Chapter 10

Simple Linear Regression and Correlation

10.1 Introduction

Aim: ⭐ To study the association among variables; ⭐ To predict outcome given covariates

Example 10.1 Altitude and Boiling point.
In the 1840s and 1850s, Forbes wanted to be able to determine the altitude from measurements of the boiling point (BP) of water. Altitude can be determined from atmospheric pressure. He collected 17 data points from different locations.

<table>
<thead>
<tr>
<th>Boiling point</th>
<th>194.5</th>
<th>194.3</th>
<th>197.9</th>
<th>198.4</th>
<th>199.4</th>
<th>199.9</th>
<th>200.9</th>
<th>201.1</th>
<th>201.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>pressure</td>
<td>20.79</td>
<td>20.79</td>
<td>22.4</td>
<td>22.67</td>
<td>23.15</td>
<td>23.35</td>
<td>23.89</td>
<td>23.99</td>
<td>24.02</td>
</tr>
<tr>
<td>100*log(pressure)</td>
<td>131.8</td>
<td>131.8</td>
<td>135</td>
<td>135.5</td>
<td>136.5</td>
<td>136.8</td>
<td>137.8</td>
<td>138</td>
<td>138.1</td>
</tr>
</tbody>
</table>

Figure 10.1: Boiling point versus pressure and 100*log(pressure)
**Questions:**

♠ How are pressure and BP related?

♠ Can pressure be predicted from BP and how well?

**Example** 10.2 *Income and Education.*

![A hypothetical data set](image.png)

Figure 10.2: Year of education versus income ($r = 0.63$)

An observational study shows the data given in Figure 10.2.
Questions:

♠ If one’s education level is 15 years, what would you estimate his/her income to be?

♠ How much is one extra year of education worth?

A professional claimed she is underpaid. How to adjust? Other possible variables include years of work, gender, rank, achievements, etc. This is a multiple regression problem.
Example 10.3 Prediction of Housing Value

Zillow.com: What is a fair market value of a house? An important variable is the size of the house or its proxy, the number of rooms.

No. Rooms versus housing value

Figure 10.3: No. of rooms versus value of house (in thousand USD) in Boston in 1970.
Questions:

— If a house has 7.5 rooms, what is the fair market value?

— What is a reasonable range for the value of the house?

— On average, how much is an extra room worth?

In reality, housing value depends on the age, distance to business center, crime rate, pollution level, tax, pupil-teacher ratio, recent sales, among others. It is a multiple regression problem.

In Ex. 10.1 — 10.3, we can see from the scatter plots that data:

— are somewhat noisy;

— have an overall linear pattern;

— are far from a defined functional form.
Example 10.4 In health science studies, many explanatory variables are collected in addition to the response. E.g.

Response: Survival time after transformation

Covariates: age, gender, blood pressure, waiting time, race etc

Purpose: Identify the risk factors and describe the association.

Purpose of regression:

♠ quantify the contribution of $X$ to $Y$

♠ summarize the association (screening variables)

♠ given $x$, predict the mean response and its associated SD
10.2 Model and Summary Statistics

**Bivariate data**: \((x_1, y_1), \ (x_2, y_2), \cdots, \ (x_n, y_n)\).

**Generic pair**: \((X, Y)\)

- \(X\) — independent variable, covariate, predictor;
- \(Y\) — dependent variable, response.

**Simple linear regression**:

\[
Y = \beta_0 + \beta_1 X + \varepsilon,
\]

- \(\beta’s\) — regression coefficients; \(\beta_0\) — intercept; \(\beta_1\) — slope.
- \(\varepsilon\) — measurement errors / part that cannot be explained by \(x\).
Data: The $i^{th}$ observation is generated from

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n.$$  

We might assume $\{\varepsilon_i\}$ are i.i.d $N(0, \sigma^2)$ (oval shape in scatter plot).
Group mean: For the linear model, we have

$E(Y|x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x,$

which is the average for the group with covariate $\approx x$ (blue line).

