

SIMPLE LINEAR REGRESSION

We consider the model for response variable, Y , as a function of the predictor, X , observed to take the value x . Specifically we consider the model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 and β_1 are the **intercept** and **slope** parameters respectively, and ϵ is a random variable with expectation zero and variance σ^2 . In this model

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x.$$

To estimate the parameters β_0 and β_1 from data $(x_i, y_i), i = 1, \dots, n$, we use the **least-squares** criterion, and choose the values $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the **sum of squared errors**

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

It can be shown that the parameter estimates depend on the following sample summary statistics:

- Sample mean of x values:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Sample mean of y values:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- Sum of Squares SS_{xx} :

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

- Sum of Squares SS_{xy} :

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The **least-squares estimates** are:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

yielding **fitted-values**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and **residual errors** (or **residuals**)

$$\hat{e}_i = y_i - \hat{y}_i.$$

An estimate of the **residual error variance** is given by

$$\hat{\sigma}^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n - 2}$$

EXAMPLE Blood Viscosity and Packed Cell Volume

The following data are measurements of packed cell volume (PCV) and blood viscosity in samples taken from 32 hospital patients. We wish to model viscosity (y) as a function of PCV (x).

Reference: Begg, C. B. and Hearn, J. B. (1966) Components of Blood Viscosity. The relative contributions of haematocrit, plasma fibrinogen and other proteins, *Clinical Science*, **31**, 87-92.

Unit	x	y	Unit	x	y	Unit	x	y	Unit	x	y
1	40.00	3.71	9	46.75	4.14	17	51.25	4.68	25	49.50	5.12
2	40.00	3.78	10	48.00	4.20	18	50.25	4.73	26	56.00	5.15
3	42.50	3.85	11	46.00	4.20	19	49.00	4.87	27	50.00	5.17
4	42.00	3.88	12	47.00	4.27	20	50.00	4.94	28	47.00	5.18
5	45.00	3.98	13	43.25	4.27	21	50.00	4.95	29	53.25	5.38
6	42.00	4.03	14	45.00	4.37	22	49.00	4.96	30	57.00	5.77
7	42.50	4.05	15	50.00	4.41	23	50.50	5.02	31	54.00	5.90
8	47.00	4.14	16	45.00	4.64	24	51.25	5.02	32	54.00	5.90

- Sample mean of x values: $\bar{x} = 47.938$; sample mean of y values: $\bar{y} = 4.646$
- Sums of Squares

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 615.75 \quad SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 75.386$$

Thus

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0.122 \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -1.223$$

The estimate of the residual error variance is

$$\hat{\sigma}^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n - 2} = \frac{2.721}{30} = 0.091$$

```
> m0.visc<-lm(viscosity~pcv,data=Viscosity)
> summary(m0.visc)

+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.4881 -0.2215  0.0469  0.1296  0.6492
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -1.2234     0.5842   -2.09   0.045 *
+ pcv           0.1224     0.0121  10.09  3.7e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.301 on 30 degrees of freedom
+ Multiple R-squared:  0.772, Adjusted R-squared:  0.765
+ F-statistic: 102 on 1 and 30 DF, p-value: 3.73e-11
```

```
> par(mar=c(4,4,1,2))
> plot(Viscosity$viscosity~Viscosity$pcv,
+      pch=19,cex=0.8, xlab='pcv', ylab='Viscosity')
> abline(coef(m0.visc),col='red')
```

