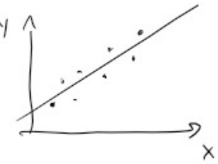
Ch.5-1

Friday, January 31, 2025 10:45 AM

ch5-1

OLS – R-square



Population model:

interest rate

inflation

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i} \quad (4.4)$$

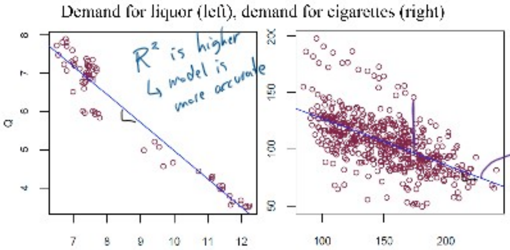
- The assumption is that changes in X lead to changes in Y.
- We are using these changes to choose the line.
- But X isn't the only reason that Y changes.
- There are things in the random error term, too.

i. How well does the estimated model explain the Y variable?
ii. or…How well do changes in X explain changes in Y?
iii. or…How well does the estimated regression line "fit" the data?
iv. or…What portion of the variance in Y can be explained by X?

R-squared is a statistic that provides a measure for all of these (equivalent) questions.

Which regression "fits" better?

Demand for liquor (left), demand for cigarettes (right)



$\bar{e} = 0$

$\sum e_i^2 \rightarrow \min_{b_0, b_1} \sum e_i^2$

$\searrow b_1 = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$

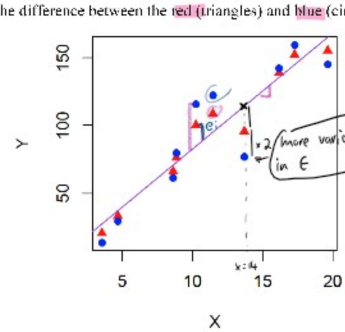
there is no other line that is closer to the data points

$lm(Q \sim P)$

$\hookrightarrow b_0, b_1$

R^2 is higher $\quad R^2$

What is the difference between the red (triangles) and blue (circles) data?



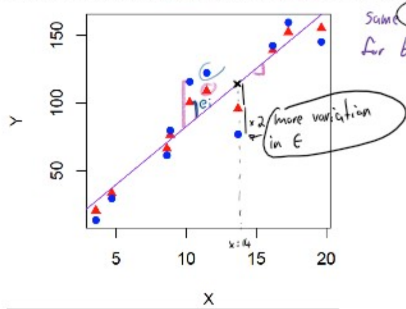
same b_0 & b_1 for both data

$\min \sum e_i^2$

$$\downarrow R^2 = \frac{ESS}{TSS} = 1 - \frac{\sum e_i^2 \uparrow}{TSS}$$

R^2 is higher $\quad R^2$

What is the difference between the red (triangles) and blue (circles) data?



Same b_0 & b_2 for both data

min $\sum e_i^2$

$$\downarrow R^2 = \frac{ESS}{TSS} = 1 - \frac{\sum e_i^2 \uparrow}{TSS}$$

- Both the red and blue data provide the same estimated line
- That is, both red and blue have the same b_1
- But, the line fits the red data better
- Changes in X account for more of the changes in Y, for red
- For the blue data, the *unobserved* factors are accounting for more of the changes (or variation) in Y

Now, we will come up with a statistic (it's just an equation using the data!), that will describe:

The portion of variance in Y that can be explained using variance in X.

Population model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (4.4)$$

lm( )

Estimated model:

$$Y_i = \boxed{b_0 + b_1 X_i} + e_i, \qquad (4.7)$$

$\hat{Y} \qquad e$

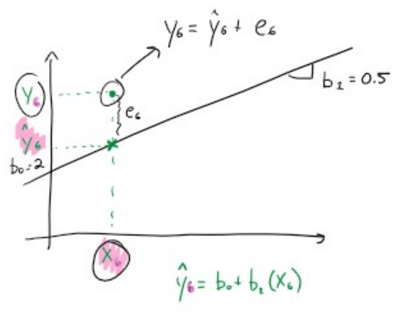
Recall:

$$\hat{Y}_i = b_0 + b_1 X_i. \qquad (4.5)$$

So:

reality = predicts + residual

$$\boxed{Y_i = \hat{Y}_i + e_i}$$



$$Y_i = \hat{Y}_i + e_i$$

Residual - unexplained part

Predicted or "explained" Y value, using the model (X data)

