OLS – R-square

Population model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4.4}$$

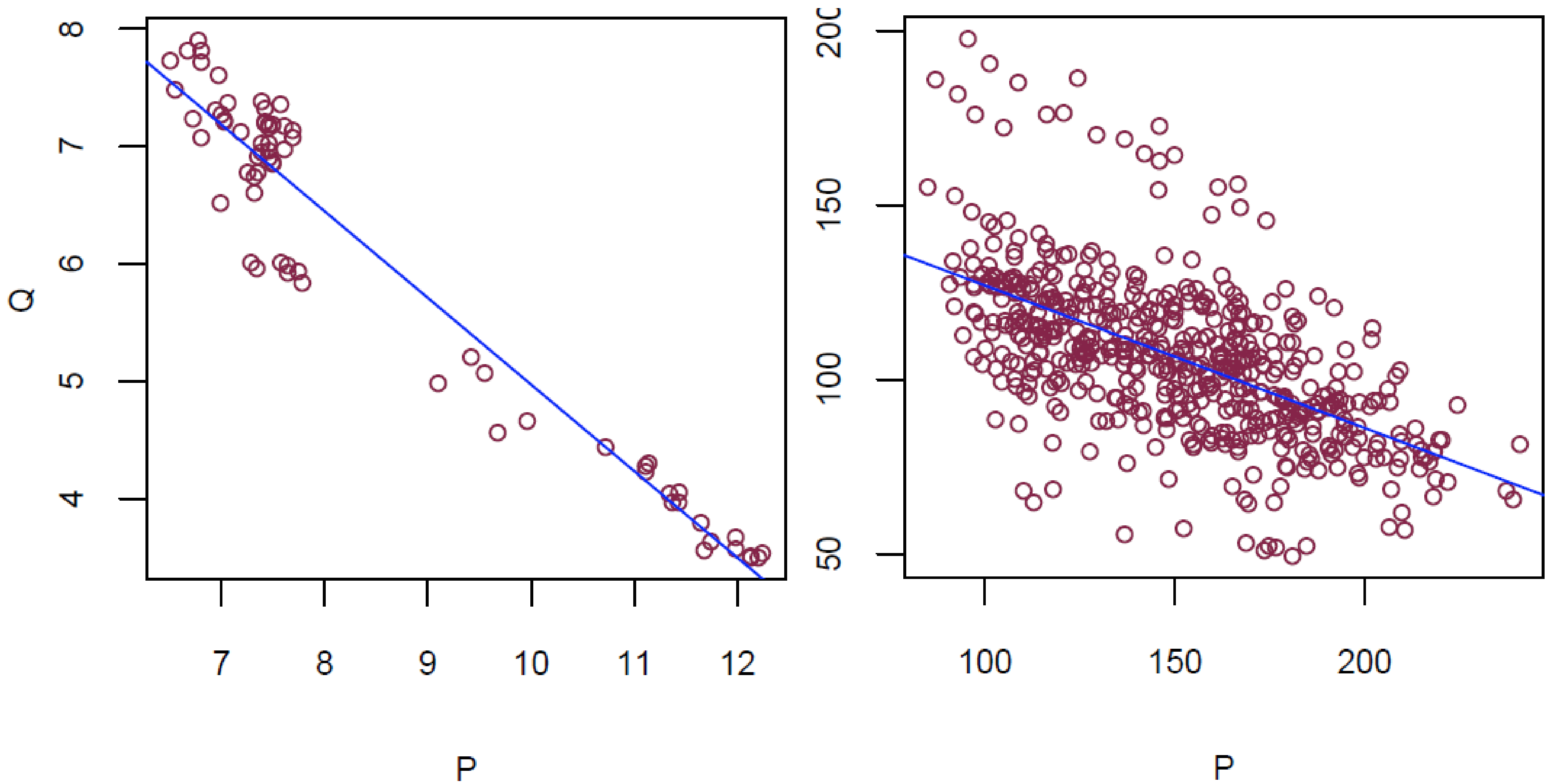
- The assumption is that changes in *X* lead to changes in *Y*.
- We are using these changes to choose the line.
- But *X* isn't the only reason that *Y* changes.
- There are things in the random error term, too.

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?

