Lesson 2 Simple Linear Regression

Introduction

In this lesson, we will look at the relationship or association between two continuous or quantitative variables. Often times, associations between more than two variables are of interest. However, the tools for examining more than two variables build on the tools that we use to evaluate the association between pairs of variables. As such, we will start building our tool box and learning how to visualize, understand, and describe the association in the setting of two continuous variables. We will expand upon these tools to understand the relationship between 3 or more variables in the next lesson.

Note: Quantitative variables (numeric variables): associated with a numeric measurement.

Outline

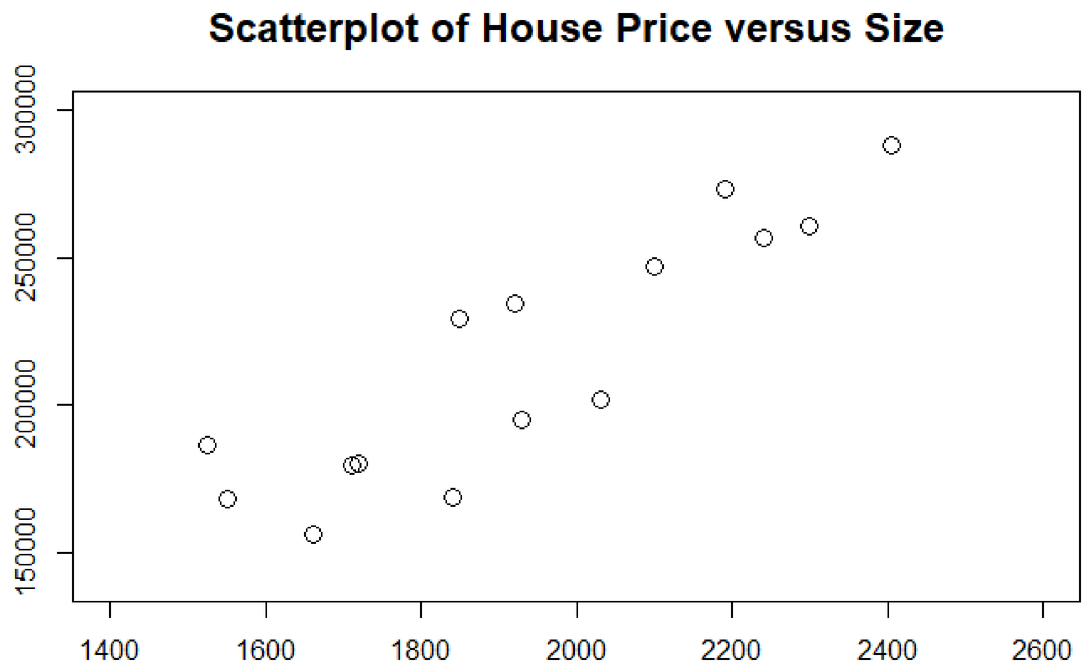
- In examining the relationship between two continuous variables, we will:
  - Visualize the relationship (plot the data via scatterplots) and describe the form, direction, and strength of the association.
  - Use numerical summaries to help us describe (if appropriate)
    - the strength of the association (via correlation)
    - the relationship between the variables (via simple linear regression)
      - Estimation of the parameters by least squares
      - Using the equation for Predictions
      - Assessing the Fit of the Regression Line

Scatterplots

- Scatterplots are the most useful way to graphically display the relationship between two continuous or quantitative variables. They show the relationship between two "paired" variables. The values of one variable are shown on the horizontal axis (x-axis) while values for the other variable are shown on the vertical axis (y-axis). Each "pair" of data is shown in the graph with one single point.

Example

Size (sqft)	House Price
1850	$229500
2190	$273300
2100	$247000
1930	$195100
2300	$261000
1710	$179700
1550	$168500
1920	$234400
1840	$168800
1720	$180400
1660	$156200
2405	$288350
1525	$186750
2030	$202100
2240	$256800



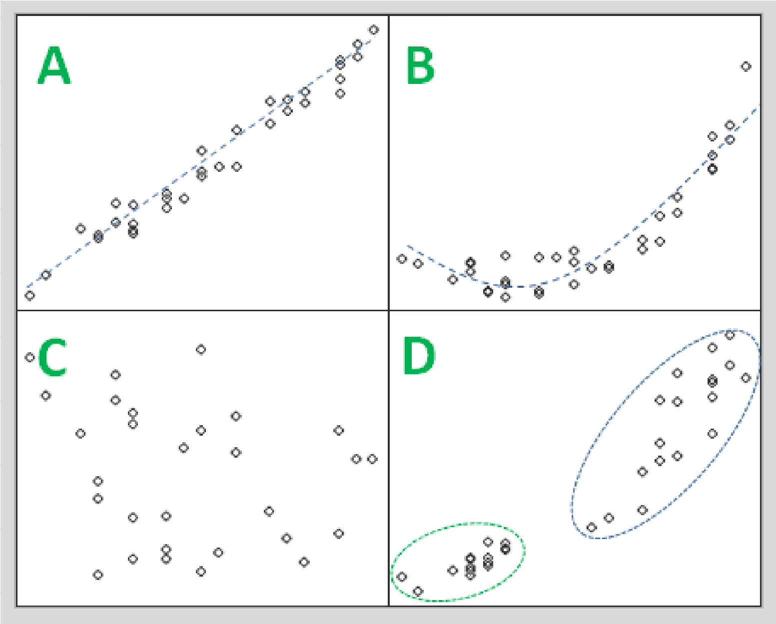
plot(housedata$Size,housedata$HousePrice)

