W02D

Benjamin Huang

2024-09-05

The coefficients in a basic linear regression can be calculated as

$$b_1 = r\frac{s_y}{s_x},$$
$$b_0 = \bar{y} - b_1\bar{x}$$

Below is an example data set and a calculation of the parameters for B~A

```
set.seed(441)
A <- 1:10
e <-  rnorm(10, 0, 1)
B <- 2*A - 1 + e
(b_1 <- cor(A, B)*sd(B)/ sd(A))
```

```
## [1] 1.660247
```

```
(b_0 <- mean(B)-b_1*mean(A))
```

```
## [1] 0.5129774
```

Q1. What are the population parameters β_0 and β_0 ? Why don't they equal the sample parameters b_0 and b_1 ?

Answer:

$$Y_1 = \beta_0 + \beta_1 X_1$$
$$Y = b_0 + b_1 X + \epsilon$$
$$\epsilon \backsim N(0, 1)$$

Below, a regression is run to regress B on A. Then a summary is produced.

```
lm <- lm(B~A)
summary(lm)
```

```
##
## Call:
## lm(formula = B ~ A)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3415 -0.7426 -0.2104  0.2461  2.0559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5130     0.8182   0.627    0.548
## A             1.6602     0.1319  12.590 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.198 on 8 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.946
## F-statistic: 158.5 on 1 and 8 DF,  p-value: 1.485e-06
```

Q2. What does "lm" stand for?

Answer: linear model.

The residual standard error is $s = 1.198$, our residual standard deviation. Note there are $n - 2 = 8$ degrees of freedom for s . The Multiple R-squared is $R^2 = 0.952$, the coefficient of determination.

Q3. What is an estimate for the variance of disturbances? What proportion of the variance of B is explained by A?

Answer:

$$
\begin{aligned}
&Estimate\, of\, disturbances : e_i = Y_i - \hat{Y}_i \\
&S_e = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} \\
&\sigma^2 \approx s^2 = 1.43. \\
&R^2 = 0.952
\end{aligned}
$$

Below is an example of an Analysis of Variance (ANOVA) table. Note that the mean square error is 1.43.

```r
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: B
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## A          1 227.405 227.405  158.52 1.485e-06 ***
## Residuals  8  11.476   1.435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q4. Compute the total sum of squares. Relate this to R^2 the variance of B.

```r
(TSS <- 227.405 + 11.476)
```

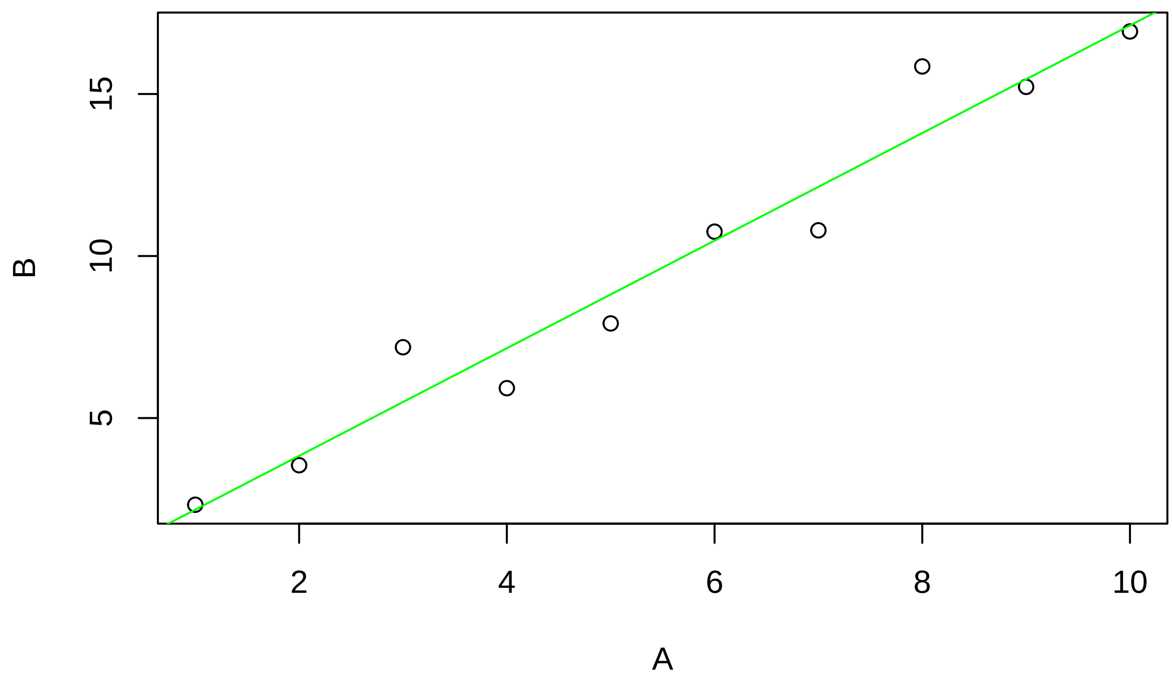
```
## [1] 238.881
```

