Ch 6

All of these can be done in proc reg and glm we just changing model statement

First-Order Model w/ two predictors: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

- $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ ↑ response plane

First-Order Model w/ more than 2 predictors: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i$

- $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_{p-1} X_{i,p-1}$
- Assumptions:
  - $\beta_0, \beta_1, \dots \beta_{p-1} \rightarrow$ parameters
  - $X_1, X_2, \dots X_{p-1} \rightarrow$ known constants
  - $\epsilon_i \sim N(0, \sigma^2)$

Qualitative Variables:

- Dummy variables: $X = \begin{cases} 0 & \text{if group1} \\ 1 & \text{if group2} \end{cases}$

Polynomial Regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \epsilon$

Transformed Regression $\ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

Generalized Linear Regression $Y_i = c_{i0}\beta_0 + c_{i1}\beta_1 + \dots c_{i,p-1}\beta_{p-1} + \epsilon$

- In matrix terms: $\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$
- least squares criterion: $Q = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_{p-1} X_{i,p-1})^2$
- coefficients $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\mathbf{Y}$
- predictions $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{Y}$ where $H = X(X'X)^{-1}X'$
- residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
- cov $\sigma^2(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$
- s^2 $s^2(\mathbf{e}) = MSE(\mathbf{I} - \mathbf{H})$

- $\underbrace{SSTO}_{\uparrow n-1 \text{ dof}} = \underline{Y}'\underline{Y} - \frac{1}{n}\underline{Y}'\underline{J}\underline{Y} = \underline{Y}'\left[\underline{I} - \left(\frac{1}{n}\right)\underline{J}\right]\underline{Y}$ $\quad J_{n\times n} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$

- $\underbrace{SSE}_{\uparrow n-p \text{ dof}} = \underline{e}'\underline{e} = (\underline{Y} - \underline{X}\underline{b})'(\underline{Y} - \underline{X}\underline{b}) = \underline{Y}'\underline{Y} - \underline{b}'\underline{X}'\underline{Y} = \underline{Y}'(\underline{I} - \underline{H})\underline{Y}$

- $\underbrace{SSR}_{\uparrow p-1 \text{ dof}} = \underline{b}'\underline{X}'\underline{Y} - \left(\frac{1}{n}\right)\underline{Y}'\underline{J}\underline{Y} = \underline{Y}'\left[\underline{H} - \frac{1}{n}\underline{J}\right]\underline{Y}$

- $MSR = \frac{SSR}{p-1}$ NOTE: $E(MSR) = \sigma^2$ when coefficients $= 0$
- $MSE = \frac{SSE}{n-p}$; otherwise $E(MSR) > \sigma^2$
- $F^* = \frac{MSR}{MSE}$ Tests $H_0: \beta_0 = \beta_1 = \dots \beta_{p-1} = 0$
- $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$ Note. Adding predictors, even just noise, can only increase R

- $R_a^2 = \dfrac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$ Note this R_a^2 makes up for the issue w/ R^2 by penalizing additional predictors

- $\sigma^2(\underline{b}) = \sigma^2(\underline{X}'\underline{X})^{-1} = \begin{bmatrix} \sigma^2(b_0) & \sigma(b_0, b_1) & \cdots & \sigma(b_0, b_{p-1}) \\ \sigma(b_1, b_0) & \sigma^2(b_1) & \cdots & \sigma(b_1, b_{p-1}) \\ \vdots & & & \vdots \\ \sigma(b_{p-1}, b_0) & \sigma(b_{p-1}, b_1) & \cdots & \sigma^2(b_{p-1}) \end{bmatrix}$

* covb option*

- $s^2(\underline{b}) = MSE(\underline{X}'\underline{X})^{-1} = \begin{bmatrix} s^2(b_0) & s(b_0, b_1) & \cdots & s(b_0, b_{p-1}) \\ s(b_1, b_0) & s^2(b_1) & \cdots & s(b_1, b_{p-1}) \\ \vdots & & & \\ s(b_{p-1}, b_0) & s(b_{p-1}, b_1) & \cdots & s^2(b_{p-1}) \end{bmatrix}$

- $\frac{b_k - \beta_k}{se(b_k)} \sim t_{(n-p)}$

* clb option*

$\therefore$ $1-\alpha$ CI for b_k: $b_k \pm t(1-\frac{\alpha}{2}; n-p)s\{b_k\}$

- Test $b_k = 0$

$t^* = \frac{b_k}{se\{b_k\}}$

Reject when $|t^*| > t(1-\alpha/2; n-p)$

- Simultaneous joint confidence intervals
  - g parameters to estimate jointly
  - $b_k \pm t(1-\alpha/2g; n-p)s\{b_k\}$ Bonferroni CI

