

Further Topics in Regression

Dr. Lee Fawcett



Quadratic Regression

Illustrative example

In analytical chemistry practice, signal-concentration plots are often used where a concentration of analysed samples is proportional to a respective analytical device's signal, for example absorbance, current, potential, peak height and peak surface.

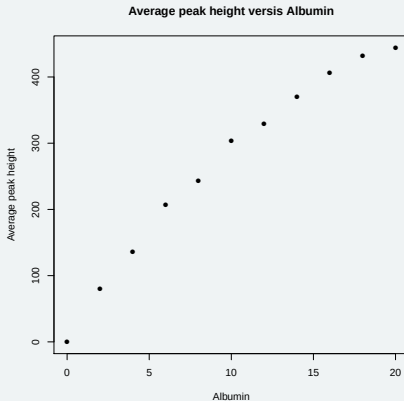
The spectrophotometric method was used to measure signal peak heights for known concentrations of albumin ($\mu\text{g mL}^{-1}$).

The data in the table overleaf show the results from the experiment; a scatterplot of these data is also shown.

Illustrative example

Albumin ($X \mu\text{g mL}^{-1}$)	0	2	4	6	8	10
Average peak height (Y)	0	80.33	135.7	207.0	243.3	303.0

Albumin ($X \mu\text{g mL}^{-1}$)	12	14	16	18	20
Average peak height (Y)	329.0	370.3	406.0	431.7	444.3



Illustrative example

- In this example, a **curved relationship** might be preferable to a linear one
- We could try a **quadratic association**, given generically by

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,$$

where once again $\epsilon \sim N(0, \sigma^2)$

Illustrative example

- In this example, a **curved relationship** might be preferable to a linear one
- We could try a **quadratic association**, given generically by

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,$$

where once again $\epsilon \sim N(0, \sigma^2)$

The linear model $Y = \beta_0 + \beta_1 X$ is **nested** within the quadratic model; for example, consider the reduction of the quadratic model when $\beta_2 = 0$

Illustrative example

```
1 > fit1 = lm(peak.height ~ albumin)
2 > summary(fit1)
3
4 Call:
5 lm(formula = peak.height ~ albumin)
6
7 Residuals:
8     Min       1Q   Median       3Q      Max
9  -47.776  -12.224    5.483   17.911   34.761
10
11 Coefficients:
12             Estimate Std. Error t value Pr(>|t|)
13 (Intercept)   47.776     16.126   2.963   0.0159 *
14 albumin       22.046      1.363  16.176 5.85e-08 ***
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
17
18 Residual standard error: 28.59 on 9 degrees of freedom
19 Multiple R-squared:  0.9667, Adjusted R-squared:  0.9631
20 F-statistic: 261.7 on 1 and 9 DF, p-value: 5.847e-08
```

Illustrative example

```
1 > albumin.sq = albumin^2
2 > fit2 = lm(peak.height ~ albumin+albumin.sq)
3 > summary(fit2)
4
5 Call:
6 lm(formula = peak.height ~ albumin + albumin.sq)
7
8 Residuals:
9     Min       1Q   Median       3Q      Max
10 -9.026 -4.103 -1.578  4.905  9.816
11
12 Coefficients:
13             Estimate Std. Error t value Pr(>|t|)
14 (Intercept)  4.95259     5.12829   0.966   0.362
15 albumin     36.32086     1.19299  30.445 1.47e-09 ***
16 albumin.sq  -0.71373     0.05745 -12.423 1.65e-06 ***
17 ---
18 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19
20 Residual standard error: 6.731 on 8 degrees of freedom
21 Multiple R-squared:  0.9984, Adjusted R-squared:  0.998
22 F-statistic: 2437 on 2 and 8 DF, p-value: 7.21e-12
```

Illustrative example

- The fitted quadratic model is

$$Y = 4.953 + 36.321X - 0.714X^2 + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$

