Lecture 15: Chapter 10

C C Moxley

UAB Mathematics

20 July 15

§10.1 Pairing Data

In Chapter 9, we talked about pairing data in a “natural” way. In this Chapter, we will essentially be discussing whether these “natural” parings are useful or not.

§10.1 Pairing Data

In Chapter 9, we talked about pairing data in a "natural" way. In this Chapter, we will essentially be discussing whether these "natural" parings are useful or not. Mainly, we'll be using the notion of correlation to do so.

§10.1 Pairing Data

In Chapter 9, we talked about pairing data in a "natural" way. In this Chapter, we will essentially be discussing whether these "natural" parings are useful or not. Mainly, we'll be using the notion of correlation to do so.

We will also look at ways of predicting values based on linear models arising from samples - this is called **linear regression**.

§10.1 Pairing Data

In Chapter 9, we talked about pairing data in a "natural" way. In this Chapter, we will essentially be discussing whether these "natural" parings are useful or not. Mainly, we'll be using the notion of correlation to do so.

We will also look at ways of predicting values based on linear models arising from samples - this is called **linear regression**. We'll also discuss methods for determining how much these predicted values may vary from the actual value.

§10.2 Correlation

Definition (Correlation)

Two variables are correlated when the values of one variable are somehow associated with the values of the other variable.

§10.2 Correlation

Definition (Correlation)

Two variables are correlated when the values of one variable are somehow associated with the values of the other variable.

Definition (Linear Correlation)

A linear correlation exists between two variables when the correlation between the two variables can be expressed as a line.

§10.2 Correlation

Definition (Correlation)

Two variables are correlated when the values of one variable are somehow associated with the values of the other variable.

Definition (Linear Correlation)

A linear correlation exists between two variables when the correlation between the two variables can be expressed as a line. This can also be seen on a scatter plot.

§10.2 Correlation

Definition (Correlation)

Two variables are correlated when the values of one variable are somehow associated with the values of the other variable.

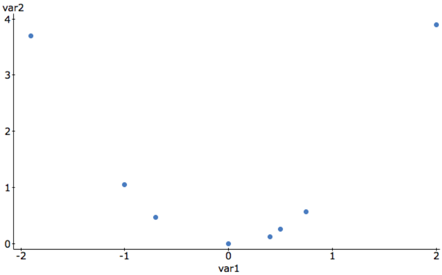
Definition (Linear Correlation)

A linear correlation exists between two variables when the correlation between the two variables can be expressed as a line. This can also be seen on a scatter plot.

Note: Correlation **does not imply causation!**

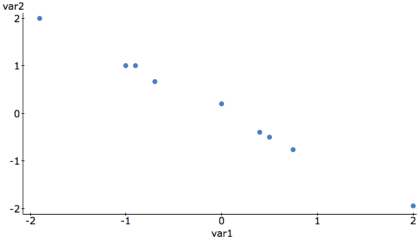
§10.2 Example

Are the following variables (one on the x-axis and the other on the y-axis) correlated?



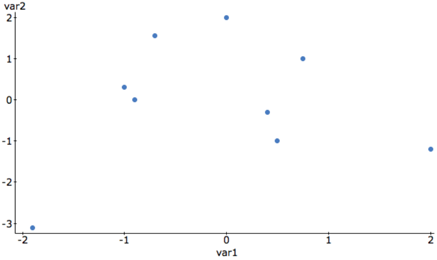
§10.2 Example

Are the following variables (one on the x-axis and the other on the y-axis) correlated?



§10.2 Example

Are the following variables (one on the x-axis and the other on the y-axis) correlated?



For a linear correlation, we can measure the "strength" of the correlation using the **linear correlation coefficient** r.

Definition (The Linear Correlation Coefficient r)

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{\sum(z_x z_y)}{n-1}$$

§10.2 Correlation

For a linear correlation, we can measure the "strength" of the correlation using the **linear correlation coefficient** r.

