

**Week 11: Model construction Case study: Constructing models**

For each of these four situations, construct an appropriate model, describe any additional X variables, and indicate how you would estimate (or test) the quantities of interest.

Warranty claims on cars. A major car manufacturer is interested in modeling the number of warranty CLAIMS per month for two different car TYPEs, A and B, as a function of TIME, the number months since the type was introduced.

1. There are no claims in month 0, but then claims increase linearly with time, perhaps at a different rate for the two car types. The quantity of interest is the difference in claims (between car types) in month 12.
2. 1 There is an initial spike in claims in month 1. After that, claims increase linearly. The quantities of interest are: a) the difference in slopes from month 1 to 12, and b) the difference in claims in month 12.

Corn yield. Farmers add nitrogen fertilizer, N, to corn to increase the YIELD. Corn produces grain without added N, but adding small or moderate amounts of N increases the yield. Large amounts of N provide no additional benefit.

3. Assume that the yield response is approximately linear from 0 lb N /acre to 100 lb N /acre. Above 100 lb N/acre, the yield is constant. You wish to estimate the yield at 100 lb N/acre and the N response (increase in yield for each additional lb N/acre) when less than 100 lb N/acre is added.
4. The linear curve described above is a simplification because most biological processes don't have sharp breaks. The common model for yield is a quadratic model with a maximum at  $N=100$  for  $N < 100$ . You want to model yield using only N values  $< 100$ . You want to estimate the difference in yield between 50 lb N/acre and 100 lb N/acre.

## bacillus2.sas

```
options ls=75 formdlim='- ' nonumber nodate;

data bacillus;
  infile 'bacillus2.txt' ;
  input trt $ pre post;

  /* reference group coding: Placebo is the ref. group */
  if trt = 'Ab1' then a1 = 1;
  else a1 = 0;

  /* a short cut way to the same thing */
  a2 = (trt = 'Ab2' );
  a3 = (trt = 'Pl' ); /* to show what happens */

  /* effects coding */
  b1 = (trt = 'Ab1' );
  b2 = (trt = 'Ab2' );
  if trt = 'Pl' then do;
    b1 = -1;
    b2 = -1;
  end;
run;

proc print;
  var trt a1-a3 b1-b2;
  title 'Leprosy study: indicator variables';
run;
```

```
proc glm;
  class trt;
  model post = trt /solution;
  lsmeans trt /stderr;
  title 'GLM solution';
run;
```

```
proc glm;
  model post = a1 a2;
  estimate 'Ab1 mean' intercept 1 a1 1;
  estimate 'Ab2 mean' intercept 1 a2 1;
  estimate 'Pl mean' intercept 1;
  title 'Indicator variable regression / reference group coding';
run;
```

```
proc glm;
  model post = b1 b2;
  estimate 'Ab1 mean' intercept 1 b1 1;
  estimate 'Ab2 mean' intercept 1 b2 1;
  estimate 'Pl mean' intercept 1 b1 -1 b2 -1;
  title 'Indicator variable regression / effects group coding';
run;
```

```
proc reg;
  model post = a1 a2 a3 /noint;
  title 'Indicator variable regression, cell means coding';
run;
```

```
proc reg;
```

```
model post = a1 a2 a3;  
title 'Indicator variable regression, overparameterized model';  
run;
```

## bacillus2.lst

## Leprosy study: indicator variables

| Obs | trt | a1 | a2 | a3 | b1 | b2 |
|-----|-----|----|----|----|----|----|
| 1   | trt | 0  | 0  | 0  | 0  | 0  |
| 2   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 3   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 4   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 5   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 6   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 7   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 8   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 9   | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 10  | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 11  | Ab1 | 1  | 0  | 0  | 1  | 0  |
| 12  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 13  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 14  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 15  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 16  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 17  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 18  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 19  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 20  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 21  | Ab2 | 0  | 1  | 0  | 0  | 1  |
| 22  | P1  | 0  | 0  | 1  | -1 | -1 |
| 23  | P1  | 0  | 0  | 1  | -1 | -1 |
| 24  | P1  | 0  | 0  | 1  | -1 | -1 |

|    |    |   |   |   |    |    |
|----|----|---|---|---|----|----|
| 25 | P1 | 0 | 0 | 1 | -1 | -1 |
| 26 | P1 | 0 | 0 | 1 | -1 | -1 |
| 27 | P1 | 0 | 0 | 1 | -1 | -1 |
| 28 | P1 | 0 | 0 | 1 | -1 | -1 |
| 29 | P1 | 0 | 0 | 1 | -1 | -1 |
| 30 | P1 | 0 | 0 | 1 | -1 | -1 |
| 31 | P1 | 0 | 0 | 1 | -1 | -1 |

