

# ST518 - Multiple linear regression module

Jason A. Osborne

N. C. State Univ.

# Multiple linear regression(MLR)

Toy example: A random sample of students taking the same exam:

IQ	Study TIME	GRADE
105	10	75
110	12	79
120	6	68
116	13	85
122	16	91
130	8	79
114	20	98
102	15	76

Consider a regression model for the GRADE of subject  $i$ ,  $Y_i$ , in which the mean of  $Y_i$  is a linear function of two predictor variables  $X_{i1} = \text{IQ}$  and  $X_{i2} = \text{Study TIME}$  for subjects  $i = 1, \dots, 8$ :

$$Y = \beta_0 + \beta_1 \text{IQ} + \beta_2 \text{TIME} + \text{error}$$

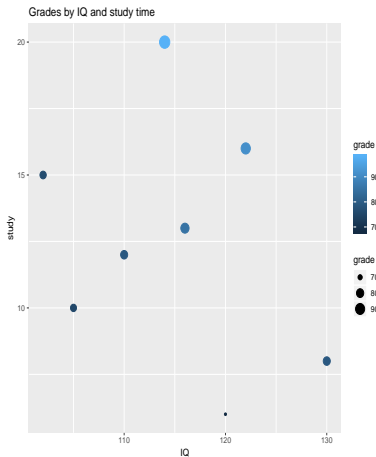
or

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i \quad \text{for } i = 1, \dots, 8$$

```

> library(ggplot2)
> iqstudy.plot <- ggplot(iqstudy.dat, aes(IQ, study))
> iqstudy.plot + geom_point(aes(color=grade, size=grade))
+ ggtitle("Grades by IQ and study time")

```



MLR model w/  $p$  independent variables

- Observed values of  $p$  independent/predictor variables for  $i^{th}$  subject from sample denoted by  $x_i. = (x_{i1}, x_{i2}, \dots, x_{ip})$
- response variable for  $i^{th}$  subject denoted by  $Y_i$
- For  $i = 1, \dots, n$ , MLR model for  $Y_i$  :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + E_i.$$

- As in SLR,  $E_1, \dots, E_n \stackrel{iid}{\sim} N(0, \sigma^2)$ , or at least  $IND(0, \sigma^2)$ .

Least squares estimates of regression parameters ( $\beta_i$ ) minimize  $SS[E]$ :

$$SS[E] = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2$$

$$\hat{\sigma}^2 = \frac{SS[E]}{n-p-1}$$

### Interpretations of regression parameters:

- $\sigma^2$  is unknown error variance parameter.
- $\beta_0, \beta_1, \dots, \beta_p$  are  $p + 1$  unknown regression parameters:
  - $\beta_0$  : average response when  $x_1 = x_2 = \dots = x_p = 0$
  - $\beta_i$  is called a slope for  $x_i$ . Represents mean change in  $y$  per unit increase in  $x_i$  **with all other independent variables held fixed**.

For the IQ/study time example, with  $p = 2$  and  $n = 8$ ,

$$\hat{\beta}_0 = 0.74, \hat{\beta}_1 = 0.47, \hat{\beta}_2 = 2.1$$

What is the uncertainty associated with these parameter estimates?

Is  $\beta_1 = 0$  and/or  $\beta_2 = 0$  consistent with data?

(Statistical inference.)

Matrix formulation of MLR

Let  $x_{i\cdot}$  be a row vector for  $p$  observed independent variables for individual  $i$

$$x_{i\cdot} = (1, x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}) \quad (1 \times (p+1)).$$

MLR model for  $Y_1, \dots, Y_n$  given by

$$Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + E_1$$

$$Y_2 = \text{_____} \quad (\text{fill this in!})$$

$$\vdots = \vdots$$

$$Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + E_n$$

System of  $n$  equations can be expressed using matrices:

$Y = X\beta + E$

- $Y$  denotes a response vector ( $n \times 1$ )
- $X$  denotes a \_\_\_\_\_ matrix ( $n \times (p+1)$ )
- $\beta$  denotes a vector of regression parameters ( $(p+1) \times 1$ )
- $E$  denotes an error vector ( $n \times 1$ ), assumed  $MVN(0, \sigma^2 I_n)$ .

To obtain matrix expressions for the LS estimates of  $\beta$ , take partial derivatives of the sum of squares function,

$$\begin{aligned} Q(\beta) &= \sum Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})^2 \\ &= (Y - X\beta)'(Y - X\beta) \\ &= \underline{\hspace{10cm}} \end{aligned}$$

Note that if  $b$  is  $p \times 1$  and  $A$  is  $(p \times p)$ , then  $\frac{\partial b'Ab}{\partial b} = (A + A')b$ .

$$\frac{\partial Q}{\partial \beta} = -2X'Y + (X'X + (X'X)')'\beta$$

The  $p + 1$  equations with  $p + 1$  unknowns obtained by setting this vector of partial derivatives are called the **normal equations**.

$$\frac{\partial Q}{\partial \beta} = 0 \implies \hat{\beta} = \underline{\hspace{10cm}}$$

### Moments of linear combinations of random vectors

Let  $Y$  denote a  $p \times 1$  random vector with mean  $\mu$  and covariance matrix  $\Sigma$ . Suppose  $a$  is a  $p \times 1$  (fixed) vector of coefficients. Then

$$\begin{aligned} E(a'Y) &= a'\mu \\ \text{Var}(a'Y) &= a'\Sigma a. \end{aligned}$$

So, let's derive  $\text{Var}(\hat{\beta}|X)$ :

$$\begin{aligned} \text{Var}(\hat{\beta}|X) &= V[(X'X)^{-1}X'Y|X] \\ &= ((X'X)^{-1}X')\text{Var}(Y|X)((X'X)^{-1}X')' \\ &= ((X'X)^{-1}X')\sigma^2 I_n((X'X)^{-1}X')' \\ &= \sigma^2((X'X)^{-1}X')((X'X)^{-1}X')' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \text{(untranspose RHS)} \\ &= \sigma^2 \underline{\hspace{10cm}} \end{aligned}$$

The variance-covariance matrix of the estimated regression coefficients.



$$\begin{aligned}
 \hat{\beta} &= (X'X)^{-1}X'Y \\
 \text{Var}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\
 &= \Sigma \\
 \widehat{\text{Var}}(\hat{\beta}) &= MS[E](X'X)^{-1} \\
 &= \hat{\Sigma} \\
 \widehat{\text{Var}}(a'\hat{\beta}) &= a'\hat{\Sigma}a
 \end{aligned}$$

- $(X'X)^{-1}$  verbalized as “x prime x inverse”
- $X$  assumed to be of full *rank*

$$\begin{aligned}
 \hat{Y} &= X\hat{\beta} = X(X'X)^{-1}X'Y = HY \\
 e &= Y - \hat{Y} = Y - X\hat{\beta} = (I - H)Y \\
 e'e &= (Y - \hat{Y})'(Y - \hat{Y}) = Y'(I - H)'(I - H)Y = Y'(I - H)Y
 \end{aligned}$$

- $\hat{Y}$  is called the vector of fitted or predicted values
- $H = X(X'X)^{-1}X'$  is called the hat matrix. It is *idempotent*.
- $e$  is the vector of residuals

