12.1 Introduction
12.2 Correlation coefficient $r$
12.3 Fitted regression line

**Linear Regression**

![Graph showing a scatter plot with a fitted regression line representing HDL (mg/dL) vs. Waist Circumference (cm).]
Two continuous variables

- We will relate $Y$ to another continuous variable $X$.
- First we will measure how linearly related $Y$ and $X$ are using the correlation.
- Then we will model $Y$ vs. $X$ using a line.
- The data arrive as $n$ pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Each pair $(x_i, y_i)$ can be listed in a table and is a point on a scatterplot.
Example 12.1.1 Amphetamine and consumption

Amphetamines suppress appetite. A pharmacologist randomly allocated $n = 24$ rats to three amphetamine dosage levels: 0, 2.5, and 5 mg/kg. She measured the amount of food consumed (gm/kg) by each rat in the 3 hours following.

<table>
<thead>
<tr>
<th>Table 12.1.1 Food consumption ($Y$) of rats (gm/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = \text{Dose of amphetamine (mg/kg)}$</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>112.6</td>
</tr>
<tr>
<td>102.1</td>
</tr>
<tr>
<td>90.2</td>
</tr>
<tr>
<td>81.5</td>
</tr>
<tr>
<td>105.6</td>
</tr>
<tr>
<td>93.0</td>
</tr>
<tr>
<td>106.6</td>
</tr>
<tr>
<td>108.3</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>SD</td>
</tr>
<tr>
<td>No. of animals</td>
</tr>
</tbody>
</table>
Example 12.1.1 Amphetamine and consumption

How does $Y$ change with $X$? Linear? How strong is linear relationship?
Environmental pollutants can contaminate food via the growing soil. Naturally occurring silicon in rice may inhibit the absorption of some pollutants. Researchers measured $Y$, amount of arsenic in polished rice ($\mu$g/kg rice), & $X$, silicon concentration in the straw (g/kg straw), of $n = 32$ rice plants.
Example 12.2.1 Length and weight of snakes

In a study of a free-living population of the snake Vipera bertis, researchers caught and measured nine adult females.

<table>
<thead>
<tr>
<th>Table 12.2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length X (cm)</strong></td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>69</td>
</tr>
<tr>
<td>66</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>54</td>
</tr>
<tr>
<td>67</td>
</tr>
<tr>
<td>59</td>
</tr>
<tr>
<td>65</td>
</tr>
<tr>
<td>63</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td><strong>SD</strong></td>
</tr>
</tbody>
</table>
Example 12.2.1 Length and weight of snakes

How strong is linear relationship?

**Figure 12.2.1** Body length and weight of nine snakes with fitted regression line
12.2 The correlation coefficient $r$

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

- $r$ is a measure of the linear dependence between two variables $X$ and $Y$.
- $-1 \leq r \leq 1$.
- If $r > 0$ then $Y$ increases with increasing $X$.
- If $r = 1$ then $Y$ increases with $X$ according to a perfect line.
- If $r < 0$ then $Y$ decreases with increasing $X$.
- If $r = 0$ then $X$ and $Y$ are not linearly associated.
- The closer $r$ is to 1 or $-1$, the more the points lay on a straight line.
Examples of $r$ for 14 different data sets

- $|r| > 0.7$, strong correlation
- $0.3 < |r| < 0.7$, moderate correlation
- $|r| < 0.3$ weak correlation
Population correlation $\rho$

