

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.

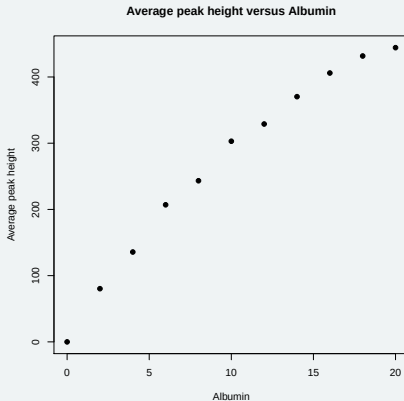
The spectrophotometric method was used to measure signal peak heights for known concentrations of albumin ($\mu\text{g mL}^{-1}$).

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

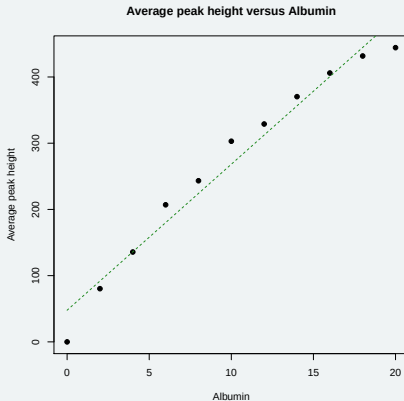
Albumin ($X \mu\text{g mL}^{-1}$)	12	14	16	18	20
Average peak height (Y)	329.0	370.3	406.0	431.7	444.3



Illustrative example

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

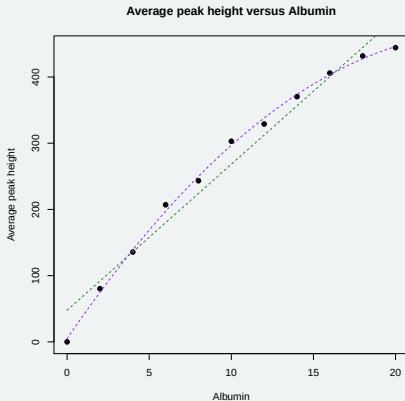
Albumin ($X \mu\text{g mL}^{-1}$)	12	14	16	18	20
Average peak height (Y)	329.0	370.3	406.0	431.7	444.3



Illustrative example

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

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 $p < 0.01\%$)

Illustrative example

- The fitted quadratic model is

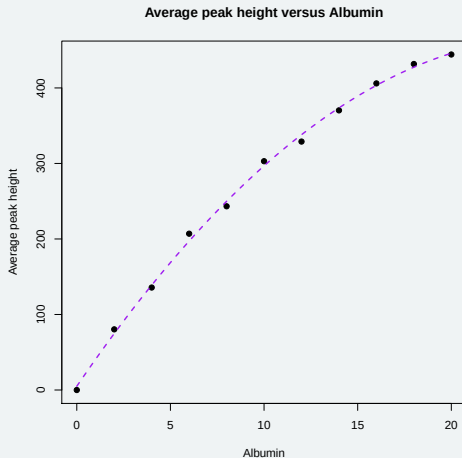
$$Y = 4.953 + 36.321X - 0.714X^2 + \epsilon,$$

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="purple", lwd=2, lty=2)
```

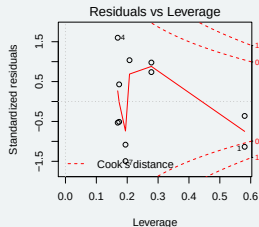
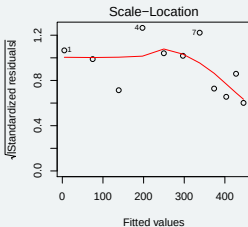
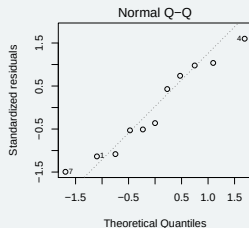
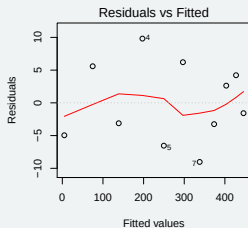


Illustrative example

Of course, we need to make sure the usual regression diagnostics hold:

```
1 par (mfrow=c(2,2))  
2 plot (fit2)
```

Illustrative example



Illustrative example

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

Illustrative example

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

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

Illustrative example

As you might imagine, we can do this much more easily in R:

```
1 predict(fit2, data.frame(albumin = 7, albumin.sq = 49))
2      1
3 224.2259
```

And for several predictions together:

