8.7 Regression

MATHEMATICAL TECHNIQUES

♠ Consider the data in the following table.

weight	yield	height	weight	yield	height
0.20	0.599	10.17	1.2	1.995	15.25
0.40	0.909	12.83	1.4	2.518	14.47
0.60	1.220	13.47	1.6	2.330	17.74
0.80	1.363	14.60	1.8	2.963	20.46
1.0	1.523	13.55	2.0	3.628	21.12

- EXERCISE **8.7.1**

  Plot yield (Y) against weight (W). Suppose we think that the line $Y = W + 0.6$ describes these data. Plot the line on your graph of yield against weight. Find and plot the residuals.
- EXERCISE **8.7.2**

  Plot height (H) against weight (W). Suppose we think that the line $H = 8W + 5$ describes the data. Plot the line on your graph of height against weight. Find and plot the residuals.
- EXERCISE **8.7.3**

  Find SSE for the model used in exercise 8.7.1
- EXERCISE **8.7.4**

  Find SSE for the model used in exercise 8.7.2
- EXERCISE **8.7.5**

  Find the null model that best fits Y as a function of W and find SST.
- EXERCISE **8.7.6**

  Find the null model that best fits H as a function of W and find SST.
- EXERCISE **8.7.7**

  Find r^2 for the model in exercise 8.7.1.
- EXERCISE **8.7.8**

  Find r^2 for the model in exercise 8.7.2.
- EXERCISE **8.7.9**

  Compute the best fit line for yield on weight. Graph the line.
- EXERCISE **8.7.10**

  Compute the best fit line for height on weight. Graph the line.
- EXERCISE **8.7.11**

  Find SSE for the line inexercise 8.7.9, find r^2 and compare with the model in exercise 8.7.1.
- EXERCISE **8.7.12**

  Find SSE for the line inexercise 8.7.10, find r^2 and compare with the model in exercise 8.7.2.
- EXERCISE **8.7.13**

  Find the correlation between weight and yield. How does it compare with r^2?
- EXERCISE **8.7.14**

  Find the correlation between weight and height. How does it compare with r^2?

♠ Linear regression has important connections with other techniques in statistics, such as testing whether two populations differ. In the following data, the independent variable takes on only two values. Find the best fit line and r^2, and then test whether the two sets of points differ in their mean distribution using the techniques in section 8.6. Assume that the data are normally distributed with known variance of 25.

Diet	Size in replicate 1	Size in replicate 2
1	9.51	25.95
1	20.40	19.26
1	14.50	20.05
1	20.76	27.31
1	17.42	21.84
1	23.34	19.45
2	28.20	22.52
2	30.12	25.32
2	24.12	24.04
2	31.43	23.53
2	33.46	28.39
2	29.70	25.17

- EXERCISE **8.7.15**
  Find the best fit line and r^2 for replicate 1 and then test whether the diet has a significant effect. Graph the regression line and the data.
- EXERCISE **8.7.16**
  Find the best fit line and r^2 for replicate 2 and then test whether the diet has a significant effect. Graph the regression line and the data. Compare with the previous problem.

♠ Best fit regression lines have many nice properties.

- EXERCISE **8.7.17**
  Show that the best fit horizontal line (used to compute SST) passes through the center of the data in the sense that the sum of the residuals is 0.
- EXERCISE **8.7.18**
  Show that the best linear fit from theorem 8.7.1 passes through the center of the data in the sense that the sum of the residuals is 0.
- EXERCISE **8.7.19**
  Consider models of the form $Y = b$. Show that the sum of the squares of the residuals is minimized when $b = \bar{Y}$, the sample mean of the y_i.
- EXERCISE **8.7.20**
  Consider models of the form $Y = aX$. Find the slope that minimizes the sum of the squares of the residuals.

♠ Consider the measurements

x	y
1.0	1.1
2.0	3.9
3.0	8.8
4.0	16.5

- EXERCISE **8.7.21**
  Find the best linear fit. Plot the line and find r^2. How good is the model?
- EXERCISE **8.7.22**
  Use the principle of least squares to write the expression you would use to fit a curve of the form $Y = aX^2 + b$. One easy way to solve is to think of a new measurement $Z = X^2$ and find the linear regression of Y on Z. Plot the linear regression of Y against Z and the curved regression of Y against X^2. Compute r^2. Which model does better?