Interpreting Scatterplots

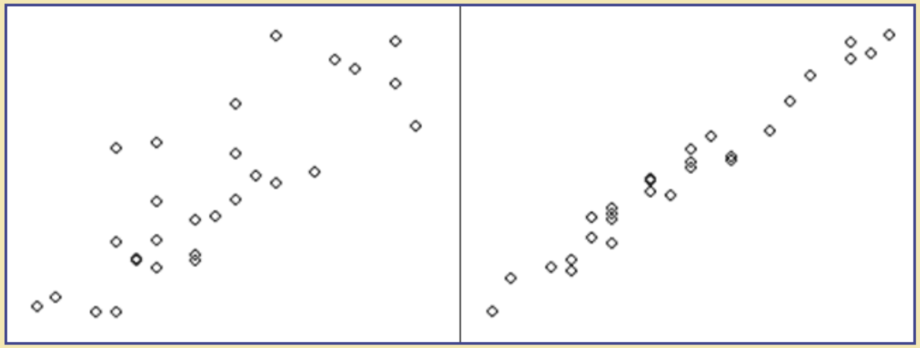
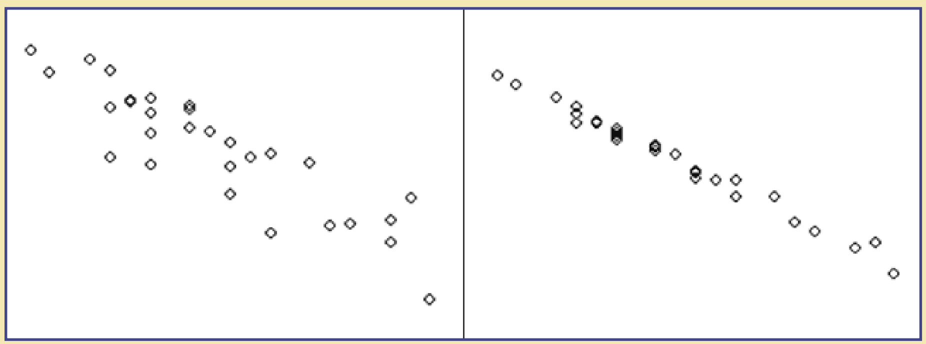
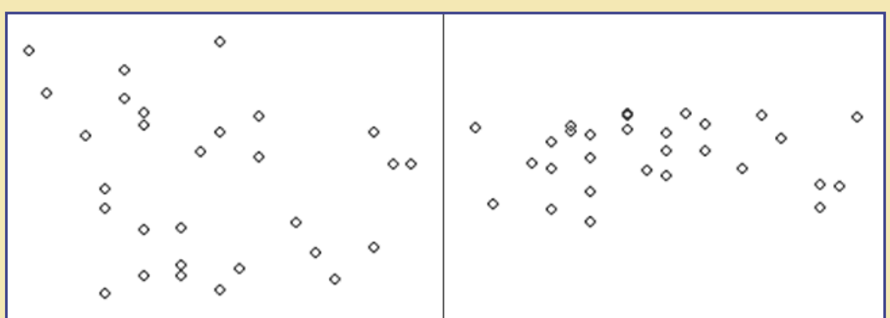
- By viewing a scatterplot, you can assess the (1) **form**, (2) **direction**, and (3) **strength** of the relationship between the two variables.

Interpreting Scatterplots - Form

- When examining a scatterplot, the form of the relationship should be noted. Relationships between variables may be described as linear (where the points tend towards a straight line pattern as in Figure A below), curved (where the points tend toward a U-shape or arced pattern as in Figure B below), or random (where the points don’t seem to follow any pattern as in Figure C below). Clusters may also be apparent (as in Figure D below).

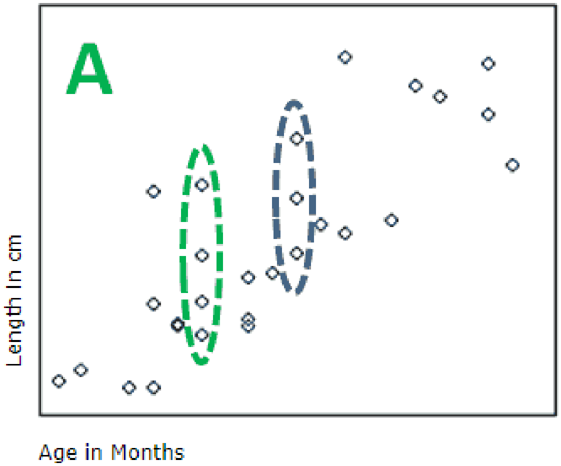


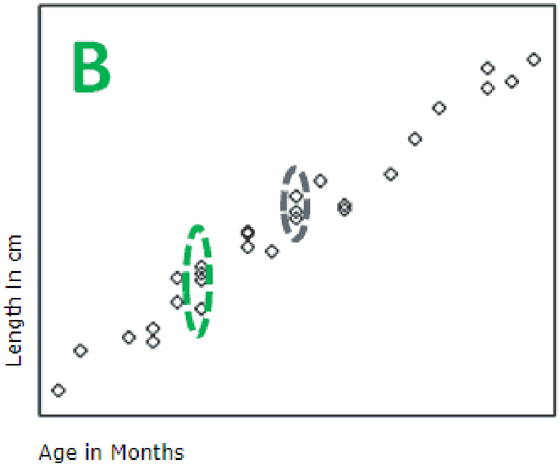
Interpreting Scatterplots - Direction

Positive Association		As one variable increases in value, the other variable also tends to increase in value
Negative Association		As one variable increases in value, the other variable tends to decrease in value
No Association		Two variables with no association tend to look like points hovering around a flat line (a horizontal line)

Interpreting Scatterplots - Strength of the Relationship

- Strength of the association between two variables is how closely the points appear to follow a clear form or pattern. For the following two scatterplots A and B, though both show variables that are positively associated, the variables in B are more strongly associated. This is apparent when looking at the length in centimeters of individual koalas with the same ages. At Zoo A, the range of values of length is quite variable for similar aged koalas. At Zoo B, the range of values is quite smaller for koalas with the same age.





