STAT 151A Final Review Sheet

Transformations:

Logit: $f(x) = \log\left(\frac{x}{1-x}\right)$ Box-Cox (with parameter p): $f(x) = \begin{cases} \frac{x^p-1}{p} & \text{if } p \neq 0 \\ \log(x) & \text{if } p = 0 \end{cases}$

Simple Linear Regression

Sample correlation: $r_{x,y} = \frac{Cov(x,y)}{SD(x)SD(y)} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{SD(x)}\frac{(y_i-\bar{y})}{SD(y)}$, where $SD(x) = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$

Regression line: $\hat{y}_i = \hat{a} + \hat{b}x_i$, where $\hat{b} = r\frac{SD(y}{SD(x)}$ and $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

Geometric approach: $n \times 1$ column vectors: $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{1}$. Project $\mathbf{y}$ onto span($\mathbf{1}$, $\mathbf{x}$).

$$\hat{\mathbf{y}} = \bar{y}\mathbf{1} + b_{y\|(x\perp 1)}(\mathbf{x} - \bar{x}\mathbf{1}), \text{where } \; b_{y\|(x\perp 1)} = \frac{\mathbf{y} \bullet (\mathbf{x} - \bar{x}\mathbf{1})}{(\mathbf{x} - \bar{x}\mathbf{1}) \bullet (\mathbf{x} - \bar{x}\mathbf{1})}$$

Multiple Linear Regression

$TotSS = \sum(y_i - \bar{y})^2 = RegSS + ErrSS$, where $RegSS = \sum(\hat{y}_i - \bar{y})^2$ and $ErrSS = \sum(y_i - \hat{y}_i)^2$.

$$R^2 = \frac{RegSS}{TotSS} \quad and \quad Adjusted\ R^2 = 1 - \frac{n-1}{n-(p+1)}\frac{ErrSS}{TotSS}$$

Linear Model Assumptions $\mathbf{y} = \mathbb{X}\boldsymbol{\beta} + \epsilon$, where ϵ_i are independent, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, and normally distributed.

Under this model, $\mathbf{y}$ is normally distributed, with $E(\mathbf{y}) = \mathbb{X}\boldsymbol{\beta}$ and $Var(\mathbf{y}) = \boldsymbol{\sigma^2}\mathbf{I_n}$.

$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y}$.

$\hat{\mathbf{y}} = H\mathbf{y}$, where $H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.

$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ and $\mathbf{e} = (I - H)\mathbf{y}$.

$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$.

$E(\hat{\mathbf{y}}) = \mathbb{X}\boldsymbol{\beta}$ and $Var(\hat{\mathbf{y}}) = \sigma^2 H$

$E(\mathbf{e}) = \mathbf{0}$ and $Var(\mathbf{e}) = \sigma^2(I - H)$

Geometric Perspective: The partial coefficient for $\mathbf{x_j}$ is

$$\hat{\beta}_j = b_{(y\perp 1, x_k, k\neq j)\|(x\perp 1, x_k, k\neq j)} \qquad and \qquad SE(\hat{\beta}_j) = \frac{\sigma}{\sqrt{(1-R_j^2)\sum(x_{j,i} - \bar{x}_j)^2}}$$

Testing For $H_o : \beta_j = 0$

$$\frac{\hat{\beta}_j}{s_e\sqrt{v_{ii}}} \sim t_{n-(p+1)}$$

Note, $SE(\hat{\beta}_j) = \sigma\sqrt{v_{ii}}$, where v_{ii} is the i^{th} diagonal element of $(\mathbb{X}^T\mathbb{X})^{-1}$.

For $H_o : \beta_1, \ldots \beta_k = 0$

$$\frac{n-(p+1)}{k}\frac{(RegSS_{full} - RegSS_{part})}{ErrSS_{full}} \sim F_{k,n-(p+1)}$$

For $H_o : L\beta = \mathbf{c}$, where L is $k \times (p+1)$, the test statistic

$(L\hat{\boldsymbol{\beta}} - c)^T[L(\mathbb{X}^T\mathbb{X})^{-1}L^T]^{-1}(L\hat{\boldsymbol{\beta}} - c)/(ks_e^2) \sim F_{k,n-(p+1)}$.