♠ We should check the dimensions for each term of the regression equation for toxin tolerance as a function of mass to show that the whole thing is consistent.

- EXERCISE **8.7.23**
  Find the dimensions of the slope $\hat{a}$.
- EXERCISE **8.7.24**
  Find the dimensions of the intercept $\hat{b}$ and check that all the parts of the equation match.

APPLICATIONS

♠ Recall the data on page 7, with the millions left out.

Colony	Old population (b_{old})	New population (b_{new})
1	0.47	0.95
2	3.3	6.4
3	0.73	1.5
4	2.8	5.6
5	1.5	3.1
6	0.62	1.2

- EXERCISE **8.7.25**
  Find r^2 for the line $b_{new} = 2b_{old}$. Graph the data and the line.
- EXERCISE **8.7.26**
  Find the best fit line, and compare with mathematically idealized model. Which makes more sense?

♠ Recall the data on page 8.

Tree	Old height (h_{old})	New height (h_{new})
1	23.1	24.1
2	18.7	19.8
3	20.6	21.5
4	16.0	17.0
5	32.5	33.6
6	19.8	20.6

- EXERCISE **8.7.27**
  Find r^2 for the line $h_{new} = h_{old} + 1$. Graph the data and the line.
- EXERCISE **8.7.28**
  Find the best fit line, and compare with mathematically idealized model. Which makes more sense?

♠ Consider the following data on the growth of two bacterial populations.

Year	Population 1	Population 2
0	100	50
1	119	68
2	168	82
3	198	141
4	259	212
5	306	399
6	421	552

- EXERCISE **8.7.29**
  Find the best fit line for population 1 as a function of time and compute r^2.
- EXERCISE **8.7.30**
  Find the best fit line for population 2 as a function of time and compute r^2.
- EXERCISE **8.7.31**
  Find the best fit line for the logarithm of population 1 as a function of time and compute r^2. Is this a better fit?
- EXERCISE **8.7.32**
  Find the best fit line for the logarithm of population 2 as a function of time and compute r^2. Is this a better fit?

♠ Consider the following data including one outlying point. Find the best fit line with and without that point. How much difference does that point make? The idea of removing one point and testing how much the fit changes is an important tool in statistics, and forms the basis of techniques called **the jacknife** and **cross-validation**.

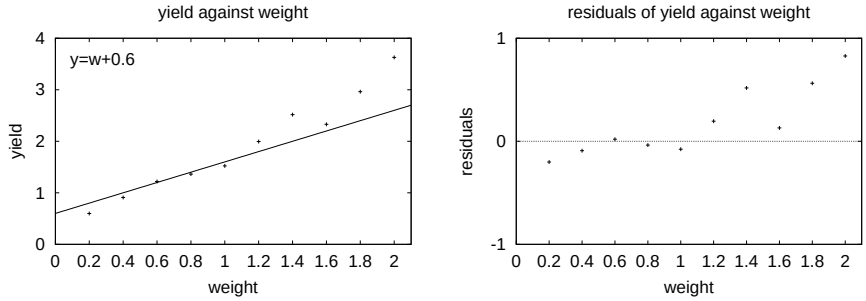
Feeding rate	Size in replicate 1	Size in replicate 2
1	11.2	9.2
1	12.1	10.2
2	19.2	21.5
2	44.2	17.0
3	31.5	33.6
3	33.4	30.6
4	38.2	43.6
4	44.3	10.6

- EXERCISE **8.7.33**
  Find the best fit line and r^2 for replicate 1 with and without the fourth point. Graph the two regression lines and the data.
- EXERCISE **8.7.34**
  Find the best fit line and r^2 for replicate 2 with and without the last point. Graph the two regression lines and the data. Why do you think the outlier affects this regression line more?

Chapter 9

Answers

8.7.1.