```r
(r_square <- 227.405 / TSS)
```

```
## [1] 0.9519593
```

Below is a plot of the points along with the best fit line.

```r
plot(A, B) # plots the points
abline(a=coef(lm)[1], b=coef(lm)[2], col = "green")
```




```r
#abline(lm, col = "blue")
```

Doubling the Explanatory Variable

Q5. Predict which of the following values will change when we double the values of A: $S_A, s_B, r_{AB}, b_0, b_1, s, R^2$

Answer:

```r
#Answer:
A2  <- 2*A
lm2 <- lm(B~A2)
summary(lm2)
```

```
##
## Call:
```

```
## lm(formula = B ~ A2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3415 -0.7426 -0.2104  0.2461  2.0559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.51298    0.81820   0.627    0.548
## A2           0.83012    0.06593  12.590 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.198 on 8 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.946
## F-statistic: 158.5 on 1 and 8 DF,  p-value: 1.485e-06
```

```r
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: B
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## A2         1 227.405 227.405  158.52 1.485e-06 ***
## Residuals  8  11.476   1.435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mathematical Explanations: If A is the only thing that changed, then s_A will change by the same factor. The correlation will stay the same because the sum is doubled and s_A double, canceling any change.

$r = \frac{1}{(n-1)s_A s_B} \sum (A_i - \bar{A})(B_i - \bar{B}) = \frac{1}{(n-1)(2 * s_A) s_B} \sum (2 * A_i - s * \bar{A})(B_i - \bar{B})$

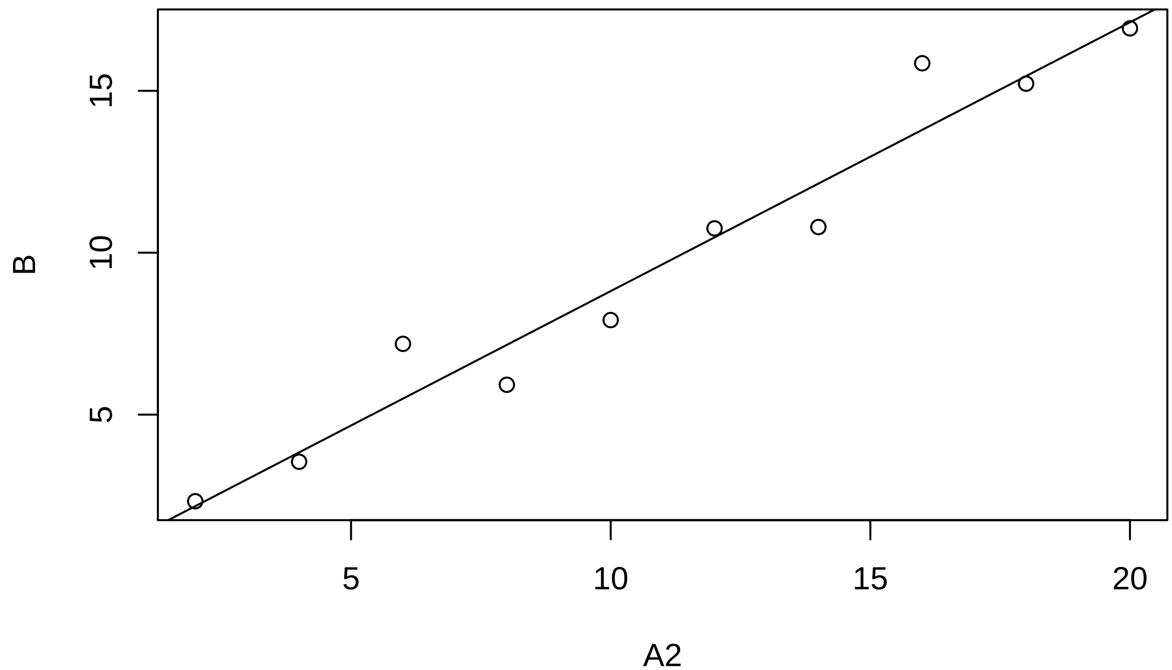
We have $b_1 = r\frac{s_B}{s_A} \rightarrow r\frac{s_B}{2 * s_A}$, so b_1 is halved.