The age in months vs length in centimeters of baby koala bears

Correlation

- Correlation (denoted as r) or the correlation coefficient is a measure of the strength and direction of a linear relationship between two quantitative variables in a sample.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

- where n is the number of data points (or pairs), x_i is the ith data point for variable x, $\bar{x}$ is the sample mean of variable x (across all data points), s_x is the sample standard deviation of variable x (across all data points), y_i is the ith data point for variable y, $\bar{y}$ is the mean of variable y, s_y is the sample standard deviation of variable y.

Exercise

- Compute the sample correlation between the ages of husbands and wives using the data below.

Age of Wife	Age of Husband
20	20
30	32
24	22
28	26
28	30
Sample mean: 26	Sample mean: 26
Sample standard deviation: 4.0	Sample standard deviation: 5.1

Compute Correlation – Using R

Size (sqft)	House Price
1850	$229500
2190	$273300
2100	$247000
1930	$195100
2300	$261000
1710	$179700
1550	$168500
1920	$234400
1840	$168800
1720	$180400
1660	$156200
2405	$288350
1525	$186750
2030	$202100
2240	$256800

```
> #Calculate Sample Correlation
> cor(housedata$Size,housedata$HousePrice)
[1] 0.8911306
> cor(housedata$HousePrice,housedata$Size)
[1] 0.8911306
```

Properties of Correlation

- The correlation takes on values between −1 and +1. The correlation is positive (>0) when there is a positive association between variables and negative (<0) then there is a negative association between variables. A correlation of 0 indicates that there is not an association between the variables. The closer the value of the correlation to the extremes (−1 or +1), the stronger the associations between the variables (the closer the points lie to a straight line). In fact, a correlation of −1 or +1 indicates that the points on a scatterplot lie perfectly along a straight line.
- The correlation between variables x and y is the same as the correlation between variables y and x.
- Correlations can be computed between paired values of two quantitative variables. You cannot use correlation to compute the correlation between gender and SAT scores, since gender is not quantitative (it is qualitative).
- Correlation measures the strength of a linear relationship only. Correlation should not be used to describe a curved relationship - even if the association is strong.

Simple Linear Regression (SLR)

- Assert a straight line on the scatterplot that represents the best fitting line to the data that captures the pattern of the relationship.
- The equation for the simple linear regression line is given by

$$y=\beta_0+\beta_1 x$$

where

  - y is the response or dependent variable
  - x is the explanatory or independent variable
  - β_0 is the intercept (the value of y when x=0)
  - β_1 is the slope (the expected change in y for each one-unit change in x)
- Regression line – the graph of the regression equation
  - Also known as the “line of best fit” or the “least square line”

Requirement for SLR

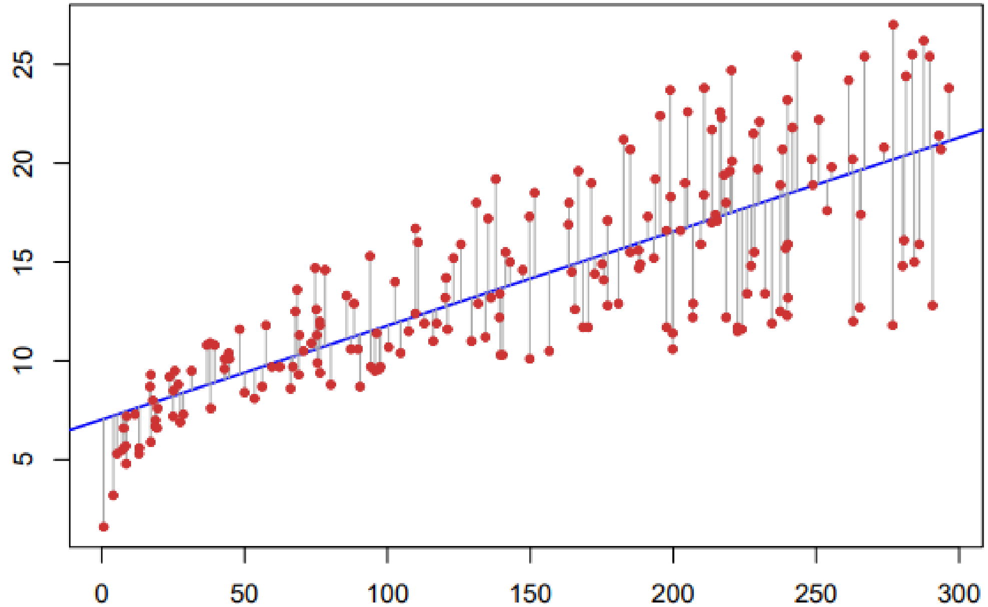
- The sample of paired data is a simple random sample of quantitative data.
- The pairs of data x, y have a **bivariate normal distribution**, meaning the following:
  - Visual examination of the scatter plot(s) confirms that the sample points follow an approximately straight line(s)
  - Because results can be strongly affected by the presence of outliers, any outliers should be removed if they are known to be errors.