```
1 > predict(fit2, data.frame(albumin=seq(10,30,2), albumin.sq=(seq(10,30,2)^2)))
2      1      2      3      4      5      6      7      8
3 296.7883 338.0259 373.5537 403.3716 427.4798 445.8780 458.5665 465.5451
4      9     10     11
5 466.8139 462.3728 452.2219
```

Price of wine

Bottle	1	2	3	4	5	6	7	8	9	10
Price (£Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00
Age (X_1 years)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rainfall (X_2 mm)	126	121	125	106	107	112	124	105	116	108
Temp. (X_3 °C)	16	20	17	18	18	22	19	15	21	20
New Zealand? (X_4)	0	1	0	0	0	1	0	0	1	0

Price of wine

Bottle	1	2	3	4	5	6	7	8	9	10
Price (£Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00
Age (X_1 years)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rainfall (X_2 mm)	126	121	125	106	107	112	124	105	116	108
Temp. (X_3 °C)	16	20	17	18	18	22	19	15	21	20
New Zealand? (X_4)	0	1	0	0	0	1	0	0	1	0

- Need to check all associations between the **dependent** and **independent** variables

Price of wine

Bottle	1	2	3	4	5	6	7	8	9	10
Price (£Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00
Age (X_1 years)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rainfall (X_2 mm)	126	121	125	106	107	112	124	105	116	108
Temp. (X_3 °C)	16	20	17	18	18	22	19	15	21	20
New Zealand? (X_4)	0	1	0	0	0	1	0	0	1	0

- Need to check all associations between the **dependent** and **independent** variables
- Should really check for **multicollinearity** too

Price of wine

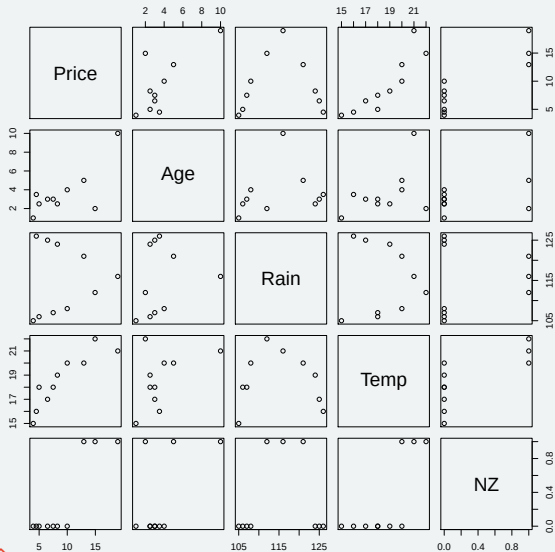
Bottle	1	2	3	4	5	6	7	8	9	10
Price (£Y)	4.50	12.95	6.50	4.99	7.50	14.95	8.25	3.95	18.99	10.00
Age (X_1 years)	$3\frac{1}{2}$	5	3	$2\frac{1}{2}$	3	2	$2\frac{1}{2}$	1	10	4
Rainfall (X_2 mm)	126	121	125	106	107	112	124	105	116	108
Temp. (X_3 °C)	16	20	17	18	18	22	19	15	21	20
New Zealand? (X_4)	0	1	0	0	0	1	0	0	1	0

- Need to check all associations between the **dependent** and **independent** variables
- Should really check for **multicollinearity** too
- Can do both of these things using **scatterplot** and **correlation matrices**

Price of wine

```
1 pairs(wine)
2
3 library("Hmisc")
4 rcorr(as.matrix(wine))
5
6      Price  Age  Rain  Temp  NZ
7 Price  1.00  0.73  0.07  0.89  0.89
8 Age    0.73  1.00  0.21  0.48  0.56
9 Rain   0.07  0.21  1.00 -0.02  0.11
10 Temp  0.89  0.48 -0.02  1.00  0.75
11 NZ    0.89  0.56  0.11  0.75  1.00
12
13 n= 10
14 P
15      Price  Age  Rain  Temp  NZ
16 Price      0.0158  0.8490  0.0006  0.0007
17 Age      0.0158      0.5518  0.1650  0.0921
18 Rain     0.8490  0.5518      0.9482  0.7645
19 Temp     0.0006  0.1650  0.9482      0.0133
20 NZ       0.0007  0.0921  0.7645  0.0133
```

Price of wine



Price of wine

- The price of wine seems most significantly associated (in a linear sense at least!) with temperature
- Generally, price seems to depend significantly on temperature, NZ and Age
- Rainfall: Not significantly associated with price, but it seems there might be a **non-linear association**
- Let's proceed in the usual way: **Backwards elimination**

Price of wine

