

Topic 8: Further Topics in Regression

Dr. Lee Fawcett

Friday, September 24th, 2021

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.

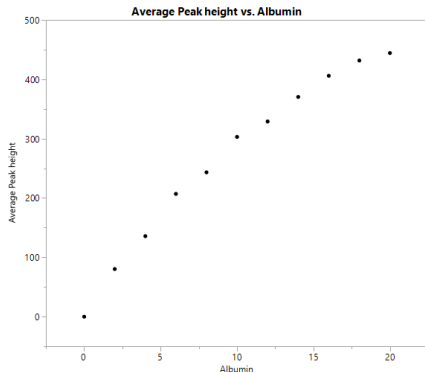
The spectrophotometric method was used to measure signal peak heights for known concentrations of albumin ($\mu\text{g mL}^{-1}$).

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

$$Y = \beta_0 + \beta_1 X + \beta_2 (X - a)^2 + \epsilon,$$

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

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	76.325548	4.43826	17.20	<.0001*
Albumin	22.046273	0.320904	68.70	<.0001*
(Albumin-10)*(Albumin-10)	-0.71373	0.057451	-12.42	<.0001*

Illustrative example

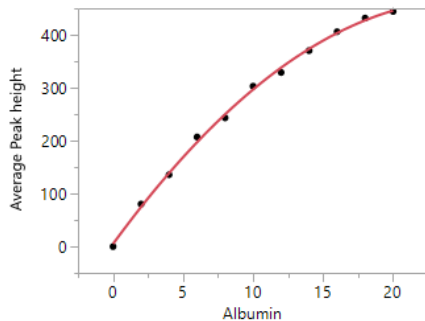
- The fitted model is

$$Y = 76.3255 + 22.0462X - 0.7137(X - 10)^2 + \epsilon,$$

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%

Illustrative example



Illustrative example

- Linear Fit
- Polynomial Fit Degree=2

Linear Fit

Average Peak height = $47.776364 + 22.046273 \cdot \text{Albumin}$

Summary of Fit

RSquare	0.966748
RSquare Adj	0.963054
Root Mean Square Error	28.5884
Mean of Response	268.2391
Observations (or Sum Wgts)	11

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	47.776364	16.12603	2.96	0.0159*
Albumin	22.046273	1.362898	16.18	<.0001*

Polynomial Fit Degree=2

Average Peak height = $76.325548 + 22.046273 \cdot \text{Albumin} - 0.7137296 \cdot (\text{Albumin} - 10)^2$

Summary of Fit

RSquare	0.998361
RSquare Adj	0.997952
Root Mean Square Error	6.731347
Mean of Response	268.2391
Observations (or Sum Wgts)	11

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	76.325548	4.43826	17.20	<.0001*
Albumin	22.046273	0.320904	68.70	<.0001*

Illustrative example

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

$$\begin{aligned}y &= 76.325548 + 22.0463273 \cdot 7 - 0.7137296 \cdot (7 - 10)^2 \\&= 224.2263.\end{aligned}$$

Binary Logistic Regression

Illustrative example

The **Leukaemia Remission dataset**¹ 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)

¹Gal *et al.*, 2019

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

Notice that the response variable here – whether a patient's disease has gone into remission – is not like the response variables in examples covered in the last session.

The response variable is now **binary** – i.e. it can only take one of two values; 0 for 'no remission' and 1 for 'remission'.

Illustrative example

We cannot perform a standard regression of Y on X here: our response variable takes the value 0 or 1, whereas the output from a standard linear regression would not be restricted to take just these values.

When our response variable is binary, we usually perform a **binary logistic regression**.

Illustrative example

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	64.2580841	74.964859	0.73	0.3913
CELL	30.8300607	52.135323	0.35	0.5543
SMEAR	24.686316	61.526122	0.16	0.6882
INFIL	-24.974474	65.280998	0.15	0.7020
LI	4.36045395	2.6579825	2.69	0.1009
BLAST	-0.0115309	2.2663458	0.00	0.9959
TEMP	-100.1734	77.753031	1.66	0.1976

Illustrative example

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-3.7771394	1.3786282	7.51	0.0061*
LI	2.89726323	1.1868229	5.96	0.0146*
For log odds of 1/0				

Illustrative example

- A binary logistic regression ensures the model returns values between 0 and 1; so not on the same scale as the response variable itself, but returning values 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

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

- The precise form of this Equation ensures the output lies between 0 and 1
- Notice that the LI coefficient is significant, since its p -value is 1.46%

Illustrative example

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

$$P(Y = 1) = \frac{e^{-3.778+2.897 \cdot 0.85}}{1 + e^{-3.778+2.897 \cdot 0.85}} = 0.212,$$

or 21.2%.

Illustrative example