IQ, Study TIME example,  $p = 2$  predictors and  $n = 8$  observations, consider  $X, Y, (X'X)^{-1}, (X'X)^{-1}X; Y, X(X'X)^{-1}X'Y$ :

$$X = \begin{pmatrix} 1 & 105 & 10 \\ 1 & 110 & 12 \\ 1 & 120 & 6 \\ 1 & 116 & 13 \\ 1 & 122 & 16 \\ 1 & 130 & 8 \\ 1 & 114 & 20 \\ 1 & 102 & 15 \end{pmatrix}, \quad X'X = \begin{pmatrix} 8 & 919 & 100 \\ 919 & 106165 & 11400 \\ 100 & 11400 & 1394 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 28.90 & -0.23 & -0.22 \\ -0.23 & 0.0018 & 0.0011 \\ -0.22 & 0.0011 & 0.0076 \end{pmatrix}$$

$$(X'X)^{-1}X'Y = \begin{pmatrix} 0.74 \\ 0.47 \\ 2.10 \end{pmatrix} = ?$$

$$SS[E] = e'e = (Y - \hat{Y})'(Y - \hat{Y}) = 45.8, \quad e'e/df = 9.15 = ?$$

$$\hat{\Sigma} = MS[E](X'X)^{-1} = \begin{pmatrix} 264.45 & -2.07 & -2.05 \\ -2.07 & 0.017 & 0.010 \\ -2.05 & 0.010 & 0.070 \end{pmatrix}$$

Distribution of parameter estimators,

- If  $E \sim N(0, \sigma^2 I)$ , then the LS estimator,  $\hat{\beta} \sim N(\beta, \Sigma)$ . and
- the  $t$ -statistics formed from  $t = (\hat{\beta}_j - \beta_j) / \sqrt{\widehat{\Sigma}_{jj}}$  follow  $t$ -distributions with  $df = n - p - 1$ .
- If  $E \sim IND(0, \sigma^2)$ , the normality of  $\hat{\beta}$  is approximate.

## Some questions - use preceding pages

- ① What is the estimate for  $\beta_1$ ? Interpretation?
- ② What is the standard error of  $\hat{\beta}_1$ ?
- ③ Is  $\beta_1 = 0$  plausible, while controlling for possible linear associations between Test Score and Study time? ( $t(0.025, 5) = 2.57$ )
- ④ Estimate the mean grade among the population of ALL students with  $IQ = 113$  who study  $TIME = 14$  hours.
- ⑤ Report a standard error for the estimate in ④
- ⑥ Report a 95% confidence interval for the quantity being estimated in ④
- ⑦ Report a 95% prediction interval for an individual student with  $IQ = 113$ ,  $TIME = 14$ .
- ⑧ Estimate the std. deviation among students whose mean estimated in ④

## Some answers

- 1  $\hat{\beta}_1 = 0.47$  (second element of  $(X'X)^{-1}X'Y$ , exam points per IQ point for students studying the same amount)
- 2  $\sqrt{0.017} = 0.13$  (square root of middle element of  $\hat{\Sigma}$ )
- 3  $H_0 : \beta_1 = 0$ , T-statistic:  $t = (\hat{\beta}_1 - 0)/SE(\hat{\beta}_1)$   
Observed value is  $t = .47/\sqrt{.017} = .47/.13 = 3.6 > 2.57$ ,  
("  $\hat{\beta}_1$  differs significantly from 0.")
- 4 Unknown population mean:  $\theta = \beta_0 + \beta_1(113) + \beta_2(14)$   
Estimate :  $\hat{\theta} = (1, 113, 14) * \hat{\beta} = 83.6$
- 5  $\text{Var}((1, 113, 14) * \hat{\beta}) = (1, 113, 14)\widehat{\text{Var}}(\hat{\beta})(1, 113, 14)'$   
or  $(1, 113, 14)\hat{\Sigma}(1, 113, 14)' = 1.3$  or  $SE(\hat{\theta}) = \sqrt{1.3} = 1.14$  (on  $df = \underline{\hspace{2cm}}$ )
- 6  $\hat{\theta} \pm t(0.025, 5)SE(\hat{\theta})$  or  $83.6 \pm 2.57(1.14)$  or  $(80.7, 86.6)$
- 7  $\hat{Y} \pm t(0.025, 5)\sqrt{MS(E) + SE(\hat{\theta})^2}$  or  $83.6 \pm 2.57\sqrt{(9.15 + 1.14^2)}$  or  $(75.3, 91.9)$
- 8  $\sqrt{MS(E)} = \sqrt{9.15} = 3.0$  (points)

```

DATA GRADES; INPUT IQ STUDY GRADE @@; CARDS;
105 10 75 110 12 79 120 6 68 116 13 85 122 16 91 130 8 79 114 20 98 102 15 76
DATA EXTRA; INPUT IQ STUDY GRADE; CARDS;
113 14 .
DATA BOTH; SET GRADES EXTRA;
PROC REG; MODEL GRADE = IQ STUDY/P CLM XPX INV COVB;

```

The SAS System  
The REG Procedure

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	IQ	STUDY	GRADE
Intercept	8	919	100	651
IQ	919	106165	11400	74881
STUDY	100	11400	1394	8399
GRADE	651	74881	8399	53617

X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	IQ	STUDY	GRADE
Intercept	28.898526711	-0.226082693	-0.224182192	0.7365546771
IQ	-0.226082693	0.0018460178	0.0011217122	0.473083715
STUDY	-0.224182192	0.0011217122	0.0076260404	2.1034362851
GRADE	0.7365546771	0.473083715	2.1034362851	45.759884688

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

(Output continued next page)

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	0.73655	16.26280	0.05	0.9656
IQ	1	0.47308	0.12998	3.64	0.0149
STUDY	1	2.10344	0.26418	7.96	0.0005

## Covariance of Estimates

Variable	Intercept	IQ	STUDY
Intercept	264.47864999	-2.069103589	-2.051710248
IQ	-2.069103589	0.016894712	0.010265884
STUDY	-2.051710248	0.010265884	0.0697933458

## Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	95% CL Predict	Residual
1	75	71.4447	1.9325	66.4770 76.4124	62.2169 80.6725	3.5553
(abbreviated)						
8	76	80.5426	1.9287	75.5847 85.5005	71.3201 89.7652	-4.5426
9	.	83.6431	1.1414	80.7092 86.5771	75.3315 91.9548	.

Sum of Residuals	0
Sum of Squared Residuals	45.75988
Predicted Residual SS (PRESS)	125.73575

(This 9<sup>th</sup> “observation” in the output data set is an illustration of the “missing y” trick to get software to generate prediction limits.)

```

> # lm() in R
> iqstudy.dat
  IQ study grade
1 105   10    75
2 110   12    79
(abbreviated)
7 114   20    98
8 102   15    76
> iqstudy.out <- lm(iqstudy.dat$grade ~ iqstudy.dat$IQ + iqstudy.dat$study)
> coef(iqstudy.out)
      (Intercept)      iqstudy.dat$IQ iqstudy.dat$study
      0.7365547      0.4730837      2.1034363
> vcov(iqstudy.out)
      (Intercept) iqstudy.dat$IQ iqstudy.dat$study
(Intercept)      264.478650      -2.06910359      -2.05171025
iqstudy.dat$IQ      -2.069104      0.01689471      0.01026588
iqstudy.dat$study  -2.051710      0.01026588      0.06979335
> summary(iqstudy.out)
Call: lm(formula = iqstudy.dat$grade ~ iqstudy.dat$IQ + iqstudy.dat$study)

Residuals:
    1     2     3     4     5     6     7     8 
3.55529  0.98300 -2.12722  2.04106 -1.10775 -0.06493  1.26318 -4.54264 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.7366    16.2628   0.045 0.965629
iqstudy.dat$IQ  0.4731     0.1300   3.640 0.014909 *
iqstudy.dat$study 2.1034     0.2642   7.962 0.000504 ***

Residual standard error: 3.025 on 5 degrees of freedom
Multiple R-squared:  0.9287,    Adjusted R-squared:  0.9002 
F-statistic: 32.57 on 2 and 5 DF,  p-value: 0.001357

```



Variable Selection

$x_1, x_2, x_3$  denote  $p$  independent variables. Consider several models:

- ①  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$
- ②  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$
- ③  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_3 x_3$
- ④  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- ⑤  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
- ⑥  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- ⑦  $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$

Language

$A$  is nested in  $B$  means model  $A$  can be obtained by placing linear restrictions on parameter values in model  $B$  (e.g.  $\beta_1 = \beta_2$ )

True or false:

- Model 1 nested in Model 4
- Model 2 nested in Model 4
- Model 3 nested in Model 4
- Model 1 nested in Model 5
- Model 4 nested in Model 1
- Model 5 nested in Model 4

$A$  nested in  $B \rightarrow A$  called *reduced*,  $B$  called *full*.