```
1 fit1 = lm(wine$Price ~ wine$Age + wine$Rain + wine$Temp + wine$NZ)
2 summary(fit1)
3
4 Call:
5 lm(formula = wine$Price ~ wine$Age + wine$Rain + wine$Temp +
6     wine$NZ)
7
8 Residuals:
9      1      2      3      4      5      6      7      8
10 -0.66440 -1.15321  0.59758 -1.93531  0.27734  0.51269  0.57420  1.05449
11      9     10
12  0.64052  0.09608
13
14 Coefficients:
15             Estimate Std. Error t value Pr(>|t|)
16 (Intercept) -11.59955   8.46381  -1.370  0.2289
17 wine$Age      0.62635   0.21159   2.960  0.0315 *
18 wine$Rain    -0.01583   0.05223  -0.303  0.7741
19 wine$Temp     1.03537   0.29267   3.538  0.0166 *
20 wine$NZ       3.77868   1.41420   2.672  0.0442 *
21
22 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
23
24 Residual standard error: 1.271 on 5 degrees of freedom
25 Multiple R-squared:  0.9636, Adjusted R-squared:  0.9345
26 F-statistic: 33.12 on 4 and 5 DF, p-value: 0.0008598
```

Price of wine

```
1 fit2 = lm(wine$Price~wine$Age+wine$Temp+wine$NZ)
2 summary(fit2)
3
4 Call:
5 lm(formula = wine$Price ~ wine$Age + wine$Temp + wine$NZ)
6
7 Residuals:
8      Min       1Q   Median       3Q      Max
9  -1.8126  -0.5658   0.3933   0.4980   1.2216
10
11 Coefficients:
12             Estimate Std. Error t value Pr(>|t|)
13 (Intercept) -13.6573     4.6543  -2.934   0.0261 *
14 wine$Age      0.6131     0.1907   3.215   0.0183 *
15 wine$Temp     1.0515     0.2651   3.966   0.0074 **
16 wine$NZ       3.7314     1.2948   2.882   0.0280 *
17 ---
18 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19
20 Residual standard error: 1.171 on 6 degrees of freedom
21 Multiple R-squared:  0.963, Adjusted R-squared:  0.9444
22 F-statistic:    52 on 3 and 6 DF, p-value: 0.0001096
```

Price of wine

- Final model does not include rainfall:

$$Y = -13.66 + 0.61X_1 + 1.05X_3 + 3.73X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Price of wine

- Final model does not include rainfall:

$$Y = -13.66 + 0.61X_1 + 1.05X_3 + 3.73X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- R^2 statistic is reassuring: Around 94% of the variability in price is explained by age, temperature and NZ together

Price of wine

- Final model does not include rainfall:

$$Y = -13.66 + 0.61X_1 + 1.05X_3 + 3.73X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- R^2 statistic is reassuring: Around 94% of the variability in price is explained by age, temperature and NZ together
- What is the estimated value of a 4 year old wine from New Zealand, if the average temperature during the growing season was 17.5°C and the total rainfall was 112mm?

Price of wine

- Final model does not include rainfall:

$$Y = -13.66 + 0.61X_1 + 1.05X_3 + 3.73X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- R^2 statistic is reassuring: Around 94% of the variability in price is explained by age, temperature and NZ together
- What is the estimated value of a 4 year old wine from New Zealand, if the average temperature during the growing season was 17.5°C and the total rainfall was 112mm?

$$y = -13.66 + 0.61 \cdot 4 + 1.05 \cdot 17.5 + 3.73 = \text{£}10.89.$$

Price of wine

- Final model does not include rainfall:

$$Y = -13.66 + 0.61X_1 + 1.05X_3 + 3.73X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- R^2 statistic is reassuring: Around 94% of the variability in price is explained by age, temperature and NZ together
- What is the estimated value of a 4 year old wine from New Zealand, if the average temperature during the growing season was 17.5°C and the total rainfall was 112mm?

$$y = -13.66 + 0.61 \cdot 4 + 1.05 \cdot 17.5 + 3.73 = \text{£}10.89.$$

- Potential issue of multicollinearity between temperature and NZ, weakening the precision of the estimated coefficients

Price of wine

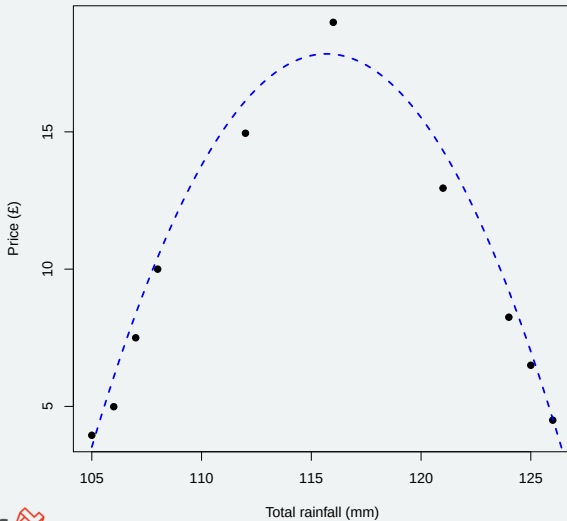
Problem: There definitely was an association between price and rainfall, just not a linear one! Let's test this:

