Topic 3
Correlation and Regression
Linear Regression I

Outline

Introduction

Covariance and correlation are measures of linear association. For the *Archeopteryx* measurements, we learn that the relationship in the length of the femur and the humerus is very nearly linear. We now turn to situations in which

- the value of the first variable x_i will be considered to be explanatory or predictive.
- The corresponding observation y_i, taken from the input x_i, is called the response.

For example, can we explain or predict the number of *de novo* mutations in an offspring from the average age of the parents? In this case, age is the explanatory variable and the number of mutations is the response.

Linear Regression

In linear regression, the response variable is linearly related to the explanatory variable, but is subject to deviation or to error. We write

$$y_i = \alpha + \beta x_i + \epsilon_i.$$

Our goal:

- given data, the x_i's and y_i's, find α and β that determines the line of best fit.

The principle of least squares regression states that

- the best choice of this linear relationship is the one that minimizes the square in the *vertical distance* from the y values in the data and the y values on the regression line
- reflecting the fact that the values of x are set by the experimenter and are thus assumed known. Thus, the "error" appears in the value of the response variable y.

Principle of Least Squares

This principle leads to a minimization problem for

$$SS(\alpha,\beta)=\sum_{i=1}^{n}\epsilon_i^2=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2.$$

Let's the denote by $\hat{\alpha}$ and $\hat{\beta}$ the value for α and β that minimize SS.

$$\frac{\partial}{\partial\alpha}SS(\alpha,\beta)=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)$$

At the values $\hat{\alpha}$ and $\hat{\beta}$, this partial derivative is 0. Consequently,

$$0=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)\qquad \sum_{i=1}^{n}y_i=\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)\qquad \bar{y}=\hat{\alpha}+\hat{\beta}\bar{x}.$$

Thus, we see that the center of mass point $(\bar{x},\bar{y})$ is on the regression line.

Principle of Least Squares

To emphasize this fact, we rewrite the line in slope-point form.

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + \epsilon_i.$$

Now, the sums of squares criterion becomes a condition on β,

$$\tilde{SS}(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} ((y_i - \bar{y}) - \beta(x_i - \bar{x}))^2.$$

Now, differentiate with respect to β and set this equation to zero for the value $\hat{\beta}$.

$$\frac{d}{d\beta}\tilde{SS}(\hat{\beta}) = -2\sum_{i=1}^{n}((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) = 0.$$

Principle of Least Squares

$$0 = \sum_{i=1}^{n}((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}).$$

Thus,

$$\begin{aligned}
\hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
\hat{\beta}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
\hat{\beta}\ \mathrm{var}(x) &= \mathrm{cov}(x, y) \\
\hat{\beta} &= \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}
\end{aligned}$$

Regression Equations

In summary, to determine the regression line $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, we have

$$\hat{\beta} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Let's begin with 6 points and derive by hand the equation for regression line. First, we find that $\bar{x} = 2.5$ and $\bar{y} = 4$.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0	7	-2.5	3	-7.5	6.25
1	5	-1.5	1	-1.5	2.25
2	5	-0.5	1	–0.5	0.25
3	4	0.5	0	-0.0	0.25
4	2	1.5	-2	–3.0	2.25
5	1	2.5	-3	–7.5	6.25
sum		0	0	$\mathrm{cov}(x, y) = -20/5$	$\mathrm{var}(x) = 17.50/5$

Linear Regression

Collecting the necessary summaries,

$$\bar{x} = 2.5 \quad \bar{y} = 4 \quad \mathrm{cov}(x, y) = -20/5 = -4 \quad \mathrm{var}(x) = 17.50/5 = 3.5$$

Thus,

$$\hat{\beta} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = -\frac{4}{3.5} = -\frac{8}{7} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 4 + \frac{8}{7} \cdot \frac{5}{2} = \frac{48}{7}$$

The equation of the regression line is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \frac{48}{7} - \frac{8}{7}x_i.$$

Exercise. Find $r(x, y)$.

Linear Regression

Exercise. Using the same data, reverse the role of explanatory and response variable and determine the regression line, $\hat{x} = \hat{\alpha}_y + \hat{\beta}_y y$.

y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$	$(y_i - \bar{y})^2$
7	0				
5	1				
5	2				
4	3				
2	4				
1	5				
sum					

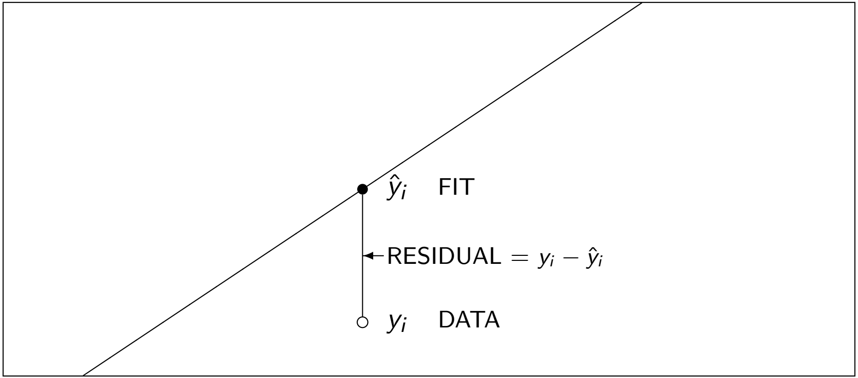
Show that these two lines are not the same. Find the square root of the product of the slopes. What do you notice?

