Correlation

The correlation coefficient $r$ is measure of the strength of the linear relationship between two variables.

$$ r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) $$

$$ = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} $$

Notice that the correlation is not affected by linear transformations of the data (such as changing the scale of measurement).
Correlation Plots

\[ r = 0.97 \]

Correlation Plots

\[ r = 0.21 \]
Correlation Plots

\[ r = -0.21 \]

\[ r = 1 \]
Correlation Plots

$r = -1$

$r = 0$
Correlation Plots

\[ r = 0.97 \]

\[ r = -0.97 \]
Summary of Correlation

- The correlation coefficient $r$ measures the strength of the linear relationship between two quantitative variables, on a scale from $-1$ to $1$.
- The correlation coefficient is $-1$ or $1$ only when the data lies perfectly on a line with negative or positive slope, respectively.
- If the correlation coefficient is near one, this means that the data is tightly clustered around a line with a positive slope.
- Correlation coefficients near 0 indicate weak linear relationships.
- However, $r$ does not measure the strength of nonlinear relationships.
- If $r = 0$, rather than $X$ and $Y$ being unrelated, it can be the case that they have a strong nonlinear relationship.
- If $|r|$ is close to 1, it may still be the case that a nonlinear relationship is a better description of the data than a linear relationship.

Simple Linear Regression

Simple linear regression is the statistical procedure for describing the relationship between an quantitative explanatory variable $X$ and a quantitative response variable $Y$ with a straight line.

In simple linear regression, the regression line is the line that minimizes the sum of the squared residuals.
Riley Larget is my son. Below is a plot of his height versus his age, from birth to 8 years.

The plot indicates that it is not reasonable to model the relationship between age and height as linear over the entire age range, but it is fairly linear from age 2 years to 8 years (24–96 months).
Finding a “best” linear fit

- Any line we can use to predict $Y$ from $X$ will have the form $Y = b_0 + b_1X$ where $b_0$ is the intercept and $b_1$ will be the slope.
- The value $\hat{y} = b_0 + b_1x$ is the predicted value of $Y$ if the explanatory variable $X = x$.
- In simple linear regression, the predicted values form a line. (In more advanced forms of regression, we can fit curves or fit functions of multiple explanatory variables.)
- For each data point $(x_i, y_i)$, the residual is the difference between the observed value and the predicted value, $y_i - \hat{y}_i$.
- Graphically, each residual is the positive or negative vertical distance from the point to the line.
- Simple linear regression identifies the line that minimizes the residual sum of squares, $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. 
Least Squares Regression

We won’t derive them, but there are simple formulas for the slope and intercept of the least squares line as a function of the sample means, standard deviations, and the correlation coefficient.

\[ Y = b_0 + b_1 X \]

\[ b_1 = r \times \frac{s_y}{s_x} \]

\[ b_0 = \bar{y} - b_1 \bar{x} \]

A Special Case

Consider the predicted value of an observation \( X = \bar{x} \).

\[ \hat{y} = b_0 + b_1 \bar{x} \]

\[ = (\bar{y} - b_1 \bar{x}) + b_1 \bar{x} \]

\[ = \bar{y} \]

So, the regression line always goes through the point \((\bar{x}, \bar{y})\).
The General Case

Let \( X = \bar{x} + zs_x \), so \( X \) is \( z \) standard deviations above the mean.

\[
\hat{y} = b_0 + b_1(\bar{x} + zs_x) \\
= (\bar{y} - b_1 \bar{x}) + b_1 \bar{x} + b_1 zs_x \\
= \bar{y} + \left( \frac{sy}{s_x} \right) z s_x \\
= \bar{y} + (rz) \times s_y
\]

Notice that if \( X \) is \( z \) SDs above the mean, we predict \( Y \) to be only \( rz \) SDs above the mean.

In the typical situation, \(|r| < 1\), so we predict the value of \( Y \) to be closer to the mean (in standard units) than \( X \). This is called the regression effect.

Riley (cont.)

```R
> n = length(age2)
> mx = mean(age2)
> sx = sd(age2)
> my = mean(height2)
> sy = sd(height2)
> r = cor(age2, height2)
> print(c(mx, sx, my, sy, r, n))
[1] 61.6625000 21.8661649 45.6718750 5.4829043 0.9990835 16.0000000
> b1 = r * sy/sx
> b0 = my - b1 * mx
> print(c(b0, b1))
[1] 30.2493290 0.2505185
```