$p$  - number of regression parameters in full model

$q$  - number of regression parameters in reduced model

$p - q$  - number of regression parameters being tested.

$$SS[Total] = SS[R] + SS[E]$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (\hat{Y}_i - Y_i)^2$$

### Variable/model Selection - concepts

In comparing two models, suppose

$\beta_1, \dots, \beta_q$  in reduced ( $r$ ) model ( $A$ )

$\beta_1, \dots, \beta_q, \beta_{q+1}, \dots, \beta_p$  in full ( $f$ ) model ( $B$ ).

Comparison of models  $A$  and  $B$  amounts to testing

$H_0 : \beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$  (model  $A$  ok)

$H_1 : \beta_{q+1}, \beta_{q+2}, \dots, \beta_p$  not all 0 (need model  $B$ )

$$\text{Let } F = \frac{(SS[E]_r - SS[E]_f)/(p - q)}{MS[E]_f} = \frac{MS[H_0]}{MS[E]}$$

Difference in the numerator called an *extra regression sum of squares*:

$$R(\beta_{q+1}, \beta_{q+2}, \dots, \beta_p | \beta_0, \beta_1, \beta_2, \dots, \beta_q) = SS[R]_f - SS[R]_r = \underline{\hspace{2cm}}.$$

(ok to suppress  $\beta_0$  in these extra  $SS$  terms.)

Theory gives that if  $H_0$  holds (model  $A$  appropriate),  $F$  behaves according to  $F$  distribution with  $p - q$  numerator,  $n - p - 1$  denominator degrees of freedom. (Write  $\sim F_{p-q, n-p-1}$ )

Extra  $SS$  terms for comparing some nested models on preceding page:

- Model 1 in model 4:  $R(\beta_2, \beta_3 | \beta_1)$
- Model 2 in model 4 ?
- Model 3 in model 4 ?
- Model 1 in model 5:  $R(\beta_3 | \beta_1)$
- Model 5 in model 4: ?

To compare Models 1 and 4, compute  $F = (R(\beta_2, \beta_3 | \beta_1)/2)/MSE_4$  on  $df = 2, n - 3 - 1$ . If observed  $F$  sufficiently large, models said to differ significantly, reduced model rejected in favor of full model. If  $F$  small, reduced model plausible.

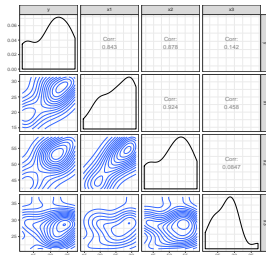
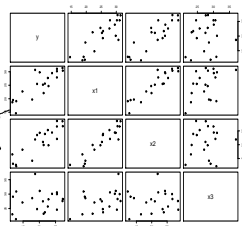
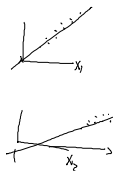
### An example: How to measure body fat?

For each of  $n = 20$  healthy individuals, the following measurements were made: bodyfat percentage  $y_i$ , triceps skinfold thickness,  $x_1$ , thigh circumference  $x_2$ , midarm circumference  $x_3$ . (See "bodyfat.txt")

```
x1    x2    x3    y
19.5  43.1  29.1  11.9
24.7  49.8  28.2  22.8
(abbreviated)
22.7  48.2  27.1  14.8
25.2  51.0  27.5  21.1
```

Summary statistics:

Symbol	Variable	mean	st. dev.
y	Body fat	20.2	5.1
x1	Triceps	25.3	5.0
x2	Thigh Circ.	51.2	5.2
x3	Midarm Circ.	27.6	3.6



```
> pairs(bodyfat.df[,c(4,1:3)])
> scatterplot(bodyfat.df, lower=list(continuous="density"), data.var=c(4,1:3), diag=list(continuous="densityDiag"))
```

Pearson Correlation Coefficients, N = 20  
Prob > |r| under H0: Rho=0

	y	x1	x2	x3
y	1.00000	0.84327 <.0001	0.87809 <.0001	0.14244 0.5491
x1	0.84327 <.0001	1.00000	0.92384 <.0001	0.45778 0.0424
x2	0.87809 <.0001	0.92384 <.0001	1.00000	0.08467 0.7227
x3	0.14244 0.5491	0.45778 0.0424	0.08467 0.7227	1.00000

Marginal associations between  $y$  and  $x_1$  and between  $y$  and  $x_2$  are highly significant, providing evidence of a strong  $r \approx 0.85$  linear association between bodyfat and triceps skinfold and between bodyfat and thigh circumference.

*Multicollinearity*: linear associations among the independent variables; causes problems such as inflated sampling variances for  $\hat{\beta}$ .

$x_1$  and  $x_2$  are particularly problematic. Imagine trying to balance a planar table top in the third dimension over “legs” that arise from the  $(x_1, x_2)$  coordinates. Highly unstable.

```

data bodyfat;
  input x1 x2 x3 y;          cards;
  19.5 43.1 29.1 11.9
  24.7 49.8 28.2 22.8
  (data abbreviated)
  22.7 48.2 27.1 14.8
  25.2 51.0 27.5 21.1
;
proc reg data=bodyfat;
  model y=x1 x2 x3; model y=x1; model y=x2; model y=x3; *usually use 4 lines of code;

```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	117.08469	99.78240	1.17	0.2578
x1	1	4.33409	<u>3.01551</u>	1.44	0.1699
x2	1	-2.85685	<u>2.58202</u>	-1.11	0.2849
x3	1	-2.18606	1.59550	-1.37	0.1896



Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-1.49610	3.31923	-0.45	0.6576
x1	1	0.85719	<u>0.12878</u>	6.66	<.0001

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-23.63449	5.65741	-4.18	0.0006
x2	1	0.85655	<u>0.11002</u>	7.79	<.0001

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	14.68678	9.09593	1.61	0.1238
x3	1	0.19943	0.32663	0.61	0.5491

Model Selection - examples

In bodyfat data, consider comparing SLR of  $Y$  on  $x_1$  with full additive model.

$$\text{Model A : } \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$$

$$\text{Model B : } \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

or the null hypothesis

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1 : \beta_2, \beta_3 \text{ not both } 0$$

after accounting for  $x_1$ .

$$F = \frac{(396.9 - 352.3)/2}{6.15} = \frac{22.3}{6.15} = 3.64$$

How many  $df$ ? The 95<sup>th</sup> percentile is  $F(0.05, \quad, \quad) = 3.63$ .

Q: Conclusion from this comparison of nested models?