    <u>Outlier</u>: in a scatter plot, an outlier is a point lying far away from the other data points.

Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for y based on the ith value of x. Then $e_i = y_i - \hat{y}_i$ represents the ith residual. We define the residual sum of squares (RSS) as

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$



Estimation of the parameters by least squares

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = r\frac{s_y}{s_x} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(See the Appendix at the end of the lecture: Derivation of least square best fit formula)

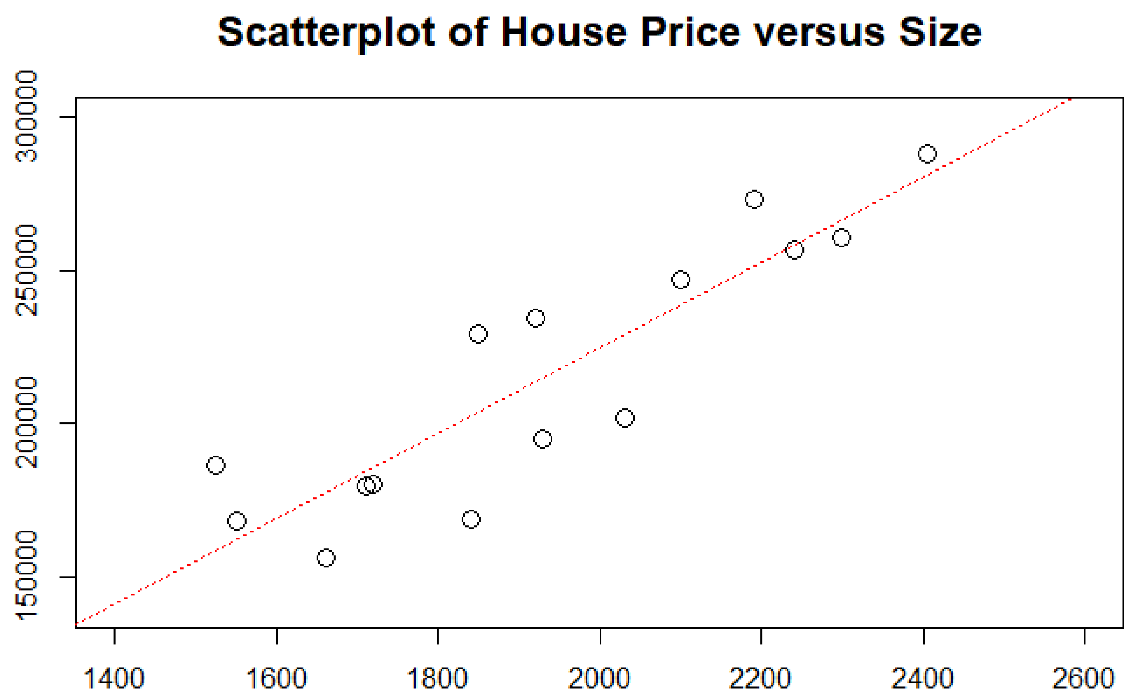
Exercise

- Previously, we calculated the correlation ($r \approx 0.93$) between the ages of husbands and wives from a small sample. Now, find the least-squares regression equation predicting husband age from the age of the wife.

Age of Wife	**Age of Husband**
20	20
30	32
24	22
28	26
28	30
Sample mean: 26	Sample mean: 26
Sample standard deviation: 4.0	Sample standard deviation: 5.1

SLR Using R

Size (sqft)	House Price
1850	$229500
2190	$273300
2100	$247000
1930	$195100
2300	$261000
1710	$179700
1550	$168500
1920	$234400
1840	$168800
1720	$180400
1660	$156200
2405	$288350
1525	$186750
2030	$202100
2240	$256800




```
m<-lm(housedata$HousePrice~housedata$Size)
#Adding regression line to the current plot
abline(m, col="red")
#Request important summary information
summary(m)
```

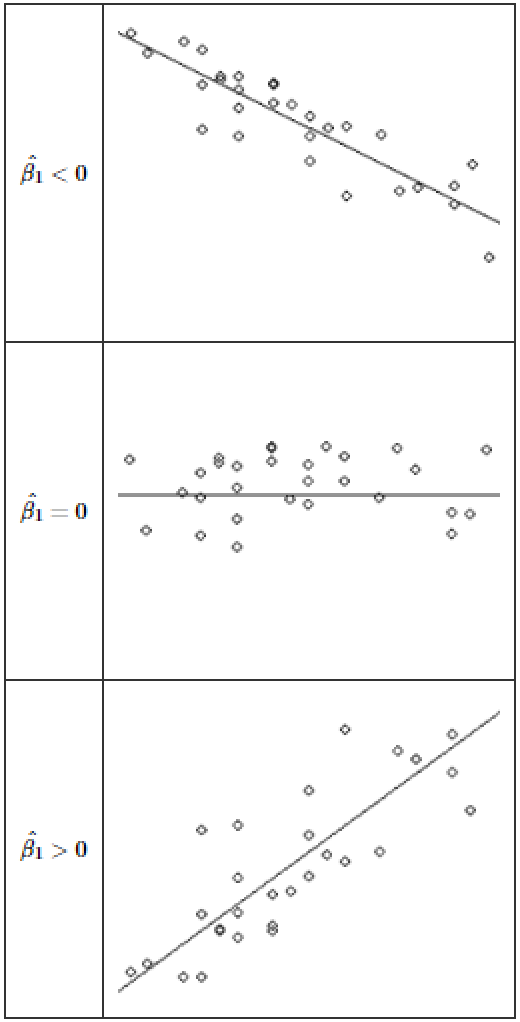
SLR Using R

```
> summary(m)

Call:
lm(formula = housedata$HousePrice ~ housedata$Size)

Residuals:
   Min     1Q Median     3Q    Max 
-33654 -12761  -1447  14534  28233 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -54191.2    38399.4  -1.411    0.182    
housedata$Size    139.5       19.7   7.081 8.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20220 on 13 degrees of freedom
Multiple R-squared:  0.7941,	Adjusted R-squared:  0.7783 
F-statistic: 50.14 on 1 and 13 DF,  p-value: 8.28e-06
```

