Covariance, Correlation and Simple Linear Regression

Notes:

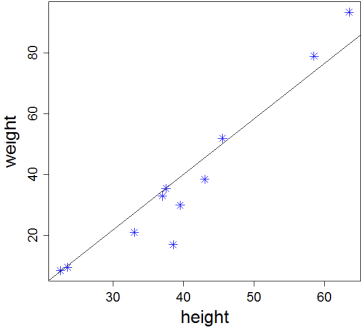
- Please fill out survey on blackboard; we will start using data next week.
- HW1 now out, due Friday next week.

Agenda:

- Review of measures of association between observations; covariance and correlation.
- Population values versus sample estimates (and notation)
- Simple linear regression
  - Framework and assumptions
  - Estimating parameters
  - How is SLR different from correlation?

Relationships Between Observations

Consider height and weight of subjects in a clinical trial:



Clearly, taller people weigh more. How do we summarize this relationship?

Variance and Covariance

We have n pairs of data:

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

- We summarize variation in X by average distance from average.

$$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

- Same thing for $\hat{\sigma}_Y^2$.
- *Covariance* is the strength of relationship:

$$\hat{\sigma}_{XY} = s_{XY} = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

Is X_i's difference from $\bar{X}$ similar to Y_i's difference from $\bar{Y}$?

Correlation

- Covariance changes with the scale of X and Y.
- *Correlation* is dimensionless:

$$\hat{\rho}_{XY} = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

  (note no squares on denominator)

- Alternatively, it is covariance of *standardized* quantities:

$$r_{XY} = \frac{1}{n-1} \sum \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

- Measures the strength of *linear* association.

Height/Weight Data

- Set X = Weight and Y = Height
- Their averages are

$$\bar{X} = 38.12500 \qquad \bar{Y} = 38.90000$$

- Variances are

$$s_X^2 = 679.0057 \qquad s_Y^2 = 163.22909$$

- Covariance is

$$s_{XY} = 297.7045$$

- Correlation is

$$r_{XY} = \frac{297.7054}{\sqrt{679.0057 * 163.22909}} = 0.8942315$$

Some Calculations in R

```
# Load in Data
> heart = read.table('heart.txt',head=TRUE)

# Now calculate Xbar
> m.weight = mean(heart$weight)
[1] 38.125

# And Ybar
> m.height = mean(heart$height)
[1] 38.9

# Var weight and height
> var.weight = var(heart$weight)
[1] 679.0057
> var.height = var(heart$height)
[1] 163.2291
```

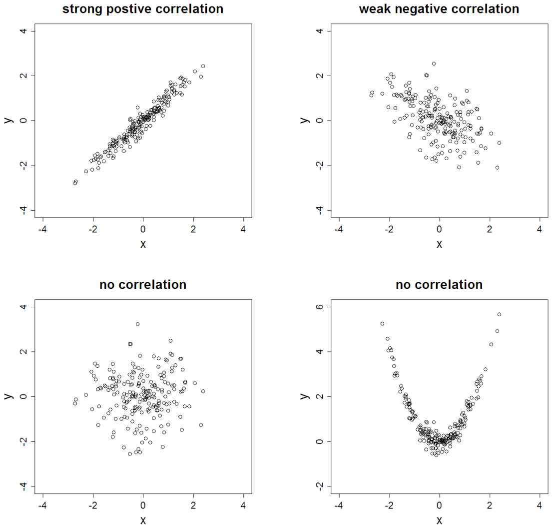
Calculations Continued

```
# Their covariance
> cov.heart = cov(heart$height,heart$weight)
[1] 297.7045

# and correlation
> cor.heart = cov.heart/sqrt( var.height * var.weight )
[1] 0.8942315

# Alternatively
> cor.heart2 = cor(heart$height,heart$weight)
[1] 0.8942315
```

Properties of r_{XY}

- $-1 \leq r_{XY} \leq 1$
- Does not depend on units of measurement for X or Y.
- $r_{XY} = \pm 1$ implies perfect *linear* association
- $r_{XY} = 0$ represents no *linear* association, (but not no nonlinear association)

Graphically

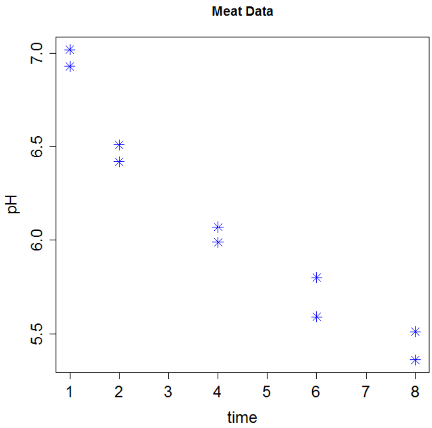


Sample and Population Values

- We calculate s_X^2 for a *sample* $X_1, \ldots, X_n$
- For a new set of data, we would get different values.
- *Population* parameters σ_X^2, σ_{XY} are parameters governing the process that produces the data.
- Usually, we can think of the population values as being what we would calculate if we had infinite data.
- Want to use *sample* values to estimate *population* values.
- Notation:
  - usually population values are given by Greek letters σ, μ.
  - sample values, are usually Roman letters (s, m) or we use hats ($\hat{\sigma}$,$\hat{\mu}$) to demonstrate that these are estimates.

