Math 321 – Chapter 7 – Linear Regression
(draft version 2019/06/20-12:52:57)

Contents

1 Paired data

We will now look at a dataset where each point has two numerical data values, X and Y, that are paired together. For example, consider that we have data on oil wells, including a depth measurement and the number of barrels produced per year. Each well has two measurements associated with it, and we would not want to mix the depth of one well with the production of another. It is clear that each depth data point has an associated production data point. We may think that production is in some way dependent on the depth of the well, e.g. maybe older deposits are deeper and more productive (or the opposite).

A paired dataset:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

1.1 Covariance and correlation

With a paired dataset, we can calculate the mean and variance of both X and Y:

$$\overline{x} = \frac{1}{n}\sum_i x_i$$

$$\overline{y} = \frac{1}{n}\sum_i y_i$$

$$s_x^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$$

$$s_y^2 = \frac{1}{n-1}\sum_i (y_i - \overline{y})^2$$

We will now define a few different sums:

$$S_{xx} = \sum_i (x_i - \overline{x})^2$$

$$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

Notice that $S_{xx} = (n-1)s_x^2$ and is always positive. However S_{xy} can be negative, and this will occur if x tends to deviate above its mean and y deviates below its mean or vice versa. S_{xy} is related to how x and y *covary*, how the variability of one variable depends on the other variable.

The *covariance* of random variables X and Y is formally defined as

$$\text{Cov}(X,Y) = E\Big[(X - E(X))(Y - E(Y))\Big].$$

If X and Y are independent, then $\text{Cov}(X,Y) = 0$.

For paired data, we define the sample covariance as

$$\text{Cov}(X,Y) = \frac{S_{xy}}{n-1}$$

In R the sample covariance can be calculated as follows assuming we have our data stored in `x` and `y`:

```
> cov(x,y)
```

or

```
> sum((x-mean(x))*(y-mean(y))
```

The *correlation* of random variables X and Y with standard deviations σ_X and σ_Y is defined as

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

The sample correlation is

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{s_x s_y}$$

In R this can be calculated by

```
> cor(x,y)
```

or

```
> cov(x,y)/sd(x)/sd(y)
```

The correlation will always be between -1 and 1. A correlation of ± 1 indicates that there is no randomness, $\sigma^2 = 0$ for the random perturbation parameter ϵ int eh linear regression equation. The relationship between X and Y is perfectly linear. A correlation of 0 indicates that X and Y are independent. The sign of the correlation indicates the sign of the slope of the line. Note that it is not identical to nthe slope of the line!

See the graph below with correlation coefficient indicated on each graph:

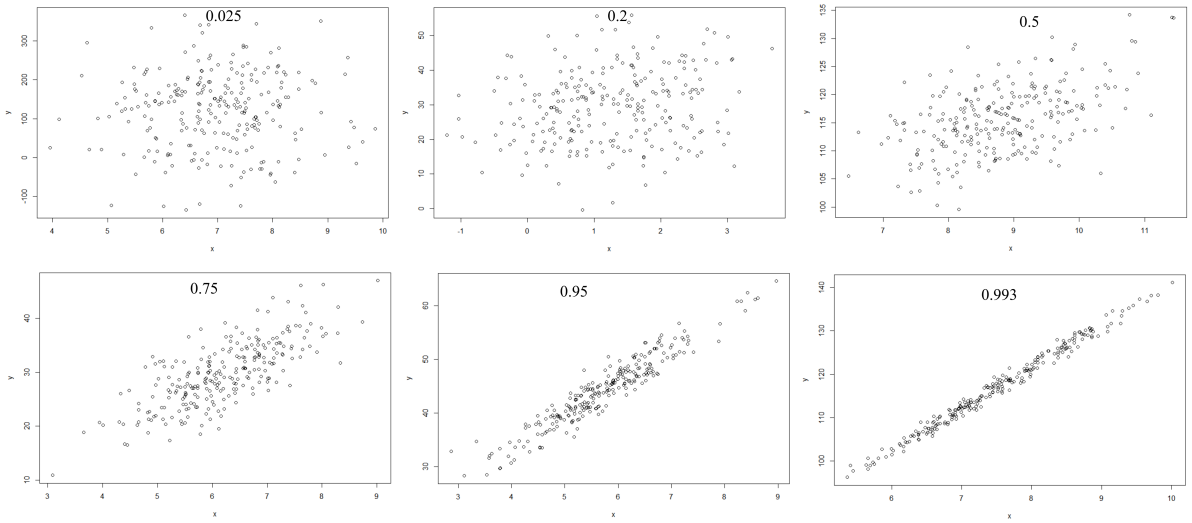


Figure 1: Various scatterplots with correlation coefficients indicated.