Definition (The Linear Correlation Coefficient r)

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{\sum(z_x z_y)}{n-1}$$

Note: The paired data must be a simple random sample of quantitative data whose scatter plot demonstrates an approximate straight-line pattern with outliers arising from known errors in sampling removed.

Determining if r is significant depends on Table A-6 on page 732 in your text. It will give you the critical values for ρ, the population parameter corresponding to r.

§10.2 Correlation

Determining if r is significant depends on Table A-6 on page 732 in your text. It will give you the critical values for ρ, the population parameter corresponding to r. The null hypothesis for determining a linear relationship is always given as below.

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Determining if r is significant depends on Table A-6 on page 732 in your text. It will give you the critical values for ρ, the population parameter corresponding to r. The null hypothesis for determining a linear relationship is always given as below.

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Always reject the null hypothesis if $|r|$ is greater than your critical value. If the sample size n falls between points in the table, you can interpolate the critical value. And if it exceeds the values in the table, you can use technology to perform a P-test on this null hypothesis.

§10.2 Example

For a sample of 12 men, the circumference of their waists (measured in inches) was found to have a correlation coefficient $r = -0.75$ when paired against the distance they could walk in five minutes. Is there evidence to support the claim that the two data points are linearly correlated if we use a significance level of 0.01?

§10.2 Example

For a sample of 12 men, the circumference of their waists (measured in inches) was found to have a correlation coefficient $r = -0.75$ when paired against the distance they could walk in five minutes. Is there evidence to support the claim that the two data points are linearly correlated if we use a significance level of 0.01?

n	$\alpha = .05$	$\alpha = .01$
4	.950	.990
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330

§10.2 Example

For a sample of 12 men, the circumference of their waists (measured in inches) was found to have a correlation coefficient $r = -0.75$ when paired against the distance they could walk in five minutes. Is there evidence to support the claim that the two data points are linearly correlated if we use a significance level of 0.01?

So yes, there is evidence to support that there is a linear correlation between the two pieces of data because $|r| > 0.708$, meaning we reject the null hypothesis that $\rho = 0$.

§10.2 Example

For the data, construct a scatter plot. And determine if there is a linear correlation.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

§10.2 Example

For the data, construct a scatter plot. And determine if there is a linear correlation.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.4961404 + 0.79907895 x
Sample size: 9
R (correlation coefficient) = 0.86611165
R-sq = 0.75014939
Estimate of error standard deviation: 1.5195458

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.4961404	1.2676204	≠ 0	7	1.1802748	0.2764
Slope	0.79907895	0.17430385	≠ 0	7	4.584402	0.0025

§10.2 Example

For the data, construct a scatter plot. And determine if there is a linear correlation.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.4961404 + 0.79907895 x
Sample size: 9
R (correlation coefficient) = 0.86611165
R-sq = 0.75014939
Estimate of error standard deviation: 1.5195458

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.4961404	1.2676204	≠ 0	7	1.1802748	0.2764
Slope	0.79907895	0.17430385	≠ 0	7	4.584402	0.0025

This yields $r = 0.8661$, and since the critical value for $n = 9$ is 0.798, we have evidence to support that there is a linear correlation between the two variables.

§10.2 Example

For the data, construct a scatter plot. And determine if there is a linear correlation.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.4961404 + 0.79907895 x
Sample size: 9
R (correlation coefficient) = 0.86611165
R-sq = 0.75014939
Estimate of error standard deviation: 1.5195458

Parameter estimates:

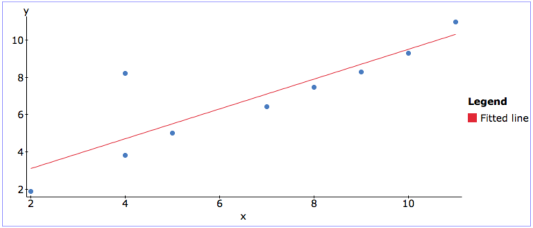
Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.4961404	1.2676204	≠ 0	7	1.1802748	0.2764
Slope	0.79907895	0.17430385	≠ 0	7	4.584402	0.0025

This yields $r = 0.8661$, and since the critical value for $n = 9$ is 0.798, we have evidence to support that there is a linear correlation between the two variables. We could also have used the P-test and seen that the P-value for the **slope** is 0.0025, which is less that $\alpha = 0.01$, telling us that we should reject our null hypothesis $\rho = 0$.