Group SD: Similarly, we have

$\text{var}(Y|x) = \sigma^2,$

which is the variance for the group with covariate $\approx x$.

The summary statistics are

♠ $x$-variable: $\bar{x}$ and $\text{SD}_x = \sqrt{\frac{S_{xx}}{n-1}}$ or $S_{xx} = \sum (x_i - \bar{x})^2$.

♠ $y$-variable: $\bar{y}$ and $\text{SD}_y = \sqrt{\frac{S_{yy}}{n-1}}$ or $S_{yy} = \sum (y_i - \bar{y})^2$.

♠ strength of linear association: $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$.
—sample correlation coefficient, where

\[ S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y}. \]

Example 10.5 Two-sample problems and regression

Let \( Y_1, \cdots, Y_m \) and \( Y_{m+1}, \cdots, Y_{m+n} \) be respectively the random sample from the first and second population. Let \( x_1 = \cdots = x_m = 0 \) and \( x_{m+1} = \cdots = x_{m+n} = 1 \) be the indicator for the first and second population. Consider the linear model

\[ Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} 
\beta_0 + \varepsilon_i & \text{if } i \leq m \\
\beta_0 + \beta_1 + \varepsilon_i & \text{if } i > m
\end{cases} \]

Thus, \( \beta_0 = \mu_1 \) and \( \beta_1 = \mu_2 - \mu_1 \).
Example 10.6 Time series and regression

Presented in Figure 10.5 is the US monthly unemployment rates. To predict this month $x_t$ using the previous month $x_{t-1}$, we have variables $y = x_t$ and $x = x_{t-1}$ and model it as (autoregressive or AR model)

$$x_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$$

Figure 10.5: (a) Unemployment rates. (b) $x_{t-1}$ versus $x_t$
10.3 Estimation of Model Parameters

Figure 10.6: Finding a line to pass through the data cloud.

**Method of least-squares**: Find $\beta_0$ and $\beta_1$ to min the part that can not be explained:

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

where $\beta_0$ and $\beta_1$ are the model parameters to be estimated.
which are the MLE’s for normal errors.

**Solution:** Setting derivatives to zero,

\[
\frac{\partial}{\partial \beta_0} \text{SSE} (\beta_0, \beta_1) = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0;
\]

\[
\frac{\partial}{\partial \beta_1} \text{SSE} (\beta_0, \beta_1) = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i)x_i = 0,
\]

we have

\[
\begin{cases}
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r \frac{SD_y}{SD_x}
\end{cases}
\]

**Fitted value:** \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \)

**Residual:** \( \hat{e}_i = y_i - \hat{y}_i \)

**SSE** = \( \sum_{i=1}^{n} \hat{e}_i^2 \), —also called **RSS** (Residual Sum of Squares).
**Regression Line:** \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \). It is used to predict the mean response \( \hat{y} \) for a given \( x \) value.

**Reg. Principle:** if \( x \) increases one SU, \( \hat{y} \) increases \( r \) SU:

\[
\frac{\hat{y} - \bar{y}}{SD_y} = r \frac{x - \bar{x}}{SD_x}.
\]

**Proof.** It is the same as the regression equation:

\[
\hat{y} = \bar{y} + r \frac{SD_y}{SD_x} (x - \bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 x.
\]
Example 10.7 A large scale study between math \((x)\) and physics \((y)\) test scores shows

\[
\begin{align*}
\text{avg math score } (\bar{x}) &= 75, \quad \text{SD}_x = 10, \quad r = 0.8 \\
\text{avg physics score } (\bar{y}) &= 70, \quad \text{SD}_y = 8.
\end{align*}
\]

The overall pattern of the data is of oval shape.

a) If a student’s math score is 80, guess his/her physics score;

b) What is the average physics score (and standard deviation) of the group having math score about 80?

