8.4 – Interaction terms

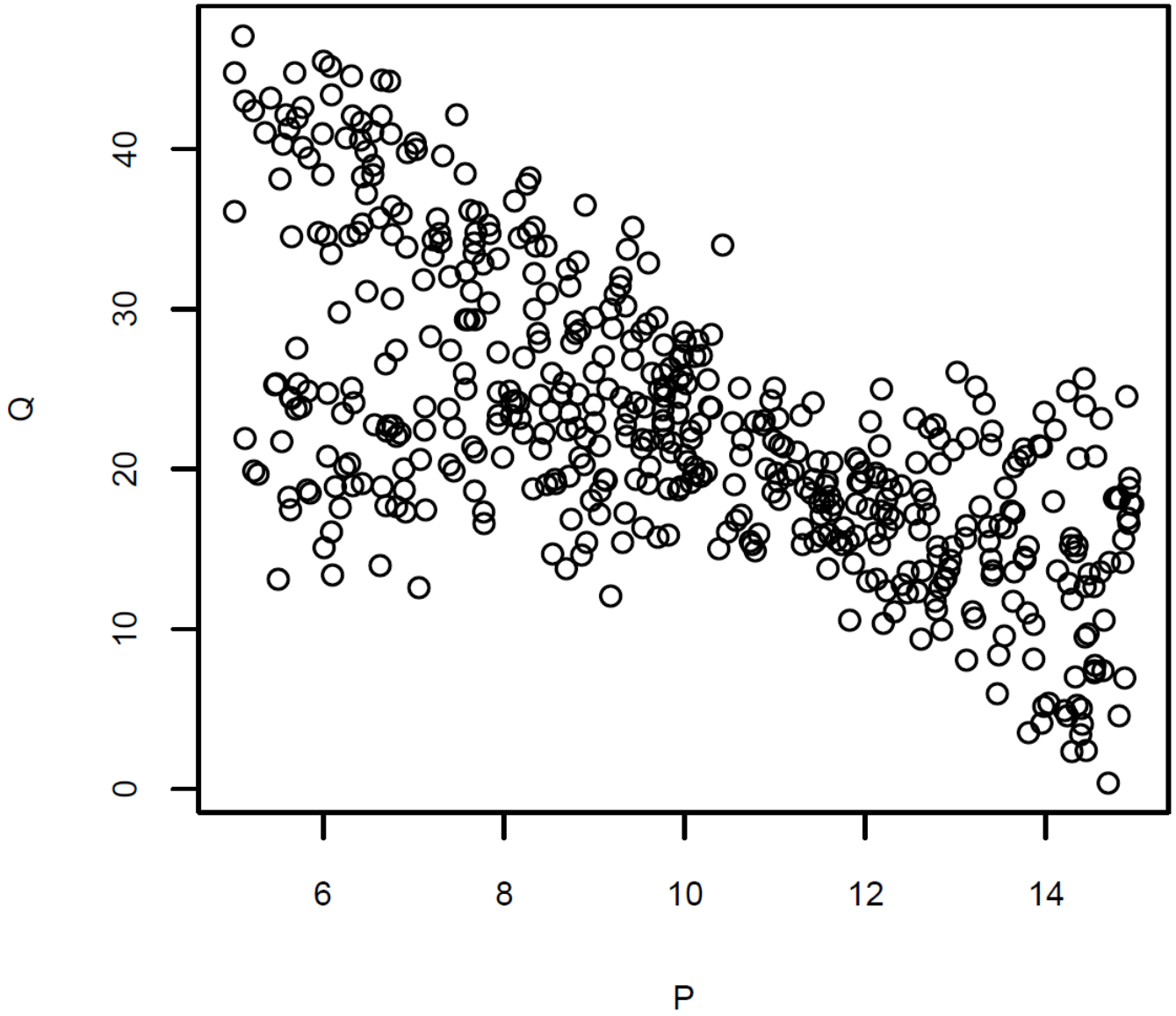
- A type of non-linear effect
- Allows for different effects for different groups (when using a dummy)

A hypothetical data set – demand for marijuana

Suppose that 500 marijuana users are surveyed in different locations, and the variables in the data are:

- Q - the quantity of marijuana consumed, in grams, per month
- P - the average price per gram in the individual's location
- $adult = 1$ if the individual is an adult, $= 0$ if the individual is a teenager

Figure 8.1: Plot of the hypothetical demand for marijuana data.



