

Multivariate Regression

Often what we're trying to model will depend on more than one variable. (The following example is taken from Prof. Howard's 442 notes.)

Example: Suppose that we are trying to estimate movie ticket sales. We will consider the problem of trying to estimate final sales based on first weekend sales. The first difficulty which we encounter is that first weekend sales often depend more on hype than quality. For example, *Silence of the Lambs* and *Dude, Where's My Car?* had similar first-weekend sales (\$13,766,814 and \$13,845,914 respectively), but the final sales weren't close: \$130,726,716 and \$46,729,374 again respectively. Somehow our model has to include more data. Maybe the quality of the movie has something to do with it? (What a thought!) We can estimate quality by critical ratings (e.g., number of stars a critic gives). Here is a table which Prof. Howard compiled, using ratings from TV Guide.

Movie	First weekend sales	Final sales	TV Guide rating
Dude, Where's My Car?	13,845,914	46,729,374	1.5
Silence of the Lambs	13,766,814	130,726,716	4.5
We Were Soldiers	20,212,543	78,120,196	2.5
Ace Ventura	12,115,105	72,217,396	3.0
Rocky V	14,073,170	40,113,407	3.5
A.I.	29,352,630	78,579,202	3.0
Moulin Rouge	13,718,306	57,386,369	2.5
A Beautiful Mind	16,655,820	170,708,996	3.0
The Wedding Singer	18,865,080	80,245,725	3.0
Zoolander	15,525,043	45,172,250	2.5

Let S represent first weekend sales, F represent final sales, and R represent ratings. Our first model will be to assume that $F = a + bS + cR$, with a , b , and c to be determined using least squares. In other words,

we seek the values which will minimize $E(a, b, c) = \sum_{i=1}^n (F_i - a - bS_i - cR_i)^2$. Let's enter these as

vectors.

```
> S := [13845914, 13766814, 20212543, 12115105, 14073170, 29352630, 13718306, 16655820,
18865080, 15525043];
S := [13845914, 13766814, 20212543, 12115105, 14073170, 29352630, 13718306, 16655820,
18865080, 15525043] (1)
```

```
> R := [1.5, 4.5, 2.5, 3.0, 3.5, 3.0, 2.5, 3.0, 3.0, 2.5];
R := [1.5, 4.5, 2.5, 3.0, 3.5, 3.0, 2.5, 3.0, 3.0, 2.5] (2)
```

```
> F := [46729374, 130726716, 78120196, 72217396, 40113407, 78579202, 57386369,
170708996, 80245725, 45172250];
F := [46729374, 130726716, 78120196, 72217396, 40113407, 78579202, 57386369,
170708996, 80245725, 45172250] (3)
```

170708996, 80245725, 45172250]

$$> E := \sum_{i=1}^{10} (F_i - a - b \cdot S_i - c \cdot R_i)^2;$$

$$E := (46729374 - a - 13845914 b - 1.5 c)^2 + (130726716 - a - 13766814 b - 4.5 c)^2 + (78120196 - a - 20212543 b - 2.5 c)^2 + (72217396 - a - 12115105 b - 3.0 c)^2 + (40113407 - a - 14073170 b - 3.5 c)^2 + (78579202 - a - 29352630 b - 3.0 c)^2 + (57386369 - a - 13718306 b - 2.5 c)^2 + (170708996 - a - 16655820 b - 3.0 c)^2 + (80245725 - a - 18865080 b - 3.0 c)^2 + (45172250 - a - 15525043 b - 2.5 c)^2 \quad (4)$$

$$> eq1 := \text{diff}(E, a) = 0; \\ eq1 := 20 a + 336260850 b + 58.0 c - 1599999262 = 0 \quad (5)$$

$$> eq2 := \text{diff}(E, b) = 0; \\ eq2 := -27234689394449004 + 336260850 a + 6117429261199902 b + 9.731625280 \cdot 10^8 c = 0 \quad (6)$$

$$> eq3 := \text{diff}(E, c) = 0; \\ eq3 := 58.0 a + 9.731625280 \cdot 10^8 b + 179.00 c - 4.911424404 \cdot 10^9 = 0 \quad (7)$$

$$> sol := \text{solve}(\{eq1, eq2, eq3\}, \{a, b, c\}); \\ sol := \{b = 0.8283738049, c = 2.528502494 \cdot 10^7, a = -7.254093221 \cdot 10^6\} \quad (8)$$

Let's see how well this works for *Moulin Rouge*, as an example. Take the values from the table and the coefficients which we found.

$$> \text{subs}(sol, a + b \cdot 13718306 + c \cdot 2.5); \\ 6.732235447 \cdot 10^7 \quad (9)$$

which is not bad: it's 67.3 million versus the actual 57.3 million.

