STAT 210
Applied Statistics and Data Analysis
Linear Regression IV:
Coefficient of Determination

Joaquin Ortega

Results from previous lectures

Results from previous lectures

Recall that the formula for the estimator of β_1 is

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)\end{aligned}$$

Coefficient of Determination

Coefficient of Determination

The analysis of variance is based on the following decomposition for the sum of squares:

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

which is usually expressed as

$$SST = SSE + SSR.$$

Since the sums are non-negative we have that $SSE \leq SST$.

Observe that they are equal only if there is no relation between the two variables: $SSR = 0$ means that $\hat{y}_i = \bar{y}$ for all i and for this to be true, the regression line must be horizontal, so $\beta_1 = 0$ and $y = \beta_0$.

Coefficient of Determination

The regression sum of squares SSR is usually interpreted as the amount of variability in Y that is explained by the regression line.

Since the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the sum of squares due to error, and SST is fixed once the sample is known, there is no better line than the regression line.

The ratio SSE/SST represents the proportion of the variability that cannot be explained by the linear regression model.

The **coefficient of determination** R^2 is defined as

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

and represents the proportion of the variation that is explained by the regression model.

Example 1

Recall the summary for the regression in the first example:

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = FL ~ CL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86395 -0.51746 -0.02826  0.50456  1.77009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15316    0.23477   0.652    0.515
## CL           0.48060    0.00714  67.313   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579
## F-statistic:  4531 on 1 and 198 DF,  p-value: < 2.2e-16
```

Coefficient of Determination

Although the model looks very good, we also fitted separate models for each species, which are `lm2` and `lm3`.

```
summary(lm2)
```

```
##
## Call:
## lm(formula = FL[sp == "B"] ~ CL[sp == "B"], data = crabs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95680 -0.17686 -0.01135  0.22143  0.82572
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.971315   0.134562   7.218 1.13e-10 ***
## CL[sp == "B"] 0.435315   0.004364  99.745  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2997 on 98 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:  0.9901
## F-statistic:  9949 on 1 and 98 DF,  p-value: < 2.2e-16
```

Coefficient of Determination

```r
summary(lm3)
```

```
##
## Call:
## lm(formula = FL[sp == "O"] ~ CL[sp == "O"], data = crabs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1344 -0.3357 -0.0249  0.2734  1.2282
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.762041   0.257726   2.957   0.0039 **
## CL[sp == "O"] 0.478668   0.007404  64.651   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4983 on 98 degrees of freedom
## Multiple R-squared:  0.9771, Adjusted R-squared:  0.9769
## F-statistic:  4180 on 1 and 98 DF,  p-value: < 2.2e-16
```

We see that the separate models are even better, accounting for 97.7 and 99% of the variability in the responses.

Coefficient of Determination

Let's look at the other two examples we have considered.

```r
summary(lm4)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

Coefficient of Determination

```r
summary(lm5)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

The first of these models has a low R^2 of 35.8% while the second has a much better value of 93.5%.

Relation with the Correlation Coefficient

We have the following proposition:

Proposition 1 Let ρ be the correlation coefficient for the sample $(x_i, y_i), i = 1, \ldots, n$. Then

$$R^2 = \rho^2.$$

Proof. The regression model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and by property 4 we know that the regression line goes through $(\bar{x}, \bar{y})$ so that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. Subtracting this equation from the first one we get

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}).$$

Squaring both sides and adding up

$$\sum_i (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2$$

Relation with the Correlation Coefficient

Recall from (3) that

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

and using (1)

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

$$= \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = \rho^2.$$

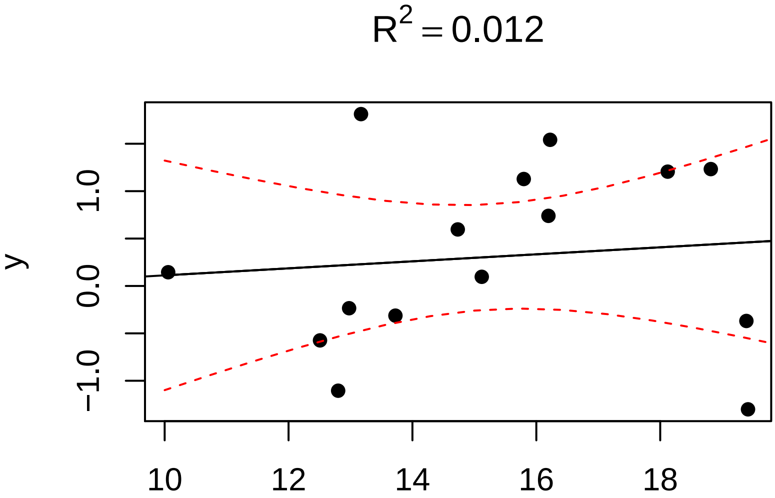
$\blacksquare$

It is important to observe that this relation is only true for simple regression. It does not hold in the multivariate case.

Coefficient of Determination

As an example let us look at some simulated data. First we look at purely random values.

