Math 321 – Chapter 7 – Linear Regression
(draft version 2019/04/24-17:49:23)

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)
```

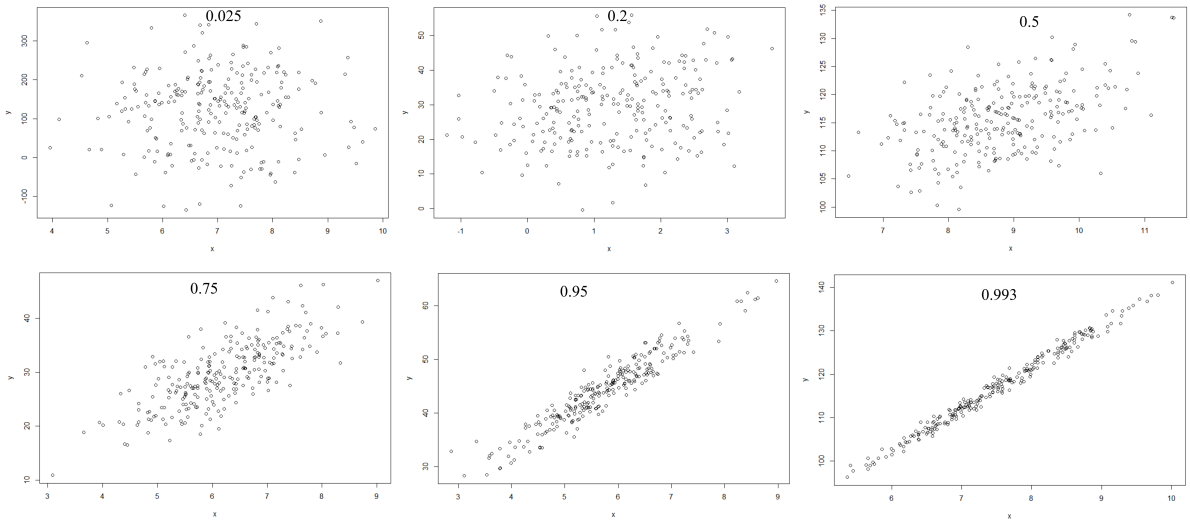
or

```
> 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:



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.

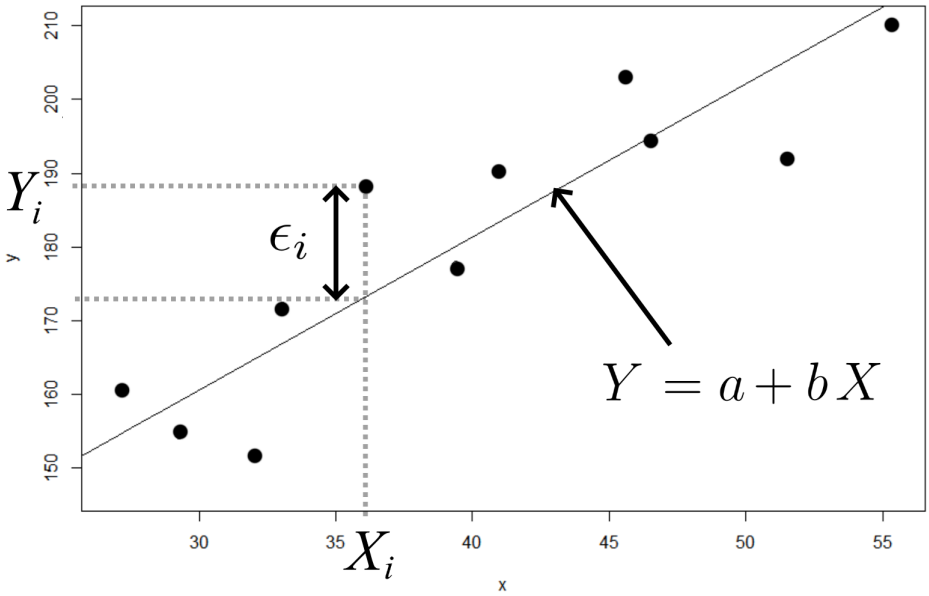
Let $\epsilon \sim N(0, \sigma^2)$ be a normal random variable wtih mean 0 and variance σ^2. This will be the random perturbation away from the linear model. Now we have:

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

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))`

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

4 Confidence interval for particular y-value

Confidence 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)))`

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

5 Confidence interval for particular mean y-value

Confidence 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)))`

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