↑ Done by choosing $\alpha = \frac{\alpha}{2g}$

SAS CODE

```
proc reg;
model y = x1 x2 x3 x4 / clb covb alpha=.05;
run;
```

$^* \alpha = \frac{\alpha}{2(g)}$ for bonferroni

OUTPUT

Source	DF	Sum of Sq	Mean Sq	F val	P val
model	# of coefficient	SSR	MSR	$\frac{MSR}{MSE}$	test $\beta_1 = \beta_2 = \dots = 0$
error	n - # of coeff.	SSE	MSE		
Corrected total	n	SSTO			

Root MSE	$\sqrt{MSE}$	R-Square	R^2
Depentent Mean	$\frac{1}{n}\sum Y_i$	Adj-Rsquare	R_a^2
Coeff Var	$(\sqrt{MSE}/\text{Dep Mean}) * 100$		

1 for cont. vars
c-1 for c level vars

Var	DF	Parameter estimate	standard error	t value	P val	$(1-\alpha)*100$ CI
int		b_0	$se(b_0)$	$b_0/se(b_0)$	Test $b_0=0$	
X_1		b_1	$se(b_1)$	$b_1/se(b_1)$	"	
X_2		b_2	$se(b_2)$	$b_2/se(b_2)$	"	
X_3		b_3	$se(b_3)$	$b_3/se(b_3)$	"	
X_4		b_4	$se(b_4)$	$b_4/se(b_4)$	"	

cov of estimates

Var	b_0	b_1	b_2	b_3	b_4
b_0	$s^2(b_0)$	$s(b_0, b_1)$	$s(b_0, b_2)$	$s(b_0, b_3)$	$s(b_0, b_4)$
b_1	$s(b_1, b_0)$	$s^2(b_1)$	$s(b_1, b_2)$	$s(b_1, b_3)$	$s(b_1, b_4)$
b_2	$s(b_2, b_0)$	$s(b_2, b_1)$	$s^2(b_2)$	$s(b_2, b_3)$	$s(b_2, b_4)$
b_3	$s(b_3, b_0)$	$s(b_3, b_1)$	$s(b_3, b_2)$	$s^2(b_3)$	$s(b_3, b_4)$
b_4	$s(b_4, b_0)$	$s(b_4, b_1)$	$s(b_4, b_2)$	$s(b_4, b_3)$	$s^2(b_4)$

How to get cov matrix?

SAS CODE

```
Proc glm;
model y = x1 x2 x3 x4 / solution clparm alpha = .05
run;
```

$*\ \alpha = \frac{\alpha}{2(g)}$ for Bonf.

Output

Source	DF	Sum of squares	Mean Square	F-value	Pr > F
Model	# of predictors	SSR	MSR	$\frac{MSR}{MSE}$	Tests all $\beta_i = 0$
Error	n − # of pred.	SSE	MSE		
Corrected Total	n	SSTO			

Rsquare	Coeff Var	Root MSE	Y Mean
R^2	$\frac{\text{Root MSE}}{\text{Y mean}} \times 100$	$\sqrt{MSE}$	$\frac{1}{n}\sum Y_i = \bar{Y}$

Type I: sequential sum of squares: Sum of squares obtained by adding the next variable to the model given previous are a part of the model

Source	DF	Type I SS	Mean Square	F value	Pr > F
x_1	1 for continuous vars	$SSR(x_1)$	$MSR(x_1)$		
x_2	c-1 for factor levels	$SSR(x_2 \mid x_1)$	$MSR(x_2 \mid x_1)$		
x_3		$SSR(x_3 \mid x_2, x_1)$	$MSR(x_3 \mid x_2, x_1)$		
x_4		$SSR(x_4 \mid x_3, x_2, x_1)$	$MSR(x_4 \mid x_3, x_2, x_1)$		

Tests $\beta_i = 0$ given previous β_i in the model

Type III: Marginal sum of squares: Sum of squares obtained by adding this variable given all others are part of the model

Source	DF	Type III SS	Mean Square	F value	Pr > F
x_1	1 for continuous variables	$SSR(x_1 \mid x_2 x_3 x_4)$	$MSR(x_1 \mid x_2 x_3 x_4)$		
x_2	c-1 for factor levels	$SSR(x_2 \mid x_1 x_3 x_4)$	$MSR(x_2 \mid x_1 x_3 x_4)$		
x_3		$SSR(x_3 \mid x_1, x_2 x_4)$	$MSR(x_3 \mid x_1 x_2 x_4)$		
x_4		$SSR(x_4 \mid x_1, x_2, x_3)$	$MSR(x_4 \mid x_1 x_2 x_3)$		

tests $\beta_i = 0$ given all other β_i in the model

Parameter	Estimate	Standard error	t value	Pr > |t|	1-α %	Conf int
int	b_0	$se(b_0)$	$\frac{b_0}{se(b_0)}$	Tests $\beta_i = 0$		
x_1	b_1	$se(b_1)$	$\frac{b_1}{se(b_1)}$			
x_2	b_2	$se(b_2)$	$\frac{b_2}{se(b_2)}$			
x_3	b_3	$se(b_3)$	$\frac{b_3}{se(b_3)}$			
x_4	b_4	$se(b_4)$	$\frac{b_4}{se(b_4)}$			