After accounting for effect of  $x_1$ , the (partial) association between  $y$  and the pair  $x_2$  and/or  $x_3$  may be declared \_\_\_\_\_ at  $\alpha = \underline{\hspace{1cm}}$ .

```
* To get this $F$-ratio in SAS, try ;
proc reg data=bodyfat;
  model y=x1 x2 x3;
  test x2=0,x3=0;
run;
```



```
proc reg;
  model y=x1 x2 x3 / ss1 ss2;
run;
```

The SAS System  
The REG Procedure

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS	Type II SS
Intercept	1	117.08469	99.78240	1.17	0.2578	8156.76050	8.46816
x1	1	4.33409	3.01551	1.44	0.1699	352.26980	12.70489
x2	1	-2.85685	2.58202	-1.11	0.2849	33.16891	7.52928
x3	1	-2.18606	1.59550	-1.37	0.1896	11.54590	11.54590

Type I - \_\_\_\_\_. Type II - \_\_\_\_\_ p-values are *partial*

Note agreement between  $p$ -values from Type II  $F$  tests and  $p$ -values from  $t$  tests from parameter estimates from MLR.

Type I sums of squares - sequential (order of selection matters)

Type II sums of squares - partial ( $\Delta$  SSE due to adding term A to model with all other terms not 'containing' A)

Type III sums of squares - partial

$$\begin{aligned}
 R(\beta_1|\beta_0) &= 352.3 \\
 R(\beta_2|\beta_0, \beta_1) &= \underline{\hspace{2cm}} \\
 R(\beta_3|\beta_0, \beta_1, \beta_2) &= \underline{\hspace{2cm}} \\
 R(\beta_1|\beta_0, \beta_2, \beta_3) &= 12.7 \\
 R(\beta_2|\beta_0, \beta_1, \beta_3) &= \underline{\hspace{2cm}}
 \end{aligned}$$

Type II test for  $\beta_j$  - test of partial association between  $y$  and  $x_j$  after accounting for all other  $x_i$

Type II  $F$ -ratios from bodyfat data for  $x_1, x_2, x_3$ , respectively:

$$F = \frac{12.7/1}{6.15} = 2.07, \quad F = \frac{7.5/1}{6.15} = 1.22, \quad F = \frac{11.5/1}{6.15} = 1.88.$$

(Partial) effects significant ? (Use  $F(0.95, 1, 16) = 4.49$ .)

Exercise: Carry out the corresponding  $F$ -tests to compare models.

In PROC REG output, which models are the type I tests comparing?

- ① Type I SS for  $x_1$  appropriate for SLR of  $y$  on  $x_1$ .
- ② Type I SS for  $x_2$  appropriate for test of association between  $y$  and  $x_2$  after accounting for  $x_1$ .
- ③ Type I test for  $x_3$  same as type II test for  $x_3$ .

In all three of these tests,  $MS[E]$  computed from full model (#4).

### Some model comparison examples

- ① Compare models 1 and 6
- ② Compare models 2 and 6

For 1. use  $R(\beta_2|\beta_0, \beta_1)$  in the  $F$  ratio:

$$\begin{aligned}
 F &= \frac{R(\beta_2|\beta_0, \beta_1)}{MS[E]_6} \\
 &= \frac{33.2}{(SS[Total] - R(\beta_1, \beta_2|\beta_0))/(20 - 2 - 1)} \\
 &= \frac{33.2}{(495.4 - 352.3 - 33.2)/(20 - 2 - 1)} \\
 &= \frac{33.2}{109.9/17} = 5.1
 \end{aligned}$$

Note that  $SS[E]_f = (SS[Total] - SS[R]_f)$  and  $SS[R]_f = SS[R]_r + R(\beta_2|\beta_0, \beta_1)$

$F(0.05, 1, 17) = 4.45$  : model 1 rejected in favor of model 6: there is evidence ( $p = 0.037$ ) of association between  $y$  and  $x_2$  after accounting for dependence on  $x_1$ .

To compare models 2 and 6, we need  $SS[R]_r = R(\beta_2|\beta_0) = 382.0$  which cannot be gleaned from preceding output. You could also get it from  $r^2_{yx_2} \times SS[Tot]$  or from running something like

```
proc reg;
  model y=x1 x2/ss1 ss2;
run;
```

## The REG Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS	Type II SS
Intercept	1	-19.17425	8.36064	-2.29	0.0348	8156.76050	34.01785
x1	1	0.22235	0.30344	0.73	0.4737	352.26980	3.47289
x2	1	0.65942	0.29119	2.26	0.0369	33.16891	33.16891

$$\begin{aligned}
 F &= \frac{R(\beta_1|\beta_0, \beta_2)/(\Delta df)}{MS[E]_f} \\
 &= \frac{(SS[R]_f - SS[R]_r)/1}{6.5} \\
 &= \frac{352.3 + 33.2 - 382.0}{6.5} = \frac{3.4}{6.5} \approx 0.5
 \end{aligned}$$

Conclusions?

- $x_2$  gives you a little when you add it to model with  $x_1$
- $x_1$  gives you nothing when you add it to model with  $x_2$
- Take model with  $x_2$ . (Has higher  $r^2$  too.)
- these comparisons of nested models easy to carry out using TEST statement in PROC REG.

## Another example, revisiting test scores and study times

Consider this sequence of analyses:

- 1 Regress GRADE on IQ.
- 2 Regress GRADE on IQ and TIME.
- 3 Regress GRADE on TIME IQ TI where  $TI = \text{TIME} \times \text{IQ}$ .

ANOVA (Grade on IQ)

SOURCE	DF	SS	MS	F	p-value
IQ	1	15.9393	15.9393	0.153	0.71
Error	6	625.935	104.32		

No evidence that IQ has anything to do with grade, but we did not look at study time. Looking at the *multiple* regression we get

The REG Procedure

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	0.73655	16.26280	0.05	0.9656
IQ	1	0.47308	0.12998	3.64	0.0149
study	1	2.10344	0.26418	7.96	0.0005

Now the test for dependence on IQ is significant  $p = 0.0149$ . Why?

## The interaction model

The REG Procedure  
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	610.81033	203.60344	26.22	0.0043
Error	4	31.06467	7.76617		
Corrected Total	7	641.87500			

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS	Type II SS
Intercept	1	72.20608	54.07278	1.34	0.2527	52975	13.84832
IQ	1	-0.13117	0.45530	-0.29	0.7876	15.93930	0.64459
study	1	-4.11107	4.52430	-0.91	0.4149	580.17582	6.41230
IQ_study	1	0.05307	0.03858	1.38	0.2410	14.69521	14.69521

Model discussion. We call the product  $I*S = IQ*STUDY$  an “interaction” term.

$$\hat{G} = 72.21 - 0.13 * I - 4.11 * S + 0.0531(I * S)$$

Now if  $IQ = 100$  we get

$$\hat{G} = (72.21 - 13.1) + (-4.11 + 5.31)S$$

and if  $IQ = 120$  we get

$$\hat{G} = (72.21 - 15.7) + (-4.11 + 6.37)S.$$

With interaction model, one extra hour of study increases expected grade by 1.20 points for someone with  $IQ = 100$  and by 2.26 points for someone with  $IQ = 120$ . Since interaction not significant, we might go back to simpler “additive” model. (example taken from Dickey’s ST512 notes.)