The residuals are

w	y	$\hat{y}$	$y-\hat{y}$
0.2	0.599	0.8	-0.201
0.4	0.909	1.0	-0.091
0.6	1.220	1.2	0.020
0.8	1.363	1.4	-0.037
1.0	1.523	1.6	-0.077
1.2	1.995	1.8	0.195
1.4	2.518	2.0	0.518
1.6	2.330	2.2	0.130
1.8	2.963	2.4	0.563
2.0	3.628	2.8	0.828

8.7.3. Adding up the squares of the residuals gives SSE=1.753.
8.7.5. The null model uses the mean yield, or $Y = \bar{Y} = 1.905$. The sum of the squares of the residuals is SST=8.259.
8.7.7. $r^2 = 1 - \dfrac{\text{SSE}}{\text{SST}} = 0.788$.
8.7.9. The best fit line is found by computing

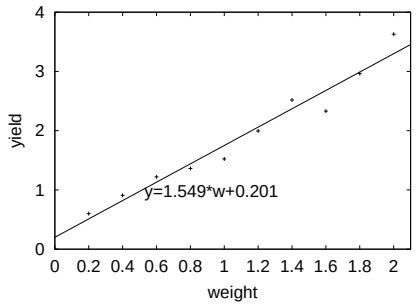
$$\begin{aligned} \widehat{\text{Cov}}(W,Y) &= 0.568 \\ \widehat{\text{Var}}(W) &= 0.367 \\ \bar{W} &= 1.10 \\ \bar{Y} &= 1.905. \end{aligned}$$

From theorem 8.7.1,

$$\hat{a} = \frac{0.568}{0.367} = 1.549$$

$$\hat{b} = 1.905 - 1.549 \cdot 1.10 = 0.201$$

The line is then $Y = 1.549W + 0.201$.



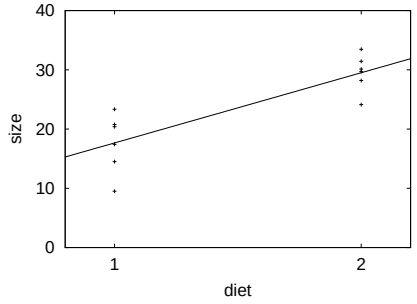
8.7.11. Using theorem 8.7.2, SSE=0.338. $r^2 = 1 - \frac{0.338}{8.25} = 0.95$, much larger than before.

8.7.13. The only additional information we need to find the correlation is the variance of Y, which is 0.918. The correlation is then

$$\rho_{YW} = \frac{\widehat{\text{Cov}}(W,Y)}{\sqrt{\widehat{\text{Var}}(W)}\sqrt{\widehat{\text{Var}}(Y)}} = \frac{0.568}{\sqrt{0.367}\sqrt{0.918}} = 0.979.$$

The correlation seems to be larger than r^2.

8.7.15. The best fit line is $S = 11.85D + 5.81$, and $r^2 = 0.70$. The mean with diet 1 is 17.66 and with diet 2 is 29.51. The variance in each of the means is 25/6, so the variance of the difference between the means is 25/3=8.333. The standard deviation of the difference between the means is then 2.89, so the means differ by 4.1 standard deviations. The associated p-value is tiny, 4.1×10^{-5}.



8.7.17. The best fit horizontal line has intercept equal to the mean. Then

$$\sum_{i=1}^{n}(y_i - \bar{Y}) = n\bar{Y} - n\bar{Y} = 0.$$

8.7.19. Let the data be $y_1, \ldots, y_n$. Then for a given b,

$$\text{SST} = \sum_{i=1}^{n}(y_i - b)^2.$$

Taking the derivative with respect to b,

$$\begin{aligned} \frac{d\text{SST}}{db} &= -2\sum_{i=1}^{n}(y_i - b) \\ &= -2(n\bar{Y} - nb) \\ &= -2n(\bar{Y} - b) \end{aligned}$$

which is 0 when $\bar{Y} = b$. Furthermore, the second derivative of SST with respect to b is equal to $2n$, which is positive. Therefore, $b = \bar{Y}$ is a minimum.

