STAT 500

Multiple Linear Regression Models
More Examples

Uncorrelated Predictors

Example: Yield of a chemical process (Myers)

Y = Yield (%)
X_1 = Temperature (^{o}F)
X_2 = Time (hours)

Data:

Y	X_1	X_2
77	160	1
79	160	2
82	165	1
83	165	2
85	170	1
88	170	2
90	175	1
93	175	2

Chemical Process Study

Full Factorial Design

$$r_{x_1,x_2} = \frac{\Sigma_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\Sigma_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \, \Sigma_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} = 0$$

Estimated Models

Model 1: $\hat{\mathbf{Y}}_i = \mathbf{-64.45 + 0.890}x_{i1}$

$R^2 = \mathbf{0.9435}$

Model 2: $\hat{\mathbf{Y}}_i = \mathbf{81.25 + 2.25}x_{i2}$

$R^2 = \mathbf{0.0482}$

Model 12: $\hat{\mathbf{Y}}_i = \mathbf{-67.825 + 0.890}x_{i1} + \mathbf{2.250}x_{i2}$

$R^2 = \mathbf{0.9918}$

Chemical Process Study

Source of variation	d.f.	SS	MS	F	p-val
reg on x_1	$\mathbf{1}$	$\mathbf{198.025}$	$\mathbf{198.025}$	574.0	.0001
reg on x_2 after x_1	$\mathbf{1}$	$\mathbf{10.125}$	$\mathbf{10.125}$	29.3	.0029
error	$\mathbf{5}$	$\mathbf{1.725}$	$\mathbf{0.345}$		
corrected total	$\mathbf{7}$	$\mathbf{209.875}$			

Source of variation	d.f.	SS	MS	F	p-val
reg on x_2	$\mathbf{1}$	$\mathbf{10.125}$	$\mathbf{10.125}$	29.3	.0029
reg on x_1 after x_2	$\mathbf{1}$	$\mathbf{198.025}$	$\mathbf{198.025}$	574.0	.0001
error	$\mathbf{5}$	$\mathbf{1.725}$	$\mathbf{0.345}$		
corrected total	$\mathbf{7}$	$\mathbf{209.875}$			

Complete Confounding

Example: Correlation between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ is one.

$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
1.95	1	5
6.25	2	10
9.85	3	15

Estimated Models

Model 1: $\hat{\mathbf{Y}}_i = -1.8833 + 3.95 x_{i1}$

$R^2 = 0.9974$

Model 2: $\hat{\mathbf{Y}}_i = -1.8833 + 0.79 x_{i2}$

$R^2 = 0.9974$

Model 12: Many choices for b_1 and b_2 in

$$\begin{aligned} \hat{\mathbf{Y}}_i &= b_0 + b_1 x_{i1} + b_2 x_{i2} = b_0 + b_1 x_{i1} + b_2(5x_{i1}) \\ &= b_0 + (b_1 + 5b_2) x_{i1} \end{aligned}$$

$R^2 = 0.9974$

Complete Confounding Example

Source of variation	d.f.	SS	MS	F	p-val
reg on x_1	$\mathbf{1}$	$\mathbf{31.205}$	$\mathbf{31.205}$	382.1	.0325
reg on x_2 after x_1	$\mathbf{0}$	$\mathbf{0.000}$	$\mathbf{0.000}$	NA	NA
error	$\mathbf{1}$	$\mathbf{0.08167}$	$\mathbf{0.08167}$		
corrected total	$\mathbf{2}$	$\mathbf{31.28667}$			

Source of variation	d.f.	SS	MS	F	p-val
reg on x_2	$\mathbf{1}$	$\mathbf{31.205}$	$\mathbf{31.205}$	382.1	.0325
reg on x_1 after x_2	$\mathbf{0}$	$\mathbf{0.000}$	$\mathbf{0.000}$	NA	NA
error	$\mathbf{1}$	$\mathbf{0.08167}$	$\mathbf{0.08167}$		
corrected total	$\mathbf{2}$	$\mathbf{31.28667}$			

Partial Confounding

Example: Correlation between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ is 0.95237

$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
1.8	1.0	5
1.7	1.1	6
5.4	1.8	11
6.1	2.0	10
7.0	2.1	9
9.6	3.0	15