## Some questions about design matrices

Recall three models under consideration for the bodyfat data

$$M_1: \hat{\mu}(x_1, x_2, x_3) = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$M_2: \mu(x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$$

$$M_6: \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{Var}(\hat{\beta}) = MS[E]_{M_1} (X'X)_{M_1}^{-1}$$

Q:  $MS[E]_{M_6} < MS[E]_{M_1}$  and  $MS[E]_{M_6} < MS[E]_{M_2}$  but the partial slopes have larger standard errors in  $M_6$ . Why? multicollinearity

$$n=20$$

Design matrices

$$X_{M6} = \begin{pmatrix} 1 & 19.5 & 43.1 \\ 1 & 24.7 & 49.8 \\ \vdots & \vdots & \vdots \\ 1 & 25.2 & 51.0 \end{pmatrix}$$

$$X_{M1} = \begin{pmatrix} 1 & 19.5 \\ 1 & 24.7 \\ \vdots & \vdots \\ 1 & 25.2 \end{pmatrix}$$

$$(X'X)_{M6} = \begin{pmatrix} 20 & 506.1 & 1023.4 \\ 13386.3 & 26358.7 & 52888.0 \end{pmatrix}$$

$$\begin{pmatrix} 20 & 506.1 \\ 13386.3 \end{pmatrix}$$

$$(X'X)_{M1} = \begin{pmatrix} ? & ? \\ ? & ? \end{pmatrix}$$

$$(X'X)_{M1}^{-1} = \begin{pmatrix} 1.39 & -0.053 \\ & 0.002 \end{pmatrix}$$

$$(X'X)_{M2} = \begin{pmatrix} ? & ? \\ ? & ? \end{pmatrix}$$

$$(X'X)_{M2}^{-1} = \begin{pmatrix} 5.08 & -0.098 \\ & 0.0019 \end{pmatrix}$$

$$(X'X)_{M6}^{-1} = \begin{pmatrix} 10.8 & 0.29 & -0.35 \\ & 0.014 & -0.012 \\ & & 0.013 \end{pmatrix}$$

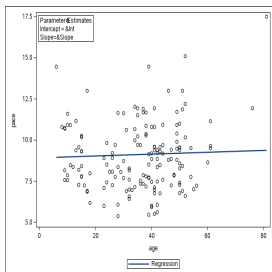
$$\begin{pmatrix} 20 & 1023 \\ 52888 \end{pmatrix}$$

Q: Why is  $\text{Var}(\hat{\beta}_0)$  bigger in  $M_2$  than in  $M_1$ ?

```

pace    age    sex
5.38333  28    M
5.46667  39    M
(abbreviated)
17.2667  10    F
17.5000  81    M

```



Resolution 5k run, Centennial campus

Symbol	Variable	mean	st. dev.	variance
y	Pace	9.1	2.2	5.0
x	Age	35.1	14.7	216.5

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	8.92271	0.45724	19.51	<.0001
age	1	0.00564	0.01203	0.47	0.6396



Let  $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$ . Quadratic model for pace ( $Y$ ) as a function of age ( $x$ ):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \quad \text{for } i = 1, \dots, 160$$

$$\hat{Y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

- $\beta = (\beta_0, \beta_1, \beta_2)'$  is a vector of unknown regression parameters
- $\sigma^2$  is the unknown error variance of paces given age  $x$ .

Compare this model with the (previously discarded) SLR model

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \dots, 160$$

Q1: Does  $\beta_1$  have the same interpretation in both models?

Q2: How can we compare the two models?

A2: Using  $F$ -ratios to compare nested models (see output next page).

$$\begin{aligned} F &= \frac{R(\beta_2 | \beta_0, \beta_1)}{MS[E]_{full}} \\ &= \frac{(SS[R]_{full} - SS[R]_{red})/1}{MS[E]_{full}} = \frac{(SS[E]_{red} - SS[E]_{full})/1}{MS[E]_{full}} \\ &= \frac{(113.6 - 1.1)/1}{4.3} = \frac{(787.0 - 674.4)/1}{4.3} \\ &= 26.2 \\ &= \left( \frac{\hat{\beta}_2}{SE} \right)^2 \end{aligned}$$

with  $F(0.05, 1, 157) = 3.90$ . Since  $26.2 \gg 3.9$ , the linear model is implausible when compared to the quadratic model. Also,  $R(\beta_1, \beta_2 | \beta_0) = 113.6$ ,  $F = 113.6/4.3 = 26.4$  so that  $H_0: \beta_1 = \beta_2 = 0$  can be rejected.

```

PROC REG DATA=one; /* age2 defined in data step as age*age */
  MODEL pace=age; /* not necessary in light of MODEL2 statement */
  MODEL pace=age age2/ss1; /* ss1 generates sequential sums of squares */
RUN;

```

The REG Procedure  
Model: MODEL1

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.09650	1.09650	0.22	0.6396
Error	158	786.99821	4.98100		
Corrected Total	159	788.09472			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	8.92271	0.45724	19.51	<.0001
age	1	0.00564	0.01203	0.47	0.6396

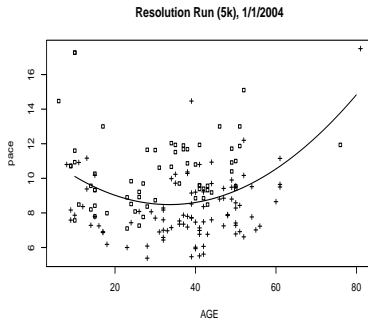
## Model: MODEL2

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64500	56.82250	13.23	<.0001
Error	157	674.44972	4.29586		
Corrected Total	159	788.09472			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type I SS
Intercept	1	11.78503	0.70216	16.78	<.0001	13310
age	1	-0.19699	0.04113	-4.79	<.0001	1.09650
age2	1	<u>0.00294</u>	<u>0.00057380</u>	<u>5.12</u>	<.0001	112.54850

$R^2(\beta_2 | \beta_0, \beta_1)$



Fitted model is

$$\hat{\mu}(x) = 11.785 - 0.197 x + 0.00294 x^2$$

or

$$\hat{\mu}(\text{age}) = 11.785 - 0.197 \text{ age} + 0.00294 \text{ age}^2.$$

Inference for response  $Y$  given predictor  $x_i$ .

Random sample of  $n = 31$  trees drawn from population of trees.  $p = 3$  variables measured on each:

- $x_{i1}$ : "girth", tree diameter in inches
- $x_{i2}$ : "height" (in feet)
- $Y_i$ : volume of timber, in cubic feet.

Given  $x_1$  and  $x_2$ , a MLR model for these data given by

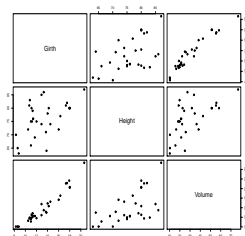
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i \text{ for } i = 1, \dots, n$$

For trees with  $x_1, x_2$  the model for mean volume is

$$\mu(x_1, x_2) = E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

A scatterplot matrix

> pairs(trees)



## Some questions involving linear combinations of regression coefficients

Consider all trees with girth  $x_{01} = 15$  in and height  $x_{02} = 80$  ft.

- Estimate the mean volume among these trees, along with a standard error and 95% confidence interval.

- Obtain a 95% prediction interval of  $y_0$ , the volume from an individual tree sampled from this population of 80 footers, with girth 15 inches.