c) If a students’s math score is 60, predict his/her physics score.
★problem a) = problem b)

**Method 1:** Regression principle

a) \( \frac{80-75}{10} = 0.5 \text{ SU in } x \implies 0.8 \times 0.5 = 0.4 \text{ SU in } y, \text{ and} \)

\[
\text{regression est.} = \bar{y} + 0.4 \text{ SD}_y = 70 + 0.4 \times 8 = 73.2
\]

c) \( \frac{60-75}{10} = -1.5 \text{ SU in } x \implies 0.8 \times (-1.5) = -1.2 \text{ SU in } y, \text{ and} \)

\[
\text{regression est.} = 70 - 1.2 \times 8 = 60.4.
\]

**Regression Effect:** In the test and retest situation, the bottom group shows overall improvement while the top group deteriorates somewhat.
E.g. Heights between two generations: tall fathers tend to associate with tall sons, but not as tall as their fathers.

**Method 2**: Regression equation

\[ \hat{\beta}_1 = r \frac{SD_y}{SD_x} = 0.8 \times \frac{8}{10} = 0.64. \]

(increasing 1 point of math increases about 0.64 in physics)

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 70 - 0.64 \times 75 = 22. \]

a) regression est. = 22 + 0.64 \times 80 = 73.2

c) regression est. = 22 + 0.64 \times 60 = 60.4
### 10.4 Estimating $\sigma^2$

The regression equation gives the mean of the group having covariate $x$. **What is the SD $\sigma$ of this group?**

A natural estimate of $\sigma^2$ is

$$
\frac{1}{n - 2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \equiv \frac{\text{SSE}}{n - 2}.
$$

due to the loss of 2 degrees of freedom.

**Computation of SSE:** It can be shown that

$$
\text{SSE} = S_{yy}(1 - r^2).
$$
**Estimator of \( \sigma \):**

\[
\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{S_{yy}(1 - r^2)}{n-2}} = \sqrt{\frac{n-1}{n-2}} SD_y \sqrt{1 - r^2},
\]

which is smaller than \( SD_y \) — the subgroup has a smaller variance.

---

**Example 10.7** (cont.) What is the likely size of the prediction error? Solution: 

\[
\hat{\sigma} \approx SD_y \sqrt{1 - r^2} = 8 \times \sqrt{1 - 0.8^2} = 4.8.
\]

(b) The group with math score about 80 has the average physics score 73.2, and \( SD 4.8 \).
10.5 Goodness-of-Fit

**Question**: How well is $x$ related to $y$?

If there is no relationship, $y_i = \beta_0 + \varepsilon_i$ and the LS estimator of $\beta_0$ minimizes

$$\sum_{i=1}^{n} (y_i - \beta_0)^2 \implies \hat{\beta}_0 = \bar{y}.$$  

The sum of squared errors (SSE) in this case is $\sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy}$.

If there is linear relationship, the unexplained variability is

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$  

Thus, the reduction in unexplained variability is

$$SS_{\text{reg}} = S_{yy} - \text{SSE} = \frac{S_{xy}^2}{S_{xx}}.$$
It is called the *sum of squares due to regression* (SSR).

**The Coefficient of Determination $R^2$:**

$$R^2 = \frac{SS_{reg}}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}.$$  

It gives the percentage of the variability of $Y$ explained by the regression on $X$. The larger, the better the fit. Note that

$$R^2 = \frac{S^2_{xy}}{S_{xx}S_{yy}} = r^2.$$
10.6 Inference of model parameters

**Standard errors:** Since the estimators are linear in \( Y \),

\[
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \sum \frac{x_i - \bar{x}}{S_{xx}} y_i
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum \left( \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right) y_i
\]

then we have:

\[
\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad \hat{\text{SE}}(\hat{\beta}_1) = \hat{\sigma} / \sqrt{S_{xx}}
\]

\[
\text{var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \quad \hat{\text{SE}}(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}
\]
Given that the estimators are normally distributed, it follows that:
\[
\frac{\hat{\beta}_0 - \beta_0}{\text{SE}(\hat{\beta}_0)} \sim t_{n-2}, \quad \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}
\]