§10.2 Example

For the data, construct a scatter plot. And determine if there is a linear correlation.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9



Notice, the scatter plot confirms this linear correlation with a single outlier. **It's important to look at the scatter plot to determine if the correlation is actually linear!**

§10.3 (Linear) Regression

Definition (Regression Line)

The regression line (or least-squares line or best fit line) is the straight line that best fits the scatter plot of the data. It's given in equation form often.

$$\hat{y} = b_0 + b_1 x$$

§10.3 (Linear) Regression

Definition (Regression Line)

The regression line (or least-squares line or best fit line) is the straight line that best fits the scatter plot of the data. It's given in equation form often.

$$\hat{y} = b_0 + b_1 x$$

The sum of the squares of the distances from this line to all the points in the scatter point is minimum, i.e. any other line will have larger distances in the sum-of-squares sense.

§10.3 (Linear) Regression

Definition (Regression Line)

The regression line (or least-squares line or best fit line) is the straight line that best fits the scatter plot of the data. It's given in equation form often.

$$\hat{y} = b_0 + b_1 x$$

The sum of the squares of the distances from this line to all the points in the scatter point is minimum, i.e. any other line will have larger distances in the sum-of-squares sense.

There are lovely equations for b_0 and b_1 on page 518 on your text, but I will be using technology to compute b_0 and b_1.

§10.3 Example

For the data, find the equation of the regression line.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

§10.3 Example

For the data, find the equation of the regression line.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.4961404 + 0.79907895 x
Sample size: 9
R (correlation coefficient) = 0.86611165
R-sq = 0.75014939
Estimate of error standard deviation: 1.5195458

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.4961404	1.2676204	≠ 0	7	1.1802748	0.2764
Slope	0.79907895	0.17430385	≠ 0	7	4.584402	0.0025

§10.3 Example

For the data, find the equation of the regression line.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.4961404 + 0.79907895 x
Sample size: 9
R (correlation coefficient) = 0.86611165
R-sq = 0.75014939
Estimate of error standard deviation: 1.5195458

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.4961404	1.2676204	≠ 0	7	1.1802748	0.2764
Slope	0.79907895	0.17430385	≠ 0	7	4.584402	0.0025

From the output table, we get that the equation is

$$\hat{y} = 1.49 + 0.799x.$$

§10.3 Example

For the data, find the equation of the regression line.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.4961404 + 0.79907895 x
Sample size: 9
R (correlation coefficient) = 0.86611165
R-sq = 0.75014939
Estimate of error standard deviation: 1.5195458

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1.4961404	1.2676204	≠ 0	7	1.1802748	0.2764
Slope	0.79907895	0.17430385	≠ 0	7	4.584402	0.0025

From the output table, we get that the equation is

$$\hat{y} = 1.49 + 0.799x.$$

Thus, $b_0 = 1.49$ and $b_1 = 0.799$.

§10.3 Concepts

You can use the regression line to extrapolate/interpolate corresponding values - simply plug in.

§10.3 Concepts

You can use the regression line to extrapolate/interpolate corresponding values - simply plug in. There are limitations to extrapolation - you want to make sure that the value you're plugging in isn't too far from the observed values.

§10.3 Concepts

You can use the regression line to extrapolate/interpolate corresponding values - simply plug in. There are limitations to extrapolation - you want to make sure that the value you're plugging in isn't too far from the observed values.

Definition (Marginal Change)

When two variables are related by a regression line, the marginal change in a variable is the amount that the output changes when the input is increased by exactly one unit.

§10.3 Concepts

You can use the regression line to extrapolate/interpolate corresponding values - simply plug in. There are limitations to extrapolation - you want to make sure that the value you're plugging in isn't too far from the observed values.