SAS generates  $\hat{\beta}$  and  $\widehat{Var}(\hat{\beta}) = MSE * (X'X)^{-1}$

$$\hat{\mu}(15, 80) = (1, 15, 80) \hat{\beta}$$

$$= -57 + 4.7(15) + .34(80)$$

$$SE(\hat{\mu}(15, 80)) = \sqrt{(1, 15, 80) \hat{\Sigma} \begin{pmatrix} 1 \\ 15 \\ 80 \end{pmatrix}}$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7684.16251	3842.08126	254.97	<.0001
Error	28	421.92136	15.06862		
Corrected Total	30	8106.08387			

Root MSE	3.88183	R-Square	0.9480
Dependent Mean	30.17097	Adj R-Sq	0.9442
Coeff Var	12.86612		

## Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-57.98766	8.63823	-6.71	<.0001
Girth	1	4.70816	0.26426	17.82	<.0001
Height	1	0.33925	0.13015	2.61	0.0145

## Covariance of Estimates

Variable	Intercept	Girth	Height
Intercept	74.6189461	0.4321713812	-1.050768886
Girth	0.4321713812	0.0698357838	-0.017860301
Height	-1.050768886	-0.017860301	0.0169393298

$$.95 = P\left(-t < \frac{\hat{\mu} - \mu_0}{\sqrt{Var(\hat{\mu} - \mu_0)}} < t\right)$$

$$= P\left(-t < \frac{\hat{\mu} - \mu_0}{\sqrt{38^2 + MSE}} < t\right)$$

$$\hat{\Sigma} = (X'X)^{-1} MSE$$

Inference for the mean response in MLR Recall that if  $Y$  a  $p \times 1$  random vector with mean  $\mu$  and covariance matrix  $\Sigma$ . and  $a$  a  $p \times 1$  (fixed) vector of coefficients.

$$\begin{aligned} E(a'Y) &= a'\mu \\ \text{Var}(a'Y) &= a'\Sigma a. \end{aligned}$$

$$.95 = P\left(-t < \frac{\hat{\mu} - \mu}{SE} < t\right)$$

Consider subpopulation of trees with Girth 15 and Height 80. To estimate mean volume among these trees, with estimated std. error, take  $x'_0 = (1, 15, 80)$  and consider  $\hat{\mu}(x_0) = x'_0 \hat{\beta}$ .

$$\begin{aligned} E(x'_0 \hat{\beta}) &= x'_0 \beta \\ \text{Var}(x'_0 \hat{\beta}) &= x'_0 \hat{\Sigma} x_0 \end{aligned}$$

Substitution of  $\hat{\beta}$  and  $\hat{\Sigma} = \text{MSE}(X'X)^{-1}$  gives the estimates:

$$\mu(x_0) = (1, 15, 80)\beta$$

no hat!

$$\hat{\mu}(x_0) = (1, 15, 80) \begin{pmatrix} -58.0 \\ 4.71 \\ 0.34 \end{pmatrix} = 39.8$$

$$P^3 \left( \hat{SE} = .85 \right)$$

$$39.8 \pm 2.05(.85)$$

a 95% CI for  $\mu(15, 80)$

$= \mu(15, 80)$

$$n=31$$

$$p=2$$

$$df_{\text{error}} = n - p - 1 = 28$$

$$\widehat{\text{Var}}(\hat{\mu}(x_0)) = (1, 15, 80) \begin{pmatrix} 74.62 & 0.43 & -1.05 \\ 0.43 & 0.070 & -0.018 \\ -1.05 & -0.018 & 0.017 \end{pmatrix} \begin{pmatrix} 1 \\ 15 \\ 80 \end{pmatrix} = 0.72$$

$$SE(\hat{\mu}(x_0)) = \sqrt{.72} = 0.849$$

which can be obtained using PROC REG and the missing y trick:

Obs	treenumber	Girth	Height	Volume	p	sepred
32	100	15	80	.	39.7748	0.84918

95% Prediction limits? Use  $\pm t(.025, 28) \sqrt{.72 + MS(E)}$ .

Partial correlations

The partial correlation coefficient for  $x_1$  in the MLR

$$E(Y|x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

is defined as the correlation coefficient between the residuals computed from the two regressions below:

$$\begin{aligned} Y &= \beta_0 + \beta_2 x_2 + \dots + \beta_p x_p + E \quad \mapsto e_{Y \cdot 2, 3, \dots, p} \\ x_1 &= \beta_0 + \beta_2 x_2 + \dots + \beta_p x_p + E \quad \mapsto e_{1 \cdot 2, 3, \dots, p} \end{aligned}$$

*residuals*

Call these sets of residuals  $e_{Y \cdot 2, 3, \dots, p}$  and  $e_{1 \cdot 2, 3, \dots, p}$  respectively. The *partial correlation* between  $y$  and  $x_1$  after accounting for the linear association between  $y$  and  $x_2, x_3, \dots, x_p$  is defined as

$$r_{y1 \cdot 2, 3, \dots, p} = \text{correlation between } e_{Y \cdot 2, 3, \dots, p} \text{ and } e_{1 \cdot 2, 3, \dots, p} \quad \frac{\sum e_{Y \cdot 2, 3, \dots, p} e_{1 \cdot 2, 3, \dots, p}}{\sqrt{\sum e_{Y \cdot 2, 3, \dots, p}^2 \sum e_{1 \cdot 2, 3, \dots, p}^2}}$$

The *partial coeff. of determination* is  $r_{y1 \cdot 2, 3, \dots, p}^2$ .

$$SS(E)$$

Note also that

$$r_{y1 \cdot 2, \dots, p}^2 = \frac{R(\beta_1 | \beta_0, \beta_2, \dots, \beta_p)}{\underbrace{SS[Total] - R(\beta_2, \beta_3, \dots, \beta_p | \beta_0)}_{SS(E)_{M1}}}$$

Bodyfat data, compare models 1,2 and 6 (ignore  $x_3$ .)

bodyfat data									
Obs	x1	x2	y	py_1	ey_1	e2_1	py_2	ey_2	e1_2
1	19.5	43.1	11.9	15.2190	-3.31903	-2.48145	13.2827	-1.38267	1.34939
2	24.7	49.8	22.8	19.6764	3.12360	-0.78756	19.0215	3.77847	0.60956
(abbreviated)									
20	25.2	51.0	21.1	20.1050	0.99500	-0.06892	20.0494	1.05061	0.04571

The partial correlation coefficient between  $y$  and  $x_1$  after accounting for  $x_2$  is  $r_{y1.2} = 0.17$  and the partial for  $x_2$  after accounting for  $x_1$  is  $r_{y2.1} = 0.48$ . The partial coefficients of determination are

$$\frac{R(\beta_1 | \beta_0, \beta_2)}{SS(\text{Tot}) - R(\beta_2 | \beta_0)} \leftarrow \text{extra } SS(1) = r_{y1.2}^2 = 0.03062 \quad \text{and} \quad \frac{R(\beta_2 | \beta_0, \beta_1)}{SS(\text{Tot}) - R(\beta_1 | \beta_0)} \leftarrow \text{extra } SS(2) = r_{y2.1}^2 = 0.23176$$

Q: If you had to choose one variable or the other from  $x_1$  and  $x_2$ , which would it be?

Q: Anything wrong with throwing both  $x_1$  and  $x_2$  in the final model?

Q: Write coefficients of determination in terms of extra sums of squares. Use  $R(\cdot|\cdot)$  notation.

Note: partial correlations obtained in SAS using PCORR2 option:

Variable	DF	Squared Partial	
		Corr	Type II
Intercept	1	.	.
x1	1	0.03062	.
x2	1	0.23176	.



