

EXAMPLE Hooker's Temperature and Pressure Data

The following data record the boiling point temperature (in degrees F) of water under different atmospheric pressures. The data were collected in a Himalayan expedition by botanist Joseph Hooker.

Reference: Forbes, J. (1957). Further experiments and remarks on the measurement of heights by boiling point of water. *Transactions of the Royal Society of Edinburgh*, 21, 235-243.

y	x	y	x	y	x	y	x
210.8	29.211	196.4	21.928	189.5	18.869	184.1	16.817
210.2	28.559	196.3	21.654	188.8	18.356	183.2	16.385
208.4	27.972	195.6	21.605	188.5	18.507	182.4	16.235
202.5	24.697	193.4	20.480	185.7	17.267	181.9	16.106
200.6	23.726	193.6	20.212	186.0	17.221	181.9	15.928
200.1	23.369	191.4	19.758	185.6	17.062	181.0	15.919
199.5	23.030	191.1	19.490	184.1	16.959	180.6	15.376
197.0	21.892	190.6	19.386	184.6	16.881		

Here, to test for significant correlation, we use the test statistic

$$T_{obs} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.996}{\sqrt{(1-0.996^2)/(31-2)}} = 59.143$$

which we must compare against the $t_{n-2} \equiv t_{29}$ distribution. For a two-tailed test at the significance level $\alpha = 0.05$, the critical values are $C_R = \pm 2.045$ (McClave and Sincich t -tables), so the hypothesis H_0 that the true correlation is zero is **rejected**.

```
> cor.test(Hooker$pressure, Hooker$temp)
+
+ Pearson's product-moment correlation
+
+ data: Hooker$pressure and Hooker$temp
+ t = 59.14, df = 29, p-value <2e-16
+ alternative hypothesis: true correlation is not equal to 0
+ 95 percent confidence interval:
+ 0.991378 0.998034
+ sample estimates:
+ cor
+ 0.99588

> m0.hooker <- lm(temp ~ pressure, data = Hooker)
> summary(m0.hooker)
+
+ Call:
+ lm(formula = temp ~ pressure, data = Hooker)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -1.67  -0.68   0.22   0.53   1.40
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  197.000    1.141    172.5 <2e-16 ***
+ pressure      0.996      0.016    61.5 <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

+ (Intercept) 146.6729      0.7764    188.9   <2e-16 ***
+ pressure      2.2526      0.0381     59.1   <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.806 on 29 degrees of freedom
+ Multiple R-squared:  0.992, Adjusted R-squared:  0.991
+ F-statistic: 3.5e+03 on 1 and 29 DF,  p-value: <2e-16

> print(summary(aov(m0.hooker)), digits = 5)

+              Df  Sum Sq Mean Sq F value    Pr(>F)
+ pressure       1 2272.47 2272.47  3497.9 < 2.2e-16 ***
+ Residuals     29   18.84    0.65
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Quadratic Model:

```

> m0.hooker<-lm(temp~pressure,data=Hooker)
> summary(m0.hooker)

+
+ Call:
+ lm(formula = temp ~ pressure, data = Hooker)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -1.67  -0.68   0.22   0.53   1.40
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) 146.6729      0.7764    188.9   <2e-16 ***
+ pressure      2.2526      0.0381     59.1   <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.806 on 29 degrees of freedom
+ Multiple R-squared:  0.992, Adjusted R-squared:  0.991
+ F-statistic: 3.5e+03 on 1 and 29 DF,  p-value: <2e-16

> m1.hooker<-lm(temp~pressure+I(pressure^2),data=Hooker)
> summary(m1.hooker)

+
+ Call:
+ lm(formula = temp ~ pressure + I(pressure^2), data = Hooker)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -0.7991 -0.2631 -0.0158  0.2514  0.7389
+
+ Coefficients:

```

```

+               Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    126.70162     2.11236   59.98 < 2e-16 ***
+ pressure        4.15763     0.19907   20.89 < 2e-16 ***
+ I(pressure^2)  -0.04375     0.00455   -9.61 2.3e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.396 on 28 degrees of freedom
+ Multiple R-squared:  0.998, Adjusted R-squared:  0.998
+ F-statistic: 7.31e+03 on 2 and 28 DF,  p-value: <2e-16

> anova(m0.hooker,m1.hooker)