(Riley’s predicted height in inches) = 30.25 + 0.25 \times (Riley’s age in months)
Using R

```r
> age2 = age[(age > 23) & (age < 97)]
> height2 = height[(age > 23) & (age < 97)]
> fit2 = lm(height2 ~ age2)
> summary(fit2)
Call:
  lm(formula = height2 ~ age2)
Residuals:
   Min      1Q  Median      3Q     Max
  -0.29911 -0.19291 -0.03355  0.21982  0.46334
Coefficients:     Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.249329   0.186745 161.98  < 2e-16 ***
age2        0.250519   0.002869   87.33  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.2429 on 14 degrees of freedom
Multiple R-Squared: 0.9982, Adjusted R-squared: 0.998
F-statistic: 7627 on 1 and 14 DF,  p-value: < 2.2e-16
```

Riley—Plot of Data

![Plot of Data](chart.png)
Riley — A Residual Plot

```r
> plot(fitted(fit2), residuals(fit2))
> abline(h = 0)
```

Riley — Interpretation

We can interpret the slope to mean that from age 2 to 8 years, Riley grew an average of about 0.25 inches per year month, or about 3 inches per year.

The intercept is the predicted value when \( X = 0 \), or Riley’s height (length) at birth. This interpretation may not be reasonable if 0 is out of the range of the data.
Residual Standard Deviation

The residual standard deviation, $s_{Y|X}$, is a measure of a typical size of a residual.

Its formula is

$$s_{Y|X} = \sqrt{\frac{SS(\text{resid})}{n-2}}$$

Notice that in simple linear regression, there are $n-2$ degrees of freedom, in contrast to our formula $n-1$.

The reason is that our model for the mean uses two parameters — it takes two points to determine a line.

Frogs

In a study on oocytes (developing egg cells) from the frog *Xenopus laevis*, a biologist injects individual oocytes from the same female with radioactive leucine, to measure the amount of leucine incorporated into protein as a function of time. (See Exercise 12.3 on page 536.)
R — Entering the data

Here is some R code to create a data frame `frog` with the time and leucine variables.

```r
> frog = data.frame(time = seq(0, 60, by = 10), leucine = c(0.02, 0.25, 0.54, 0.69, 1.07, 1.5, 1.74))
> attach(frog)
> frog
  time leucine
  1    0    0.02
  2   10    0.25
  3   20    0.54
  4   30    0.69
  5   40    1.07
  6   50    1.50
  7   60    1.74
```

R — Plotting the data

```r
> plot(time, leucine)
```

![Plot of time vs. leucine](image-url)


R — Fitting a model

```r
> fit = lm(leucine ~ time)
> summary(fit)
Call:
  lm(formula = leucine ~ time)

Residuals:
   1       2       3       4       5       6       7
  0.0675  0.0050  0.0025 -0.1400 -0.0525  0.0850  0.0325

Coefficients:
            EstimateStd. Error t valuePr(>|t|)
(Intercept) -0.047500 0.057192  -0.831 0.444
           time 0.029250 0.001586 18.440 8.63e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 0.08393 on 5 degrees of freedom
Multiple R-Squared: 0.9855,Adjusted R-squared: 0.9826
F-statistic: 340 on 1 and 5 DF,  p-value: 8.628e-06
```

R — Verifying textbook formulae

```r
> n = nrow(frog)
> mx = mean(time)
> sx = sd(time)
> my = mean(leucine)
> sy = sd(leucine)
> r = cor(time, leucine)
> c(mx, sx, my, sy, r)
[1] 30.0000000 21.6024690 0.8300000 0.6365009 0.9927280
> b1 = r * sy/sx
> b0 = my - b1 * mx
> syx = sqrt(sum(residuals(fit)^2)/(n - 2))
> seb1 = syx/sx/sqrt(n - 1)
> ts = b1/seb1
> c(b0, b1, syx, seb1, ts)
[1] -0.047500000 0.029250000 0.083934498 0.001586213 18.440147363
```
### R — A Residual Plot

```r
> plot(fitted(fit), residuals(fit), xlab = "Fitted Values", ylab = "Residuals")
> abline(h = 0)
```

There might be some nonlinearity....

---

### R — Plotting the regression line

```r
> plot(time, leucine, pch = 16)
> abline(fit)
```

---
Fitting a Quadratic Model

We add an additional parameters to fit a curve. Fitting a quadratic curve is a reasonable choice (unless there was a scientific reason to fit an exponential curve, for example).

Here is a quadratic model.

\[ (\text{leucine}) = \beta_0 + \beta_1(\text{time}) + \beta_2(\text{time})^2 \]

The residual sum of squares will cannot be larger when fitting a quadratic model instead of a linear model — lines are special cases of parabolas. But, we want to be careful not to overfit.