8.7.21. The best fit line is found by computing

$$\begin{aligned} \widehat{\mathrm{Cov}}(X,Y) &= 6.388 \\ \widehat{\mathrm{Var}}(X) &= 1.250 \\ \bar{X} &= 2.50 \\ \bar{Y} &= 7.575. \end{aligned}$$

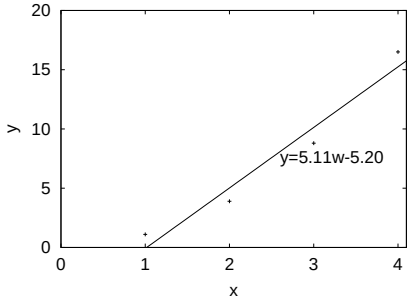
From theorem 8.7.1,

$$\begin{aligned} \hat{a} &= \frac{6.388}{1.250} = 5.110 \\ \hat{b} &= 7.575 - 5.11 \cdot 2.50 = -5.20. \end{aligned}$$

The line is then

$$Y = 5.11X - 5.20.$$

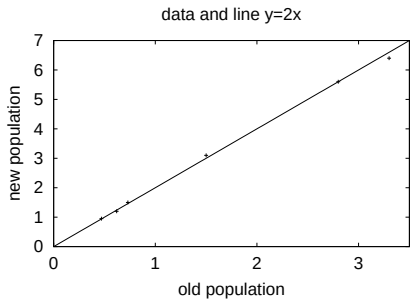
Using theorem 8.7.2, SSE=6.027. Furthermore, the sum of the squared differences of Y from the mean 7.575 is 136.6. Therefore, $r^2 = 1 - \frac{6.207}{136.6} = 0.96$, which looks pretty good.



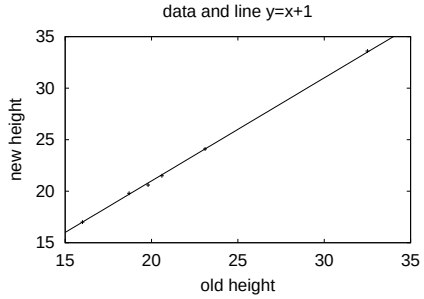
8.7.23. $\hat{a}$ has dimensions

$$\begin{aligned} \hat{a} &= \frac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\mathrm{Var}}(X)} \\ &= \frac{\mathrm{mass} \cdot \mathrm{tolerance}}{\mathrm{mass}^2} \\ &= \frac{\mathrm{tolerance}}{\mathrm{mass}} \end{aligned}$$

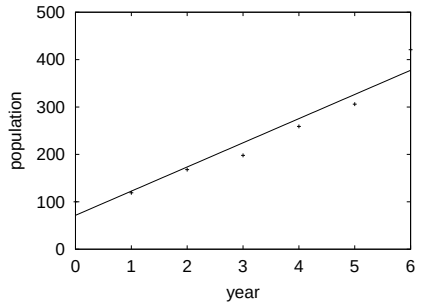
8.7.25. The residuals are 0.01, -0.2, 0.04, 0.0, 0.1 and -0.04, giving SSE=0.0533. Using the horizontal line at the mean of b_{new}, 3.125, gives residuals of -2.175, 3.275, -1.625, 2.475, -0.025 and -1.925, so SST=27.92. This gives $r^2 = 0.998$.



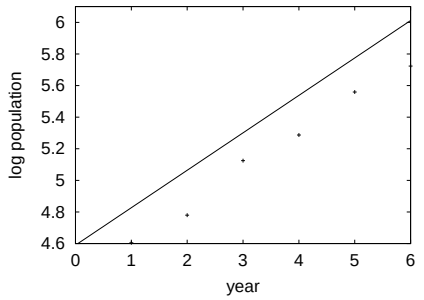
8.7.27. The residuals are 0.0, 0.1, -0.1, 0.0, 0.1 and -0.2, giving SSE=0.07. Using the horizontal line at the mean of b_{new}, 22.77, gives residuals of 1.33, -2.97, -1.27, -5.77, 10.83 and -2.17, and SST=167.49. This gives $r^2 = 0.9996$.



8.7.29. The best fit line has slope 51 and intercept 71.5 with SSE=4148. SST uses the horizontal line 224.5, and gives SST=77046. Then $r^2 = 0.946$.



8.7.31. Using the logarithms of the population sizes, we find the best fit line $\ln(P) = 4.59 + 0.237Y$. The r^2 value is 0.99. This fit looks a lot better.



8.7.33. The line with all the data is $S = 8.95F + 6.875$, and $r^2 = 0.63$. Removing the fourth point, the best fit line is $S = 10.1F + 1.16$ with $r^2 = 0.97$. The outlier really messes things up.