+ Analysis of Variance Table
+
+ Model 1: temp ~ pressure
+ Model 2: temp ~ pressure + I(pressure^2)
+   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
+ 1       29 18.840
+ 2       28  4.382  1      14.46 92.39 2.29e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

> xv<-seq(15,30,by=0.01)
> fit.lin<-lm(temp~pressure, data=Hooker)
> fit.quad<-lm(temp~pressure+I(pressure^2), data=Hooker)
>
> hooker.fit.lin<-predict(fit.lin, newdata=data.frame(pressure=xv))
> hooker.fit.quad<-predict(fit.quad, newdata=data.frame(pressure=xv))
>
> par(mar=c(4,4,2,2))
> plot(temp~pressure, data=Hooker, pch=19, cex=0.6)
> lines(xv,hooker.fit.lin,col='black')
> lines(xv,hooker.fit.quad,col='black', lty=2)
> legend(16,210, c('Straight-line','Quadratic'), lty=c(1,2))

```

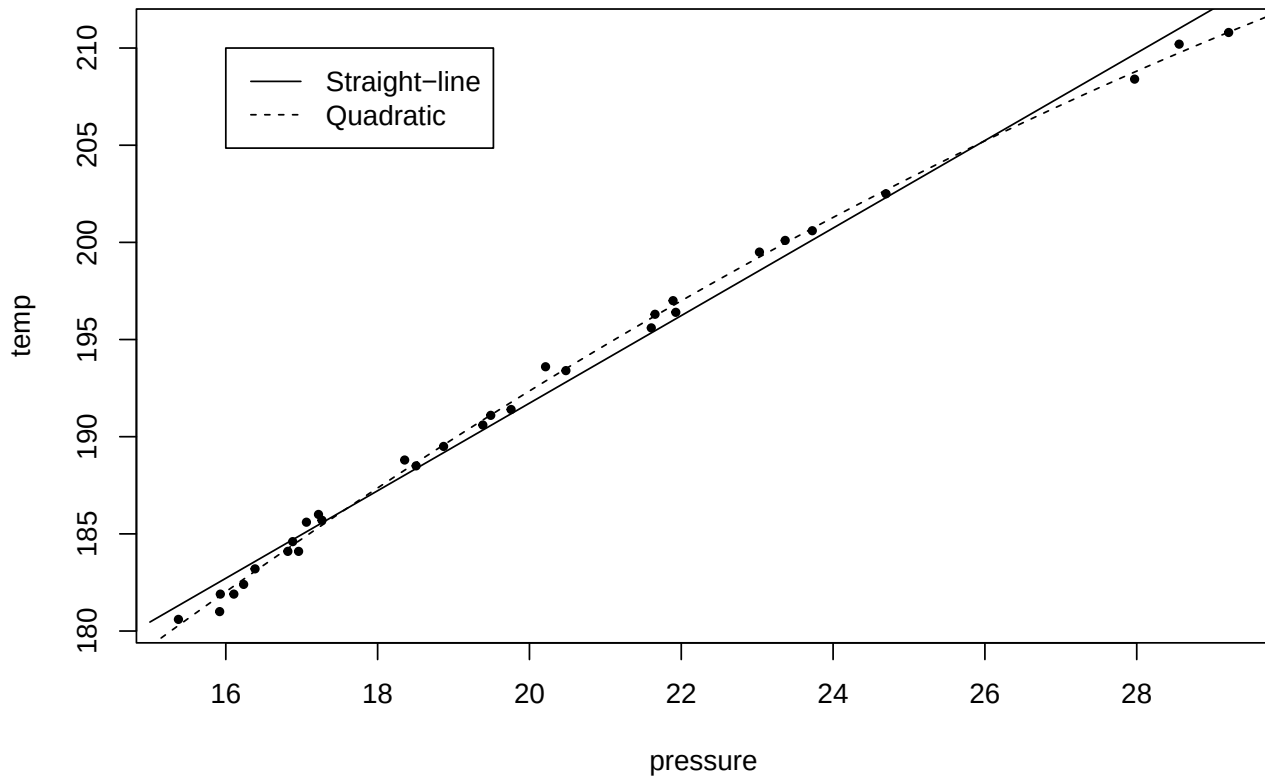


Figure 23: Hooker data: straight-line and quadratic fits