Categorical Variables: For $\mathbf{v}$ a categorical variable with levels a, b, c and $\mathbf{x}$ a numeric variable, we have the linear model: $y = \alpha + \beta_x x + \gamma_a D_a + \gamma_b D_b + \epsilon$, where $D_a = 1$ when $v =$ "a" and $D_a = 0$ otherwise. We also have the model with an interaction: $y = \alpha + \beta_x x + \gamma_a D_a + \gamma_b D_b + \tau_a x D_a + \tau_b x D_b + \epsilon$.

Bootstrap:
Normal-theory interval: $\widehat{\beta} \pm \Phi^{-1}(1-\alpha/2) \cdot \widehat{SE}^*(\widehat{\beta})$
Percentile interval: if $\widehat{\beta}^*_{(q)}$ is qth quantile of the bootstrap estimates,

$$\left[\widehat{\beta}^*_{(\alpha/2)},\ \widehat{\beta}^*_{(1-\alpha/2)}\right]$$

Studentized interval:

$$[\hat{\beta} - q^*_{(1-\alpha/2)}\hat{SE}^*(\hat{\beta}), \hat{\beta} - q^*_{(\alpha/2)}\hat{SE}^*(\hat{\beta})]$$

where $q_{(p)}$ is pth quantile of studentized bootstrap statistics

$$\frac{\hat{\beta}^* - \hat{\beta}}{\hat{SE}(\hat{\beta}^*)}$$

.

Leverage:
The Hat matrix $H = X(X^TX)^{-1}X^T$ is a projection matrix. The diagonal elements $h_i \equiv h_{ii}$ are the hat values. In simple linear regression they are $h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_j-\bar{x})^2}$. Note $\bar{h} = (p+1)/n$ and this can be compared to h_i.

Outliers:
Standardized residuals: $\tilde{e}_i = \frac{e_i}{s_e\sqrt{1-h_i}}$ and Studentized residuals: $e_i^* = \frac{e_i}{s_{e_{(-i)}}\sqrt{1-h_i}}$

Influence on $\widehat{\beta}$: Cook's Distance $D_i = \frac{\tilde{e}_i^2}{p+1} \times \frac{h_i}{1-h_i}$

Influence on SE: $COVRATIO_i = \left[(1-h_{ii})\left(\frac{n-p-2+e_i^{*2}}{n-p-1}\right)^{p+1}\right]^{-1}$

Model Selection
Adjusted R^2: $1 - \frac{n-1}{n-(p(m)+1)}\frac{ErrSS(m)}{TotSS}$
Mallows C_p: $p(m)+1+(k-p(m))(F-1) = 2(p(m)+1) - n + \frac{ErrSS(m)}{s_e^2}$
Generalized Cross-Validation: $GCV(m) = \frac{nErrSS(m)}{(n-(p(m)+1))^2}$
Leave-one-out cross-validation: $\sum_{i=1}^n e_{-i}^2(m)/n$, where $e_{-i}(m)$ is the residual obtained for y_i when fitting without i^{th} observation.
AIC: $-2\log\mathcal{L} + 2(p(m)+1)$. For normal errors: $n\log(ErrSS(m)/n) + 2(p(m)+1)$
BIC: $-2\log\mathcal{L} + \log(n)(p(m)+1)$. For normal errors: $n\log(ErrSS(m)/n) + \log(n)(p(m)+1)$.

Shrinkage Methods Standardized variables so they are on the same scale, i.e., $\mathbf{z}_i = (\mathbf{x}_i - \bar{x}_i)/s_i$, where $s_i^2 = \frac{1}{n-1}\sum_j(x_{ji}-\bar{x}_i)^2$.

Ridge Regression: $\min_{\boldsymbol{\beta}}\{|\mathbf{y} - Z\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|^2\}$. Solution: $\hat{\boldsymbol{\beta}}_R = (X^TX + \lambda I_p)^{-1}X^T\mathbf{y}$

Lasso Regression: $\min_{\boldsymbol{\beta}} |\mathbf{y} - Z\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|_1$, where $|\boldsymbol{\beta}|_1 = \sum|\beta_j|$.