---

GLM solution

The GLM Procedure

Class Level Information

| Class | Levels | Values         |
|-------|--------|----------------|
| trt   | 4      | Ab1 Ab2 P1 trt |

|                             |    |
|-----------------------------|----|
| Number of Observations Read | 31 |
| Number of Observations Used | 30 |

---

GLM solution

The GLM Procedure

Dependent Variable: post

| Source          | DF | Sum of Squares | Mean Square | F Value |
|-----------------|----|----------------|-------------|---------|
| Model           | 2  | 293.600000     | 146.800000  | 3.9     |
| Error           | 27 | 995.100000     | 36.855556   |         |
| Corrected Total | 29 | 1288.700000    |             |         |

| R-Square | Coeff Var | Root MSE | post Mean |
|----------|-----------|----------|-----------|
| 0.227826 | 76.84655  | 6.070878 | 7.900000  |

| Source | DF | Type I SS   | Mean Square | F Value |
|--------|----|-------------|-------------|---------|
| trt    | 2  | 293.6000000 | 146.8000000 | 3.9     |

| Source | DF | Type III SS | Mean Square | F Value |
|--------|----|-------------|-------------|---------|
| trt    | 2  | 293.6000000 | 146.8000000 | 3.9     |

| Parameter | Estimate | Standard Error | t Value |
|-----------|----------|----------------|---------|
|-----------|----------|----------------|---------|

|           |     |             |   |            |       |
|-----------|-----|-------------|---|------------|-------|
| Intercept |     | 12.30000000 | B | 1.91978008 | 6.41  |
| trt       | Ab1 | -7.00000000 | B | 2.71497903 | -2.58 |
| trt       | Ab2 | -6.20000000 | B | 2.71497903 | -2.28 |
| trt       | P1  | 0.00000000  | B | .          | .     |

NOTE: The X'X matrix has been found to be singular, and a general inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely e

---

GLM solution

The GLM Procedure  
Least Squares Means

| trt | post LSMEAN | Standard Error | Pr >  t |
|-----|-------------|----------------|---------|
| Ab1 | 5.30000000  | 1.9197801      | 0.0102  |
| Ab2 | 6.10000000  | 1.9197801      | 0.0037  |
| P1  | 12.30000000 | 1.9197801      | <.0001  |

---

Indicator variable regression / reference group coding

The GLM Procedure

Number of Observations Read 31

Number of Observations Used 30

-----  
 Indicator variable regression / reference group coding

The GLM Procedure

Dependent Variable: post

| Source          | DF | Sum of Squares | Mean Square | F Value |
|-----------------|----|----------------|-------------|---------|
| Model           | 2  | 293.600000     | 146.800000  | 3.9     |
| Error           | 27 | 995.100000     | 36.855556   |         |
| Corrected Total | 29 | 1288.700000    |             |         |

| R-Square | Coeff Var | Root MSE | post Mean |
|----------|-----------|----------|-----------|
| 0.227826 | 76.84655  | 6.070878 | 7.900000  |

| Source | DF | Type I SS  | Mean Square | F Value |
|--------|----|------------|-------------|---------|
| a1     | 1  | 101.400000 | 101.400000  | 2.7     |
| a2     | 1  | 192.200000 | 192.200000  | 5.2     |

| Source | DF | Type III SS | Mean Square | F Value |
|--------|----|-------------|-------------|---------|
| a1     | 1  | 245.0000000 | 245.0000000 | 6.6     |
| a2     | 1  | 192.2000000 | 192.2000000 | 5.2     |

| Parameter | Estimate   | Standard Error | t Value |
|-----------|------------|----------------|---------|
| Ab1 mean  | 5.3000000  | 1.91978008     | 2.76    |
| Ab2 mean  | 6.1000000  | 1.91978008     | 3.18    |
| Pl mean   | 12.3000000 | 1.91978008     | 6.41    |

| Parameter | Estimate    | Standard Error | t Value | Pr |
|-----------|-------------|----------------|---------|----|
| Intercept | 12.30000000 | 1.91978008     | 6.41    | <  |
| a1        | -7.00000000 | 2.71497903     | -2.58   | 0  |
| a2        | -6.20000000 | 2.71497903     | -2.28   | 0  |