Illustrative example

- The fitted quadratic model is

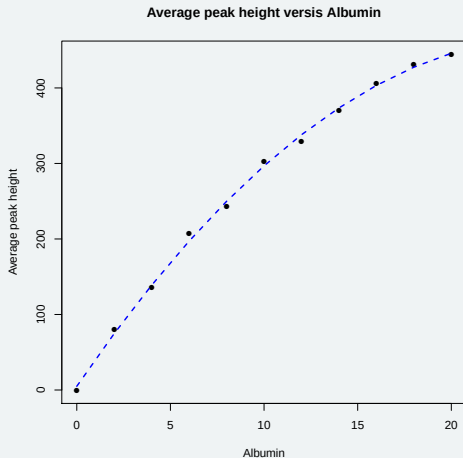
$$Y = 4.953 + 36.321X - 0.714X^2 + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$

- Both the linear and quadratic terms are **statistically significant** (both have coefficients with $p < 0.01\%$)
- The R^2 statistic is 99.8% – an improvement over that from the linear model (96.3%)

Illustrative example

```
1 > lines(fitted(fit2)~albumin, col="blue", lwd=2, lty=2)
```



Illustrative example

What might we expect the average peak height to be, for an albumin concentration of $7\mu\text{g mL}^{-1}$?

$$\begin{aligned}y &= 4.95259 + 36.32086 \cdot 7 - 0.71373 \cdot 7^2 \\&= 224.2258.\end{aligned}$$

Binary Logistic Regression

Illustrative example

The **Leukaemia Remission dataset** (Gal *et al.*, 2019) has a **binary response variable** (REMISS, Y) taking the value 1 if leukaemia remission occurred in a patient and zero otherwise.

Illustrative example

The **Leukaemia Remission dataset** (Gal *et al.*, 2019) has a **binary response variable** (REMISS, Y) taking the value 1 if leukaemia remission occurred in a patient and zero otherwise.

The **predictor variables** are:

1. Cellularity of the marrow clot section (CELL, X_1)
2. Smear differential percentage of blasts (SMEAR, X_2)
3. Percentage of absolute marrow leukaemia cell infiltrate (INFIL, X_3)
4. Percentage labelling index of the bone marrow leukaemia cells (LI, X_4)
5. Absolute number of blasts in the peripheral blood (BLAST, X_5)
6. The highest temperature prior to start of treatment (TEMP, X_6)

Illustrative example

REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP
1	0.8	0.83	0.66	1.9	1.1	1
1	0.9	0.36	0.32	1.4	0.74	0.99
0	0.8	0.88	0.7	0.8	0.18	0.98
0	1	0.87	0.87	0.7	1.05	0.99
1	0.9	0.75	0.68	1.3	0.52	0.98
0	1	0.65	0.65	0.6	0.52	0.98
1	0.95	0.97	0.92	1	1.23	0.99
0	0.95	0.87	0.83	1.9	1.35	1.02
0	1	0.45	0.45	0.8	0.32	1
0	0.95	0.36	0.34	0.5	0	1.04
0	0.85	0.39	0.33	0.7	0.28	0.99
0	0.7	0.76	0.53	1.2	0.15	0.98
0	0.8	0.46	0.37	0.4	0.38	1.01
0	0.2	0.39	0.08	0.8	0.11	0.99
0	1	0.9	0.9	1.1	1.04	0.99
1	1	0.84	0.84	1.9	2.06	1.02
0	0.65	0.42	0.27	0.5	0.11	1.01
0	1	0.75	0.75	1	1.32	1
0	0.5	0.44	0.22	0.6	0.11	0.99
1	1	0.63	0.63	1.1	1.07	0.99
0	1	0.33	0.33	0.4	0.18	1.01
0	0.9	0.93	0.84	0.6	1.59	1.02
1	1	0.58	0.58	1	0.53	1
0	0.95	0.32	0.3	1.6	0.89	0.99
1	1	0.6	0.6	1.7	0.96	0.99
1	1	0.69	0.69	0.9	0.4	0.99
0	1	0.73	0.73	0.7	0.4	0.99