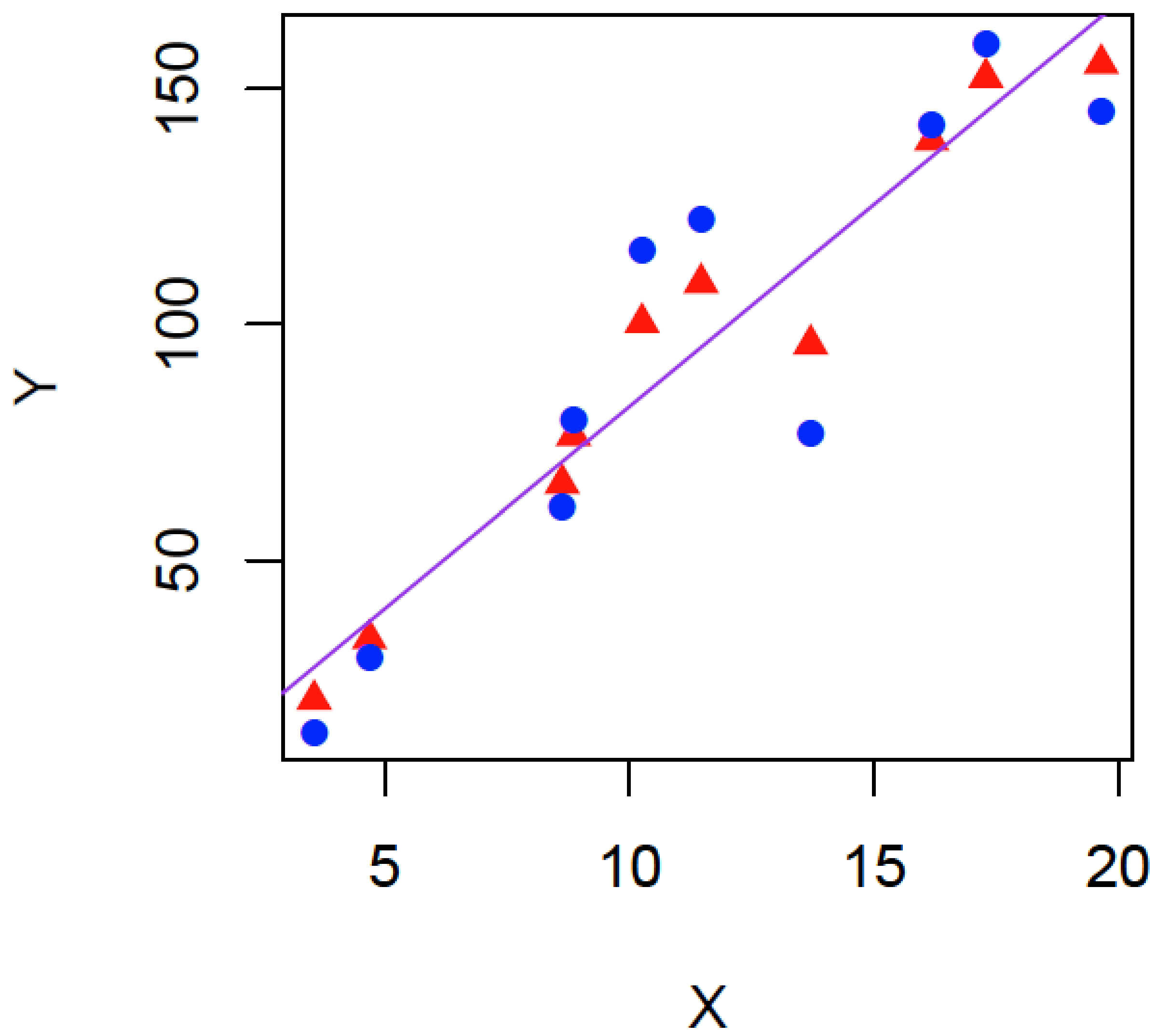
R-squared is a statistic that provides a measure for all of these (equivalent) questions.

Which regression "fits" better?

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



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



- 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:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4.4}$$

Estimated model:

$$Y_i = b_0 + b_1 X_i + e_i, \tag{4.7}$$

Recall:

$$\hat{Y}_i = b_0 + b_1 X_i. \tag{4.5}$$

So:

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

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

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: TSS = ESS + RSS

TSS – total sum of squares

ESS – explained sum of squares

RSS – residual sum of squares

R-squared will then be defined as:

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

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

- no fit
- perfect fit

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).

```
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
```