## Partial regression plots

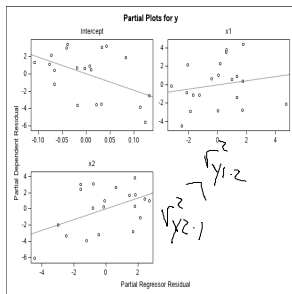
A plot of the residuals from the regression

$$Y = \beta_0 + \beta_2 x_2 + \cdots + \beta_p x_p + E$$

versus the residuals from the regression

$$X_1 = \beta_0 + \beta_2 x_2 + \cdots + \beta_p x_p + E$$

is called a partial regression or leverage plot for  $x_1$  or in the MLR. Can be generated using ODS GRAPHICS ON and the PARTIAL command in the MODEL statement of PROC REG:



Q: What can these plots tell us?

A1: They convey info. about linear associations between  $y$  and candidate variable  $x_i$  after accounting for linear dependence of  $y$  on other variables  $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p$

A2: can convey info about nonlinear associations between  $y$  and  $x_i$  after accounting for other linear associations

A3: They can illuminate possible outliers.

Some exercises (hint: use matrix algebra or SAS).

- ➊ Regarding the `resrun` data as randomly sampled a population of interest. Consider the sub-population of 32 year old males. Fit a quadratic regression function and use it to obtain an estimate of the mean 5k pace in this cohort of all 32 yr-old male runners. Report a standard error and 95% confidence interval.
- ➋ Obtain a 95% prediction interval for one such runner.
- ➌ Explain the difference between the two intervals in questions 1 and 2.
- ➍ At what rate is  $\mu(x)$  changing with age? Estimate the appropriate function.
- ➎ Estimate  $\theta$ , the peak age to run a 5k in the fastest time. Is  $\theta$  a linear function of regression parameters? Can you obtain an unbiased estimate of the standard error of  $\theta$ ?

# Cook's D

Cook's D for an observation  $i$  is a measure of influence on predictions:

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{pMS(E)}$$

*Handwritten note: #vars + 1*

where  $\hat{y}_{j(i)}$  is the fitted value for the  $j^{th}$  observation if the  $i^{th}$  observation is held out.

cdp from ODS output

Obs	pace	age	sexf	CooksD
1	17.5000	81	0	0.52957
2	11.9333	76	1	0.45148
3	17.2667	10	1	0.13515

$$p = 3 + 1 = 4$$

The MEANS Procedure

Analysis Variable : diff

Mean	Corrected SS	USS
0.0271229	6.6410416	6.7587459

mycooksD computed just for age=81 subject

Obs	_TYPE_	_FREQ_	uss	mycooksD
1	0	160	6.75875	0.52957

```

*resruncooksds.sas;
*Cook's D calculated by PROC REG and from definition;
*inspect code carefully;
data one; set one;
    race=1*scan(crace,1,':')+1/60*scan(crace,2,':');
    pace=1*scan(cpace,1,':')+1/60*scan(cpace,2,':');
    /* The SCAN command extracts characters from a character string */
    age2=age*age; sexf=(sex="F"); agef=age*sexf; age2f=age2*sexf;
    pace2=pace; if age=81 then pace2=.;
run;
ods graphics on; ods trace on; ods listing close;
proc reg data=one plots=cooksds;
    id name pace age sexf;
    model pace=age age2 sexf/influence;
    ods output cooksplot=cdp(rename=(id1=name id2=pace id3=age id4=sexf)) ;
    output out=preds1 p=p1;
run;
proc reg data=one; *where age<81;
    model pace2=age age2 sexf/influence;
    output out=preds2 p=p2;
run;
ods listing ;
proc sort data=cdp; by descending cooksds;
proc print data=cdp (obs=3); title "cdp from ODS output";
    var pace age sexf cooksds;
run;

```

Cook's D for every observation may be “delivered” using PROC REG and SAS ODS with the keyword “cooksplot” discovered using [ods trace on](#).

cdp from ODS output

Obs	pace	age	sexf	CooksD
1	17.5000	81	0	0.52957
2	11.9333	76	1	0.45148
3	17.2667	10	1	0.13515

## Calculating Cook's D from the definition ...

```

proc sort data=preds1; by name; run;
proc sort data=preds2; by name; run;
data both;
    merge preds1 preds2;
    by name;
    diff=p1-p2;
run;
proc sort data=both; by descending age;run;
proc means data=both mean css uss;
    var diff;
    output out=uss uss=uss;
run;
data uss; set uss; mycooksD=uss/(4*3.19068); run;
/* MSE=3.19068 was hard-coded after inspection of output from MODEL1*/
proc print data=uss;
    title "mycooksD computed just for age=81 subject";
run;

```

## The MEANS Procedure

Analysis Variable : diff

Mean	Corrected SS	USS
0.0271229	6.6410416	6.7587459

mycooksD computed just for age=81 subject

3

Obs	_TYPE_	_FREQ_	uss	mycooksD
1	0	160	6.75875	0.52957

### More about model selection: $R^2$ , $R_a^2$ and Mallows's $C_p$

In statistical modelling in general, one goal is often to identify a model explains variability in some response ( $y$ ) of interest through its association with explanatory factors or variables. The principle of model parsimony dictates that it is best to construct a model which explains things, but with as few variables as possible.

Suppose that the true regression function underlying observed data is given by

$$E(Y|x_1, \dots, x_{q+1}) = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q \quad (1)$$

but that an analysis leads to the model

$$E(Y|x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{q-2} x_{q-2} \quad (2)$$

Model (2) is said to be *underspecified*. On the other hand, suppose another analysis leads to the model

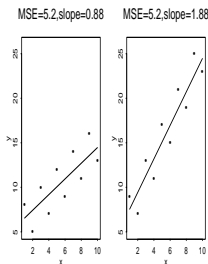
$$E(Y|x) = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \dots + \hat{\beta}_q x_q + \hat{\beta}_{q+1} x_{q+1} \quad (3)$$

Then model (3) is said to be *overspecified*.

- Underspecified models introduce bias ( $E(\hat{\mu}(x)) \neq E(Y|x)$ .)
- overspecified models inflate the sampling variances of estimators.

Which mistake is worse? (In the sense that underspecification is equivalent to fewer type I errors for tests of the form  $H_0 : \beta_i = 0$ , it may be preferable, as a rule of thumb.)  
The coefficient of multiple determination in MLR,  $R^2$ :

- proportion of variability accounted for by a linear model, also the squared correlation between observed  $(y_1, y_2, \dots)$  and predicted  $(\hat{y}_1, \hat{y}_2, \dots)$  values
- a reasonable criterion for model selection, but not infallible.



For two datasets with “equal” variability unexplained by SLR, the model with larger absolute slope will have higher  $R^2$ . To see this, recall that

$$R^2 = \frac{SS[R]}{SS[Tot]} = \frac{SS[Tot] - SS[E]}{SS[Tot]} = 1 - \frac{SS[E]}{SS[Tot]}.$$

bigger spread in  $y \implies$  bigger  $R^2$

Q: Which line yields a higher  $r^2$ ? Is this a better fit?

$R_a^2$ , or the *adjusted coefficient of multiple correlation* is given by

$$R_a^2 = 1 - \left( \frac{n-1}{n-p-1} \right) \frac{SS[E]}{SS[Tot]}$$

It imposes a penalty on added independent variables.

#### Mallow's $C_p$ statistic

Suppose  $m$  denotes number of independent variables in full model,  $p \leq m$  denotes number of candidates under consideration in reduced model and  $n$  denotes sample size.

$$C_p = p + 1 + \frac{(MS[E](p) - MS[E](m)) * (n - p - 1)}{MS[E](m)}$$

Subset models for which  $C_p \leq p + 1$  are preferred.



In addition to  $R^2$ ,  $C_p$  adjusted  $R^2$ , we have AIC, AICc, BIC, ...