**Confidence intervals:**

— Intercept $\beta_0$: \[\hat{\beta}_0 \pm t_{\alpha/2,n-2} \text{SE}(\hat{\beta}_0).\]

— Slope $\beta_1$: \[\hat{\beta}_1 \pm t_{\alpha/2,n-2} \text{SE}(\hat{\beta}_1).\]

The same principle applies to the hypotheses tests.

---

**Example 10.1** (cont.) For Forbes’ data, take $y = \log(Pressure)$.

\[n = 17, \bar{x} = 202.95, \quad \bar{y} = 139.60\]
\[S_{xx} = 530.78, \quad S_{yy} = 427.91, \quad S_{xy} = 475.38\]
(a) Construct the 95% CI for $\beta_1$.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{475.38}{530.78} = .8956$$

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2}} = \sqrt{\frac{2.148}{15}} = .3784$$

$$\text{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{.3784}{\sqrt{530.78}} = 0.0164$$

$$\text{df} = n - 2 = 15, \quad t_{0.025,15} = 2.13.$$ 

Thus 95% CI for $\beta_1$ is $0.896 \pm 2.13 \times 0.0164 = (.86, .93)$.

(b) Test $H_0 : \beta_1 = 0.95 \longleftrightarrow H_1 : \beta_1 \neq 0.95$

— Method 1: 0.95 is not in 95% CI, reject $H_0$;

— Method 2: $t = \frac{0.896 - 0.95}{0.0164} = -3.29$, reject $H_0$ as $|t| \geq 2.13$;

— Method 3: P-value = 2 $P(T_{15} > 3.29) = .5\%$, evidence is very
strong, reject $H_0$.

```r
> logP = 100*log10(pressure)  # define log-pressure
> fit = lsfit(BoilingPoint, logP)  # least-square fits
> ls.print(fit)  # print the summary result
Residual Standard Error=0.3792
R-Square=0.995
F-statistic (df=1, 15)=2961.547
p-value=0

|             | Estimate | Std.Err | t-value | Pr(>|t|) |
|-------------|----------|---------|---------|----------|
| Intercept   | -42.1642 | 3.3414  | -12.6189| 0        |
| X           | 0.8956   | 0.0165  | 54.4201 | 0        |
```

10.7 Predictions

Given the new value $x^*$, we would like to predict its response

$$Y^* = \beta_0 + \beta_1 x^* + \varepsilon^*, \text{ with } \text{var}(\varepsilon^*) = \sigma^2.$$
The expectation $E(Y^* | x^*) = \beta_0 + \beta_1 x^*$ is estimated as

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* = \bar{y} + \hat{\beta}_1 (x^* - \bar{x}) = \sum \left( \frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{S_{xx}} \right) y_i$$

The variance of the prediction error $(Y^* - \hat{y}^*)$ is

$$\text{var}(Y^* - \hat{y}^*) = \text{var}(Y^*) + \text{var}(\hat{y}^*) = \sigma^2 + \sigma^2 \left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right],$$

coming from two sources: $\varepsilon^*$ and estimated coefs: $\hat{\beta}_0$ and $\hat{\beta}_0$.

**SE of prediction**

$$\hat{\text{SE}}_{\text{pred}}(\hat{y}^* | x^*) = \hat{\sigma} \left[ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]^{1/2}.$$ 

A $(1 - \alpha)\%$ predictive interval of $Y^*$ is

$$\hat{y}^* \pm t_{\alpha/2, n-2} \hat{\text{SE}}_{pred}(\hat{y}^* | x^*).$$
Ex. 10.1 (cont.) Construct the 95% predictive interval for the log-pressure at $x^* = 205$.