```
## [1] 0.012
```



Coefficient of Determination

```r
set.seed(98765)
xx <- runif(15,10,20)
zz <- rnorm(15)
(r.sq <-round(summary(lm(zz~xx))$r.squared,3))
plot(xx,zz,pch=16, xlab='x', ylab='y')
abline(lm(zz~xx))
title(main= bquote(R^2 == .(r.sq)))
xx.new <- data.frame(xx=seq(10,20, length.out = 15))
pc <- predict(lm(zz~xx),xx.new, int='c')
matlines(xx.new$xx, pc, lty=c(1,2,2),
         col=c('black','red','red'))
pp <- predict(lm(zz~xx),xx.new, int='p')
matlines(xx.new$xx, pp, lty=c(1,2,2),
         col=c('black','red','red'))
```

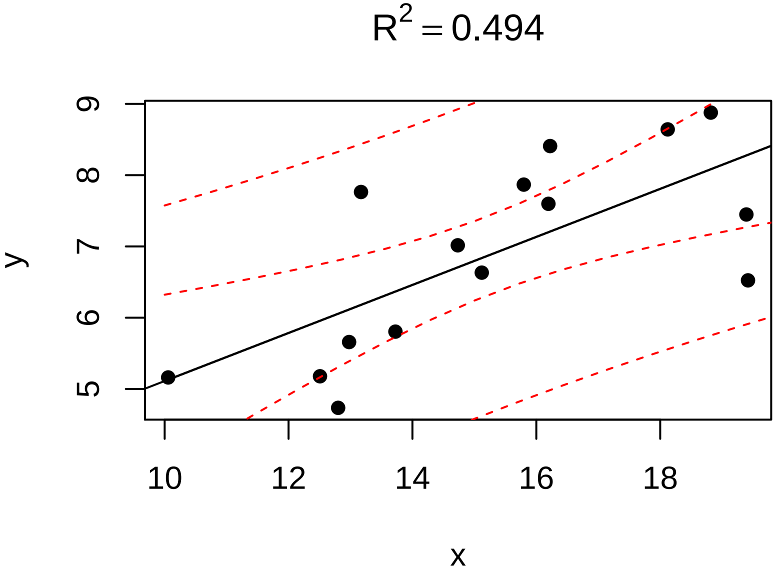
Coefficient of Determination

In this case there is no relation between x and y, the dependent variable takes purely random values, which are independent of the values of x. This is reflected in a low R^2 value of 0.012.

As a second example, let us use the same normal variables we just generated as noise in a linear relation between x and y.

Coefficient of Determination

```
## [1] 0.494
```



Coefficient of Determination

```
yy1 <- 2 + 0.3*xx + zz
plot(xx,yy1,pch=16, xlab='x', ylab='y')
abline(lm(yy1~xx))
(r.sq <-round(summary(lm(yy1~xx))$r.squared,3))
title(main= bquote(R^2 == .(r.sq)))
pc <- predict(lm(yy1~xx),xx.new, int='c')
matlines(xx.new$xx, pc, lty=c(1,2,2),
         col=c('black','red','red'))
pp <- predict(lm(yy1~xx),xx.new, int='p')
matlines(xx.new$xx, pp, lty=c(1,2,2),
         col=c('black','red','red'))
```

Coefficient of Determination

Now there is a linear relation between y and x but the variability due to the variance of the noise makes the explained variability to be only about 50%. As a third and final example, let us reduce noise variability by rescaling it.

```
yy2 <- 2 + 0.3*xx + zz/10
plot(xx,yy2,pch=16, xlab='x', ylab='y')
abline(lm(yy2~xx))
(r.sq <-round(summary(lm(yy2~xx))$r.squared,3))
title(main= bquote(R^2 == .(r.sq)))
pc <- predict(lm(yy2~xx),xx.new, int='c')
matlines(xx.new$xx, pc, lty=c(1,2,2),
         col=c('black','red','red'))
pp <- predict(lm(yy2~xx),xx.new, int='p')
matlines(xx.new$xx, pp, lty=c(1,2,2),
         col=c('black','red','red'))
```

Coefficient of Determination

```
## [1] 0.988
```