- Just like $\bar{y}$ estimates $\mu$ and $s_y$ estimates $\sigma$, $r$ estimates the unknown population correlation $\rho$.
- If $\rho = 1$ or $\rho = -1$ then all points in the population lie on a line.
- Sometimes people want to test $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$, or they want a 95% confidence interval for $\rho$.
- These are easy to get in R with the `cor.test(sample1,sample2)` command.
R code for amphetamine data

```r
> cons=c(112.6,102.1,90.2,81.5,105.6,93.0,106.6,108.3,73.3,84.8,67.3,55.3,
> + 80.7,90.0,75.5,77.1,38.5,81.3,57.1,62.3,51.5,48.3,42.7,57.9)
> amph=c(0,0,0,0,0,0,0,0,2.5,2.5,2.5,2.5,2.5,2.5,2.5,2.5,5.0,5.0,5.0,5.0,5.0,5.0)
> cor.test(amph,cons)

Pearson's product-moment correlation

data:  amph and cons
t = -7.9003, df = 22, p-value = 7.265e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9379300 -0.6989057
sample estimates:
cor
-0.859873

r = -0.86, a strong, negative relationship.
P-value = 0.0000000073 < 0.05 so reject $H_0 : \rho = 0$ at the 5% level. There is a significant, negative linear association between amphetamine intake and food consumption. We are 95% confident that the true population correlation is between -0.94 and -0.70.
\[ r = 0.94 \], a strong, positive relationship.
Comments

- Order doesn’t matter, either (X, Y) or (Y, X) gives the same correlation and conclusions. Correlation is “symmetric.”
- Significant correlation, rejecting $H_0 : \rho = 0$ doesn’t mean $\rho$ is close to 1 or $-1$; it can be small, yet significant.
- Rejecting $H_0 : \rho = 0$ doesn’t mean X causes Y or Y causes X, just that they are linearly associated.
12.1 Introduction
12.2 Correlation coefficient $r$
12.3 Fitted regression line
12.3 Fitting a line to scatterplot data

- $b_0$: intercept
- $b_1$: slope
- determine $b_0$ and $b_1$ by minimizing $\sum_{i=1}^{n}[y_i - (b_0 + b_1x_i)]^2$
12.3 Fitting a line to scatterplot data

We will fit the line

\[ Y = b_0 + b_1 X \]

to the data pairs.

- \( b_0 \) is the **intercept**, the fitted value of \( y \) when \( x=0 \).
- \( b_1 \) is the **slope**, the amount of change in \( y \) that occurs with one unit change in \( x \).

The values for \( b_0 \) and \( b_1 \) we use gives the **least squares** line.

These are the values that make \( \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2 \) as small as possible.

They are

\[ b_1 = r \left( \frac{s_y}{s_x} \right) \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}. \]
> fit=lm(cons~amph)
> plot(amph,cons)
> abline(fit)
> summary(fit)

Call:
lm(formula = cons ~ amph)

Residuals:
     Min      1Q  Median       3Q      Max
-21.512  -7.031   1.528   7.448  27.006

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99.331      3.680  26.99   < 2e-16 ***
amph        -9.007      1.140  -7.90  7.27e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 11.4 on 22 degrees of freedom
Multiple R-squared:  0.7394,    Adjusted R-squared:  0.7275
F-statistic:   62.41 on 1 and 22 DF,  p-value: 7.265e-08

For now, just pluck out $b_0 = 99.331$ and $b_1 = -9.007$
cons = 99.33 – 9.01 amph.
Here, \( b_0 = -301.1 \) and \( b_1 = 7.19 \).
weight = $-301.1 + 7.19 \text{ length}$. 
The $i$th residual is $e_i = y_i - \hat{y}_i$. This gives the vertical amount that the line missed $y_i$ by.

\[ SS(\text{resid}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2. \]

$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$ is sample standard deviation of the $Y$’s. Measures the “total variability” in the data.
$s_e$, $s_y$, and $r^2$

- $s_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = \sqrt{SS(\text{resid})/(n-2)}$ is “residual standard deviation” of the $Y$s. Measures variability around the regression line.

- If $s_e \approx s_y$ then the regression line isn’t doing anything!

- If $s_e < s_y$ then the line is doing something.

- $r^2 \approx 1 - \frac{s_e^2}{s_y^2}$ is called the multiple R-squared, and is the percentage of variability in $Y$ explained by $X$ through the regression line.

- R calls $s_e$ the residual standard error.
\( s_e \) is just average length of residuals

\[ s_e = 12.5 \text{ and } s_y = 35.3. \quad r^2 = 0.89 \text{ so 89\% of the variability in weight is explained by length.} \]
Roughly 68% of observations are within $s_e$ of the regression line (shown above); 95% are within $2s_e$. 

68%-95% rule for regression lines