  TA's can be picky about this.

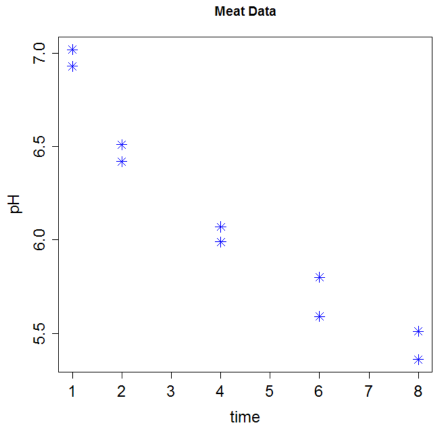
Acidity Data



- Processing requires packaged food to have pH less than 6.0 (declines from 7.2).
- Experimental data obtained at 5 time intervals.

What's the correlation?

Acidity Data



- Processing requires packaged food to have pH less than 6.0 (declines from 7.2).
- Experimental data obtained at 5 time intervals.

What's the correlation?

Meaningless! We fixed the times before the experiment.

Simple Linear Regression (SLR)

Correlation summarizes relationship between two *random* quantities. What about controlled experiments?

- Data are X (= time) and Y (= pH) for each sample.
- We would like to know if X and Y are related.
- We have chosen the values of X, want to use this to predict Y.
- If X is not controlled, we can also ask:

Does knowing X tell us anything about Y?

  - X = *independent* variable
  - Y = *dependent* variable

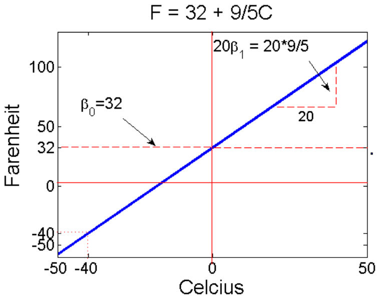
We will still consider the linear relationship between X and Y.

The Linear Model

A simple model:

$$Y = \beta_0 + \beta_1 X$$

For every unit increase in X, Y increases by β_1 units.



- β_0 = the *intercept*: the value of Y when $X = 0$
- β_1 = the *slope*: how Y changes with X.

The *Statistical* Linear Model

- The deterministic model is rarely exact (and there are no interesting statistics when it is).
- Instead, we account for deviations from the linear model by including an error term:

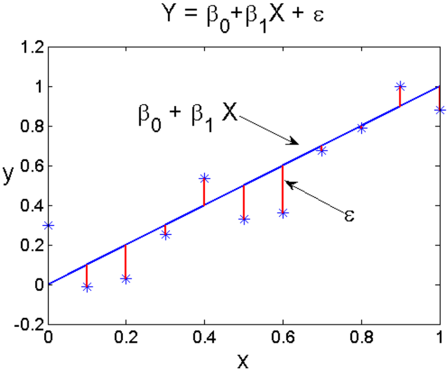
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $\epsilon =$ "random error", assumed different each time.
- Assume that $E\epsilon = 0$, that is the linear model holds on average:

$$EY = \beta_0 + \beta_1 X$$

- Here we think of EY as averaging Y repeatedly measured *at the same value of* X.

The Statistical Model Illustrated




```
beta0 = 0;    beta1 = 1            # regression parameters
X = seq(0,1,by=0.1)                # values of X
epsilon = rnorm(11,mean=0,sd=0.2)  # only random piece

Y = beta0 + beta1*X + epsilon
```

Sample Data

In practise we observe n pairs of data

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

The SLR model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the errors satisfy

1. $E(\epsilon_i) = 0$
2. $\text{var}(\epsilon_i) = \sigma^2$ (homogenous variances)
3. All the ϵ_i are independent
4. All the ϵ_i are normally distributed (Gaussian)

In other words, $\epsilon_1, \ldots, \epsilon_n$ are an independent random sample from $N(0, \sigma^2)$.

But now I need to know β_0, β_1 and σ^2.