Illustrative example

```
1 > leukaemia = read.table("https://www.mas.ncl.ac.uk/~nlf8/sellafield/leukaemia.txt")
2 > attach(leukaemia)
3 > fit3 = glm(REMISS~CELL+SMEAR+INFIL+LI+BLAST+TEMP, family = binomial)
4 > summary(fit3)
5
6 Call:
7 glm(formula = REMISS ~ CELL + SMEAR + INFIL + LI + BLAST + TEMP,
8     family = binomial)
9
10 Deviance Residuals:
11      Min       1Q   Median       3Q      Max
12 -1.95404  -0.66259  -0.02516   0.78184   1.57465
13
14 Coefficients:
15             Estimate Std. Error z value Pr(>|z|)
16 (Intercept)   64.25808    74.96480   0.857   0.391
17 CELL         30.83006    52.13520   0.591   0.554
18 SMEAR        24.68632    61.52601   0.401   0.688
19 INFIL       -24.97447    65.28088  -0.383   0.702
20 LI           4.36045     2.65798   1.641   0.101
21 BLAST        -0.01153     2.26634  -0.005   0.996
22 TEMP       -100.17340    77.75289  -1.288   0.198
23
24 (Dispersion parameter for binomial family taken to be 1)
25
26 Null deviance: 34.372 on 26 degrees of freedom
27 Residual deviance: 21.594 on 20 degrees of freedom
28 AIC: 35.594
29
30 Number of Fisher Scoring iterations: 8
```

Illustrative example

```
1 > fit4 = glm(REMISS ~ LI, family = binomial)
2 > summary(fit4)
3
4 Call:
5 glm(formula = REMISS ~ LI, family = binomial)
6
7 Deviance Residuals:
8     Min       1Q   Median       3Q      Max
9  -1.9448  -0.6465  -0.4947   0.6571   1.6971
10
11 Coefficients:
12             Estimate Std. Error z value Pr(>|z|)
13 (Intercept)  -3.777      1.379  -2.740  0.00615 **
14 LI           2.897      1.187   2.441  0.01464 *
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
17
18 (Dispersion parameter for binomial family taken to be 1)
19
20 Null deviance: 34.372  on 26  degrees of freedom
21 Residual deviance: 26.073  on 25  degrees of freedom
22 AIC: 30.073
23
24 Number of Fisher Scoring iterations: 4
```

Illustrative example

- A binary logistic regression ensures the model returns values between 0 and 1; not on the same scale as the response variable, but on a restricted scale
- In fact, the values returned from a binary logistic regression model are **probabilities**
- Specifically, the model form in our leukaemia example is

$$P(Y = 1) = \frac{e^{-3.777+2.897X_4}}{1 + e^{-3.778+2.897X_4}}.$$

Illustrative example

- A binary logistic regression ensures the model returns values between 0 and 1; not on the same scale as the response variable, but on a restricted scale
- In fact, the values returned from a binary logistic regression model are **probabilities**
- Specifically, the model form in our leukaemia example is

$$P(Y = 1) = \frac{e^{-3.777+2.897X_4}}{1 + e^{-3.778+2.897X_4}}.$$

- Notice that the X_4 coefficient is significant, since its p -value is 1.46%

Illustrative example

What is the chance of remission in a patient with an LI value of 0.85?

$$P(Y = 1) = \frac{e^{-3.777+2.897 \cdot 0.85}}{1 + e^{-3.777+2.897 \cdot 0.85}} = 0.212,$$

or 21.2%.

Illustrative example

```
1 > plot(LI, REMISS, pch=19,lwd=3)
2 > curve(exp(fit4$coef[1]+fit4$coef[2]*x)/(1+exp(fit4$coef[1]+fit4$coef[2]*x)), from=
    min(LI), to = max(LI), col="blue", lty=2, lwd=2,add=TRUE)
```