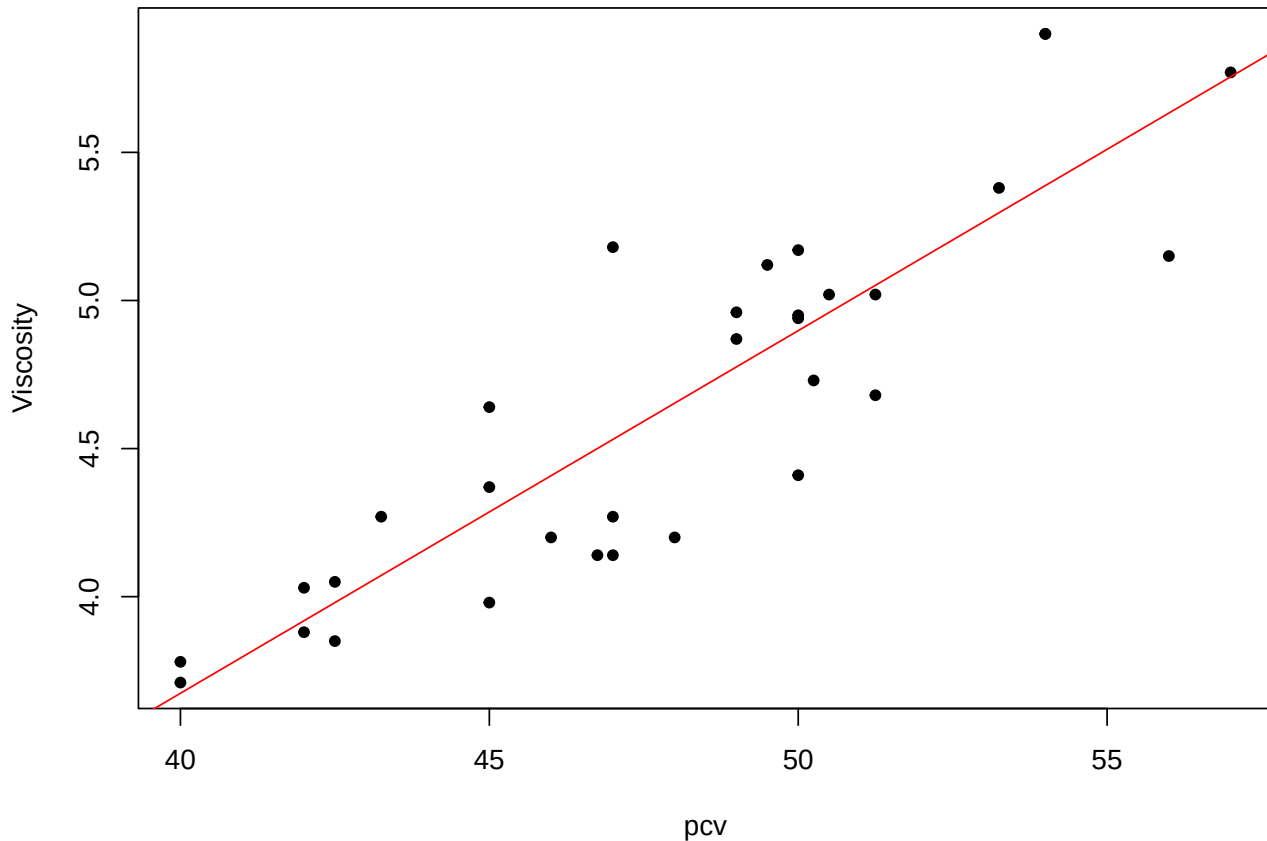


Figure 21: Viscosity data: straight-line fit

```
> sqrt(sum(residuals(m0.visc)^2)/30)
+ [1] 0.301156
```

```
> cor.test(Viscosity$pcv,Viscosity$viscosity)
+
+ Pearson's product-moment correlation
+
+ data: Viscosity$pcv and Viscosity$viscosity
+ t = 10.09, df = 30, p-value = 3.73e-11
+ alternative hypothesis: true correlation is not equal to 0
+ 95 percent confidence interval:
+ 0.764350 0.939586
+ sample estimates:
+ cor
+ 0.878818
```

EXAMPLE Coleman Report Data

Data were collected at 20 US schools, and used to examine the relationship between performance of students in the school in a verbal reasoning test and the socioeconomic status of the catchment area.

Reference: Mosteller and Tukey (1977) *Data Analysis and Regression*

School	Score	Status	School	Score	Status
	<i>y</i>	<i>x</i>		<i>y</i>	<i>x</i>
1	7.20	37.01	11	-12.86	23.30
2	-11.71	26.51	12	0.92	35.20
3	12.32	36.51	13	4.77	34.90
4	14.28	40.70	14	-0.96	33.10
5	6.31	37.10	15	-16.04	22.70
6	6.16	33.90	16	10.62	39.70
7	12.70	41.80	17	2.66	31.80
8	-0.17	33.40	18	-10.99	31.70
9	9.85	41.01	19	15.03	43.10
10	-0.05	37.20	20	12.77	41.01

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

$$T_{obs} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.927}{\sqrt{(1-0.927^2)/(20-2)}} = 10.499$$

which we must compare against the $t_{n-2} \equiv t_{18}$ distribution. For a two-tailed test at the significance level $\alpha = 0.05$, the critical values are $C_R = \pm 2.101$ (from t -tables), so the hypothesis H_0 that the true correlation is zero is **rejected**.

```
> cor.test(Coleman$status, Coleman$testscore)
+
+ Pearson's product-moment correlation
+
+ data: Coleman$status and Coleman$testscore
+ t = 10.5, df = 18, p-value = 4.2e-09
+ alternative hypothesis: true correlation is not equal to 0
+ 95 percent confidence interval:
+ 0.821826 0.971207
+ sample estimates:
+ cor
+ 0.927161
```