Definition (Marginal Change)

When two variables are related by a regression line, the marginal change in a variable is the amount that the output changes when the input is increased by exactly one unit.

Definition (Outlier/Influential Point)

In a scatter plot, an outlier is a point lying far from the others. An influential point is one which greatly affects the regression line.

§10.3 Example

Use the equation of the regression line to estimate the value of $\hat{y}$ when x is 6 (using our previous example).

§10.3 Example

Use the equation of the regression line to estimate the value of $\hat{y}$ when x is 6 (using our previous example).

Well, we simply plug in!

$$\hat{y} = 1.49 + 0.799(6) = 6.284.$$

§10.3 Example

Use the equation of the regression line to estimate the value of $\hat{y}$ when x is 6 (using our previous example).

Well, we simply plug in!

$$\hat{y} = 1.49 + 0.799(6) = 6.284.$$

What's the marginal change?

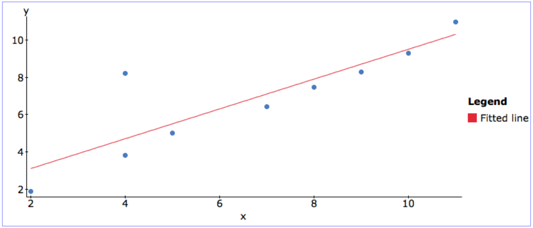
§10.3 Example

Use the equation of the regression line to estimate the value of $\hat{y}$ when x is 6 (using our previous example).

Well, we simply plug in!

$$\hat{y} = 1.49 + 0.799(6) = 6.284.$$

What's the marginal change? It's the slope! Were there any influential points?



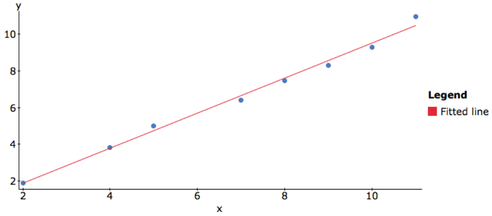
§10.3 Example

Use the equation of the regression line to estimate the value of $\hat{y}$ when x is 6 (using our previous example).

Well, we simply plug in!

$$\hat{y} = 1.49 + 0.799(6) = 6.284.$$

What's the marginal change? It's the slope! Were there any influential points?



§10.4 Prediction Intervals and Variation

Definition (Prediction Interval)

A range of values used to estimate a dependent variable is called a prediction interval.

§10.4 Prediction Intervals and Variation

Definition (Prediction Interval)

A range of values used to estimate a dependent variable is called a prediction interval.

To construct a prediction interval from a regression line, we simply calculate the following.

$$\hat{y} - E < y < \hat{y} + E,$$

where $\hat{y}$ is the point estimate obtained from the regression line and E is given by

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}.$$

The standard error of estimate s_e can be calculated in StatCrunch or by using formulas 10-5 or 10-6 on page 532.

§10.4 Example

A survey conductor wants to know if a certain drug can be transferred to a child by its nursing mother. She sampled seven pairs of mothers and children. The concentration of the drug in the mother's body x was related to the concentration in the child's body by the regression line $\hat{y} = 0.033 + 0.91x$ with $\bar{x} = 1.046$, $\sum x = 6.05$, $\sum x^2 = 9.6$, and $s_e = 0.11$. Calculate the prediction interval with $\alpha = 0.1$ for $x = 1.3$.

§10.4 Example

A survey conductor wants to know if a certain drug can be transferred to a child by its nursing mother. She sampled seven pairs of mothers and children. The concentration of the drug in the mother's body x was related to the concentration in the child's body by the regression line $\hat{y} = 0.033 + 0.91x$ with $\bar{x} = 1.046$, $\sum x = 6.05$, $\sum x^2 = 9.6$, and $s_e = 0.11$. Calculate the prediction interval with $\alpha = 0.1$ for $x = 1.3$.

$$E = (4.0322)(0.11)\sqrt{1 + \frac{1}{7} + \frac{7(1.3 - 1.046)^2}{7(9.6) - (6.05)^2}} = 0.477.$$