```
1 > fit3 = lm(wine$Price ~ wine$Rain+rainsq)
2 > summary(fit3)
3
4 Call:
5 lm(formula = wine$Price ~ wine$Rain + rainsq)
6
7 Residuals:
8     Min       1Q   Median       3Q      Max
9  -0.8627  -0.5816  -0.2001   0.3655   1.6486
10
11 Coefficients:
12             Estimate Std. Error t value Pr(>|t|)
13 (Intercept) -1.658e+03  1.026e+02  -16.16  8.46e-07 ***
14 wine$Rain    2.897e+01  1.784e+00   16.23  8.19e-07 ***
15 rainsq      -1.252e-01  7.721e-03  -16.22  8.26e-07 ***
16 ---
17 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18
19 Residual standard error: 0.9047 on 7 degrees of freedom
20 Multiple R-squared:  0.9742, Adjusted R-squared:  0.9668
21 F-statistic: 132.1 on 2 and 7 DF, p-value: 2.761e-06
```

Price of wine

```
1 plot(wine$Rain, wine$Price, xlab="Total rainfall (mm)", ylab="Price", main="", pch  
   =19)  
2 curve((-1.658e+3)+(2.897e1)*x-(1.252e-1)*x^2, from=105, to=127, col="blue", lty=2, lwd  
   =2, add=TRUE)
```

Price of wine



Note from this fit the R^2 value of 96.68%.

What about including the other variables too, in a multiple regression model?

Price of wine

```
1 > fit4 = lm(wine$Price ~ wine$Age + wine$Rain + +rainsq + wine$Temp + wine$NZ)
2 > summary(fit4)
3
4 Call:
5 lm(formula = wine$Price ~ wine$Age + wine$Rain + +rainsq + wine$Temp +
6     wine$NZ)
7
8 Residuals:
9      1      2      3      4      5      6      7      8      9     10
10  0.3294 -0.7861 -0.1134 -0.2418 -0.7155  0.4448  0.1090  0.4118  0.3413  0.2205
11
12 Coefficients:
13             Estimate Std. Error t value Pr(>|t|)
14 (Intercept) -2.089e+03  5.671e+02  -3.684  0.0211 *
15 wine$Age      1.872e-01  1.650e-01   1.135  0.3199
16 wine$Rain     3.666e+01  1.001e+01   3.662  0.0215 *
17 rainsq       -1.585e-01  4.328e-02  -3.663  0.0215 *
18 wine$Temp     -5.981e-01  4.727e-01  -1.265  0.2744
19 wine$NZ       -8.247e-01  1.467e+00  -0.562  0.6041
20
21 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
22
23 Residual standard error: 0.6808 on 4 degrees of freedom
24 Multiple R-squared:  0.9916, Adjusted R-squared:  0.9812
25 F-statistic:  95 on 5 and 4 DF, p-value: 0.0003025
```

Price of wine

```
1 > fit5 = lm(wine$Price ~ wine$Age + wine$Rain + +rainsq + wine$Temp)
2 > summary(fit5)
3
4 Call:
5 lm(formula = wine$Price ~ wine$Age + wine$Rain + +rainsq + wine$Temp)
6
7 Residuals:
8      1          2          3          4          5          6          7
9  0.1946739 -0.9264540 -0.0001244 -0.4366023 -0.5511771  0.3823755  0.2441489
10      8          9         10
11  0.4168738  0.3665303  0.3097555
12
13 Coefficients:
14             Estimate Std. Error t value Pr(>|t|)
15 (Intercept) -1.815e+03  2.703e+02  -6.716  0.00111 **
16 wine$Age      2.344e-01  1.319e-01   1.777  0.13572
17 wine$Rain     3.184e+01  4.800e+00   6.633  0.00117 **
18 rainsq       -1.377e-01  2.076e-02  -6.633  0.00117 **
19 wine$Temp     -4.137e-01  3.160e-01  -1.309  0.24748
20
21 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
22
23 Residual standard error: 0.6325 on 5 degrees of freedom
24 Multiple R-squared:  0.991, Adjusted R-squared:  0.9838
25 F-statistic: 137.5 on 4 and 5 DF, p-value: 2.68e-05
```

Price of wine