Indicator variable regression / effects group coding

The GLM Procedure

Number of Observations Read 31  
 Number of Observations Used 30

-----

Indicator variable regression / effects group coding

The GLM Procedure

Dependent Variable: post

| Source          | DF | Sum of Squares | Mean Square | F Value |
|-----------------|----|----------------|-------------|---------|
| Model           | 2  | 293.600000     | 146.800000  | 3.9     |
| Error           | 27 | 995.100000     | 36.855556   |         |
| Corrected Total | 29 | 1288.700000    |             |         |

| R-Square | Coeff Var | Root MSE | post Mean |
|----------|-----------|----------|-----------|
| 0.227826 | 76.84655  | 6.070878 | 7.900000  |

| Source | DF | Type I SS  | Mean Square | F Value |
|--------|----|------------|-------------|---------|
| b1     | 1  | 245.000000 | 245.000000  | 6.6     |
| b2     | 1  | 48.600000  | 48.600000   | 1.3     |

| Source | DF | Type III SS | Mean Square | F Value |
|--------|----|-------------|-------------|---------|
| b1     | 1  | 101.4000000 | 101.4000000 | 2.7     |
| b2     | 1  | 48.6000000  | 48.6000000  | 1.3     |

  

| Parameter | Estimate   | Standard Error | t Value |
|-----------|------------|----------------|---------|
| Ab1 mean  | 5.3000000  | 1.91978008     | 2.76    |
| Ab2 mean  | 6.1000000  | 1.91978008     | 3.18    |
| Pl mean   | 12.3000000 | 1.91978008     | 6.41    |

| Parameter | Estimate     | Standard Error | t Value | Pr |
|-----------|--------------|----------------|---------|----|
| Intercept | 7.900000000  | 1.10838555     | 7.13    | <  |
| b1        | -2.600000000 | 1.56749387     | -1.66   | 0  |
| b2        | -1.800000000 | 1.56749387     | -1.15   | 0  |

-----

Indicator variable regression, cell means coding

The REG Procedure

Model: MODEL1

Dependent Variable: post

Number of Observations Read

31

|  |    |
|--|----|
| Number of Observations Used                | 30 |
| Number of Observations with Missing Values | 1  |

NOTE: No intercept in model. R-Square is redefined.

### Analysis of Variance

| Source            | DF | Sum of Squares | Mean Square | F Value |
|-------------------|----|----------------|-------------|---------|
| Model             | 3  | 2165.90000     | 721.96667   | 19.59   |
| Error             | 27 | 995.10000      | 36.85556    |         |
| Uncorrected Total | 30 | 3161.00000     |             |         |

|                |          |          |        |
|----------------|----------|----------|--------|
| Root MSE       | 6.07088  | R-Square | 0.6852 |
| Dependent Mean | 7.90000  | Adj R-Sq | 0.6502 |
| Coeff Var      | 76.84655 |          |        |

### Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr |
|----------|----|--------------------|----------------|---------|----|
| a1       | 1  | 5.30000            | 1.91978        | 2.76    |    |
| a2       | 1  | 6.10000            | 1.91978        | 3.18    |    |
| a3       | 1  | 12.30000           | 1.91978        | 6.41    |    |

Indicator variable regression, overparameterized model

The REG Procedure

Model: MODEL1

Dependent Variable: post

|  |    |
|--|----|
| Number of Observations Read                | 31 |
| Number of Observations Used                | 30 |
| Number of Observations with Missing Values | 1  |

Analysis of Variance

| Source          | DF | Sum of Squares | Mean Square | F Value |
|-----------------|----|----------------|-------------|---------|
| Model           | 2  | 293.60000      | 146.80000   | 3.98    |
| Error           | 27 | 995.10000      | 36.85556    |         |
| Corrected Total | 29 | 1288.70000     |             |         |

|                |          |          |        |
|----------------|----------|----------|--------|
| Root MSE       | 6.07088  | R-Square | 0.2278 |
| Dependent Mean | 7.90000  | Adj R-Sq | 0.1706 |
| Coeff Var      | 76.84655 |          |        |

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A report of 0 or B means that the estimate is biased.

NOTE: The following parameters have been set to 0, since the variable is a linear combination of other variables as shown.

$$a3 = \text{Intercept} - a1 - a2$$

## Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr |
|-----------|----|--------------------|----------------|---------|----|
| Intercept | B  | 12.30000           | 1.91978        | 6.41    |    |
| a1        | B  | -7.00000           | 2.71498        | -2.58   |    |
| a2        | B  | -6.20000           | 2.71498        | -2.28   |    |
| a3        | 0  | 0                  | .              | .       |    |