- Notice anything?
- Ignore the *adult* dummy variable, estimate a regression

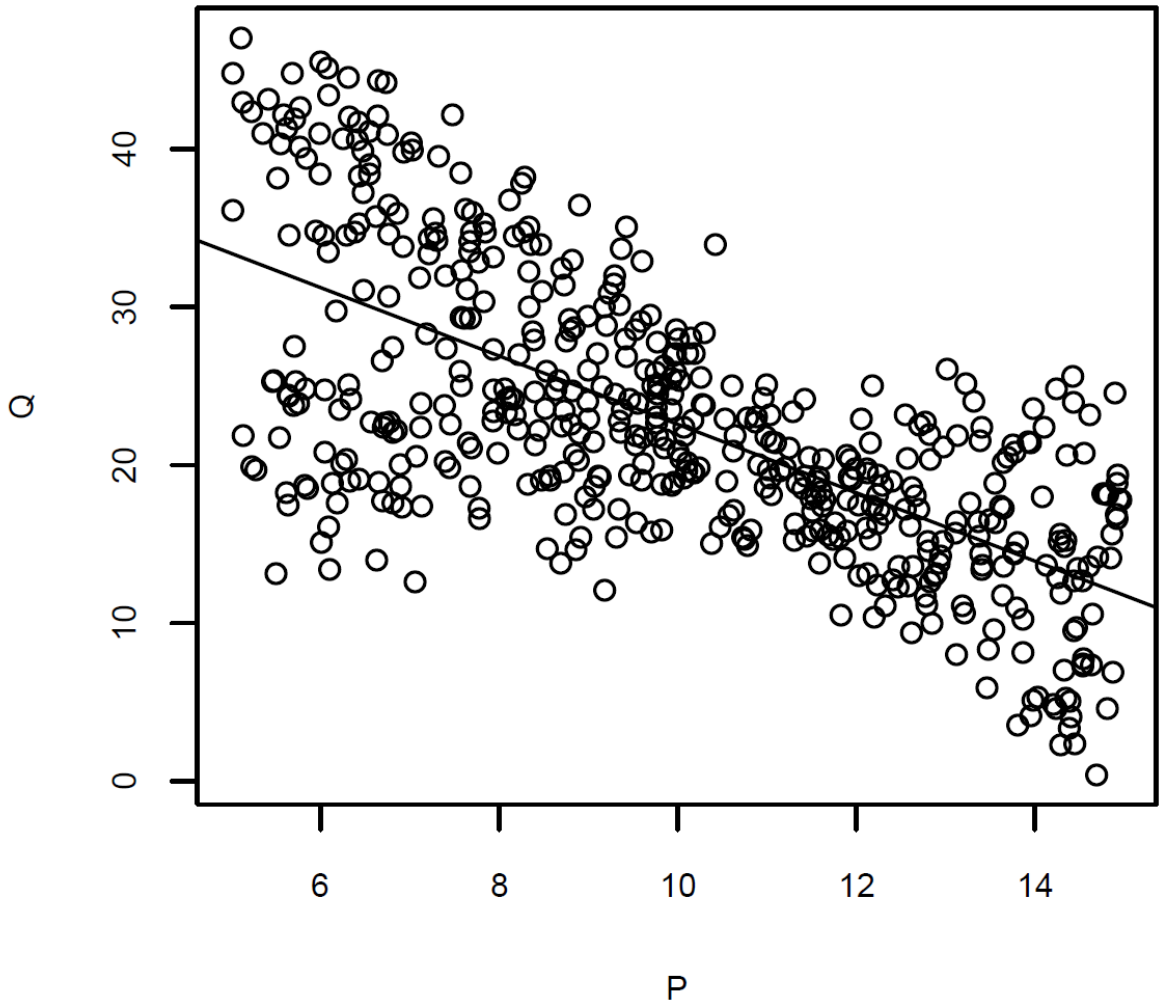
```
summary(lm(Q ~ P))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.2152     1.0776   41.03   <2e-16 ***
P            -2.1634     0.1041  -20.78   <2e-16 ***
```

Increase in price of $1 leads to decrease in consumption of 2.16 grams/month.

Add the line:

