

# 19. SIMPLE LINEAR REGRESSION IV

## The Coefficient of Determination, $R^2$

Once we have decided that  $\beta$  is not zero, so that a linear relationship seems to exist between  $x$  and  $y$ , it is useful to measure the *strength* of this linear relationship. Such a measure is provided by the **coefficient of determination**,  $R^2$ .

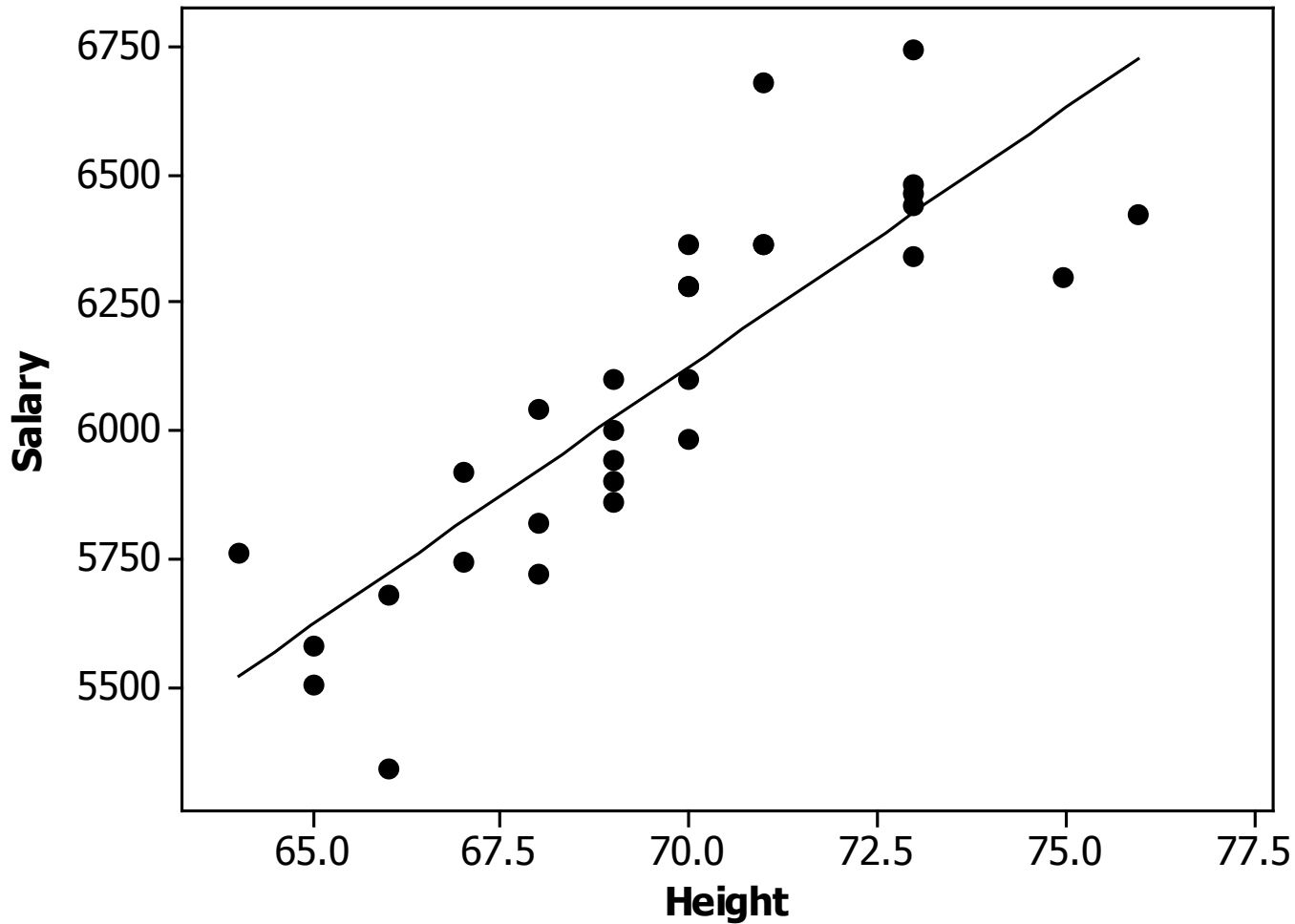
To understand  $R^2$ , note that one of the aims of regression analysis is to study the relationship between  $x$  and  $y$ , i.e., to try to use the value of  $x$  to "explain"  $y$ .

- Keep in mind, though, that this "explanation" may not be one of cause and effect.

Recall the Salary vs. Height data.