```
1 fit6 = lm(wine$Price~wine$Age+wine$Rain++rainsq)
2 summary(fit6)
3
4 Call:
5 lm(formula = wine$Price ~ wine$Age + wine$Rain + +rainsq)
6
7 Residuals:
8      Min       1Q   Median       3Q      Max
9 -0.78306 -0.38636  0.05188  0.36464  0.92325
10
11 Coefficients:
12             Estimate Std. Error t value Pr(>|t|)
13 (Intercept) -1.484e+03  1.010e+02 -14.699  6.23e-06 ***
14 wine$Age      3.182e-01  1.220e-01   2.608   0.0403 *
15 wine$Rain     2.594e+01  1.756e+00  14.772  6.05e-06 ***
16 rainsq       -1.122e-01  7.579e-03 -14.804  5.97e-06 ***
17 ---
18 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
19
20 Residual standard error: 0.6691 on 6 degrees of freedom
21 Multiple R-squared:  0.9879, Adjusted R-squared:  0.9819
22 F-statistic: 163.3 on 3 and 6 DF, p-value: 3.855e-06
```

- The **final model** includes: Age, Rainfall and $(\text{Rainfall})^2$

Price of wine

- The **final model** includes: Age, Rainfall and $(\text{Rainfall})^2$
- Notice the change in the R^2 statistic:

Model	Included variables	R^2
fit1	Age, Rainfall, Temp., NZ	0.9345
fit2	Age Temp., NZ	0.9444
fit3	Rainfall, $(\text{Rainfall})^2$	0.9668
fit4	Age, Rainfall, $(\text{Rainfall})^2$, Temp., NZ	0.9812
fit5	Age, Rainfall, $(\text{Rainfall})^2$, Temp.	0.9838
fit6	Age, Rainfall, $(\text{Rainfall})^2$	0.9819

- Now, our final model is

$$Y = -1484 + 0.3182X_1 + 25.94X_2 - 0.1122X_2^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

Price of wine

- Now, our final model is

$$Y = -1484 + 0.3182X_1 + 25.94X_2 - 0.1122X_2^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

- What is the estimated value of a 4 year old wine from New Zealand, if the average temperature during the growing season was 17.5°C and the total rainfall was 112mm?

Price of wine

- Now, our final model is

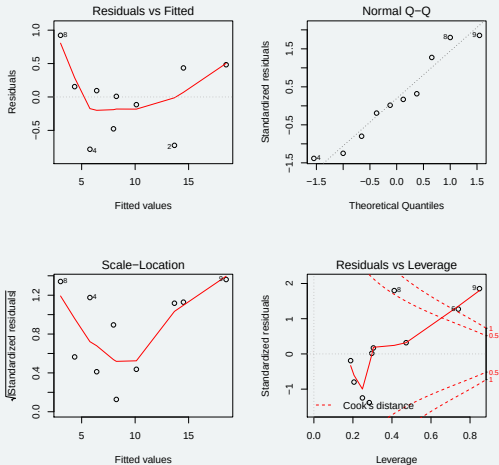
$$Y = -1484 + 0.3182X_1 + 25.94X_2 - 0.1122X_2^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

- What is the estimated value of a 4 year old wine from New Zealand, if the average temperature during the growing season was 17.5°C and the total rainfall was 112mm?

$$y = -1484 + 0.3182 \cdot 4 + 25.94 \cdot 112 - 0.1122 \cdot 112^2 = \text{£}15.12.$$

Price of wine

```
1 par(mfrow=c(2,2))  
2 plot(fit6)
```



Quadratic regression

We will now break for a problem session covering this section.

Please pause the video here and go to **Problem Session 1** of **Lesson 8**.

See you in a bit!

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

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

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 so far.

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.

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

```
1 > attach(leukaemia)
2 > fit1 = glm(REMISS~CELL+SMEAR+INFIL+LI+BLAST+TEMP, family = binomial)
3 > summary(fit1)
4
5 Call:
6 glm(formula = REMISS ~ CELL + SMEAR + INFIL + LI + BLAST + TEMP,
7     family = binomial)
8
9 Deviance Residuals:
10      Min       1Q   Median       3Q      Max
11 -1.95404  -0.66259  -0.02516   0.78184   1.57465
12
13 Coefficients:
14             Estimate Std. Error z value Pr(>|z|)
15 (Intercept)  64.25808   74.96480   0.857   0.391
16 CELL        30.83006   52.13520   0.591   0.554
17 SMEAR       24.68632   61.52601   0.401   0.688
18 INFIL      -24.97447   65.28088  -0.383   0.702
19 LI          4.36045    2.65798   1.641   0.101
20 BLAST       -0.01153    2.26634  -0.005   0.996
21 TEMP      -100.17340   77.75289  -1.288   0.198
22
23 (Dispersion parameter for binomial family taken to be 1)
24
25     Null deviance: 34.372  on 26  degrees of freedom
26 Residual deviance: 21.594  on 20  degrees of freedom
27 AIC: 35.594
28
29 Number of Fisher Scoring iterations: 8
```

Illustrative example

```
1 > fit2 = glm(REMISS ~ LI, family = binomial)
2 > summary(fit2)
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 LI 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?

