Assignment15

John Bute

2024-10-15

R Markdown

```
set.seed (1)
x= rnorm(100)
y=2*x+rnorm(100)
```

Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate β, the standard error of this coefficient estimate, and the t−statistic and p−value associated with the null hypothesis H0 : β = 0. Comment on these results. (You can perform regression without an intercept using the command lm(x ∼ y + 0))

```
model <- lm(y ~ x + 0)

summary(model)

##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The coefficient estimate B in Y = Bx (the slope passes through the origin since the intercept = 0) is 1.9939. Furthermore, the standard error is 0.1065. Since the standard error is extremly small, we can assume that the model is very confident in the accuracy of the coefficient estimate. The T-value is 18.79, which details that it is far away from the null hypothesis that our coefficient is 0, as it is 18.73 standard errors away from 0. The p-value is < 2.2e-16, which means that it is extremly unlikely that there is a strong relationship between x and y by random chance. Therefore, we can safely reject the null hypothesis H0: B = 0.

Finally we can fully confirm the rejection of the null hypothesis by doing the following calculations:

$$\hat{\beta} \pm t_{\text{critical}} \times SE(\hat{\beta})$$

Where

$$t_{\text{critical}}$$

is the critical value from the t-distribution table for 99 degrees of freedom, which is approximately 1.984 for a confidence level of 95%.

Thus, the confidence interval is approximately: $$

1.9939 = 1.9939 This gives a confidence interval of:

$$(1.7826,2.2052)$$

2. Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis H0 : β = 0. Comment on these results

```
model <- lm(x ~ y + 0)

summary(model)

##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.8699 -0.2368  0.1030  0.2858  0.8938 
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)    
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776 
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The coefficient estimate is 0.39111, while the standard error is 0.02089, meaning that the estimate of the model is precise. The t-value is 18.73, which suggests that the coefficient is significantly different from our null hypothesis of 0, suggesting there is a significant relationship between y and x. Finally the p-value is less than 2e-16, which is basically 0. This means that we can be certain that the predictor y is significantly associated with the response x and therefore the null hypothesis of B = 0 must be rejected.

What is the relationship between the results obtained in 1. and 2.?

I would say that the relationship of y onto x is 1.9939, which is roughly 2 times. Meanwhile the regression of x onto y is 0.391, which is approximately 1/2.5. Therefore, both results share a reciprocal nature regarding their coefficients, as it is expected since both are linear models of y onto x and vice-versa. Swapping the dependent and independent variables will result in a relationship between coefficients that is reciprocal.

Furthermore, the R-squared is of same value, since the variability of a dependent variable based on an independent one will stay the same, regardless if they switch roles. Both regressions explain the same % of variability in the data.

The t-statistic and p-values are identical too, as the significance of the predictor is high when explaining the response variables.

4. The coefficient's formula, when there is a 0 intercept, is the following

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}$$

Meanwhile, the formula for the standard error $SE(\hat{\beta})$ is:

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}x_i\right)^2}{(n-1)\sum_{i=1}^{n} x_i^2}}$$

The t-statistic for testing the null hypothesis $H_0\!: \beta = 0$ is:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Expanding further, we get:

$$t = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2} \times \sqrt{\frac{(n-1)\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}\left(y_i - \hat{\beta}x_i\right)^2}}$$

$$t = \sqrt{\frac{(\sum_{i=1}^{n} x_i\, y_i)^2 (n-1)\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}\left(y_i - \hat{\beta}x_i\right)^2\,(\sum_{i=1}^{n} x_i^2)^2}}$$

$$t = \sqrt{\frac{(\sum_{i=1}^{n} x_i\, y_i)^2 (n-1)}{\sum_{i=1}^{n}\left(y_i^2 - 2\hat{\beta}x_i y_i + \hat{\beta}^2 x_i^2\right)\sum_{i=1}^{n} x_i^2}}$$

$$t = \sqrt{\frac{(\sum_{i=1}^{n} x_i\, y_i)^2 (n-1)}{\left(\sum_{i=1}^{n} y_i^2 - 2\hat{\beta}\sum_{i=1}^{n} x_i\, y_i + \hat{\beta}^2\sum_{i=1}^{n} x_i^2\right)\sum_{i=1}^{n} x_i^2}}$$

$$t = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i\, y_i\right)^2 (n-1)}{\left(\sum_{i=1}^{n} y_i^2 - 2\left(\frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}\right)\sum_{i=1}^{n} x_i\, y_i + \left(\frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \sum_{i=1}^{n} x_i^2\right)\sum_{i=1}^{n} x_i^2}}$$

$$t = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i\, y_i\right)^2 (n-1)}{\left(\sum_{i=1}^{n} y_i^2 - 2\left(\frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}\right)\sum_{i=1}^{n} x_i\, y_i + \frac{\left(\sum_{i=1}^{n} x_i\, y_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right)\sum_{i=1}^{n} x_i^2}}$$

$$t = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i\, y_i\right)^2 (n-1)}{\left(\sum_{i=1}^{n} y_i^2 - \left(\frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}\right)\sum_{i=1}^{n} x_i\, y_i\right)\sum_{i=1}^{n} x_i^2}}$$

$$t = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i\, y_i\right)^2 (n-1)}{\sum_{i=1}^{n} x_i^2\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)\frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}\sum_{i=1}^{n} x_i\, y_i\right)}}$$

$$t = \frac{\sum_{i=1}^{n} x_i\, y_i \sqrt{n-1}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_i\, y_i\right)^2\right)}}$$

We proved it algebraically, now will will confirm numerically in R

```
set.seed(1)

x <- rnorm(100)
y <- 2 * x + rnorm(100)
model <- lm(y ~ x + 0)

beta <- coef(model)[1]
se_beta <- summary(model)$coefficients[2]

t_stat_manual <- beta/se_beta

n <- length(x)

sum_xy <- sum(x*y)
sum_x2 <- sum(x^2)
sum_y2 <- sum(y^2)

numerator <- sqrt(n-1) * sum_xy
denominator <- sqrt(sum_x2 * sum_y2 - sum_xy^2)

t_stat_derived <- numerator/denominator
```

```
t_stat_manual

##        x
## 18.72593

t_stat_derived

## [1] 18.72593
```

5. Using the results from 4., argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

y onto [

t =]

x onto y

$$t = \frac{(\sum_{i=1}^{n} y_i\, x_i)\sqrt{(n-1)}}{\sqrt{(\sum_{i=1}^{n} y_i^2)(\sum_{i=1}^{n}(x_i^2) - (\sum_{i=1}^{n} y_i\, x_i)^2}}$$

As you can see, both formulas are symmetric, therefore the t-statistic for y onto x and x onto y is the same.

6. In R, show that when regression is performed with an intercept, the t-statistic for H0 : β1 = 0 is the same for the regression of y onto x as it is for the regression of x onto y

```
set.seed(1)

x <- rnorm(100)
y <- 2 * x + rnorm(100)
model_y_x <- lm(y ~ x)
model_x_y <- lm(x ~ y)


beta_y_x <- coef(model_y_x)["x"]
se_beta_y_x <- summary(model_y_x)$coefficients["x", "Std. Error"]

t_stat_y_x <- beta_y_x/se_beta_y_x

beta_x_y <- coef(model_x_y)["y"]

se_beta_x_y <- summary(model_x_y)$coefficients["y", "Std. Error"]

t_stat_x_y <- beta_x_y/se_beta_x_y

t_stat_x_y
```

```
##       y
## 18.5556
```

```
t_stat_y_x
```

```
##       x
## 18.5556
```