1.2 A graphical interpretation of correlation

Now we will consider a graphical interpretation of correlation. If you have taken a Linear Algebra or Vector Calculus course (usually Calculus III), then you should recall the concept of *dot product* and how it relates to the cosine of the angle between two vectors. Consider our X and Y data as n-dimensional vectors. Let's shift each by their respective mean and define $\mathbf{u} = X - \overline{x}$ and $\mathbf{v} = Y - \overline{y}$.

$$\mathbf{u} = \langle x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x} \rangle$$
$$\mathbf{v} = \langle y_1 - \overline{y}, y_2 - \overline{y}, \ldots, y_n - \overline{y} \rangle$$

We have the dot product formula $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\theta$ where teh magnitude of a vector is given by $|\mathbf{u}| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sum u_i^2$. In Calculus III, one normally only works up to three dimensions, but in Linear Algebra, the dot product is generalized up to any dimensions.

Note that given the definitions of vectors $\mathbf{u}$ and $\mathbf{v}$ above in terms of our X and Y paired data, we have that

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = (n-1)\text{Cov}(X, Y)$$

$$|\mathbf{u}| = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{(n-1)\text{Var}(X)}$$

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \sqrt{(n-1)\text{Var}(Y)}$$

Thus

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \frac{(n-1)\text{Cov}(X, Y)}{\sqrt{(n-1)\text{Var}(X)}\sqrt{(n-1)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{s_X s_Y} = \text{Cor}(X, Y)$$

so we can think of the correlation as the cosine of the angle between the (shifted) datasets.

2 Linear regression

We will assume there is a linear relationship between the paired data X and Y. The assumed linear relationship is

$$Y = a + bX.$$

The slope is b and the y-intercept is a.

This equation is deterministic though int he sense that if you plug in an X value, you will get a precise y value. We will introduce a term that will cause Y to randomly deviate from the expected value given by this equation.

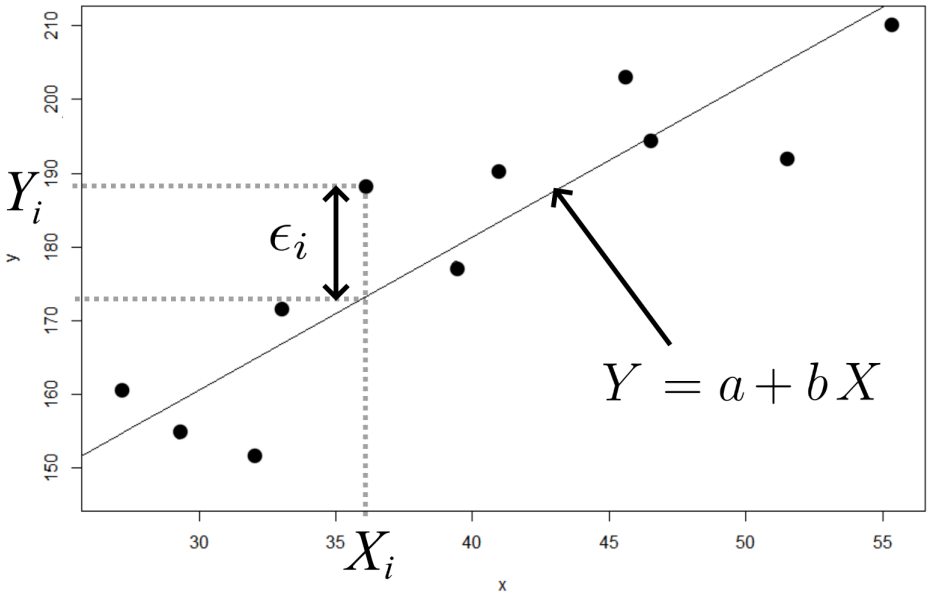
Let $\epsilon \sim N(0, \sigma^2)$ be a normal random variable wtih mean 0 and variance σ^2. This will be the random perturbation away from the linear model. Now we have:

$$Y = a + bX + \epsilon.$$

Furthermore, each Y value may be perturbed differently away from the expected $Y = aX + b$:

$$Y_i = a + bX_i + \epsilon_i.$$

For a given dataset, our goal now will be to find a and b in order to create a line that "best fits" the data. See the figure.



