Stat 500: Formula sheet for Midterm 2 - page 1 of 2

Simple Linear Regression

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
b_0 &= \overline{Y} - b_1 \overline{X} \\
E\, b_0 &= \beta_0 \\
Var\, b_0 &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2} \right) \\
b_1 &= \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_j (x_j - \overline{x})^2} \\
E b_1 &= \beta_1 \\
Var\, b_1 &= \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2} \\
\hat{Y}_i &= b_0 + b_1 x_i \\
e_i &= Y_i - \hat{Y}_i \\
Var\, e_i &= \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right) \\
\hat{X}_{new} &= (Y_{new} - b_0)/b_1 \\
s_e^2 &= \frac{1}{n-2} \sum_i e_i^2 \\
s_{b_0} &= s_e \sqrt{1/n + \overline{x}^2 / \sum_j (x_j - \overline{x})^2} \\
s_{b_1} &= s_e \sqrt{1 / \sum_j (x_j - \overline{x})^2} \\
s_{\hat{Y}} &= s_e \sqrt{1/n + (x_i - \overline{x})^2 / \sum_j (x_j - \overline{x})^2} \\
s_p &= \sqrt{s_e^2 + s_{\hat{Y}}^2} \\
&= s_e \sqrt{1 + 1/n + (x_i - \overline{x})^2 / \sum_j (x_j - \overline{x})^2} \\
s_{x_{new}} &= s_p / \mid b_1 \mid \\
R^2 &= \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}
\end{aligned}
$$

Model Comparison

$$
F = \frac{(SS_{red} - SS_{full})/(df_{red} - df_{full})}{SS_{full}/df_{full}}
$$

Multiple Regression

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon \\
Y &= X\beta + \epsilon \\
b = \hat{\beta} &= (X'X)^{-1}(X'Y) \\
\widehat{\text{Var}}\, b &= MSE\, (X'X)^{-1} \\
s_e^2 &= MSE = (Y - Xb)'(Y - Xb)/(n-k-1) \\
\hat{Y} &= Xb = X(X'X)^{-1}X'Y = P_x\, Y \\
\hat{Y}_i &= x_i b \\
s_{\hat{Y}_i} &= \sqrt{MSE\, x_i (X'X)^{-1} x_i'} \\
\hat{y} &\pm \sqrt{(k+1) F_{k+1, df_{error}, 1-\alpha}}\, s_{\hat{y}} \\
s_p &= \sqrt{s_{\hat{Y}_i}^2 + MSE} \\
b_{std\, j} &= b_j \frac{SD(X_j)}{SD(Y)} \\
F &= \frac{(Cb - m)'[C(X'X)^{-1}C']^{-1}(Cb - m)}{r\, MSE_{full}} \\
b_w = \hat{\beta}_w &= (X'WX)^{-1}(X'WY) \\
\widehat{\text{Var}}\, b_w &= MSE\, (X'WX)^{-1} \\
s_w^2 &= (Y - Xb)'W(Y - Xb)/(n-k-1) \\
F_{lack\, of\, fit} &= \frac{MS_{lof}}{MS_{error}}
\end{aligned}
$$

ANOVA tables

Reduced (k_R parameters) and Full Model (k_F parameters)

Source	d.f.	sums of squares
Reduced model	$k_R - 1$	SS_{red}
Diff. between models	$k_F - k_R$	$SS_{red} - SS_{full}$
Error	$n - k_F$	SS_{full}

Regression

Source	d.f.	sums of squares	E MS
Regression	k	$\sum_i (\hat{Y}_i - \overline{Y})^2$	$\sigma^2 + \beta_1^2 \sum_i (x_i - \overline{x})^2)$
Error	$n - k - 1$	$\sum_i (Y_i - \hat{Y}_i)^2$	σ^2
Total	$n - 1$	$\sum_i (Y_i - \overline{Y})^2$	