Categorical responses
Logistic Regression $y_i \sim$ Bernoulli(π_i), $\quad \log(\frac{\pi_i}{1-\pi_i}) = \mathbf{x}_i^T\boldsymbol{\beta}$ or $\pi_i = \frac{1}{1+\exp(-\mathbf{x}_i^T\boldsymbol{\beta})}$.
Likelihood: $\mathcal{L}(\boldsymbol{\beta}, y_1, \ldots, y_n) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \prod_{i=1}^n \left(\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right)^{y_i}\frac{1}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}$.
Estimating equations: $X^T\mathbf{y} = X^T\hat{\boldsymbol{\pi}}$
Asymptotic distribution: $\sqrt{nI(\boldsymbol{\beta}_o)}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o)$ converges to $N(0, I)$, where $I(\boldsymbol{\beta_o}) = X^TVX$ and $V = Diag(\pi_i(1-\pi_i))$.

Polytomous/multinomial logit: For m categories, $\pi_{j,i} = \frac{\exp(\gamma_{j0}+\gamma_{j1}x_{1,i}+\cdots+\gamma_{jp}x_{p,i})}{1+\sum_{l=1}^{m-1}\exp(\gamma_{l,0}+\gamma_{l,1}x_{1,i}+\cdots+\gamma_{l,p}x_{p,i})}$ for $j = 1, \ldots, m-1$, and $\pi_{m,i} = 1 - (\pi_{1,i} + \cdots + \pi_{m-1,i})$.

Proportional Odds For ordered categories, we can use a simpler model: for $j = 1, \dots, m-1$,

$$\log\left(\frac{\mathbb{P}(y_i > j)}{\mathbb{P}(y_i \le j)}\right) = \alpha_j + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i},$$

Likelihood Ratio and Deviance $LR = \frac{\max_{\beta \in \mathcal{B}_o} \mathcal{L}(\beta)}{\max_{\beta \in \Omega} \mathcal{L}(\beta)}$, where $H_o : \beta \in \mathcal{B}_o$, $H_A : \beta \in \mathcal{B}_A$, and $\Omega = \mathcal{B}_o \cup \mathcal{B}_A$.

Likelihood Ratio Test: Under the null hypothesis $G^2_{partial} := -2LLR = -2[l(\hat{\beta}_{partial}) - l(\hat{\beta}_{full})]$ has an asymptotic χ^2_k distribution.

Residual Deviance: $D(\hat{\beta}) = -2[l(\hat{\beta}) - l(\hat{\beta}_{sat})]$, where β_{sat} is for the saturated model.
Multiple Correlation Coefficient: $R^2 := 1 - \frac{D_F}{D_0}$, where D_0 is the deviance for the constant model.
Standardized Pearson Residual: $R_{P,i} = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}} \cdot \frac{1}{\sqrt{1-h_{ii}}}$.
Standardized Deviance Residual: $d_i = \frac{\pm\sqrt{-2(y_i \log \widehat{\pi}_i + (1-y_i)\log(1-\widehat{\pi}_i))}}{\sqrt{1-h_i}}$, where the sign matches the sign of $y_i - \widehat{\pi}_i$.

Cubic Regression Splines Different local cubic polynomial between each of m knots:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4((x_i - k_1)_+)^3 + \cdots + \beta_{m+3}((x_i - k_m)_+)^3 + \epsilon_i,$$

where $k_1 < \cdots < k_m$.

Natural Cubic Splines Same as cubic regression splines but with extra constraint that regions outside of knots are linear functions.

Generalized Additive Models With p explanatory variables $x_1, \dots, x_p$:

$$y_i = f_1(x_{1i}) + \ldots + f_p(x_{pi}) + \epsilon_i$$

Each $f_j(x_{ji})$ is a one-variable nonlinear regression such as a spline. To fit, combine bases for each function f_j into one giant basis.

Classification and Regression Trees

Regression trees: minimize ErrSS $= \sum_{\text{leaves } j} \sum_{i \in \text{leaf } j} (y_{ij} - \overline{y}_j)^2$ at each split, predict at leaves by taking average of remaining observations.

Classification trees: minimize misclassification rate, or Gini index $= \sum_{\text{leaves } j} n_j \sum_{k=1}^{K} \overline{p}_{jk}(1 - \overline{p}_{jk})$, or entropy $= -\sum_{\text{leaves } j} n_j \sum_{k=1}^{K} \overline{p}_{jk} \log(\overline{p}_{jk})$ at each split, predict at leaves by taking most popular category.