### Fitted Line Plot for Salary vs. Height

$$\text{Salary} = -902.2 + 100.4 \text{ Height}$$



S	192.702
R-Sq	71.4%
R-Sq(adj)	70.3%

## Regression Analysis: Salary versus Height

### Analysis of Variance

Source	DF	SS	MS	F-Value	P-Value
Regression	1	2590433	2590433	69.76	0.000
Error	28	1039754	37134		
Total	29	3630187			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
192.702	71.36%	70.34%	65.65%

### Coefficients

Term	Coef	SE	Coef	T-Value	P-Value
Constant	-902		837	-1.08	0.290
Height	100.4		12.0	8.35	0.000

### Regression Equation

$$\text{Salary} = -902 + 100.4 \text{ Height}$$

If we look at the  $y$ 's (the salaries) as a data set, we note that they are not all the same; the  $y$ 's exhibit variability. A rough measure of this variability is the total sum of squares,

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 .$$

Note that SST is  $(n-1)$  times the sample variance of the  $y$ 's.

If there is a linear relationship between  $x$  and  $y$ , then the variability of the  $y$ 's is not due entirely to chance fluctuations.

Instead, the fact that the salaries are different can be partially "explained" by the fact that the heights ( $x$ ) are different. Of course, salary is not completely explained by height, so part of the variability in the salaries remains unexplained.

- Interestingly, the variability in salaries can be broken into two parts, the first attributed to differences in height, and the second attributed to other factors not yet accounted for.

- We have the following important formula:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 ,$$

or  $SST = SSR + SSE$ .

### **Interpretation:**

The variability of the  $y$ 's ( $SST$ ) can be broken into two parts,  $SSR + SSE$ .

- The first part is the regression sum of squares,  $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ .

This is simply  $(n-1)$  times the sample variance of the fitted values.

Since  $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$  the fluctuation in the  $\hat{y}_i$ 's is completely "explained" by the fluctuation in the  $x_i$ 's.

Thus, *SSR* is the part of the variability of  $y$  which can be "explained" by  $x$  (or, more precisely, by the *regression* of  $y$  on  $x$ ).

• The second part is the residual sum of squares,  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .

The residuals represent the part of  $y$  which remains after we try to "explain"  $y$  based on  $x$ .

(This is clear, since  $y_i = \hat{y}_i + e_i$ .)

Thus, *SSE* represents the part of the variability of  $y$  which is "unexplained" by  $x$ .

This unexplained variability is attributed to chance, or to other factors not yet considered.

Now, we define the **coefficient of determination** by  $R^2 = \frac{SSR}{SST}$ .

- We see that  $R^2$  measures the proportion of the variability of  $y$  that is "explained" by  $x$ .

An equivalent definition is  $R^2 = 1 - \frac{SSE}{SST}$ .

- It can be shown that  $0 \leq R^2 \leq 1$ .

We will get  $R^2 = 1$  if, and only if, all points lie exactly on a straight (non-horizontal) line.

The closer  $R^2$  is to 1, the stronger the linear relationship between  $x$  and  $y$ .

If  $R^2$  is near zero, then almost none of the variability of  $y$  is explained by  $x$ , so the linear relationship is weak.

We will get  $R^2 = 0$  if, and only if,  $\hat{\beta} = 0$ .

This can happen in a variety of ways, including:

(1) All  $y$ 's lie on a horizontal line;

(2) The data points lie on a parabola  $y = a + b x^2$ , which peaks in the middle of the range of the equally-spaced  $x$ 's.

- Note that in (2), there is a clear nonlinear relationship but no linear relationship whatsoever! So keep in mind that  $R^2$  only measures the strength of the *linear* relationship.

If  $R^2$  is large, we say that  $x$  and  $y$  are "highly correlated". In this case, there is a strong linear relationship between  $x$  and  $y$ .

If  $R^2$  is near zero, we say that  $x$  and  $y$  are nearly "uncorrelated". In this case, the linear relationship is weak.



Note that  $R^2$  is the square of the **correlation coefficient**  $r$ .  
But the use of  $r$  is potentially misleading.

For example, if  $r = 0.8$ , then only 64% of the variability in  $y$  is "explained" by  $x$ .

The coefficient of determination  $R^2$  contains the same information as  $r$  (except for the sign of the slope), and has the interpretation as the proportion of explained variability.

Note that a high correlation should not be taken as evidence of a causal relationship. Consider the Rotten Tomatoes example. The ultimate explanation of a high Audience Score is presumably that the movie was popular. This is presumably what causes the Critics Score to also be high. So it's not that high Critics Scores are the *cause* (or the explanation, in any meaningful sense) of high Audience Scores.

**Eg:** Since 71.36% of the variability in salary is “explained” by height, the linear relationship is strong. Height is a good predictor of salary. The other 28.64% of the variability in salary is unexplained, but we could try to include more variables in our regression. This would definitely improve the  $R^2$ .  
(We will return to this point later.)

**Eg:** For the Stock Market example, the Minitab output shows that  $R^2 = 0.0506$ . Only 5.06% of the variability in Today's returns is "explained" by Yesterday's returns.

Although the linear relationship is statistically significant (low  $p$ -value), it is still quite weak (low  $R^2$ ).