Recall that $\hat{\beta}_1 = 0.896$. The estimated slope is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 139.60 - 0.896 \times 202.95 = -42.24$$

The predicted value is

$$\hat{y}^* = -42.24 + 0.896 \times 205 = 141.44.$$

The SE of the prediction is given by

$$\widehat{SE}_{pred}(\hat{y}^*|x^*) = \hat{\sigma} \left[ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]^{1/2}$$

$$= 0.3784 \times \left[ 1 + \frac{1}{17} + \frac{(205 - 202.95)^2}{530.78} \right]^{1/2} = 0.391$$
The 95% predictive interval at \( x^* = 205 \) is

\[
141.44 \pm 2.13 \times .391 = (140.6, 142.3)
\]

```r
> fit = lm(logP ~ BoilingPoint) # a different way of fit model
> summary(fit) # a different way of summary

Call:
lm(formula = logP ~ BoilingPoint)

Residuals:
     Min      1Q  Median      3Q     Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.16418   3.34136  -12.62   2.17e-09 ***
BoilingPoint  0.89562   0.01646   54.42   < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.9946
F-statistic: 2962 on 1 and 15 DF,  p-value: < 2.2e-16
```
> new = data.frame(BoilingPoint = seq(200, 210, 5))
   # variable name has to be the same
> predict(fit, new, interval = "prediction")  # prediction interval
  fit  lwr  upr
  1 136.9594 136.1214 137.7974
  2 141.4375 140.6028 142.2721  # This is for BoilingPoint = 205
  3 145.9155 145.0480 146.7831

### try also the following
predict(fit, new, se.fit = TRUE)  # give the SE of the fit
predict(fit, new, interval = "confidence")
  # confidence interval for the group mean

**Ex. 10.7** (cont.)

avg math score ($\bar{x}$) = 75, \(SD_x = 10\), \(r = 0.8\)

avg physics score ($\bar{y}$) = 70, \(SD_y = 8\). \(n = 900\)

(a) Predict a student’s physics score if his math score is 85 and attach the size of the prediction error.

\[
\frac{85 - 75}{10} = 1 \text{ SU in } x \rightarrow 0.8 \times 1 = 0.8 \text{ SU in } y
\]

regression estimate: \(y = 70 + 0.8 \times 8 = 76.4\).
\[ \hat{\sigma} = \sqrt{\frac{n-1}{n-2}} \text{SD}_y \sqrt{1 - r^2} = \sqrt{\frac{899}{898}} \times 8 \times \sqrt{1 - 0.8^2} = 4.80 \]

\[ \text{SE}_{pred} = \hat{\sigma} \left[ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]^{1/2} = 4.80 \left[ 1 + \frac{1}{n} + \frac{(85 - 75)^2}{(n - 1) \text{SD}_x^2} \right]^{1/2} \]

\[ = 4.80 \left[ 1 + \frac{1}{900} + \frac{1}{899} \right] = 4.81. \]

Thus, the prediction is 76.4, give or take 4.81.

(b) Of those whose math score is about 85, what percentage of them scored above average in physics?

The physics score of the students in this subgroup is \( \approx N(76.4, 4.80^2) \). The percentage of this subgroup above the average is given by:

\[ P\{ PS > 70 \} = 1 - \Phi \left( \frac{70 - 76.4}{4.80} \right) = 1 - \Phi(-1.333) = 90.9\%. \]

(note that the overall percentage of students who scored over 70 in physics was 50%).