Illustrative example

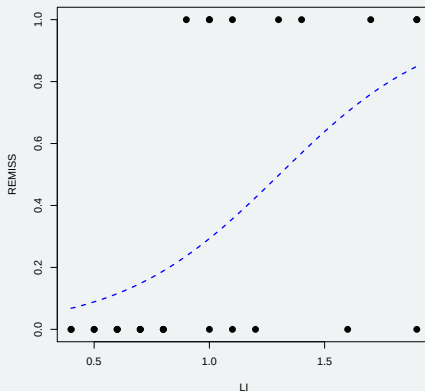
What is the chance of remission in a patient with an LI value of 0.85?

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

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)
```



Binary logistic regression

We will now break for a problem session covering this section.

Please pause the video here and go to **Problem Session 2** of **Lesson 8**.

See you in a bit!

Log-Linear Regression

Illustrative example

The data below, stored in the dataframe `speed` in R, shows the number of road traffic accidents between April 2018–March 2019 at 67 locations in the Northumbria Police force area.

Also shown is the average recorded speed (miles per hour), the percentage of vehicles being driven in excess of the speed limit, and the daily average annual traffic flow.

```
1 > head(speed)
2   Site Average.Speed Over.limit Flow Accidents
3 1 103          36.6      88.0 3733          2
4 2 104          34.6      10.8 13448         6
5 3 190          20.6       4.0 16625        23
6 4 194          34.0      74.1 10379         7
7 5 417          37.2      90.5 10949         2
8 6 451          46.8      23.7 22663         8
```

We might expect Accidents to be related to all three of 'Average Speed', 'Over Limit' and 'Flow', but the nature of these associations might be surprising...

Illustrative example

```
1 > pairs(speed[-1])
2 > library("Hmsic")
3 > rcorr(as.matrix(speed[-1]))
4
```

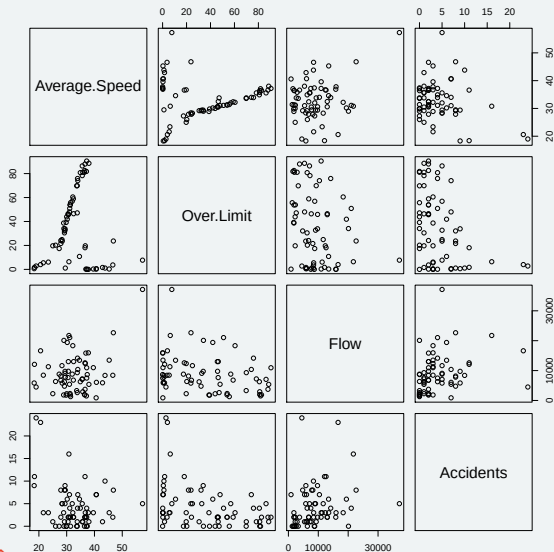
	Average.Speed	Over.Limit	Flow	Accidents
Average.Speed	1.00	0.06	0.27	-0.27
Over.Limit	0.06	1.00	-0.28	-0.39
Flow	0.27	-0.28	1.00	0.26
Accidents	-0.27	-0.39	0.26	1.00

```
9
10 n= 67
11
12 P
13
```

	Average.Speed	Over.Limit	Flow	Accidents
Average.Speed		0.6491	0.0256	0.0256
Over.Limit	0.6491		0.0222	0.0010
Flow	0.0256	0.0222		0.0356
Accidents	0.0256	0.0010	0.0356	

```
14
15
16
17
```

Illustrative example



Illustrative example

- In a linear sense, accident counts are significantly associated with all three independent variables

Illustrative example

- In a linear sense, accident counts are significantly associated with all three independent variables
- There is **multicollinearity**: Remove 'Flow'?