Regression Equation: House Price ($) = 139.5 * Size(sqft) – 54191.2

Interpretation

- The estimate of the slope parameter $\hat{\beta}_1$ gives the expected (average) or predicted change in the response variable (y) for a one-unit increase in the explanatory variable (x). The estimate of the slope parameter $\hat{\beta}_1$ also gives insight into the direction of the relationship between the variables.
- What is the relationship between r and $\hat{\beta}_1$?
  - Answer: r and $\hat{\beta}_1$ have same sign.



Using the equation for Predictions

- The equation of the least-squares regression line can be used to predict the expected value of the response variable for new values of the explanatory variable. This is done by substituting the new value of x into the regression equation and calculating the associated value for $\hat{y}$. The predicted value $\hat{y}$ for a given value of x can be interpreted as the average value of the response for the given value of x.
- Example:
  - The least-squares regression line that describes the relationship between size and price of house is given by

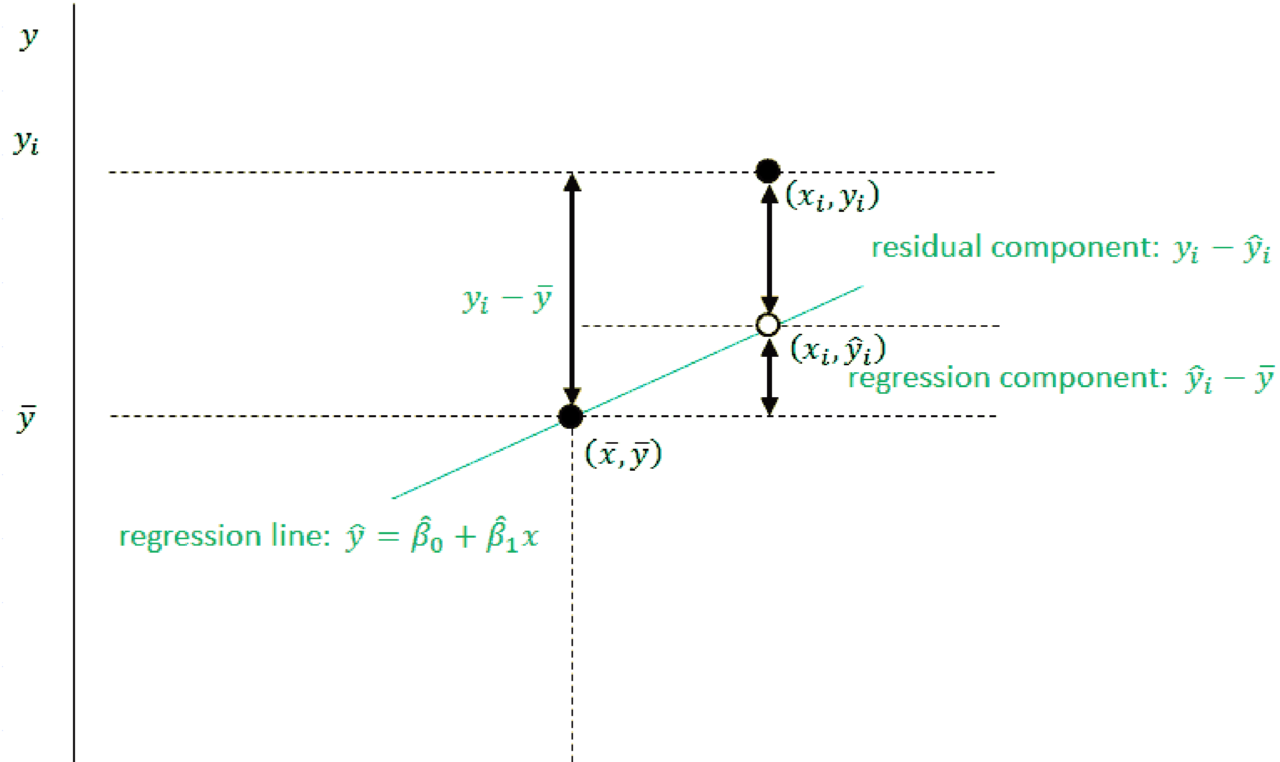
$$\hat{y} = 139.5 * x - 54191.2$$

  - If a house is 2500 sqft, what is the expected price?
  - We can predict the expected house price by plugging in x=2500 into the regression equation. That is, the expected house price is

$$\hat{y} = 139.5 * 2500 - 54191.2 \approx 294559.$$

Assessing the Fit of the Regression Line

- Notice the regression line always goes through the point$(\bar{x},\ \bar{y})$
- For any given data point, the difference between the mean response $\bar{y}$ and the observed response value y_i can be split into two parts: (1) the regression component and the (2) residual component. For any sample point (x_i, y_i)
  - the regression component = $(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y} = \hat{y}_i - \bar{y}$
  - the residual component = $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$