(c) On average, how much each point increase in math contributes to the physics score? Answer this question through a 95% CI.

\[ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r \frac{\text{SD}_y}{\text{SD}_x} = 0.8 \times \frac{8}{10} = 0.64. \]
\[
\hat{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{4.80}{SD_x \sqrt{n - 1}} = \frac{4.80}{10 \sqrt{899}} = 0.0160
\]

Thus, the 95\% CI for \( \beta_1 \) is

\[
0.64 \pm t_{0.025, 898} \times 0.160 = 0.64 \pm 0.0314 = [0.61, 0.67].
\]

(d) Would increasing 10 points of the math score increase more than 6 points of the physics score?

\[
H_0 : \beta_1 \leq 0.6 \quad \leftrightarrow \quad H_1 : \beta_1 > 0.6.
\]

\[
t = \frac{\hat{\beta}_1 - \beta_{10}}{SE} = \frac{0.64 - 0.6}{0.0160} = 2.50
\]

P-value = 1 - pt(2.5, 898) = .63%.

Strong evidence against \( H_0 \) \( \implies \) the answer is ”yes.”

(e) Is it reasonable to assume that increasing 10 points of the math score increases on average about 6.5 points of the physics score?

\[
H_0 : \beta_1 = 0.65 \quad \leftrightarrow \quad H_1 : \beta_1 \neq 0.65
\]
\[ t = \frac{0.64 - 0.65}{0.0160} = -0.625, \text{ P-value} = 53.2\%. \]

Weak evidence against \( H_0 \) \( \implies \) accept \( H_0 \) \( \implies \) the answer is "yes."

### 10.8 Correlation

Recall that the covariance and correlation between \( X \) and \( Y \) are

\[ \text{cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) \quad \text{and} \quad \rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}. \]

**Sample covariance**

\[ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{S_{xy}}{n-1}. \]

**Sample correlation:**

\[ r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \]

**Properties:** Like population corr., \( r \) has the following properties:

\[ -1 \leq r \leq 1; \]
The larger the $|r|$, the stronger the linear association;

$r > 0$: positive association; $r < 0$: negative association;

$r$ is independent of unit.

Figure 10.7: Simulated data with sample correlations 0.026, -0.595, -0.990, 0.241, 0.970, 0.853.
Some Caveats:

— $r \approx 0$ does not mean no association.

— $r$ does not provide the evidence of causation.

♣ correlation between shoe size and reading skills is high for school kids. Buying larger shoes? Confounding factor: age.
beer consumption and lake water level. **Confounding factor:** temperature.

Correlation is sometimes computed based on averages. Such a correlation is called **ecological correlation.** It overstates the strength of the association.

Figure 10.9: Left panel: Income vs level of education averaged in groups (3, $r=0.902$) instead of individuals (30, $r=0.572$). Right panel: Overall cancer risk versus per capita daily food intake (from G.Dallal c1999).
When the bivariate data are not normal, the correlation and other summary statistics are not sufficient. In addition, the correlation might not be very meaningful.

Figure 10.10: The scatter plot for the hypothetical data presented in Figure 10.2, with education between 12 years and 16 years removed. The correlation increases from 0.63 to 0.73. The marginal distributions are not normal either.
Figure 10.11: Each data set has a correlation coefficient of 0.7 (from G.Dallal c1999).
Questions: Are $X$ and $Y$ correlated? The null hypothesis is $H_0 : \rho = 0$. There are three kinds of alternative hypotheses:

(a) $H_1 : \rho > 0$.  
(b) $H_1 : \rho < 0$.  
(c) $H_1 : \rho \neq 0$.

Exact test: Intuitively, reject $H_0$ in (a) when $r$ is large and reject $H_0$ in (c) when $|r|$ is large. This is equivalent to using

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which is monotonic in $r$.

Null distribution: When the bivariate data are normal, under $H_0 : \rho = 0$, $T \sim t_{n-2}$.

Example 10.8 A random sample of 45 measurements shows the correlation between the $x$ (expression of a protein) and $y$ (survival time) variables to be 0.29.

(a) Is there substantial evidence that $x$ and $y$ are correlated at the 5% significance level?