A more sophisticated model might include the product *SR*. Let's see if that helps.

$$> EI := \sum_{i=1}^{10} (F_i - d - e \cdot S_i - f \cdot R_i - (g \cdot S_i) \cdot R_i)^2;$$

$$EI := (46729374 - d - 13845914 e - 1.5 f - 2.07688710 \cdot 10^7 g)^2 + (130726716 - d - 13766814 e - 4.5 f - 6.19506630 \cdot 10^7 g)^2 + (78120196 - d - 20212543 e - 2.5 f - 5.05313575 \cdot 10^7 g)^2 + (72217396 - d - 12115105 e - 3.0 f - 3.63453150 \cdot 10^7 g)^2 + (40113407 - d - 14073170 e - 3.5 f - 4.92560950 \cdot 10^7 g)^2 + (78579202 - d - 29352630 e - 3.0 f - 8.80578900 \cdot 10^7 g)^2 + (57386369 - d - 13718306 e - 2.5 f - 3.42957650 \cdot 10^7 g)^2 + (170708996 - d - 16655820 e - 3.0 f - 4.99674600 \cdot 10^7 g)^2 + (80245725 - d - 18865080 e - 3.0 f - 5.65952400 \cdot 10^7 g)^2 + (45172250 - d - 15525043 e - 2.5 f - 3.88126075 \cdot 10^7 g)^2 \quad (10)$$

$$> eq1 := \text{diff}(EI, d) = 0; \\ eq1 := 336260850 e + 58.0 f + 9.731625280 \cdot 10^8 g + 20 d - 1599999262 = 0 \quad (11)$$

$$> eq2 := \text{diff}(EI, e) = 0; \\ eq2 := -27234689394449004 + 6117429261199902 e + 9.731625280 \cdot 10^8 f \quad (12)$$

$$+ 1.770602362 \cdot 10^{16} g + 336260850 d = 0$$

$$\begin{aligned} &> \text{eq3} := \text{diff}(E1, f) = 0; \\ \text{eq3} &:= 9.731625280 \cdot 10^8 e + 179.00 f + 2.968649326 \cdot 10^9 g + 58.0 d - 4.911424404 \cdot 10^9 = 0 \end{aligned} \quad (13)$$

$$\begin{aligned} &> \text{eq4} := \text{diff}(E1, g) = 0; \\ \text{eq4} &:= 1.770602362 \cdot 10^{16} e + 2.968649325 \cdot 10^9 f + 5.341274358 \cdot 10^{16} g + 9.731625280 \cdot 10^8 d \\ &\quad - 8.265907673 \cdot 10^{16} = 0 \end{aligned} \quad (14)$$

$$\begin{aligned} &> \text{sol1} := \text{solve}(\{\text{eq1}, \text{eq2}, \text{eq3}, \text{eq4}\}, \{d, e, f, g\}); \\ \text{sol1} &:= \{d = -3.464250518 \cdot 10^7, g = -.6521158975, e = 2.757657921, f = 3.448571939 \cdot 10^7\} \end{aligned} \quad (15)$$

See how this works for *Moulin Rouge* :

$$\begin{aligned} &> \text{subs}(\text{sol1}, d + e \cdot S_7 + f \cdot R_7 + (g \cdot S_7) \cdot R_7); \\ &\quad \quad \quad 6.703737493 \cdot 10^7 \end{aligned} \quad (16)$$

A bit better. Now, as a challenge, let's see how these formulas work on the legendary stinker "Gigli". This had a rating of 1.0 from TV Guide, and initial weekend sales of \$3,753,518.

$$\begin{aligned} &> \text{subs}(\text{sol}, a + b \cdot 3753518 + c \cdot 1.0); \\ &\quad \quad \quad 2.114024771 \cdot 10^7 \end{aligned} \quad (17)$$

$$\begin{aligned} &> \text{subs}(\text{sol1}, d + e \cdot 3753518 + f \cdot 1.0 + g \cdot 3753518 \cdot 1.0); \\ &\quad \quad \quad 7.746404091 \cdot 10^6 \end{aligned} \quad (18)$$

The actual final sales were \$6,087,542. Neither prediction was great, but certainly the second model seems closer.