Actual Y data

To get R-squared:

- we'll start by taking the sample variance of both sides.
- This will break the variance in Y up into two parts:
- variance that we can explain ($\hat{Y}$),
- and variance that we can't explain (e).
- After some algebra, we'll write: $\boxed{TSS = ESS + RSS}$

$R^2 = \frac{ESS}{TSS}$

$E[S_Y^2] = \sigma_Y^2$ unbiased

$$\boxed{S_Y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}} \quad (\sigma_Y^2)$$

$$S_Y^2 = S_{\hat{Y}}^2 + S_e^2 + \cancel{2cov(\hat{y}, e)}$$

$$R^2 = \frac{S_{\hat{Y}}^2}{S_Y^2}$$

$y = \hat{y} + e \quad$ take sample variance of both sides $\quad \frac{\sum(y_i - \bar{y})^2}{n-1} = s_Y^2$

$$\boxed{S_Y^2 = S_{\hat{y}}^2 + S_e^2} \text{ (no covariance)}$$

$\hookrightarrow \hat{y}$ & e are independent.

Ch. 2

$Var(X + Y)$

$= var(X) + var(Y) + 2cov\cancel{(X, Y)}$

$\bar{y} \qquad 0$

are independent.

$$\boxed{\frac{\sum (y_i-\bar{y})^2}{\cancel{n-1}} = \frac{\sum (\hat{y}-\bar{\hat{y}})^2}{\cancel{n-1}} + \frac{\sum (e_i-\bar{e})^2}{\cancel{n-1}}}$$

$Var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$

0 if X & Y independent

TSS – total sum of squares

ESS – explained sum of squares

RSS – residual sum of squares

R-squared will then be defined as:

$$\boxed{R^2 = \frac{ESS}{TSS}}$$

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y}$$

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum e_i^2$$

$$\boxed{TSS = ESS + RSS}$$

~~total sum~~ of squares, explained, residual

~~in other texts: SST = SSR + SSE~~

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{y}-\bar{y})^2 / n-1}{\sum (y-\bar{y})^2 / n-1} = \frac{S^2_{\hat{Y}}}{S^2_Y}$$

$$\boxed{R^2 = 1 - \frac{RSS}{TSS}}$$

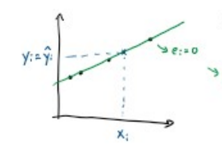
Two extremes will bound R^2 between 0 and 1:

- no fit → $R^2 = 0$
- perfect fit → all residuals are Ø, and all $y_i = \hat{y}_i$

To get R^2 in R, use the `summary()` command:
`summary(lm(y ~ x))`

It provides a lot of information (we'll figure out the rest later).

$R^2 = 1$



$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

$R^2 = 1$ if $ESS = TSS$ and $RSS = 0$

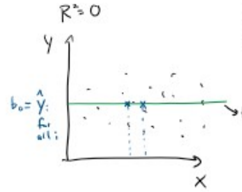
→ $R^2 = 1$

$ESS = TSS$

$\sum (\hat{y}_i - \bar{y})^2 = \sum (y_i - \bar{y})^2$

Same

$R^2 = 0$



$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$ $TSS > 0$

$ESS = 0$ → $\sum (\hat{y}_i - \bar{y})^2$ → $\hat{y}_i = \bar{y}$

$RSS = TSS$


```
summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-37.114 -12.570  -0.226  12.739  31.249

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.284     17.866  -0.184 0.858736
x              8.583      1.431   5.999 0.000324 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.75 on 8 degrees of freedom
Multiple R-squared:  0.8181,	Adjusted R-squared:  0.7954
F-statistic: 35.98 on 1 and 8 DF,  p-value: 0.0003239
```

→ the model explains 82% of the variation in Y