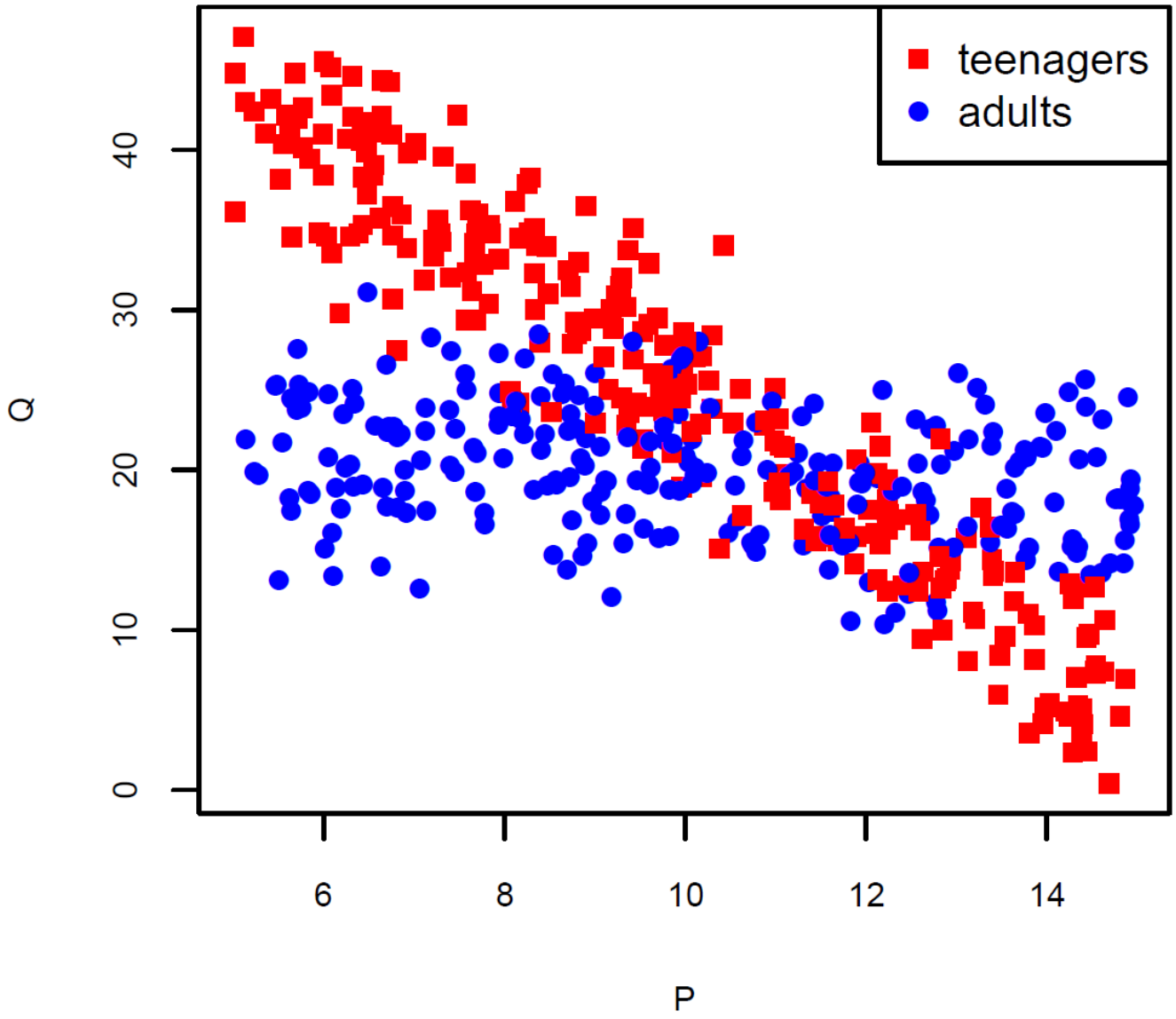
Figure 8.2: Marijuana data, with estimated regression line from $Q = \beta_0 + \beta_1 P + \epsilon$ added to the plot.



- We’re getting an “average” regression line for the two groups
- Ideally, we would like a separate regression slope for each
- Why might the slope (marginal effect) be different between groups

Plot the data by group (teenagers and adults):

Figure 8.3: Marijuana data plotted by age group.



Let’s add the dummy variable to the regression:

```
summary(lm(Q ~ P + adult))
```