```
> res.lin<-residuals(fit.lin)
> res.quad<-residuals(fit.quad)
> par(mar=c(4,4,1,0), mfrow=c(1,2))
> plot(Hooker$pressure, res.lin, ylim=range(-2,2),
+      pch=19, cex=0.6, xlab='pressure', ylab='Residuals')
> abline(h=0, lty=2)
> text(22,2, 'Straight-line fit')
> plot(Hooker$pressure, res.quad, ylim=range(-2,2),
+      pch=19, cex=0.6, xlab='pressure', ylab='Residuals')
> abline(h=0, lty=2)
> text(22,2, 'Quadratic fit')
```

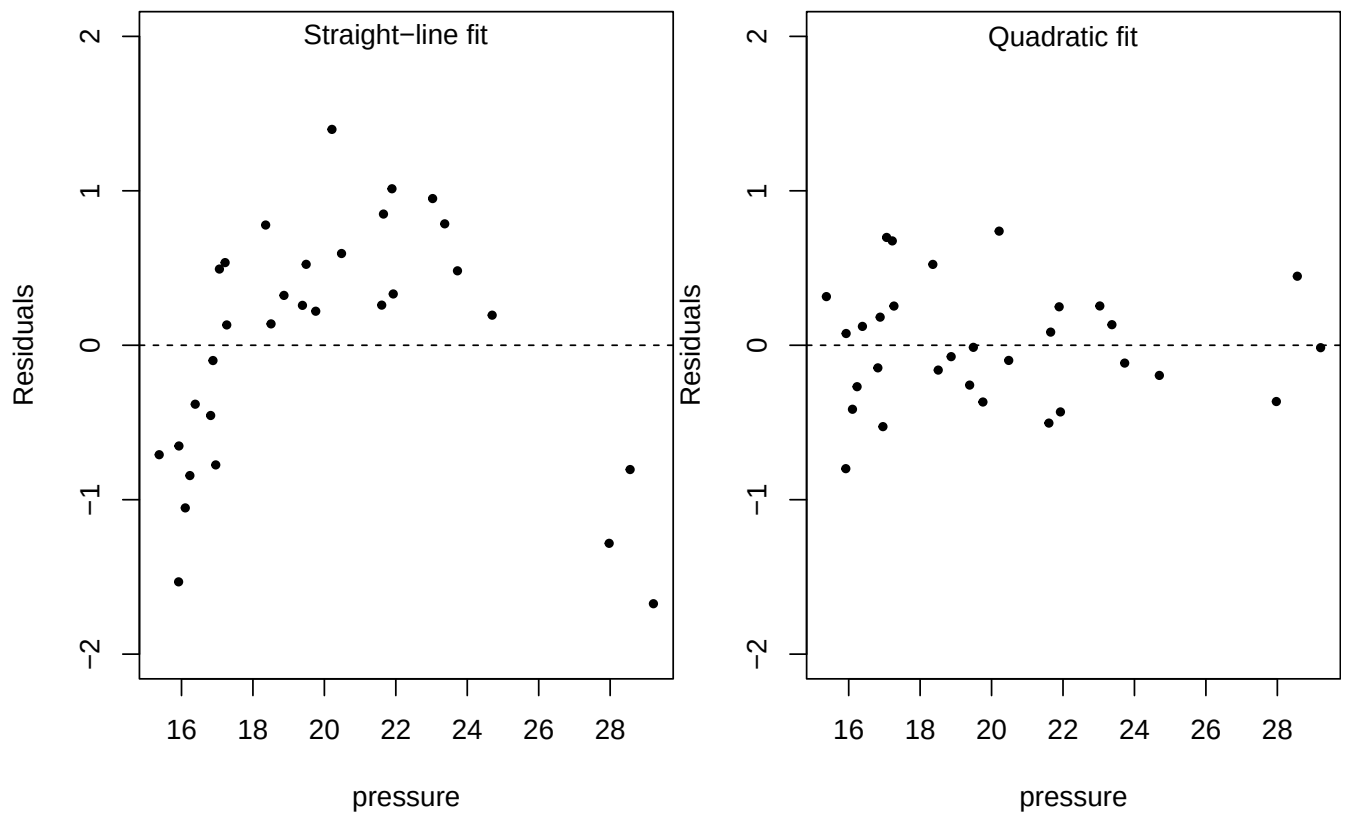


Figure 24: Hooker data: residuals from straight-line and quadratic fits

MATRIX MANIPULATIONS FOR ESTIMATION (MATERIAL NOT EXAMINABLE)

An $r \times c$ **matrix** $\mathbf{A}$ is a rectangular arrangement of numbers with r rows and c columns;

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix}$$

Recall the following rules for manipulating matrices:

- **Transpose:** the transpose operator ^{T} means “flipping” a $r \times c$ matrix into a $c \times r$ matrix. That is

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix} \iff \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{r1} \\ a_{12} & a_{22} & \cdots & a_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1c} & a_{2c} & \cdots & a_{rc} \end{bmatrix}$$

For example, if $r = 2$ and $c = 4$

$$\mathbf{A} = \begin{bmatrix} 5 & -4 & 0 & 1 \\ 3 & 5 & -2 & 0 \end{bmatrix} \iff \mathbf{A}^T = \begin{bmatrix} 5 & 3 \\ -4 & 5 \\ 0 & -2 \\ 1 & 0 \end{bmatrix}$$

A square matrix $\mathbf{A}$ is termed **symmetric** if $\mathbf{A} = \mathbf{A}^T$.

- **Matrix Multiplication:** If $\mathbf{A}$ and $\mathbf{B}$ are two matrices, where $\mathbf{A}$ is a $r_1 \times c$ matrix and $\mathbf{B}$ is a $c \times r_2$ matrix, then the product $\mathbf{A.B}$ (also written $\mathbf{AB}$) is an $r_1 \times r_2$ matrix, with (i, j) th element

$$\sum_{k=1}^c a_{ik} b_{kj} \quad i = 1, \dots, r_1, \quad j = 1, \dots, r_2.$$

For example,

$$\begin{bmatrix} 5 & -4 & 0 & 1 \\ 3 & 5 & -2 & 0 \end{bmatrix} \begin{bmatrix} 3 & 3 & -3 \\ -1 & 2 & -2 \\ 0 & -2 & 0 \\ -5 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 5 & -6 \\ 4 & 23 & -19 \end{bmatrix}$$

That is, for the first entry in the result matrix, we multiply the **first row** of the first matrix by the **first column** of the second matrix:

$$(5 \times 3) + (-4 \times -1) + (0 \times 0) + (1 \times -5) = 15 + 4 - 5 = 14$$

Note that for matrix multiplication to work, we need the first matrix to have the same number of columns as the number of rows in the second matrix. If this holds, the matrices are termed **conformable**. In general, for rectangular matrices

