Chapter 12: Linear regression I

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

So far...

- One sample continuous data (Chapters 6 and 8).
- Two sample continuous data (Chapter 7).
- One sample categorical data (Chapter 9).
- Two sample categorical data (Chapter 10).
- More than two sample continuous data (Chapter 11).
- Now: continuous predictor X instead of group.

Two continuous variables

- Instead of relating an outcome Y to "group" (e.g. 1, 2, or 3), 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.

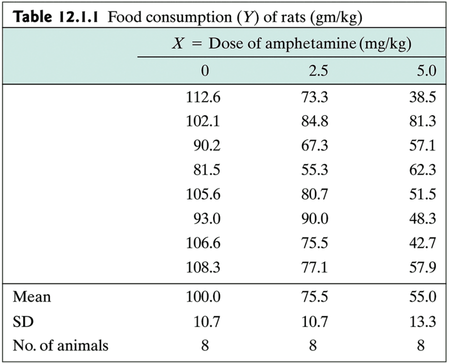
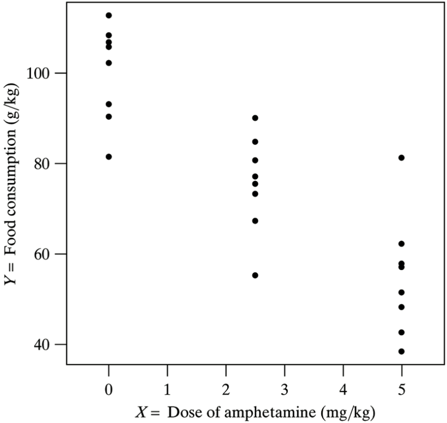


Table 12.1.1 Food consumption (Y) of rats (gm/kg)

	X = Dose of amphetamine (mg/kg)		
	0	2.5	5.0
	112.6	73.3	38.5
	102.1	84.8	81.3
	90.2	67.3	57.1
	81.5	55.3	62.3
	105.6	80.7	51.5
	93.0	90.0	48.3
	106.6	75.5	42.7
	108.3	77.1	57.9
Mean	100.0	75.5	55.0
SD	10.7	10.7	13.3
No. of animals	8	8	8

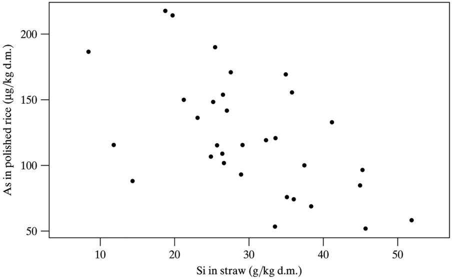
Example 12.1.1 Amphetamine and consumption



How does Y change with X? Linear? How strong is linear relationship?

Example 12.1.2 Arsenic in rice

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 (μg/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.

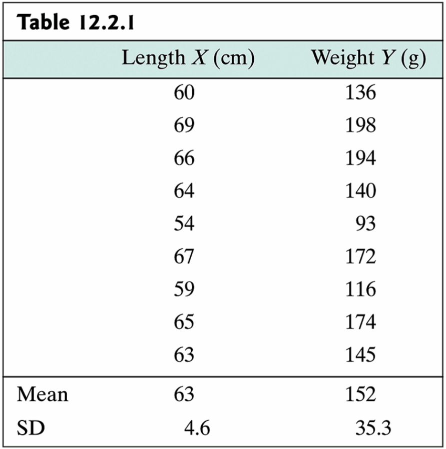


Table 12.2.1

	Length X (cm)	Weight Y (g)
	60	136
	69	198
	66	194
	64	140
	54	93
	67	172
	59	116
	65	174
	63	145
Mean	63	152
SD	4.6	35.3

Example 12.2.1 Length and weight of snakes

How strong is linear relationship?