The forecast of today's return based on yesterday's return will not be very accurate.

## Regression Analysis: Today versus Yesterday

### Analysis of Variance

Source	DF	SS	MS	F-Value	P-Value
Regression	1	100.0	100.042	71.32	0.000
Error	1338	1876.7	1.403		
Total	1339	1976.8			

### Model Summary

S	R-sq	R-sq(adj)
1.18433	5.06%	4.99%

### Coefficients

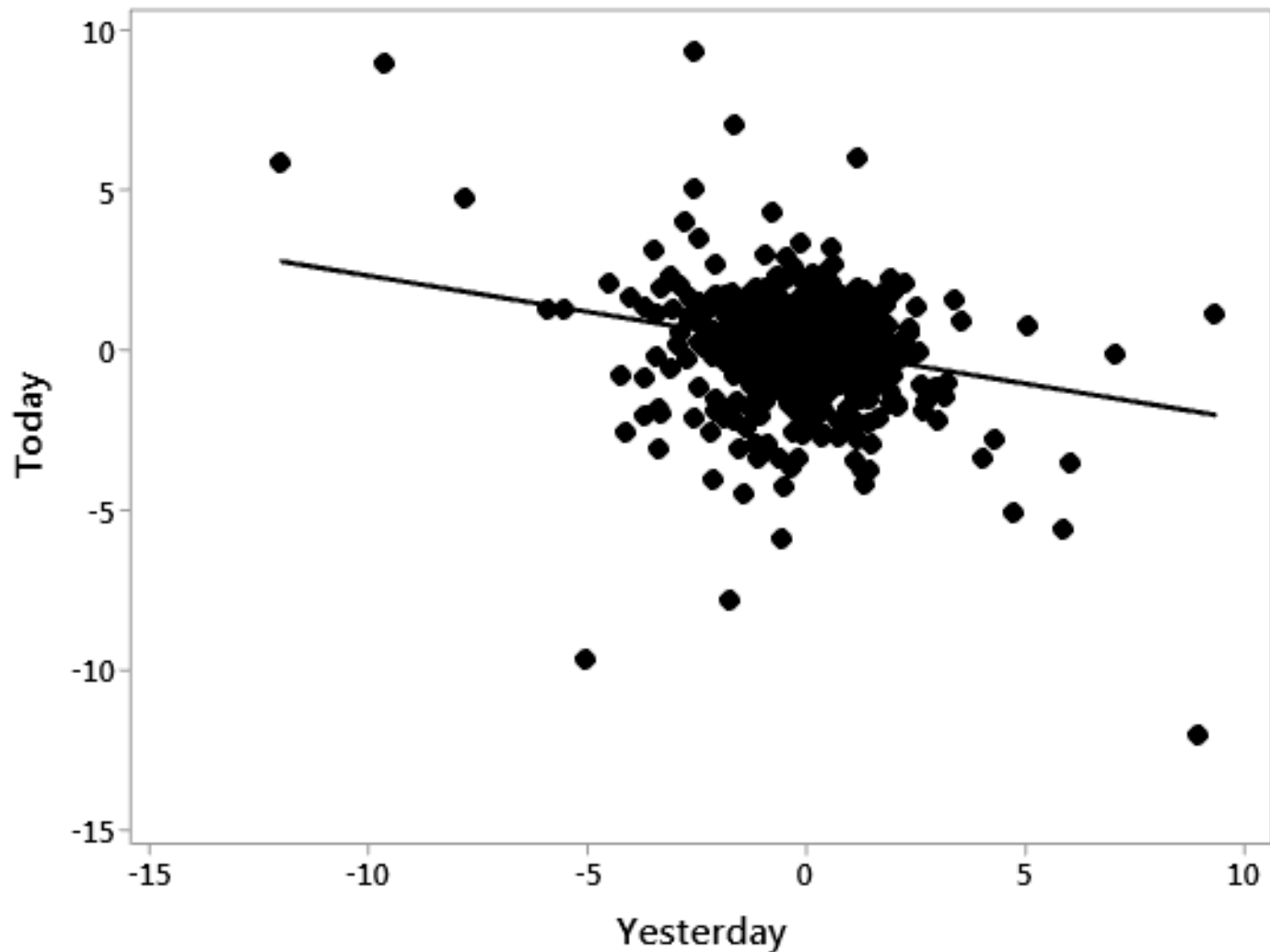
Term	Coef	SE Coef	T-Value	P-Value
Constant	0.0846	0.0324	2.61	0.009
Yesterday	-0.2249	0.0266	-8.45	0.000

### Regression Equation

Today = 0.0846 - 0.2249 Yesterday

# Market Returns

Today = 0.08459 - 0.2249 Yesterday



S	1.18433
R-Sq	5.1%
R-Sq(adj)	5.0%

**[R Demo: LeastSquaresFitWithRsquare]**