```
1 > speed2 = speed[-4]
2 > head(speed2)
3   Site Average.Speed Over.Limit Accidents
4 1 103          36.6         88.0         2
5 2 104          34.6         10.8         6
6 3 190          20.6          4.0        23
7 4 194          34.0         74.1         7
8 5 417          37.2         90.5         2
9 6 451          46.8         23.7         8
10
11 > rcorr(as.matrix(speed2[-1]))
12                Average.Speed Over.Limit Accidents
13 Average.Speed                1.00         0.06      -0.27
14 Over.Limit                   0.06         1.00      -0.39
15 Accidents                   -0.27        -0.39         1.00
16
17 n= 67
18 P
19                Average.Speed Over.Limit Accidents
20 Average.Speed                0.6491         0.0256
21 Over.Limit                   0.6491         0.0010
22 Accidents                    0.0256         0.0010
```

Illustrative example

We could now proceed to a model that regresses 'Accidents' on both 'Average Speed' and 'Over Limit'.

However, notice that the dependent variable consists of **counts**.

As with the leukaemia dataset, we cannot use standard linear regression techniques here: For example, we cannot have negative accidents, but a standard linear regression would not restrict regression estimates to be non-negative.

Illustrative example

We could now proceed to a model that regresses 'Accidents' on both 'Average Speed' and 'Over Limit'.

However, notice that the dependent variable consists of **counts**.

As with the leukaemia dataset, we cannot use standard linear regression techniques here: For example, we cannot have negative accidents, but a standard linear regression would not restrict regression estimates to be non-negative.

Thus, a **log-linear** model is often used, with a **Poisson** assumption for the dependent variable.

Illustrative example

In our example, we assume that accident counts at site j , say y_j , are Poisson-distributed with mean μ_j – that is

$$y_j \sim \text{Poisson}(\mu_j),$$

and that

$$\log \mu_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j},$$

where x_1 and x_2 represent ‘Average Speed’ and ‘Over Limit’ respectively.

Illustrative example

In our example, we assume that accident counts at site j , say y_j , are Poisson-distributed with mean μ_j – that is

$$y_j \sim \text{Poisson}(\mu_j),$$

and that

$$\log \mu_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j},$$

where x_1 and x_2 represent ‘Average Speed’ and ‘Over Limit’ respectively.

The right-hand-side of the above equation is often termed the **linear predictor**; the log function ensures that μ_j is always positive, since

$$\mu_j = \exp \{ \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} \}.$$

Illustrative example

As for binary logistic regression, we use the `glm` function in R, with a switch of **model family** from **binomial** to **Poisson**.

Illustrative example

```
1 > head(speed2)
2   Site Average.Speed Over.Limit Accidents
3   1    103          36.6       88.0        2
4   2    104          34.6       10.8        6
5   3    190          20.6        4.0       23
6 > fit1 = glm(Accidents ~ Average.Speed + Over.Limit, family=poisson)
7 > summary(fit1)
8 Call:
9 glm(formula = Accidents ~ Average.Speed + Over.Limit, family = poisson)
10
11 Deviance Residuals:
12     Min       1Q   Median       3Q      Max
13 -3.4853 -1.6948 -0.5456  0.8828  3.6985
14
15 Coefficients:
16             Estimate Std. Error z value Pr(>|z|)
17 (Intercept)  3.010612   0.246506  12.213  < 2e-16 ***
18 Average.Speed -0.033659   0.007741  -4.348  1.37e-05 ***
19 Over.Limit   -0.016537   0.002430  -6.804  1.02e-11 ***
20 ---
21 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
22
23 (Dispersion parameter for poisson family taken to be 1)
24
25     Null deviance: 293.10  on 66  degrees of freedom
26 Residual deviance: 215.02  on 64  degrees of freedom
27 AIC: 399.95
28
29 Number of Fisher Scoring iterations: 5
```

Illustrative example

- The fitted model is:

$$\mu_j = \exp \{3.0106 - 0.0337 \cdot \text{Average.Speed} - 0.0165 \cdot \text{Over.Limit}\}$$

Illustrative example

- The fitted model is:

$$\mu_j = \exp \{3.0106 - 0.0337 \cdot \text{Average.Speed} - 0.0165 \cdot \text{Over.Limit}\}$$

- Both independent variables are significant predictors of accident counts, with significance beyond the 0.1% level

Illustrative example

- The fitted model is:

$$\mu_j = \exp \{3.0106 - 0.0337 \cdot \text{Average.Speed} - 0.0165 \cdot \text{Over.Limit}\}$$

- Both independent variables are significant predictors of accident counts, with significance beyond the 0.1% level
- Note the **Akaike Information Criterion**, $AIC = 399.95$. This is a measure of prediction error (and thereby a measure of relative quality) of the fitted model

Illustrative example

- How many accidents might we expect in a one year period, for a new location which has an average observed speed of 40.5 mph and with around 15% of vehicles driving in excess of the speed limit?

Illustrative example

- How many accidents might we expect in a one year period, for a new location which has an average observed speed of 40.5 mph and with around 15% of vehicles driving in excess of the speed limit?

$$\mu_j = \exp \{3.0106 - 0.0337 \cdot 40.5 - 0.0165 \cdot 15\} = 4.12 \text{ Accidents.}$$

Illustrative example

Note the Poisson assumption requires that $E(Y) = Var(Y)$.

- We have
 - $\bar{y} = 4.284$
 - $s_y^2 = 22.752$

There is evidence to suggest that $Var(Y) \gg E(Y)$!

Illustrative example

Note the Poisson assumption requires that $E(Y) = Var(Y)$.

- We have
 - $\bar{y} = 4.284$
 - $s_y^2 = 22.752$

There is evidence to suggest that $Var(Y) \gg E(Y)$!

- The **Negative Binomial** distribution can thus be used in place of the Poisson, which has an **over-dispersion parameter** to account for a relatively large variance

Illustrative example