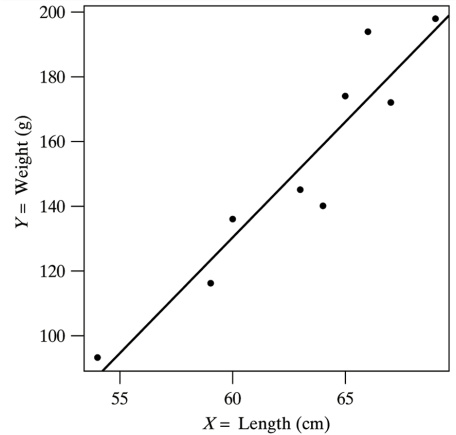


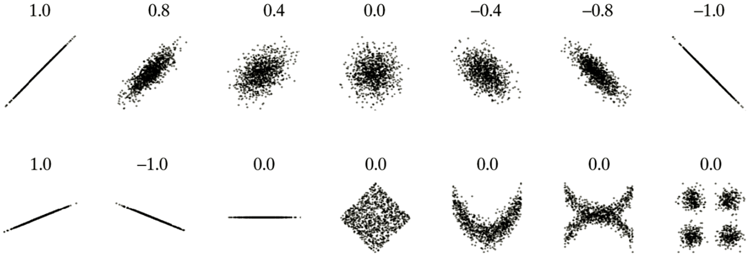
Figure 12.2.1 Body length and weight of nine snakes with fitted regression line

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 measures the strength and direction (positive or negative) of how *linearly* related Y is with X.
- $-1 \leq r \leq 1$.
- If $r = 1$ then Y increases with X according to a perfect line.
- If $r = -1$ then Y decreases with X according to a perfect line.
- 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



Population correlation ρ

- Just like $\bar{y}$ estimates μ and s_y estimates σ, r estimates the unknown *population correlation* ρ.
- 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 ρ.
- These are easy to get in R with the `cor.test(sample1,sample2)` command.

R code for amphetamine data

```
> 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(0,0,0,0,0,0,0,0,2.5,2.5,2.5,2.5,2.5,2.5,2.5,2.5,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0)
> 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$=0.000000073 < 0.05$ so reject $H_0 : \rho = 0$ at the 5% level.
There is a signficant, 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 code for snake data

```
> length=c(60,69,66,64,54,67,59,65,63)
> weight=c(136,198,194,140,93,172,116,174,145)
> cor.test(length,weight)

        Pearson's product-moment correlation

data:  length and weight
t = 7.5459, df = 7, p-value = 0.0001321
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7489030 0.9883703
sample estimates:
      cor
0.9436756
```

$r = 0.94$, a strong, positive relationship. What else do we conclude?

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 ρ 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.

12.3 Fitting a line to scatterplot data

We will fit the line

$$Y = b_0 + b_1 X$$

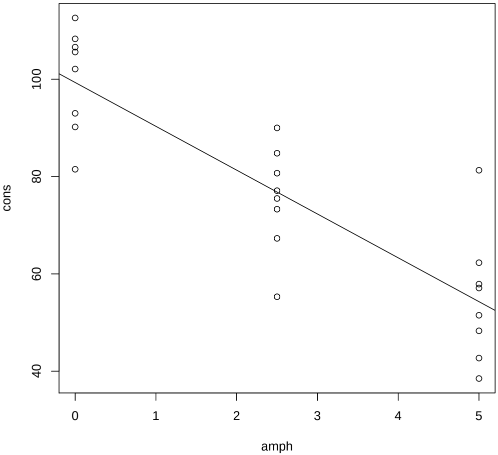
to the data pairs.

- b_0 is the **intercept**, how high the line is on the Y-axis.
- b_1 is the **slope**, how much the line changes when X is increase by one unit.
- 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

$$b_1 = r\left(\frac{s_y}{s_x}\right) \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