b_0 stays the same because the *2 in $\bar{x}$ cancels with the 1/2 in b_1.

$b_0 = \bar{y} - b_1 * \bar{x} = \bar{y} - \frac{1}{2} b_1 * (2 * \bar{x})$

Below is the plot of (A2, B) and the best fit line.

```r
plot(A2, B) # plots the points
abline(a=coef(lm2)[1], b=coef(lm2)[2])
```



Doubling the Variable of Interest

Q6. Predict which of the following values will change when we double the values of B: $S_A, s_B, r_{AB}, b_0, b_1, s, R^2$

Answer:

```r
#Answer:
B2 <- 2*B
lm3 <- lm(B2~A)
summary(lm3)
```

```
##
## Call:
## lm(formula = B2 ~ A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6830 -1.4851 -0.4208  0.4922  4.1118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0260     1.6364   0.627    0.548
## A             3.3205     0.2637  12.590 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.395 on 8 degrees of freedom
```

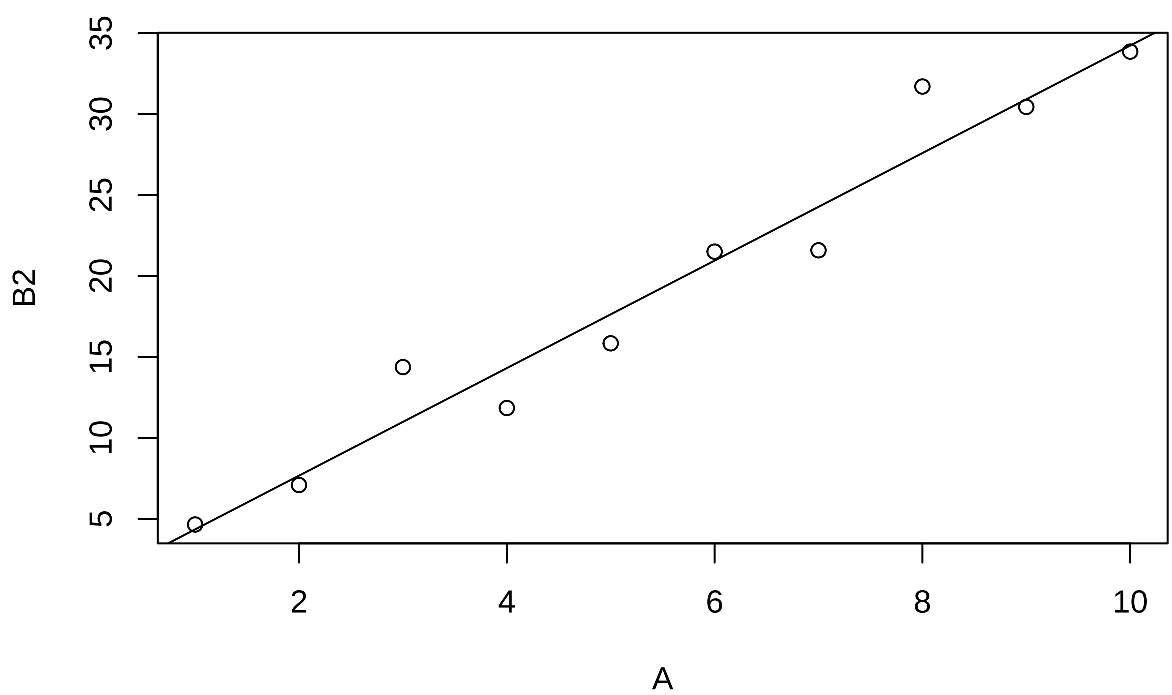
```
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.946
## F-statistic: 158.5 on 1 and 8 DF,  p-value: 1.485e-06
```

```r
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: B2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## A          1 909.62  909.62  158.52 1.485e-06 ***
## Residuals  8  45.91    5.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below is a plot of (A, B2) and the best fit line.