Let the likelihood function for a model parameterized by  $k$ -dimensional  $\theta$  (including intercept, if any), using a sample of size  $n$  be denoted  $\mathcal{L}$

$$\mathcal{L}(\theta) = f(y_1, \dots, y_n; \theta)$$

Let the maximum likelihood estimator of  $\theta$  be denoted  $\hat{\theta}$ . Then

$$AIC = -2 \log \mathcal{L}(\hat{\theta}) + 2k$$

When sample size  $n$  is small there is a corrected version of AIC:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$$

Schwarz or Bayesian Information Criterion (SIC/BIC)

$$BIC = -2 \log \mathcal{L}(\hat{\theta}) + \log(n)k$$

(penalty in BIC larger than in AIC)

Mallow's  $C_p$  for reduced model of dimension  $q$ .

$$\begin{aligned} C_p &= \frac{SS[E]_r}{MS[E]_f} + 2p - n \\ &= p + \frac{MS(E)_r - MS(E)_f}{MS(E)_f} (n - p) \end{aligned}$$

Values small  $MS(E)_r$  but also small  $p$ .

```

/* models with potentially nonlinear age effects allowed to vary with sex */
proc reg data=one;
  model pace=age age2 sexf agef age2f/selection=cp aic bic rsquare;
run;

```

The SAS System

The REG Procedure

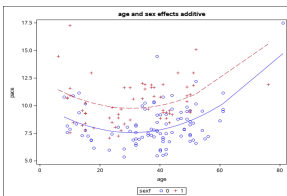
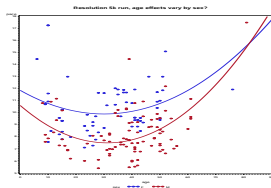
Dependent Variable: pace

C(p) Selection Method

Number of Observations Used

Number in

Model	C(p)	R-Square	Adjusted R-Square	AIC	BIC	MSE	Variables in Model
3	2.9901	0.3684	0.3563	189.5866	191.8431	3.19068	age age2 sexf
4	4.8893	0.3688	0.3525	191.4825	193.8103	3.20918	age age2 sexf age2f
4	4.9883	0.3684	0.3521	191.5848	193.9061	3.21123	age age2 sexf agef
4	5.0408	0.3682	0.3519	191.6390	193.9568	3.21232	age age2 agef age2f
5	6.0000	0.3725	0.3521	192.5612	195.0257	3.21147	age age2 sexf agef age2f
3	11.7363	0.3328	0.3199	198.3700	200.1864	3.37073	age age2 agef
(abbreviated)							
1	89.0585	0.0014	-.0049	258.8883	259.2590	4.98100	age



For MLR with i.i.d. normal errors, and  $\beta((p+1) \times 1)$

$$\mathcal{L}(\beta, \sigma^2) = \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\sigma^2[(y_i - x_i\beta)/\sigma]^2\right\}$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\sigma}^2 = \frac{n - (p+1)}{n} MS(E)$$

$$\log \mathcal{L}(\beta, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta)$$

$$\begin{aligned} \log \mathcal{L}(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= -\frac{n}{2} \log \hat{\sigma}^2 + \text{constant} \end{aligned}$$

PROC REG reports  $AIC = n \log \left( \frac{SSE}{n} \right) + 2(p+1)$ .

[link to SAS PROC REG DOC](#)

For the full model,  $MS(E) = 3.21147$  and

```
> 160*(log(3.21147*(160-6)/160)) + 2*6
[1] 192.5612
```

So, check your software's computations!

### Residual diagnostics

- Residuals can be plotted against independent/predictor variables to check for model inadequacy. (e.g. if relationship is quadratic, but only a linear model was fit, this plot will reveal a pattern between residuals and predictor.)
- Residuals can be plotted against predicted values to look for inhomogeneity of variance (heteroscedasticity). Look for residuals for which variability increases or “fans out” as one looks left-to-right in this plot (or vice-versa).
- The sorted residuals can be plotted against the normal inverse of the empirical CDF of the residuals in a normal plot to assess the normal distributional assumption. A nonlinear association in such a q-q plot indicates nonnormality. If data-rich, a histogram of residuals can also be used.

## Normal plots of residuals

- 1 Obtain the observed quantiles by ordering the residuals:

$$e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}.$$

- 2 For each  $i = 1, \dots, n$  compute the expected quantile from

$$q_{(i)} = z\left(1 - \frac{i}{n+1}\right).$$

- 3 Plot the (ordered) residuals on the vertical axis versus the (ordered) theoretical quantiles on the horizontal axis.

The *empirical cumulative probability* associated with  $e_{(i)}$  is

$$p_{(i)} = \frac{\text{Rank of } e_{(i)}}{n+1}.$$

Corresponding theoretical quantiles obtained via

$$q_{(i)} = z(1 - p_{(i)}).$$

e.g. suppose  $n = 9$  then for  $i = 1$  we look up the 10<sup>th</sup> percentile of  $N(0, 1)$  which is -1.282  
 ... for  $i = 9$  we look up  $q_{(9)} = +1.282$ . Plot the ordered residuals (empirical quantiles) against the theoretical quantiles and expect linearity.

```

ods listing close;
ods graphics on;
proc reg data=running;
    model pace=sexf age age2;    *general linear model.  will discuss soon;
    output out=resids p=yhat r=resid;
run;
proc rank data=resids out=resids2;
    ranks rankresid;
    var resid;
run;
data resids2;
    set resids2;
    ecdf=rankresid/(160+1);      *160 runners;
    q=probit(ecdf);
run;
ods listing ;
proc print data=resids2 ;
    var age pace yhat resid rankresid ecdf q;
run;
proc gplot data=resids;
    plot resid*q;
run;

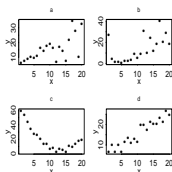
```

## The SAS System

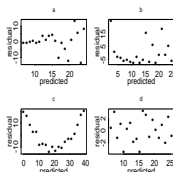
1

Obs	age	pace	yhat	resid	rankresid	ecdf	q
1	28	5.3833	7.5837	-2.20040	14.0	0.08696	-1.35974
2	39	5.4667	7.7671	-2.30046	10.0	0.06211	-1.53728
3	41	5.5167	7.8735	-2.35681	6.0	0.03727	-1.78332
4	42	5.6167	7.9351	-2.31841	9.0	0.05590	-1.59015
5	40	5.9333	7.8175	-1.88416	18.0	0.11180	-1.21700
(abbreviated)							
156	6	14.4667	11.4534	3.01324	155.0	0.96273	1.78332
157	52	15.1000	11.0579	4.04215	157.0	0.97516	1.96263
158	10	17.2667	10.9473	6.31937	158.5	0.98447	2.15636
159	10	17.2667	10.9473	6.31937	158.5	0.98447	2.15636
160	81	17.5000	14.7178	2.78223	152.0	0.94410	1.59015

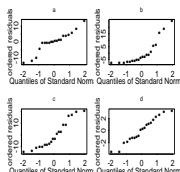
# A fun exercise: Match up letters a,b,c,d with the model violation



$y \text{ v } x$



resid v predicted ( $y \text{ v } \hat{y}$ )



normal plots of resids

- 1 Heteroscedasticity (nonconstant \_\_\_\_\_)
- 2 Nonlinearity ( $\mu(x)$  not linear in \_\_\_\_\_)
- 3 Nonnormality (vertical variation in  $y$  about  $\mu(x)$  not \_\_\_\_\_-shaped)
- 4 Model fits (hurray!)