Ch 7

Extra Sums of squares: The marginal reduction in Sum of squared errors when one or several variables are added to a model

i.e. $SSR(X_2|X_1)$ = The marginal reduction in SSE when we add X_2 to a model containing X_1

$= SSE(X_1) - SSE(X_1, X_2)$

$= SSR(X_1, X_2) - SSR(X_1)$

i.e. $SSR(X_3|X_1,X_2) = SSE(X_1,X_2) - SSE(X_1,X_2,X_3)$

$= SSR(X_1,X_2,X_3) - SSE(X_1, X_2)$

i.e. $SSR(X_2,X_3|X_1) = SSE(X_1) - SSE(X_1,X_2,X_3)$

$= SSR(X_1,X_2,X_3) - SSR(X_1)$

↑ We can get these in either proc reg or proc glm through the linear regression ANOVA Table

$\boxed{SSTO} = SSR + SSE$

$\boxed{SSR(X_1,X_2)} = SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$

Note: $STTO(X_1,X_2) = SSR(X_1) + SSR(X_2|X_1) + SSE(X_1,X_2)$

* Proc glm output Type I and Type III page 25

$\boxed{MSR(X_2|X_1)} = \dfrac{SSR(X_2|X_1)}{1}$

$\boxed{MSR(X_2,X_3|X_1)} = \dfrac{SSR(X_2,X_3|X_1)}{2}$

$\underline{\text{T test } \beta_k = 0}$ $H_0: \beta_k = 0$ simulataneous

$H_a: \beta_k \neq 0$

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

* Given in parameter estimate tables in glm & reg

$\underline{\text{F test } \beta_k = 0}$ $H_0: \beta_k = 0$ given $X_1, X_2 \ldots X_{k-1}$ are already in the model

$H_a: \beta_k \neq 0$

$$F^* = \frac{SSR(X_3|X_1,X_2)}{1} \div \frac{SSE(X_1,X_2,X_3)}{n-4} = \frac{MSR(X_3|X_1,X_2)}{MSE(X_1,X_2,X_3)}$$

* Given in type I and type III test tables in glm

$\underline{\text{F test } \beta_1 = \beta_j = \beta_k = 0}$ $H_0: \beta_2 = \beta_3 = 0$

H_a: At least one $\neq 0$

$$F^* = \frac{SSE(X_1) - SSE(X_1,X_2,X_3)}{(n-2) - (n-4)} \div \frac{SSE(X_1,X_2,X_3)}{n-4}$$

$$= \frac{SSR(X_2,X_3|X_1)}{2} \div \frac{SSE(X_1,X_2,X_3)}{n-4} = \frac{MSR(X_2,X_3|X_1)}{MSE(X_1,X_2,X_3)}$$

* Add 'test x2=x3=0;'

$\boxed{R^2_{Y1|23}} = \frac{SSR(X_1|X_2,X_3)}{SSR(X_2,X_3)}$ = The proportionate reduction in the variation in Y remaining in a model including X_2 and X_3 when X_1 is added

$\boxed{r_{Y1|23}} = \pm\sqrt{R^2_{Y1|23}}$

Multicollinearity - when predictors carry the same information
ie when corr(x,y) is high, r_{12}^2 is high
* To see whether or not this is an issue look at the scatter plot matrix and correlation tables on pg 28.

Effects
1) Estimated coefficients vary greatly over samples
2) Coefficients may be deemed statistically not significant even though a statistical relationship.
3) Our general w/ 'one unit increase in x' interpretation might not make sense if effect is spread across correlated variables
4) On Extra Sums of Squares
$SSR(X_1) = 350$ } if X_1 & X_2 are highly correlated
$SSR(X_2|X_1) = 2$ } ie multicollinearity exists
5) On R^2
$R^2_{Y1} = \frac{SSR(X_1)}{SSTO} = .75$ } if X_1 & X_2 are highly correlated
$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = .02$ } ie multicollinearity exists
6) On $se(b_k)$

Variables in Model	$se(b_1)$	$se(b_2)$
X_1	.10	—
X_2	—	.12
X_1, X_2	.3	.31
$X_1\ X_2\ X_3$	3	3.2

*se inflates as we add variables to the model

Test for multicollinearity - VIF $\boxed{VIF} > 10 \rightarrow$ Multicollinearity
$\boxed{\text{Tolerance}} = \frac{1}{VIF} < .1 \rightarrow$ Multicollinearity

* Sas code
```
proc reg;
  model y = x1 x2 x3 / VIF tol;
```