The Least Squares Principle

Choose β_0 and β_1 to minimize the squared distance between the observed Y and the value predicted from X:

$$\text{Minimize: } \text{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To do this, calculate

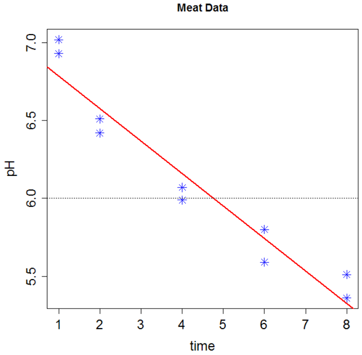
$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) \text{ and } S_{XX} = \sum (X_i - \bar{X})^2$$

Note we do not divide by n – referred to as "sums of squares".

The minimizing values of the parameters are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

pH Data



We have

$$\begin{aligned}\bar{X} &= 4.20\\ \bar{Y} &= 6.12\\ S_{XX} &= 65.6\\ S_{XY} &= -13.69\\ \hat{\beta}_0 &= 7.00\\ \hat{\beta}_1 &= -0.21\end{aligned}$$

Quick calculation of time to reach pH 6.0 is

$$\hat{\beta}_0 + \hat{\beta}_1 t = 6 \Rightarrow t = \frac{6.0 - \hat{\beta}_0}{\hat{\beta}_1} = 4.8 hours$$

Predicted Values and Residuals

We have estimated $\hat{\beta}_0$ and $\hat{\beta}_1$.

- Best prediction for Y with a new X (X_{new})is

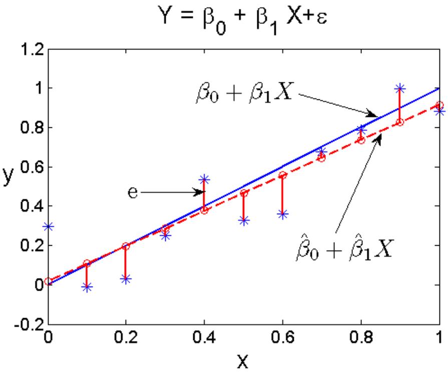
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{new}$$

- For the X_i that we already have, the *fitted values* are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Then our estimates of the errors are the *residuals*:

$$e_i = Y_i - \hat{Y}_i$$



Variance

Full specification of the model is

$$Y = \beta_0 + \beta_1 X + \epsilon, \ \epsilon \sim N(0, \sigma^2)$$

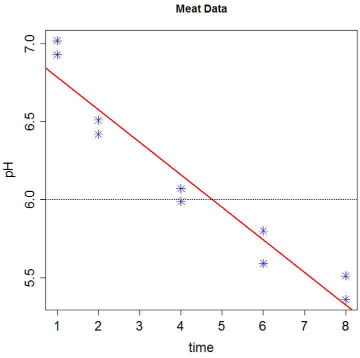
Would still like to know about σ^2 (ie, does knowing X tell us much at all?)

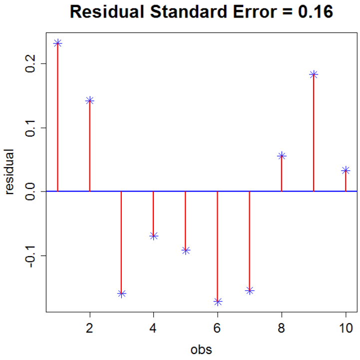
Estimate variance by Mean Squared Error

$$\hat{\sigma}^2 = \frac{\text{Error SS}}{\text{Error DF}} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

Error DF = "effective sample size" for estimating σ^2; accounts for estimating β_0 and β_1.

Residuals


Meat Data
pH
7.0
6.5
6.0
5.5
1
2
3
4
5
6
7
8
time


Residual Standard Error = 0.16
residual
0.2
0.1
0.0
-0.1
2
4
6
8
10
obs

Example Inference

```
> food.mod = lm(pH~time,data=food)
> summary(food.mod)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17174 -0.13870 -0.01805  0.12056  0.23220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.99649    0.09691   72.20 1.51e-12 ***
time        -0.20869    0.01970  -10.59 5.51e-06 ***
---
Signif. codes:  '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1595 on 8 degrees of freedom
Multiple R-squared: 0.9335,     Adjusted R-squared: 0.9251
F-statistic: 112.2 on 1 and 8 DF,  p-value: 5.509e-06
```

Summary

- Correlation and covariance = measure of linear association between two *random* quantities.
- Simple linear regression = linear dependence of Y *given* X. X need not be random.
- Statistical model: randomness in Y is in *error* about a linear trend.
- Least squares estimates for regression lines.

Next

- Precision of the least squares estimate.
- Inference, confidence intervals and prediction intervals.
- Readings: Fox 5.1/5.2, 11