```r
plot(A, B2) # plots the points
abline(a=coef(lm3)[1], b=coef(lm3)[2])
```



Double Both

Q6. Predict which of the following values will change when we double the values of A and B

Answer:

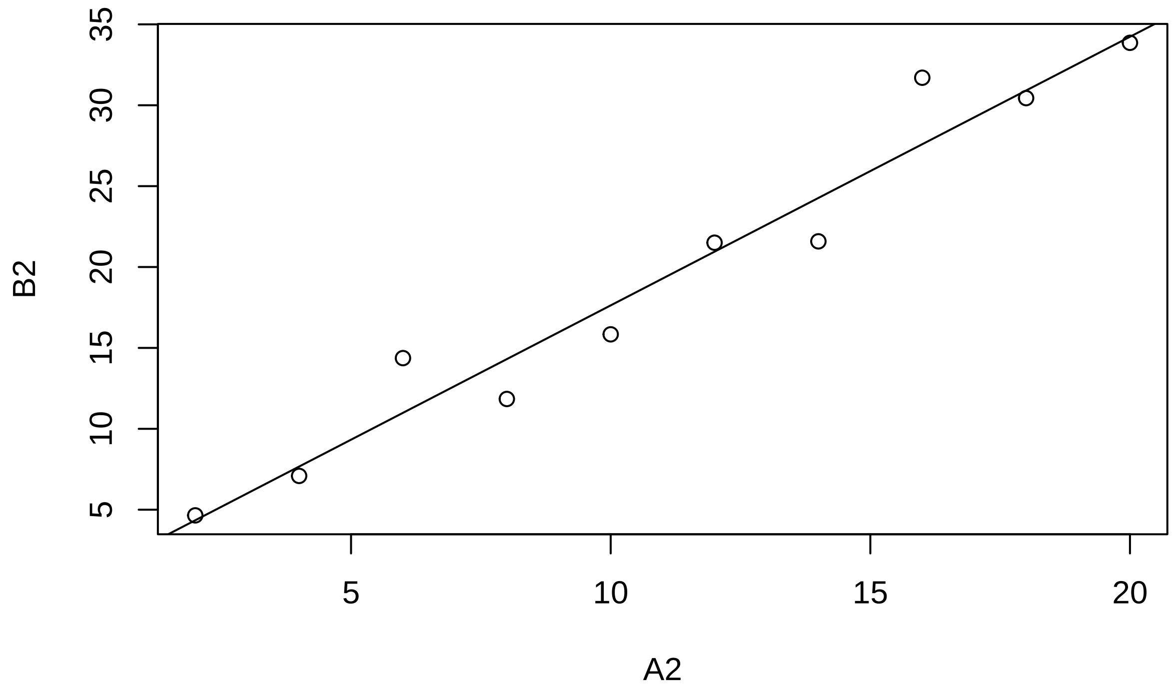
```r
#Answer:
lm4 <- lm(B2 ~ A2)
summary(lm4)
```

```
##
## Call:
## lm(formula = B2 ~ A2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6830 -1.4851 -0.4208  0.4922  4.1118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0260     1.6364   0.627    0.548
## A2            1.6602     0.1319  12.590 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.395 on 8 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.946
## F-statistic: 158.5 on 1 and 8 DF,  p-value: 1.485e-06
```

```r
anova(lm4)
```

```
## Analysis of Variance Table
##
## Response: B2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## A2         1 909.62  909.62  158.52 1.485e-06 ***
## Residuals  8  45.91    5.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(A2, B2) # plots the points
abline(a=coef(lm4)[1], b=coef(lm4)[2])
```


B2
A2
5
10
15
20
25
30
35
5
10
15
20