```
> 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
```

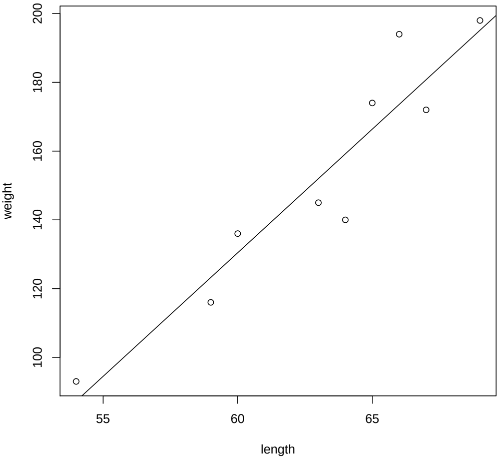
For now, just pluck out $b_0 = 99.331$ and $b_1 = -9.007$



$$\text{cons} = 99.33 - 9.01\ \text{amph}.$$

```
> fit=lm(weight~length)
> plot(length,weight)
> abline(fit)
> summary(fit)

Call:
lm(formula = weight ~ length)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.192  -7.233   2.849   5.727  20.424 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -301.0872    60.1885  -5.002 0.001561 ** 
length         7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared: 0.8905,     Adjusted R-squared: 0.8749 
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

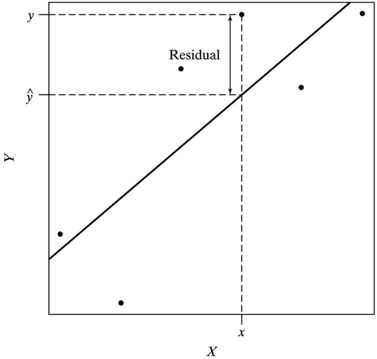
Here, $b_0 = -301.1$ and $b_1 = 7.19$



$$\text{weight} = -301.1 + 7.19\ \text{length}.$$

Residuals

- The ith fitted value is $\hat{y}_i = b_0 + b_1 x_i$, the point on the line above x_i.
- The ith residual is $e_i = y_i - \hat{y}_i$. This gives the vertical amount that the line missed y_i by.



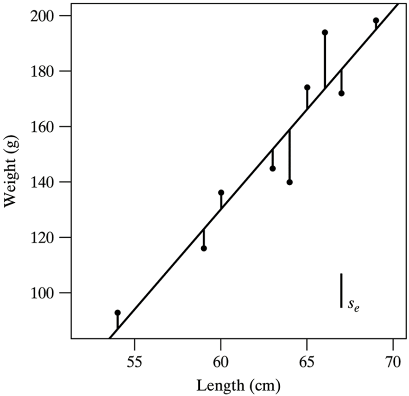
Residual sum of squares and s_e

- SS(resid)$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$.
- (b_0, b_1) make SS(resid) as small as possible.
- $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ is sample standard deviation of the Y's. Measures the "total variability" in the data.

s_e, s_y, and r^2

- $s_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\mathrm{SS(resid)}/(n-2)}$ is "residual standard deviation" of the Ys. 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*.

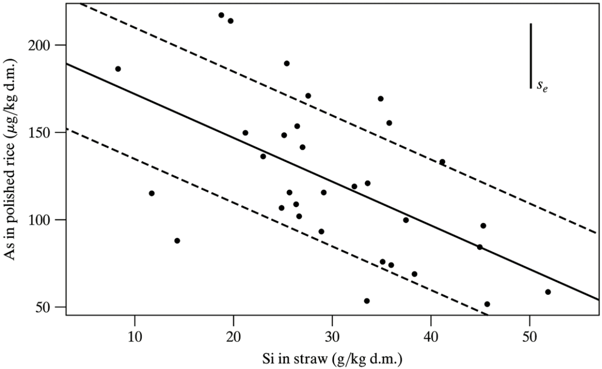
s_e is just average length of residuals



```
> sd(weight)
[1] 35.33766
```

$s_e = 12.5$ and $s_y = 35.3$. $r^2 = 0.89$ so 89% of the variability in weight is explained by length.

68%-95% rule for regression lines



Roughly 68% of observations are within s_e *of the regression line* (shown above); 95% are within 2 s_e.