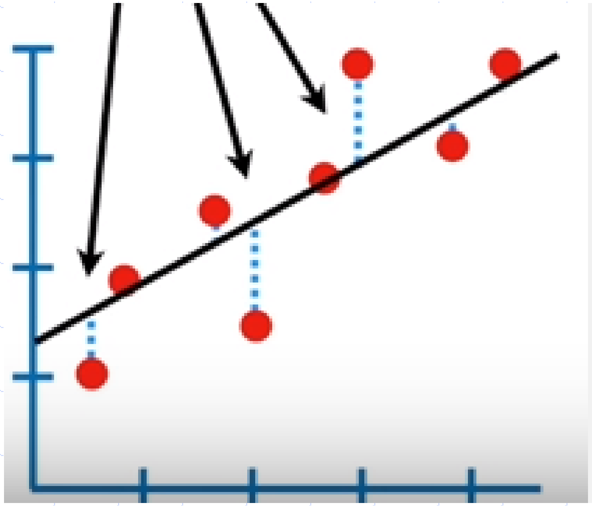


Assessing the Fit of the Regression Line

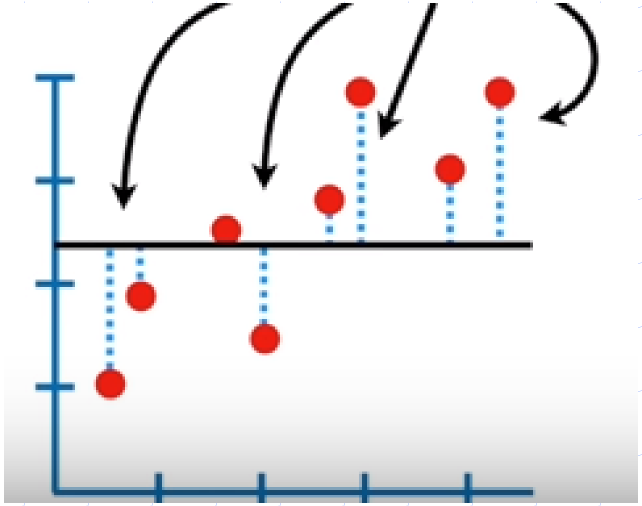
- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ (called the total sum of squares or Total SS) represents the sum of squares of the deviations of the individual sample points from the sample mean
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (called the residual sum of squares or Res SS) represents the sum of squares of the residual components
- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (called the regression sum of squares or Reg SS) represents the sum of squares of the regression components
- That is, Total SS = Res SS + Reg SS.
- One of the measures that we use to assess the fit of the data is the coefficient of variation (R^2, read "R-squared").
  - $R^2 = \frac{Reg\ SS}{Total\ SS}$

Visualize the sums

- Total Sum of Squares: $\sum_{i=1}^{n}(y_i - \bar{y})^2$ represents the sum of squares of the deviations of the individual sample points from the sample mean (as shown in the picture on the right)
- Residual Sum of Squares: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ represents the sum of squares of the residual components (as shown in the picture on the left)



Good fit



Bad fit

Analysis of R-Squared

- $R^2 = \frac{Reg\ SS}{Total\ SS} = \frac{Total\ SS - Res\ SS}{Total\ SS} = 1 - \frac{Res\ SS}{Total\ SS}$
- R^2=1 implies ResSS = 0. (In that case, all data points fell perfectly along the regression line).
- On the other hand, R^2=0 implies RegSS = 0. (In that case, the regression line is a bad fit as in previous slide.)
- The value of R^2 can range between 0 and 1, and the higher its value the more accurate the regression model is.

Summary

- In this lesson, we have tried to understand the relationship or association between two continuous or quantitative variables. Scatterplots help visualize the relationship between two quantitative variables. Using a scatterplot, we can describe the form, strength and direction of the relationship between variables. Correlation helps us quantify the strength and direction between two quantitative variables that are linearly associated. SLR allowed us a convenient form to characterize the relationship.

Appendix: (Optional) Derivation of least square best fit formula

Calculus Review

1. *Finding local max and min of function of one variable.* If $f(x)$ is twice differentiable and f has a (local) minimum at $(a, f(a))$, then $f'(a) = 0$. Moreover, if it happens that $f'(a) = 0$, then $(a, f(a))$ *must be* a minimum (rather than a maximum or neither max nor min) if, in addition, we have $f''(a) > 0$. (Example: If $f(x) = (x-1)^2$, then $f'(x) = 2(x-1)$ and $f''(x) = 2$, so we have $f'(1) = 0$ and $f''(1) > 0$. It follows that $(1, 0)$ is a minimum for f.)

2. *Finding local min of a function of two variables.* Given a function $f(x, y)$, a *minimum* for f is a pair (a, b) such that $f(a, b) \leq f(x, y)$ for all x, y. Suppose f has all first and second partial derivatives. If f does have a minimum at (a, b), then it must be the case that

$$\frac{\partial f}{\partial x}(a, b) = 0 = \frac{\partial f}{\partial y}(a, b). \tag{1}$$

Moreover, if (1) holds and the following *second derivative criteria* are satisfied, then f *must* have a minimum at (a,b):

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2}(a,b) \cdot \frac{\partial^2 f}{\partial y^2}(a,b) &> \left(\frac{\partial f}{\partial x \partial y}\right)^2 (a,b) \\ \frac{\partial^2 f}{\partial x^2}(a,b) &> 0. \end{aligned}$$

Standard Deviation and Variance

A formula often used for deriving the variance of a theoretical distribution is as follows:

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = E(X - \mu^2).$$

where X is a random variable and E is expected value.