Lack of Fit

Source	d.f.	sums of squares
Regression	k	$\sum_{ij} (\hat{Y}_{ij} - \overline{Y}_{..})^2$
Lack of Fit	$r - k - 1$	$\sum_{ij} (\overline{Y}_{i.} - \hat{Y}_{ij})^2$
Pure error	$n - r$	$\sum_{ij} (Y_{ij} - \overline{Y}_{i.})^2$
Total	$n - 1$	$\sum_{ij} (Y_{ij} - \overline{Y}_{..})^2$

Correlation

$$\rho = \frac{\text{Cov } X,Y}{\sqrt{\text{Var } X \text{ Var } Y}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$
$$r = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_j (x_j - \overline{x})^2 \sum_i (Y_i - \overline{Y})^2}} = b_1 s_x / s_y$$
$$T = r\sqrt{n-2}/\sqrt{1-r^2}$$
$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$
$$Z = \sqrt{n-3}(z_r - z_\rho)$$
$$r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$$
$$\rho\frac{\sigma_y}{\sigma_x} = \beta_1$$
$$\mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x = \beta_o$$
$$(1-\rho^2)\sigma_y^2 = \sigma^2$$
$$\mu_x = \mu_x$$
$$\sigma_x^2 = \sigma_x^2$$

Fun with models

$$Y_i = \beta_o + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$
$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \epsilon_i$$
$$Y_i = \beta_o + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$
$$Y_i = \beta_o + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i$$
$$Y_i = \beta_o + \beta_1 x_i + \beta_2 (x_i - 100) z_i + \epsilon_i$$
$$Y_i = \beta_o + \beta_1 x_i + \beta_2 (x_i - 100) z_i + \beta_3 z_i + \epsilon_i$$
$$Y_i = \beta_o + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$
$$\hat{X}_{max} = -\hat{\beta}_1/(2\hat{\beta}_2)$$
$$\text{Var } \hat{X}_{max} \approx \left(\frac{1}{2\hat{\beta}_2}\right)^2 \text{Var } \hat{\beta}_1 + \left(\frac{\hat{\beta}_1}{2\hat{\beta}_2^2}\right)^2 \text{Var } \hat{\beta}_2 - \frac{\hat{\beta}_1}{2\hat{\beta}_2^3}\text{Cov } \hat{\beta}_1\hat{\beta}_2$$
$$H_o : X_{max} = c \leftrightarrow H_o : \beta_1 + 2c\beta_2 = 0$$

Diagnostics

$$VIF_j = (1-R_j^2)^{-1}$$
$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$
$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\boldsymbol{I} - \boldsymbol{H})\mathbf{Y}$$
$$\text{Var}(\mathbf{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$
$$\sum_{i=1}^n h_{ii} = k+1$$
$$r_i = e_i/\sqrt{MSE(1-h_{ii})}$$
$$t_i = e_i/\sqrt{MSE_{(i)}(1-h_{ii})} = e_i\left[\frac{n-(k+1)-1}{SSE(1-h_{ii})-e_i^2}\right]^{0.5}$$
$$t_i \sim t_{n-k-2}$$
$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} = t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5}$$
$$\text{D}_i = \frac{\sum_j(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)MSE} = \left(\frac{r_i^2}{k+1}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)$$
$$\text{DFBETAS}_{k,i} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}$$
$$\sigma_i^2 = \sigma^2 e^{\lambda^T \mathbf{z}_i}$$
$$DW = \sum_i (e_i - e_{i-1})^2 / \sum_i e_i^2 = 2 - 2\,Cor(e_i, e_{i-1})$$
$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \tau_i + \varepsilon_{ij}, \quad \text{Var } \tau_i = \sigma_s^2,\ \text{Var } \varepsilon_{ij} = \sigma_e^2$$
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i = \rho\varepsilon_{i-1} + u_i$$

Guidelines:

VIF	10
h_{ii}	$2\frac{k+1}{n}$ or $3\frac{k+1}{n}$
DFFITS	
small/med n	1
large n	$2\sqrt{\frac{k+1}{n}}$
Cook's D	
no concern	$F_{k+1,n-k-1,0.2}$
substantial	$F_{k+1,n-k-1,0.5}$
DFBETA	
small/med n	1
large n	$\frac{2}{\sqrt{n}}$