① Estimation & CI of mean of $\hat{Y}$ given observation vector $\underline{X}$

Interval Estimation of Mean Response

- Given $\boxed{\underline{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$, $\boxed{E(Y_h)} = \underline{X}_h' \underline{\beta}$ and $\boxed{\hat{Y}_h} = \underline{X}_h' \underline{b}$

$\boxed{E(\hat{Y}_h)} = E(Y_h)$

$\boxed{\sigma^2(\hat{Y}_h)} = \sigma^2 \underline{X}_h' (\underline{X}'\underline{X})^{-1} \underline{X}_h$

$\boxed{s^2(\hat{Y}_h)} = MSE(\underline{X}_h' (\underline{X}'\underline{X})^{-1} \underline{X}_h)$

$= \underline{X}_h' s^2(\underline{b}) \underline{X}_h$

- 1-α Confidence Interval for $E\{Y_h\}$

$$\hat{Y}_h \pm t(1-\alpha/2, n-p)\, s(\hat{Y}_h)$$

$p = \#$ of predictors $+1$

② Estimation & CI of prediction of $\hat{Y}_{new}$ given new observation

- Interval Estimation of Prediction
- Given $\boxed{\underline{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$, $s^2(\hat{Y}_h) = MSE(\underline{X}_h' (\underline{X}'\underline{X})^{-1} \underline{X}_h) = \underline{X}_h' s^2(\underline{b}) \underline{X}_h$

$s^2(pred) = MSE + s^2(\hat{Y}_h) = MSE(1 + \underline{X}_h' (\underline{X}'\underline{X})^{-1} \underline{X}_h)$

- 1-α Confidence Interval for Prediction of new observation

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)\, s\{pred\}$$

③ Estimation & CI of mean of m prediction $\hat{Y}_h$ given new observation

Interval Estimation of Mean of m Predictions

- Given $\underline{X}_h = \begin{bmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$, $s^2(\hat{Y}_h) = MSE(\underline{X}_h' (\underline{X}'\underline{X})^{-1} \underline{X}_h) = \underline{X}_h' s^2(\underline{b}) \underline{X}_h$

$s^2(predmean) = \frac{MSE}{m} + s^2\{\hat{Y}_h\} = MSE(\frac{1}{m} + \underline{X}_h' (\underline{X}'\underline{X})^{-1} \underline{X}_h)$

- 1-α Confidence Interval for Mean of m Predictions

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)\, s\{predmean\}$$

SAS CODE Reg

```
proc reg;
 model y=x1 x2 x3 x4/ clm cli;
run;
```

(① clm, ② cli)

SAS CODE GLM

```
proc reg;
   model y= x1 x2 x3 x4 /clm cli;
run;
```

(① clm, ② cli)

Diagnostics & Remedial Measures

Scatter Plot Matrix

SAS CODE

```
proc sgscatter;
    matrix y x1 x2;
run;
```

	Y	X_1	X_2
Y	ρ_{YY}	ρ_{YX_1} p value for $\rho_{YX_1}=0$	ρ_{YX_2} pvalue for $\rho_{YX_2}=0$
X_1	ρ_{X_1Y} pvalue for $\rho_{X_1Y}=0$	$\rho_{X_1X_1}$	$\rho_{X_1X_2}$ pvalue for $\rho_{X_1X_2}=0$
X_2	ρ_{X_2Y} pvalue $\rho_{X_2Y}=0$	$\rho_{X_2X_1}$ pvalue for $\rho_{X_2X_1}=0$	$\rho_{X_2X_2}$

SAS CODE

```
proc corr pearson;
 var y x1 x2;
run;
```

Note: We use these to check for multicollinearity among predictors

Residual Plots

Code, same as before page 11

- Residuals vs. predictors - same as before, check marginal plots to check if curvature effect is needed
- Residuals vs. new predictors - See correlation with response to gague worthiness of model
- Residuals vs. fitted values, Abs(Residuals) vs. fitted values → Check for megaphone shape → non constant variance
  - If this is an issue, plot abs(residuals) vs predictors and find those causing the issue marginally

- Same Normality tests on pg 14 apply
- Same constant variance tests apply pg 13
- Same F Test for Lack of fit applys pg 14

- Same remedies on pages 15-16 apply