The *population variance* is variance of a finite population. This is computed by translating the general variance formula (second expression above) into the context of finite populations

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

The *sample variance* is a way to estimate population variance by treating a relatively small sample as representative of the whole population. If one translates the population variance, replacing population values with sample values, one obtains:

Proof of equivalence of two variance fmlas

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2\end{aligned}$$

The *sample variance* is a way to estimate population variance by treating a relatively small sample as representative of the whole population. We take a sample with replacement of n values $y_1, \ldots, y_n$ from the population, where $n < N$ (where N is number of individuals in whole population) and estimate the variance on the basis of this sample. Directly taking the variance of the sample data gives the average of the squared deviations:

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

This estimate of population variance tends, on average, to be slightly too small – it is said to be *downward biased.* This is shown by the calculation of the expected value of σ_y^2 (see https://en.wikipedia.org/wiki/Variance)

$$E[\sigma_y^2] = \frac{n-1}{n}\sigma^2$$

A proof that uncorrected formula for variance based on a sample is downward biased

Since the y_i are selected randomly, both $\bar{y}$ and σ_y^2 are random variables. Their expected values can be evaluated by averaging over the ensemble of all possible samples $\{y_i\}$ of size n from the population. For σ_y^2 this gives:

$$\begin{aligned}
E[\sigma_y^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i^2 - \frac{2}{n} y_i \sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{k=1}^{n} y_k\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}E[y_i^2] - \frac{2}{n}\sum_{j\neq i} E[y_i y_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}^{n} E[y_j y_k] + \frac{1}{n^2}\sum_{j=1}^{n} E[y_j^2]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\sigma^2+\mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}(\sigma^2+\mu^2)\right] \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Hence σ_y^2 gives an estimate of the population variance that is biased by a factor of $\frac{n-1}{n}$. For this reason, σ_y^2 is referred to as the *biased sample variance*. Correcting for this bias yields the *unbiased sample variance*:

$$s^2 = \frac{n}{n-1}\sigma_y^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Either estimator may be simply referred to as the *sample variance* when the version can be determined by context.

Correlation Formulas

Population correlation coefficient is usually denoted ρ ("rho").

The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where E is the expected value operator, *cov* means covariance, and *corr* is a widely used alternative notation for the correlation coefficient.

A *biased* estimate of population correlation coefficient based on a given sample (that is, uncorrected, as discussed in slides on variance) is given by

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n s'_x s'_y}$$
$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where s'_x and s'_y are the *uncorrected* sample standard deviations of X and Y.

For the purpose of estimating population correlation coefficient, usually the *corrected* or *unbiased* standard deviation is used. This involves replacing n by n - 1 in the usual formula for variance (as described earlier). We have these unbiased correlation coefficient formulas:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and s_x and s_y are the corrected sample standard deviations of X and Y.

Derivation of least square best fit formula

Let e be the squared residual function defined by $e(y, y') = (y - y')^2$. Recall that the equation of a line with (finite) slope β_1 and y-intercept β_0 is given by $y = \beta_0 + \beta_1 x$.

Given data points $(x_1, y_1), \ldots, (x_n, y_n)$, $(n > 1)$, we let $y = \beta_0 + \beta_1 x$ denote the "best-fit" line for these data points. We attempt to compute β_0 and β_1 using calculus, as we describe next. (Note that we do not expect a uniquely determined best fit line if $n = 1$, so we assume $n > 1$. We also assume that the x_i are all distinct.) We will think of m and b as variables for the possible slope and y-intercept of this optimal line. Denote the RSS function by $E(m, b)$; that is,

$$E(m, b) = \sum_{i=1}^{n} e(y_i, mx_i + b). \tag{2}$$

We obtain the value $(m, b) = (\beta_1, \beta_0)$ as the pair for which $E(m, b)$ is minimum. The values m, b are such that the "distances" between the y-value of the ith data point and the y-value of the line $y = mx + b$ at $x = x_i$ have been minimized.

Derivation (continued)

This approach implies that the "best fit" line has been taken to be the line for which slope and y-intercept value minimizes the sum of the squares of differences between these y-values; there are other types of minimization that could be considered (such as minimizing sum of the *absolute values* of differences between y values).

Expanding (2), we obtain

$$
\begin{aligned}
E(m,b) &= \sum_{i=1}^{n}(y_i - mx_i - b)^2 \\
&= \sum_{i=1}^{n}\left(y_i^2 + m^2x_i^2 + b^2 - 2mx_iy_i - 2by_i + 2bmx_i\right) \\
&= \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} m^2x_i^2 + \sum_{i=1}^{n} b^2 - \sum_{i=1}^{n} 2mx_iy_i - \sum_{i=1}^{n} 2by_i + \sum_{i=1}^{n} 2bmx_i
\end{aligned}
$$

Derivation (continued)

To make the derivative computations more readable, we define the following constants:

$$X = \sum_{i=1}^{n} x_i, \quad Y = \sum_{i=1}^{n} y_i, \quad A = \sum_{i=1}^{n} x_i^2, \quad B = \sum_{i=1}^{n} x_i y_i, \quad C = \sum_{i=1}^{n} y_i^2.$$

Rewriting the last line of the expansion of $E(m, b)$ given above, we obtain:

$$E(m, b) = C + m^2 A + b^2 n - 2mB - 2bY + 2bmX.$$

We compute the necessary partial derivatives (the first two are *first partials* and the last three are *second partials*).

$$\begin{aligned}
\frac{\partial E}{\partial m} &= 2mA - 2B + 2bX \\
\frac{\partial E}{\partial b} &= 2bn - 2Y + 2mX \\
\frac{\partial^2 E}{\partial m^2} &= 2A \\
\frac{\partial^2 E}{\partial b^2} &= 2n \\
\frac{\partial^2 E}{\partial m \partial b} &= 2X.
\end{aligned}$$

