Correlation

The correlation coefficient $r$ is a 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

Regression

Bret Larget

Department of Statistics

University of Wisconsin - Madison

December 10, 2003
Correlation Plots

$r = 1$

Correlation Plots

$r = 0$

Correlation Plots

$r = -0.21$

Correlation Plots

$r = -1$
**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.
Riley

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 $Y = b_0 + b_1 X$ where $b_0$ is the intercept and $b_1$ will be the slope.
- The value $\hat{y} = b_0 + b_1 x$ 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, $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. 

Riley

Riley Largent is my son. Below is a plot of his height versus his age, from birth to 8 years.
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 \((\bar{x}, \bar{y})\).

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.

\[
Y = b_0 + b_1 X \\
b_1 = r \frac{s_y}{s_x} \\
b_0 = \bar{y} - b_1 \bar{x}
\]

Riley (cont.)

\[
> n = length(age2) \\
> mx = mean(age2) \\
> sx = sd(age2) \\
> my = mean(height2) \\
> sy = sd(height2) \\
> r = cor(age2, height2) \\
> print(c(mx, sx, my, sy, r, n)) \\
[1] 61.5625000 21.8661649 5.4829043 0.9990835 16.0000000 \\
> b1 = r * sy/sx \\
> b0 = my - b1 * mx \\
> print(c(b0, b1)) \\
[1] 30.2493290 0.2505185
\]

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

The General Case

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

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

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—Plot of Data

![Plot of Data](image)

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
-0.2991  0.1929  0.0336  0.2198  0.4633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.24932    0.18674   161.98  < 2e-16 ***
age2         0.25052    0.00286     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—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.

Riley—A Residual Plot

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

![Residual Plot](image)
**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.)

---

**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.
**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)
```

**R — Fitting a model**

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

**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)
> b1 = r * sy/sx
> b0 = my - b1 * mx
> sx2 = sqrt(sum(residuals(fit)^2)/(n - 2))
> seb1 = sx2/sx/sqrt(n - 1)
> ts = b1/seb1
> c(b0, b1, sx2, seb1, ts)
```

---

Statistics 371, Fall 2003
R — Plotting the regression line

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

R — A Residual Plot

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

There might be some nonlinearity ....

R — Fitting a Curve

```r
> fit2 = lm(leucine ~ time + I(time^2))
> summary(fit2)
```

```
Call:
  lm(formula = leucine ~ time + I(time^2))

Residuals:
   Min     1Q   Median     3Q    Max
-0.004524  0.005000  0.045714 -0.082381 -0.009286  0.085000 -0.039524

Coefficients:                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.02452    0.05822    0.421  0.69522
  time           0.02061    0.00454    4.534  0.01050 *
I(time^2)      -0.00144    0.00073    1.979  0.11889
---
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, Adj. R-squared: 0.989
F-statistic: 271.2 on 2 and 4 DF,  p-value: 5.359e-05
```

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 — Residual Plot for Quadratic Model

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

Tree Example

```r
> amazon = read.table("amazon.txt", header = T)
> attach(amazon)
> amazon
   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
```

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)
```

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.)
> 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  5.148    -3.63    0.001 **
 diameter     4.392  1.948     2.25    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

> plot(diameter, age, pch = 16)

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

> qqnorm(residuals(fit))

Normal Q-Q Plot

Sample Quantiles

Theoretical Quantiles

> plot(fitted(fit), residuals(fit), xlab = "Fitted Values", ylab = "Residuals")
> abline(h = 0)
Amazon — Residual Plot

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

![Residual Plot](image)

Statistics 371, Fall 2003

Amazon — Log Transformation

```r
> fit2 = lm(log(age) ~ diameter)
> summary(fit2)
Call: lm(formula = log(age) ~ diameter)
Residuals:
    Min 1Q Median 3Q Max
-0.84455 -0.29119 -0.01418 0.25096 0.98499
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.055228   0.465050 10.870   2.44e-09 ***
diameter   0.008516   0.003417  2.492   0.0227 *
---
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
```

Statistics 371, Fall 2003

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.

Amazon — Normal Probability Plot

```r
> qqnorm(residuals(fit2))
```

![Normal Q-Q Plot](image)

Statistics 371, Fall 2003
One Last Example

Statistics 371, Fall 2003