The regression analysis is contained below:

```
> m0.cole<-lm(testscore~status, data=Coleman)
> summary(m0.cole)
+
+ Call:
+ lm(formula = testscore ~ status, data = Coleman)
+
+ Residuals:
```

```

+      Min      1Q Median      3Q      Max
+ -8.941 -1.219 -0.055  1.951  6.989
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -50.682      5.193   -9.76  1.3e-08 ***
+ status        1.534      0.146   10.50  4.2e-09 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 3.71 on 18 degrees of freedom
+ Multiple R-squared:  0.86, Adjusted R-squared:  0.852
+ F-statistic: 110 on 1 and 18 DF, p-value: 4.2e-09

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

+              Df Sum Sq Mean Sq F value    Pr(>F)
+ status         1 1513.2 1513.21  110.23 4.199e-09 ***
+ Residuals     18  247.1   13.73
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> par(mar=c(4,4,1,2))
> plot(Coleman$testscore~Coleman$status,
+       pch=19,cex=0.8, xlab='status', ylab='testscore')
> abline(coef(m0.cole),col='red')

```

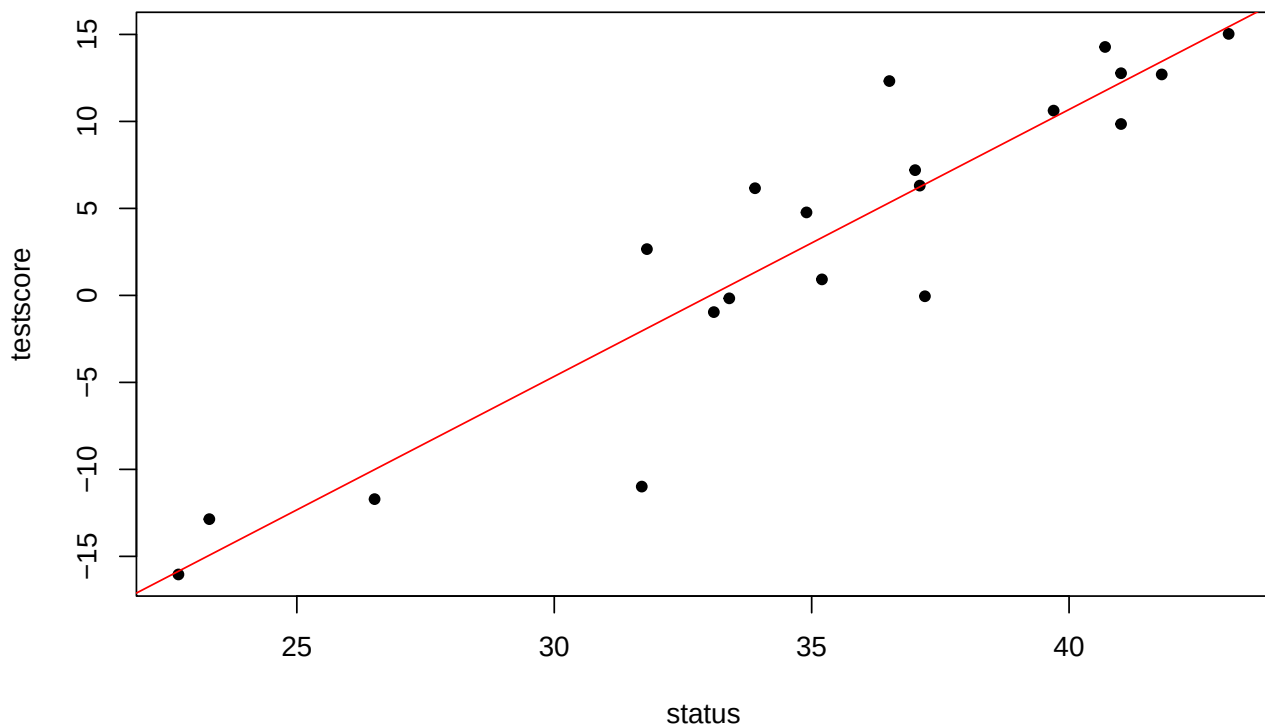


Figure 22: Coleman data: straight-line fit