Derivation (continued)

We set the first partials to 0 and solve (this will involve solving two equations with two unknowns). Solving the second partial derivative equation $2bn - 2Y + 2mX = 0$ yields

(3). $$b = \frac{Y - mX}{n}$$

Solving the first partial derivative equation $2mA - 2B + 2bX = 0$ yields the following chain of equations:

$$\begin{aligned}
mA - B + bX &= 0 \\
mA - B + \left(\frac{Y - mX}{n}\right) X &= 0 \\
mA - B + \frac{XY - mX^2}{n} &= 0 \\
mnA - nB + XY - mX^2 &= 0 \\
m(nA - X^2) &= nB - XY \\
m &= \frac{nB - XY}{nA - X^2}.
\end{aligned}$$

Derivation (continued)

We verify that (m, b) satisfies the second derivative criteria for being a minimum of f. We must show that

$$\begin{aligned}
\frac{\partial^2 E}{\partial m^2}(m,b) \cdot \frac{\partial^2 f}{\partial b^2}(m,b) &> \left(\frac{\partial f}{\partial m \partial b}\right)^2 (m,b) \\
\frac{\partial^2 f}{\partial m^2}(m,b) &> 0.
\end{aligned}$$

Therefore, substituting values, we must verify

$$\begin{aligned}
2A \cdot 2n &> (2X)^2 \\
2A &> 0,
\end{aligned}$$

that is

$$\begin{aligned}
2(\sum_{i=1}^{n} x_i^2) \cdot 2n &> 4\left(\sum_{i=1}^{n} x_i\right)^2 \\
2\sum_{i=1}^{n} x_i^2 &> 0.
\end{aligned}$$

For the second inequality, notice that sum is non-negative since all terms are non-negative. But the sum must actually be positive: Since $n > 1$ and all the x_i are distinct, at least one of the x_i is nonzero, and so, for this i, $x_i^2 > 0$.

Derivation (continued)

For the first inequality, we invoke the Cauchy-Schwartz inequality. (See https://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality.) Recall the general formula tells us that for any reals $a_1, \ldots, a_n$, $b_1, \ldots, b_n$,

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \tag{4}$$

with *strict* inequality if the following condition holds:

(5) some $a_i \neq 0$ and, for every real r, there is i so that $a_i r + b_i \neq 0$

In the present context, we let $a_i = 1$ and $b_i = x_i$. Using these substitutions (and reading (4) from right to left) yields

$$\sum_{i=1}^{n} 1^2 \sum_{i=1}^{n} x_i^2 \geq \left(\sum_{i=1}^{n} x_i\right)^2 .$$

To obtain the required strict inequality, we apply (5): The requirement $a_i \neq 0$ follows since $a_i = 1$. Given a real r, we must find i so that $r + x_i \neq 0$. But such an i must exist because $n > 1$ and the x_i are distinct (if for all i we had $r + x_i = 0$, then for all i, $x_i = -r$, which implies the x_i are not distinct).

Least Square Formula Using Correlation Coefficient

Using techniques from calculus, we showed that the intercept and slope of the least squares best fit line $y = \beta_0 + \beta_1 x$ can be given by

$$\beta_0 = \frac{Y - \beta_1 X}{N} \qquad \beta_1 = \frac{NB - XY}{NA - X^2},$$

where

$$\begin{aligned} X &= \sum x_i \qquad & B = \sum x_i y_i \\ Y &= \sum y_i \qquad & C = \sum y_i^2 \\ A &= \sum x_i^2 & \end{aligned}$$

That is,

$$\beta_0 = \frac{1}{N} \sum y_i - \beta_1 \sum x_i$$

and

$$\beta_1 = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

We wish to show that this formula is equivalent to

$$\beta_0 = \bar{y} - \beta_1 \bar{x}, \qquad \beta_1 = \rho \frac{s_y}{s_x}.$$

Assuming β_1 values are the same, we have immediately:

$$\bar{y} - \beta_1 \bar{x} = \frac{1}{N} \sum y_i - \beta_1 \left(\frac{1}{N} \sum x_i\right)$$

Equivalent form for variance

Claim. $n \sum (x_i - \bar{x})^2 = n \sum x_i^2 - (\sum x_i)^2$

Proof.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$= \frac{\sum_{i=1}^{N} \left(x_i^2 - 2\mu x_i + \mu^2\right)}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 \sum_{i=1}^{N} 1}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu \frac{\sum_{i=1}^{N} x_i}{N} + \frac{\mu^2 N}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu^2 + \mu^2$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{N^2}$$

Multiplying both sides by n^2 gives the result.

The β_1 Values Are Equal

We use the biased population correlation coefficient and biased std deviations. (The result does not hold when the unbiased versions are used.) We must show:

$$r \cdot \frac{s_y}{s_x} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Proof using the biased population correlation coefficient and biased std deviations:

$$\begin{aligned}
r \cdot \frac{s_y}{s_x} &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2}\sqrt{n \sum y_i^2 - (\sum y_i)^2}} \cdot \sqrt{\frac{1/n \cdot \sum (y_i - \bar{y})^2}{1/n \cdot \sum (x_i - \bar{y})^2}} \\
&= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2}\sqrt{n \sum y_i^2 - (\sum y_i)^2}} \cdot \sqrt{\frac{n \cdot \sum (y_i - \bar{y})^2}{n \cdot \sum (x_i - \bar{y})^2}} \\
&= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}
\end{aligned}$$

$\Leftarrow$ uses the Claim on the previous slide