R — Fitting a Curve

```r
> fit2 <- lm(leucine ~ time + I(time^2))
> summary(fit2)
Call:
  lm(formula = leucine ~ time + I(time^2))
Residuals:
     1       2       3       4       5       6       7
-0.004524  0.005000  0.045714 -0.082381 -0.009286  0.085000 -0.039524
Coefficients:  
    Estimate  Std. Error  t value  Pr(>|t|)  
(Intercept) 2.452e-02  5.822e-02   0.421 0.6952  
time 2.061e-02  4.545e-03   4.534 0.0105 *  
I(time^2) 1.441e-04  7.278e-05   1.979 0.1189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0667 on 4 degrees of freedom
Multiple R-Squared: 0.9927,     Adjusted R-squared: 0.989
F-statistic: 271.2 on 2 and 4 DF,  p-value: 5.359e-05
```
**R — Plotting the Curve**

```r
> plot(time, leucine, pch = 16)
> abline(fit)
> new = predict(fit2, data.frame(time = 0:60))
> lines(0:60, new, lty = 2)
```

**R — Residual Plot for Quadratic Model**

```r
> plot(fitted(fit2), residuals(fit2), xlab = "Fitted Values", ylab = "Residuals")
> abline(h = 0)
```
Tree Example

We want to be able to estimate the age of trees in the Amazon by their diameter. Use the carbon dating as a proxy for the actual age. (See exercise 12.43 on page 575.)

```r
> amazon = read.table("amazon.txt", header = T)
> attach(amazon)
> amazon
given
diameter   age
 1   180 1372
 2   120 1167
 3   100  895
 4   225  842
 5   140  722
 6   142  657
 7   139  582
 8   110  562
 9   150  562
10   150  552
11   115  512
12   140  512
13   180  455
14   100  352
15   112  352
16    82  249
17    97  227
18   118  249
19   130  227
20   110  172
```
Amazon — Plot

```R
> plot(diameter, age, pch = 16)
```

![Scatter plot of diameter vs. age](chart.png)

Amazon — Fitting a Line

```R
> fit = lm(age ~ diameter)
> summary(fit)
```

```
Call:
  lm(formula = age ~ diameter)

Residuals:
     Min      1Q  Median      3Q     Max
-325.22  -140.67   -81.10    63.47   658.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.770     265.148   -0.071    0.9443
  diameter    4.392      1.948    2.254    0.0369 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 288.7 on 18 degrees of freedom
Multiple R-Squared: 0.2202,   Adjusted R-squared: 0.1769
F-statistic: 5.082 on 1 and 18 DF,  p-value: 0.03687
```
Amazon — Residual Plot

```r
> plot(fitted(fit), residuals(fit), xlab = "Fitted Values", ylab = "Residuals")
> abline(h = 0)
```

![Residual Plot](image)

Amazon — Normal Probability Plot

```r
> qqnorm(residuals(fit))
```

![Normal Probability Plot](image)
Amazon — Log Transformation

\[ \texttt{fit2 = lm(log(age) ~ diameter)} \]
\[ \texttt{summary(fit2)} \]
Call:
\[ \text{lm(formula = log(age) ~ diameter)} \]

Residuals:

\[
\begin{array}{c|c|c|c|c|c}
\text{Min} & \text{1Q} & \text{Median} & \text{3Q} & \text{Max} \\
-0.84455 & -0.29119 & -0.01418 & 0.25096 & 0.98499 \\
\end{array}
\]

Coefficients:

\[
\begin{array}{c|c|c|c|c|c|c|c|c}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
\text{(Intercept)} & 5.055228 & 0.465050 & 10.870 & 2.44e-09 & *** \\
diameter & 0.008516 & 0.003417 & 2.492 & 0.0227 & * \\
\end{array}
\]

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5063 on 18 degrees of freedom
Multiple R-Squared: 0.2566, Adjusted R-squared: 0.2153
F-statistic: 6.212 on 1 and 18 DF, p-value: 0.02267

Amazon — Residual Plot

\[ \texttt{plot(fitted(fit2), residuals(fit2), xlab = "Fitted Values", ylab = "Residuals")} \]
\[ \texttt{abline(h = 0)} \]
Amazon — Normal Probability Plot

> qnorm(residuals(fit2))

Normal Q–Q Plot

One Last Example

FEV (forced expiratory volume) is a measure of lung capacity and strength. It increases as children grow larger.

The next page has two separate residual plots, one of the residuals versus height from a fit of FEV versus height and one of the residuals versus height from a fit of log(FEV) versus height.

Notice that the residuals in the first plot have a fan shape but those in the second are more evenly spread out. Also, the first plot shows a nonlinear trend that is not present in the second plot.
One Last Example

Original Data

Transformed Data

Statistics 371, Fall 2004