$$\mathbf{A.B} \neq \mathbf{B.A} \quad \text{and} \quad \mathbf{A.B.C} = \mathbf{A.(B.C)} = (\mathbf{A.B}).\mathbf{C}$$

- **Matrix Identity :** A square $k \times k$ with ones along the main diagonal, and zeros elsewhere, is termed the **identity** matrix, and denoted $\mathbf{I}_k$

$$\mathbf{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{so that} \quad \mathbf{I}_k \cdot \mathbf{A} = \mathbf{A} \quad \text{for any } k \times k \text{ matrix } \mathbf{A}$$

- **Matrix Inversion :** A square $k \times k$ matrix $\mathbf{A}$ has an **inverse**, denoted $\mathbf{A}^{-1}$ if

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}_k$$

MATRICES IN LINEAR REGRESSION

- $n \times 1$ vector $\mathbf{y} = [y_1, \dots, y_n]^\top$
- $n \times 2$ matrix $\mathbf{X}$ given by

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$$

- 2×1 Parameter estimate vector $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]^\top$

It can be shown that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

The other quantities of interest in statistical inference for the simple linear regression are also available in matrix form.

- SSE:

$$SSE = S(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})$$

- Residual error variance estimate, $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{S(\hat{\beta})}{n-2} = \frac{1}{n-2} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})$$

- Variance/Standard Errors of the Parameter estimates:

$$\text{Var}[\hat{\beta}] = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

This is a 2×2 matrix, with diagonal entries equal to the squared estimated standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$, $s_{\hat{\beta}_0}^2$ and $s_{\hat{\beta}_1}^2$ respectively.

- Fitted-values :

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{H}\mathbf{y}$$

say, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$

- Residuals:

$$\hat{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

- Prediction: if $\mathbf{x}_p = [1, x_p]^\top$, then the prediction is at the value x_p is

$$y_p = \mathbf{x}_p^\top \hat{\beta}$$

and the prediction error variances are

$$\begin{aligned} \text{Expected Value} &: \hat{\sigma}^2 \mathbf{x}_p^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_p \\ \text{Individual Value} &: \hat{\sigma}^2 (1 + \mathbf{x}_p^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_p) \end{aligned}$$