Estimated Models

Model 1: $\hat{\mathbf{Y}}_i = \mathbf{-2.328 + 4.142}\boldsymbol{x_{i1}}$

$\boldsymbol{R^2 = 0.978}$

Model 2: $\hat{\mathbf{Y}}_i = \mathbf{-2.114 + 0.791}\boldsymbol{x_{i2}}$

$\boldsymbol{R^2 = 0.865}$

Model 12: $\hat{\mathbf{Y}}_i = \mathbf{-2.247 + 4.655}\boldsymbol{x_{i1}} \mathbf{- 0.109}\boldsymbol{x_{i2}}$

$\boldsymbol{R^2 = 0.980}$

Partial Confounding Example

Source of variation	d.f.	SS	MS	F	p-val
reg on x_1	$\mathbf{1}$	$\mathbf{46.215}$	$\mathbf{46.215}$	146.6	0.0012
reg on x_2 after x_1	$\mathbf{1}$	$\mathbf{0.073}$	$\mathbf{0.073}$	0.23	0.6639
error	$\mathbf{3}$	$\mathbf{0.946}$	$\mathbf{0.315}$		
corrected total	$\mathbf{5}$	$\mathbf{47.233}$			

Source of variation	d.f.	SS	MS	F	p-val
reg on x_2	$\mathbf{1}$	$\mathbf{40.859}$	$\mathbf{40.859}$	129.6	0.0015
reg on x_1 after x_2	$\mathbf{1}$	$\mathbf{5.428}$	$\mathbf{5.428}$	17.21	0.0254
error	$\mathbf{3}$	$\mathbf{0.946}$	$\mathbf{0.315}$		
corrected total	$\mathbf{5}$	$\mathbf{47.233}$			

Multiple Regression
Interpreting Regression Coefficients

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- β_j is the jth regression coefficient or the jth *partial* regression coefficient
- β_j is the change in the mean of Y for a unit change in X_j **with all other variables held constant**
- Sometimes this is not possible and the values of other explanatory variables change when X_j changes: (e.g., polynomial terms (X_j, X_j^2) or interaction terms (X_i, X_j, X_iX_j) or other highly correlated predictors)

Multiple Regression
Interpreting Regression Coefficients

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- An alternative interpretation: β_j is the linear effect of X_j on Y after adjusting for the linear effect of the other predictors on Y and the linear effects of the other predictors on X_j

- Let P_{-x_j} represent the projection matrix without variable X_j (delete column $j+1$ from the model matrix X). Then, $\hat{\beta}_j$ is found from the regression of $(I - P_{-x_j})Y$ on $(I - P_{-x_j})X_j$

Multiple Regression
Interpreting Regression Coefficients

- Example: Brain size data (an observational study)
  Question: Do species with longer gestation times have bigger brains?

- Plots, biology $\Rightarrow$ linear in log variables

- Model 1: $\boldsymbol{log(brain)_i = \beta_0 + \beta_1 log(gest_i) + \epsilon_i}$
  $\hat{\boldsymbol{\beta}}_1 = \mathbf{2.23} \Rightarrow$ Species differing by 1 unit log gestation time (e.g. log(gest) = 2 and log(gest)=1) differ in log(brain size) by 2.23 units, on average.

- Biology $\Rightarrow$ body size associated with both

Multiple Regression
Interpreting Regression Coefficients

- Model 2:

  $$log(brain_i) = \beta_0 + \beta_1 log(gest_i) + \beta_2 log(body_i) + \epsilon_i$$

  $$\hat{\beta}_1 = 0.668$$

  Two species with the same body size but differing by 1 unit log gestation time differ in log brain size by 0.668 units, on average.

- So, when is β_j in multiple regression equal to β_j from simple linear regression?

  Answer: When X_j is uncorrelated with the rest of the explanatory variables.

Multiple Regression
Interpreting Regression Coefficients

- Consider the regression of one set of residuals $(\boldsymbol{I} - \boldsymbol{P}_{-x_j})\boldsymbol{Y}$ on another set of residuals $(\boldsymbol{I} - \boldsymbol{P}_{-x_j})\boldsymbol{X_j}$
  - Regress log(brain) on log(body):
    residual $= e_i = (\boldsymbol{I} - \boldsymbol{P}_{-x_j})\boldsymbol{Y}$
  - Regress log(gest) on log(body):
    residual$= g_i = (\boldsymbol{I} - \boldsymbol{P}_{-x_j})\boldsymbol{X_j}$
  - β_2 is regression coefficient for regression of e_i on g_i:

$$e_i = \beta_2 g_i + \eta_i$$