Residuals

The residual, the difference between the fit and the data is an estimate $\hat{\epsilon}_i$ for the error.

$$\hat{\epsilon}_i = \text{RESIDUAL}_i = \text{DATA}_i - \text{FIT}_i = y_i - \hat{y}_i.$$

By rearranging terms, $\text{DATA}_i = \text{FIT}_i + \text{RESIDUAL}_i$, or $y_i = \hat{y}_i + \hat{\epsilon}_i$.



Residuals

The regression line is

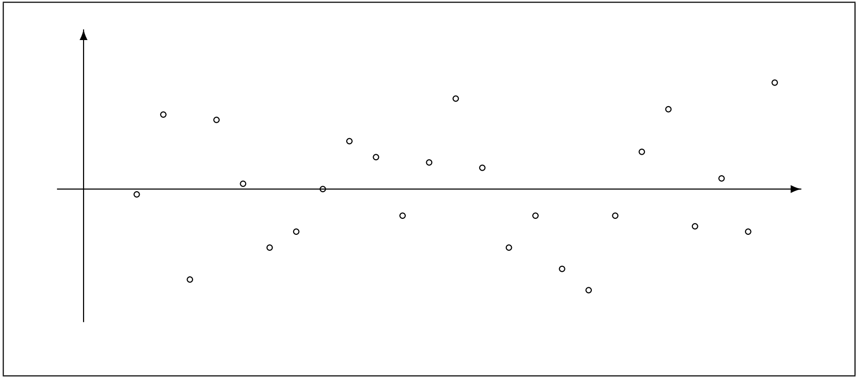
$$\hat{y}_i = \frac{48}{7} - \frac{8}{7}x_i$$

	DATA	FIT	RESIDUAL
x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
0	7	$48/7$	$1/7$
1	5	$40/7$	-$5/7$
2	5	$32/7$	$3/7$
3	4	$24/7$	$4/7$
4	2	$16/7$	-$2/7$
5	1	$8/7$	-$1/7$
	sum		0

Notice that the sum of the residuals is 0. This follows from $\frac{\partial}{\partial \alpha} SS(\hat{\alpha}, \hat{\beta}) = 0$.

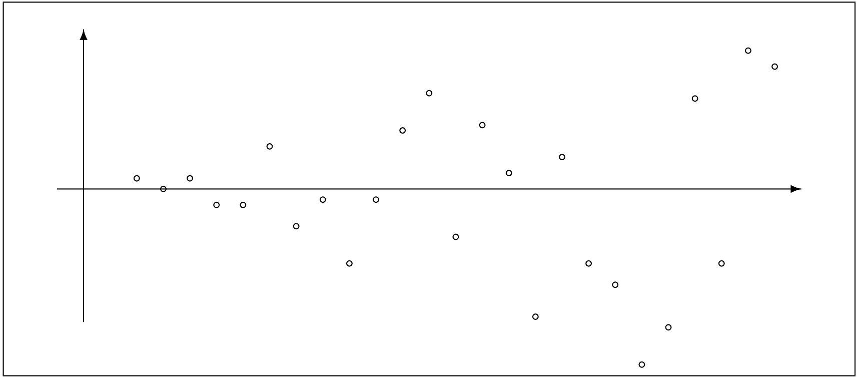
Residuals

We next show three examples of the residuals plotting against the value of the explanatory variable.



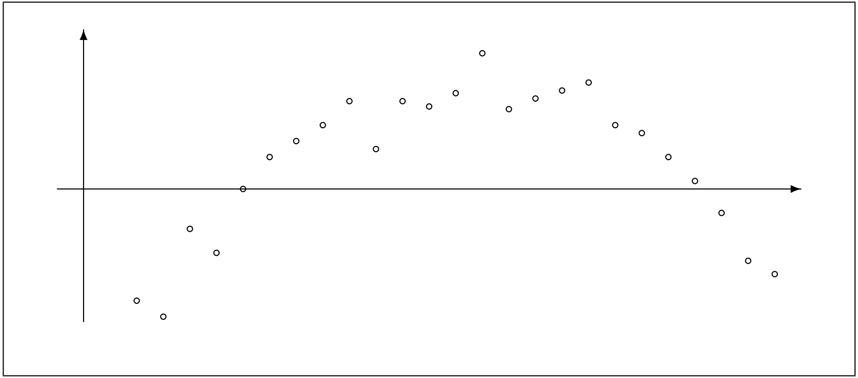
Regression fits the data well - homoscedasticity.

Residuals



Prediction is less accurate for large x, an example of heteroscedasticity

Residuals



Data has a curve. A straight line fits the data poorly.