§10.4 Example

A survey conductor wants to know if a certain drug can be transferred to a child by its nursing mother. She sampled seven pairs of mothers and children. The concentration of the drug in the mother's body x was related to the concentration in the child's body by the regression line $\hat{y} = 0.033 + 0.91x$ with $\bar{x} = 1.046$, $\sum x = 6.05$, $\sum x^2 = 9.6$, and $s_e = 0.11$. Calculate the prediction interval with $\alpha = 0.1$ for $x = 1.3$.

$$E = (4.0322)(0.11)\sqrt{1 + \frac{1}{7} + \frac{7(1.3 - 1.046)^2}{7(9.6) - (6.05)^2}} = 0.477.$$

Also $\hat{y} = 0.033 + 0.91(1.3) = 1.216$.

§10.4 Example

A survey conductor wants to know if a certain drug can be transferred to a child by its nursing mother. She sampled seven pairs of mothers and children. The concentration of the drug in the mother's body x was related to the concentration in the child's body by the regression line $\hat{y} = 0.033 + 0.91x$ with $\bar{x} = 1.046$, $\sum x = 6.05$, $\sum x^2 = 9.6$, and $s_e = 0.11$. Calculate the prediction interval with $\alpha = 0.1$ for $x = 1.3$.

$$E = (4.0322)(0.11)\sqrt{1 + \frac{1}{7} + \frac{7(1.3 - 1.046)^2}{7(9.6) - (6.05)^2}} = 0.477.$$

Also $\hat{y} = 0.033 + 0.91(1.3) = 1.216$. So our prediction interval is $(0.739, 1.693)$.

§10.4 Coefficient of Determination

Definition (Coefficient of Determination)

The coefficient of determination is given by r^2. But it can also be calculated by the ratio $r^2 = \dfrac{\text{explained variation}}{\text{total variation}}$.

§10.4 Coefficient of Determination

Definition (Coefficient of Determination)

The coefficient of determination is given by r^2. But it can also be calculated by the ratio $r^2 = \dfrac{\text{explained variation}}{\text{total variation}}$.

The total deviation $(y - \bar{y})$ is the vertical distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$. The explained deviation $(\hat{y} - \bar{y})$ is the distance between the point $\hat{y}$ and $\bar{y}$. The unexplained deviation $(y - \hat{y})$ is the distance between the point $\hat{y}$ and the y. See Formula 10-7 on page 535.

§10.4 Example

For the paired data below, calculate the explained variation, the total variation, and the prediction interval for $x = 3$.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

§10.4 Example

For the paired data below, calculate the explained variation, the total variation, and the prediction interval for $x = 3$.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

We get the explained variation as 48.53, the total variation as 64.69, and the prediction interval as (-0.1845, 7.9713).

§10.4 Example

For the paired data below, calculate the explained variation, the total variation, and the prediction interval for $x = 3$.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

We get the explained variation as 48.53, the total variation as 64.69, and the prediction interval as (-0.1845, 7.9713). Also, $r^2 = 0.750$, so 75% of the variation can be explained by the linear relationship between the two variables.

§10.4 Example

For the paired data below, calculate the explained variation, the total variation, and the prediction interval for $x = 3$.

x	7	8	10	5	11	9	4	4	2
y	6.42	7.48	9.3	5	10.98	8.29	3.82	8.22	1.9

We get the explained variation as 48.53, the total variation as 64.69, and the prediction interval as (-0.1845, 7.9713). Also, $r^2 = 0.750$, so 75% of the variation can be explained by the linear relationship between the two variables. When going from r^2 to r, make sure to give r the sign of the slope of the regression line!

§A Note

It's important to remember that the regression line does not **always** give the best predicted value $\hat{y}$ for a value x. This is only the case when the hypothesis test suggests that the paired data is linearly correlated! If they are not linearly correlated, the best predicted value for **any** value x is the mean of the y-values, i.e. $\hat{y} = \bar{x}$.