Testing problem: $H_0 : \rho = 0 \iff H_1 : \rho \neq 0$. The test statistic

$$t = \frac{\sqrt{45 - 2} \times 0.29}{\sqrt{1 - 0.29^2}} = 2.0.$$
Table 10.1: **Summary of exact test for** $H_0: \rho = 0$

<table>
<thead>
<tr>
<th>Problem</th>
<th>Reject region</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>$t &gt; t_{\alpha,n-2}$</td>
<td>$P(T_{n-2} &gt; t) =$</td>
</tr>
<tr>
<td>(b)</td>
<td>$t &lt; -t_{\alpha,n-2}$</td>
<td>$P(T_{n-2} &lt; t) =$</td>
</tr>
<tr>
<td>(c)</td>
<td>$</td>
<td>t</td>
</tr>
</tbody>
</table>

Hence, the P-value $= P(|T_{43}| > 2.0) = 5.2\%$. We don’t have strong enough evidence against $H_0$, namely, $x$ and $y$ are uncorrelated.

(b) What **assumption(s)** did you make in the above inference?

The data are a random sample from the bivariate normal dist.
**Fisher transformation**: When \((X_1, Y_1), \ldots, (X_n, Y_n)\) is a sample from a bivariate normal distribution, then

\[
V = \frac{1}{2} \ln \frac{1 + r}{1 - r} \sim \mathcal{N}(\mu_r, \frac{1}{n - 3}), \quad \text{where } \mu_r = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}.
\]

\((1 - \alpha)\)-CI for \(\mu_r\): \(v \pm z_{\alpha/2}/\sqrt{n - 3} = (c_1, c_2)\), or for \(\rho\):

\[
\left(\frac{\exp(2c_1) - 1}{\exp(2c_1) + 1}, \frac{\exp(2c_2) - 1}{\exp(2c_2) + 1}\right).
\]

Note that \(V\) is a monotonic function in \(r\), the *sample correlation*. One can test \(H_0 : \rho = \rho_0\) against

\((a)H_1 : \rho > \rho_0, \quad (b)H_1 : \rho < \rho_0, \quad (c)H_1 : \rho \neq \rho_0.\)

The test statistic is

\[
Z = \sqrt{n - 3}(V - \mu_0); \quad \mu_0 = \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0}.
\]

**Example 10.9** Based on a sample of size 103, the sample correlation between the income and education was found to be 0.4.
(a) Construct the 90% CI for the population correlation.

The Fisher transformation

\[ v = \frac{1}{2} \ln \frac{1 + .4}{1 - .4} = 0.4236 \]

\( \alpha = 0.1, \ z_{\alpha/2} = 1.645 \), 90% CI for the Fisher transform of \( \rho \) is

\[ 0.4236 \pm 1.645/\sqrt{100} = (0.2591, 0.5881) \].

Hence, by converting this into \( \rho \), the 90% CI for \( \rho \) is

\[ \exp(2 \times 0.2591) - 1, \exp(2 \times 0.5881) - 1 \]

\[ \exp(2 \times 0.2591) + 1, \exp(2 \times 0.5881) + 1 \]

\[ = (0.2535, 0.5285) \].

(b) Is there any substantial evidence that the population correlation between the income and education is at least .2?

The problem: \( H_0 : \rho \leq .20 \quad \leftrightarrow \quad H_1 : \rho > .20 \). Note that \( \mu_0 = \frac{1}{2} \ln \frac{1+.2}{1-.2} = 0.2027 \). Hence, the test statistic

\[ z = (0.4236 - 0.2027)\sqrt{103 - 3} = 2.209. \]

Thus, the P-value is \( P(Z > 2.209) = 1 - \Phi(2.209) = 1.36\% \). We have substantial evidence that the population correlation > .2.
Ex.10.1. (cont.) Test the null hypothesis that BoilingPoint and logP are uncorrelated and construct 95% confidence interval for the population correlation.

```
> cor.test(BoilingPoint, logP)

Pearson's product-moment correlation

data:  BoilingPoint and logP
t = 54.42, df = 15, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
  0.9928242 0.9991143
sample estimates:
   cor
  0.9974771
```