Let $\hat{y}_i = a + bx_i$. This is our estimated y_i-value for the given x_i-value. Our "squared deviation" from the regression line is $(\hat{y}_i - y_i)^2$. We will sum these up to get our *summed square errors*

$$SSE = \sum_i (\hat{y}_i - y_i)^2$$

and divide by $n - 2$ to get the *mean squared error* (MSE):

$$MSE = \frac{SSE}{n-2}.$$

We perform a minimization procedure to minimize the SSE, and we solve for a and b to get

$$\hat{b} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

The parameters a and b are unknown, and we use $\hat{a}$ and $\hat{b}$ as our estimates of them.

In R we can calculate these as follows:

```
> b = cov(x,y)/var(x)
  a = mean(y)-b*mean(x)
  linefit = a+b*x
  SSE = sum((y-linefit)^2)
  MSE = SSE/(n-2)
```

To plot your data in R with the regression line overlayed on it:

```
> plot(x,y)
  lines(x,linefit)
```

Alternatively in R we can generate a regression line using the `lm()` method:

```
> regline = lm(y~x)
  SSE = sum((y-regline$fitted.values)^2)
  MSE = SSE/(length(x)-2)
  plot(x,y)
  abline(regline)
```

3 Confidence interval for slope

Confidence interval for the slope of the regression line:

> `b+c(-1,1)*qt(1-`α`/2,length(x)-2)*sqrt(MSE/(length(x)-1)/var(x))`

$$\hat{b} \pm t_{1-\alpha/2,n-2}\sqrt{\frac{MSE}{(n-1)\text{Var}(X)}}$$

4 Prediction interval for particular y-value

Prediction interval for an individual y-value given a particular x-value:

> `n=length(x)`

> `xval=`x (input your x-value here)

> `a+b*xval+c(-1,1)*qt(1-`α`/2,n-2)*sqrt(MSE*(1+1/n+(xval-mean(x))^2/(n-1)/var(x)))`

$$\hat{y}_i \pm t_{1-\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(X_i-\overline{X})^2}{(n-1)\text{Var}(X)}\right)}$$

5 Prediction interval for particular mean y-value

Prediction interval for mean y-value given a particular x-value: (note that this is the same as above but without the "1+"

> `n=length(x)`

> `xval=`x (input your x-value here)

> `a+b*xval+c(-1,1)*qt(1-`α`/2,n-2)*sqrt(MSE*(1/n+(xval-mean(x))^2/(n-1)/var(x)))`

$$\hat{y}_i \pm t_{1-\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{(X_i-\overline{X})^2}{(n-1)\text{Var}(X)}\right)}$$

6 Complete code for full linear regression analysis

The following is a complete code for constructing and plotting a linear regression line and scatterplot and constructing $(1-\alpha)100\%$ confidence intervals for the slope. Items in red

need to be filled in.

```
> x = c(x_1, x_2, ..., x_n)
  y = c(y_1, y_2, ..., y_n)
  alpha = α
  xbar = mean(x)
  sx = sd(x)
  ybar = mean(y)
  sy = sd(y)
  n = length(x)
  bhat = cov(x,y)/sx^2
  ahat = ybar-bhat*xbar
  yfit = ahat+bhat*x
  plot(x,y)
  lines(x,ahat+bhat*x)
  SSE = sum((yfit-y)^2)
  MSE = SSE/(n-2)
  bci = bhat+c(-1,1)*qt(1-alpha/2,n-2)*sqrt(MSE/(n-1)/sx^2)
```

7 Summary

R commands:

```
lm(y~x)
cov(x,y)
cor(x,y)
b=cov(x,y)/var(x)
a=mean(y)-b*mean(x)
plot(x,y)
abline(lm(y~x))
SST=sum((y-mean(y))^2)
SSR=sum((lm(y~x)$fitted.values-mean(y))^2)
SSE=sum((lm(y~x)$fitted.values-y)^2)
```

Notation and formulas:

$Y = a + bX$ is assumed real relationship between X and Y, for a given X value, the observed Y value deviates from this line randomly.

$\hat{y}_i$ = predicted y-value from linear regression formula, given x-value x_i
$\hat{b}$ = point estimate of slope of regression line
$\hat{a}$ = point estimate of y-intercept of regression line

$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$ sum of squared total deviations
(sum of squared distances from data to mean)
$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ sum of squared deviations due to regression
(sum of squared distances from predicted point on regression line to mean)
$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i - \overline{y})^2$ sum of squared deviations due to random error
(sum of squared distances from data to predicted point on regression line)

$SST = SSR + SSE$