y_i = 1] = E[y_i]$
- $g(p) = \log p/(1-p);\ g^{-1}(\theta) = \exp(\theta)/(1 + \exp(\theta))$

Normal Linear Model

- $y_i \in \mathbb{R};\ y_i \sim N(\mu_i, \sigma^2),\ \mu_i = E[y_i]$
- $g(\mu) = \mu$

Poisson Regression

- $y_i \in \{0, 1, 2, 3, \ldots\};\ y_i \sim$ Poisson$(\mu_i),\ \mu_i = E[y_i]$
- $g(\mu) = \log\mu;\ g^{-1}(\theta) = \exp(\theta)$

Causal Inference

- ACE $= \frac{1}{N}\sum_{i=1}^{N}(y_i^1 - y_i^0) = \frac{1}{N}\sum_{i=1}^{N} y_i^1 - \frac{1}{N}\sum_{i=1}^{N} y_i^0 = E[y^1] - E[y^0]$; but can't obsertve both y_i^1 and y_i^0.
- $\widehat{\text{ACE}} = \hat{\beta}_1$ in $y_i = \beta_0 + \beta_1 T_i + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i$; x_{ik}'s are confounders and/or pre-treatment covariates.
- *Obs. Studies:* What can a "cause" be?
- *Obs. Studies:* Instrumental Variables, Propensity Scores, Regression Discontinuity, Bounding Confounders

Multilevel Models – Example: Random Intercept

Hierarchical Form

$$\begin{aligned} \text{Level 1:}\quad y_i &\overset{indep}{\sim} N(\alpha_{0j[i]}, \sigma^2) \\ \text{Level 2:}\quad \alpha_{0j} &\overset{iid}{\sim} N(\beta_0, \tau^2) \end{aligned}$$

Multi-Level Form

$$\begin{aligned} y_i &= \alpha_{0j[i]} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\ \alpha_{0j} &= \beta_0 + \eta_{0j},\ \eta_{0j} \overset{iid}{\sim} N(0, \tau^2) \end{aligned}$$

Variance-Components Form

$$\begin{aligned} y_i &= \beta_0 + \eta_{0j[i]} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\ &\qquad\qquad \eta_{0j} \overset{iid}{\sim} N(0, \tau^2) \end{aligned}$$

"lmer" Form

```
lmer(y ~ 1 + (1 | group))
```

Laird-Ware Form

$$y = X\beta + Z\eta + \epsilon$$

Model selection / model criticism

- AIC, BIC, DIC: 2, 3, 10
- Marginal Residuals, Conditional Residuals, Rand. Eff Residuals