```
1 > install.packages("MASS")
2 > library(MASS)
3 > fit2 = glm.nb(Accidents ~ Average.Speed + Over.Limit)
4 > summary(fit2)
5 Call:
6 glm.nb(formula = Accidents ~ Average.Speed + Over.Limit, init.theta = 1.749647079,
7         link = log)
8
9 Deviance Residuals:
10      Min       1Q   Median       3Q      Max
11  -2.2607  -0.9353  -0.3327   0.4284   1.5974
12
13 Coefficients:
14             Estimate Std. Error z value Pr(>|z|)
15 (Intercept)   2.68483    0.53712   4.999 5.78e-07 ***
16 Average.Speed -0.02484    0.01582  -1.570 0.116494
17 Over.Limit    -0.01502    0.00402  -3.735 0.000188 ***
18 ---
19 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
20
21 (Dispersion parameter for Negative Binomial(1.7496) family taken to be 1)
22 Null deviance: 96.646 on 66 degrees of freedom
23 Residual deviance: 76.607 on 64 degrees of freedom
24 AIC: 334.04
25
26 Number of Fisher Scoring iterations: 1
27
28             Theta: 1.750
29             Std. Err.: 0.479
```

Illustrative example

```
1 > fit3 = glm.nb(Accidents ~ Over.Limit)
2 > summary(fit3)
3 Call:
4 glm.nb(formula = Accidents ~ Over.Limit, init.theta = 1.604253113,
5         link = log)
6
7 Deviance Residuals:
8     Min       1Q   Median       3Q      Max
9  -2.3066  -0.8682  -0.3391   0.4202   1.9419
10
11 Coefficients:
12             Estimate Std. Error z value Pr(>|z|)
13 (Intercept)  1.919639   0.173263  11.079 < 2e-16 ***
14 Over.Limit  -0.016024   0.004093  -3.915 9.04e-05 ***
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
17
18 (Dispersion parameter for Negative Binomial(1.6043) family taken to be 1)
19
20 Null deviance: 91.893 on 66 degrees of freedom
21 Residual deviance: 75.779 on 65 degrees of freedom
22 AIC: 334.83
23
24 Number of Fisher Scoring iterations: 1
25
26             Theta:  1.604
27             Std. Err.:  0.419
```

Illustrative example

- Notice the change in AIC :

Mode	Poisson (fit1)	Negative Binomial (fit2)	Negative Binomial (fit3)
AIC	399.95	334.04	334.83
Over-dispersion parameter	1	1.75 (0.479)	1.604 (0.419)

Illustrative example

- Notice the change in AIC :

Mode	Poisson (fit1)	Negative Binomial (fit2)	Negative Binomial (fit3)
AIC	399.95	334.04	334.83
Over-dispersion parameter	1	1.75 (0.479)	1.604 (0.419)

- The log-linear predictor is:

$$\mu_j = \exp \{1.9196 - 0.0160 \cdot \text{Over.Limit}\}$$

- In our last example, we have 15% of vehicles being driven over the speed limit, giving:

$$\mu_j = \exp \{1.9196 - 0.0160 \cdot 15\} = 5.36 \text{ Accidents}$$

Log-linear regression

We will now break for a problem session covering this section.

Please pause the video here and go to **Problem Session 3** of **Lesson 8**.

See you in a bit!

Conclusions

1 Quadratic regression

- ▶ Simple way to capture apparent **non-linear** associations
- ▶ Should compare the fit to that of a simple linear regression, to check for improvement

Conclusions

1 Quadratic regression

- ▶ Simple way to capture apparent **non-linear** associations
- ▶ Should compare the fit to that of a simple linear regression, to check for improvement

2 Binary logistic regression

- ▶ A form of **Generalised Linear Model**. Should be used when the response is **binary**
- ▶ Most common model form is the **logistic** model, though others (e.g. probit) are available

Conclusions

1 Quadratic regression

- ▶ Simple way to capture apparent **non-linear** associations
- ▶ Should compare the fit to that of a simple linear regression, to check for improvement

2 Binary logistic regression

- ▶ A form of **Generalised Linear Model**. Should be used when the response is **binary**
- ▶ Most common model form is the **logistic** model, though others (e.g. probit) are available

3 Log-linear regression

- ▶ Another example of a GLM, used whe the response consists of **counts**
- ▶ Most common assumption is that of a **Poisson** distribution
- ▶ A **Negative binomial** model can be used when there is evidence of **over-dispersion**