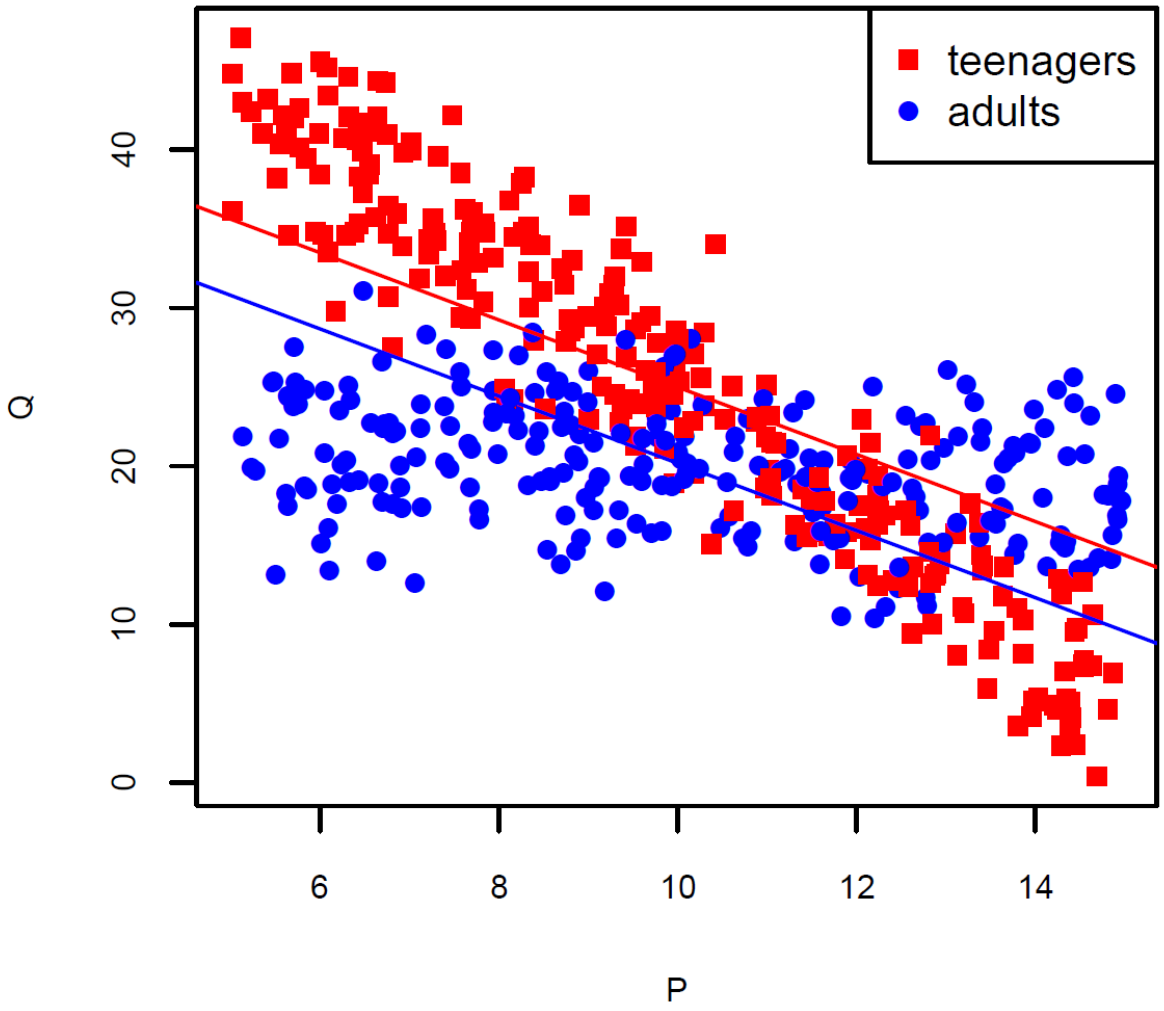
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.21319    1.02971  44.880   <2e-16 ***
P           -2.12242    0.09712 -21.854   <2e-16 ***
adult       -4.81124    0.54975  -8.752   <2e-16 ***
```

Interpretation?

- Adults consume 4.81 g less
- Slope?

Does the dummy variable do the trick? See the regression lines plotted:

Figure 8.4: With the addition of the dummy variable, each group has a different intercept, but the same slope.



Two separate regression lines, but only the intercepts differ (slope the same). In order to get what we want, we need an *interaction term*. In this case, it will be a *dummy-continuous* interaction.

Ideally, we want to allow the effect of *P* on *Q* to be different for adults and teenagers. How to do this?

Estimate the population model:

$$Q = \beta_0 + \beta_1 P + \beta_2 adult + \beta_3 (adult \times P) + \epsilon \qquad (8.2)$$

where $adult \times P$ is the interaction term, and is a new variable that is created by multiplying the other two variables together. To see how model 8.2 allows for two separate lines, consider what the population model is for teenagers ($adult = 0$), and for adults ($adult = 1$).

Population model for teenagers

Let's substitute in the value $adult = 0$ into equation 8.2 and get the population model for teenagers:

$$\begin{aligned} Q &= \beta_0 + \beta_1 P + \beta_2(0) + \beta_3(0 \times P) + \epsilon \\ &= \beta_0 + \beta_1 P + \epsilon \end{aligned} \tag{8.3}$$

From equation 8.3, we can see that the intercept is β_0 and the slope is β_1.

Population model for adults

Substituting in the value $adult = 1$ into equation 8.2, we get the population model for adults:

$$\begin{aligned} Q &= \beta_0 + \beta_1 P + \beta_2(1) + \beta_3(1 \times P) + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)P + \epsilon \end{aligned} \tag{8.4}$$

For adults, the intercept is $\beta_0 + \beta_2$ and the slope is $\beta_1 + \beta_3$. The marginal effect of price on consumption differs by β_3 between the two groups.

Estimation with an interaction term

To include a dummy-continuous interaction term in our regression, we simply create a new variable by multiplying the dummy variable (*adult*) and the continuous variable P together:

```
adult_P <- adult*P
```

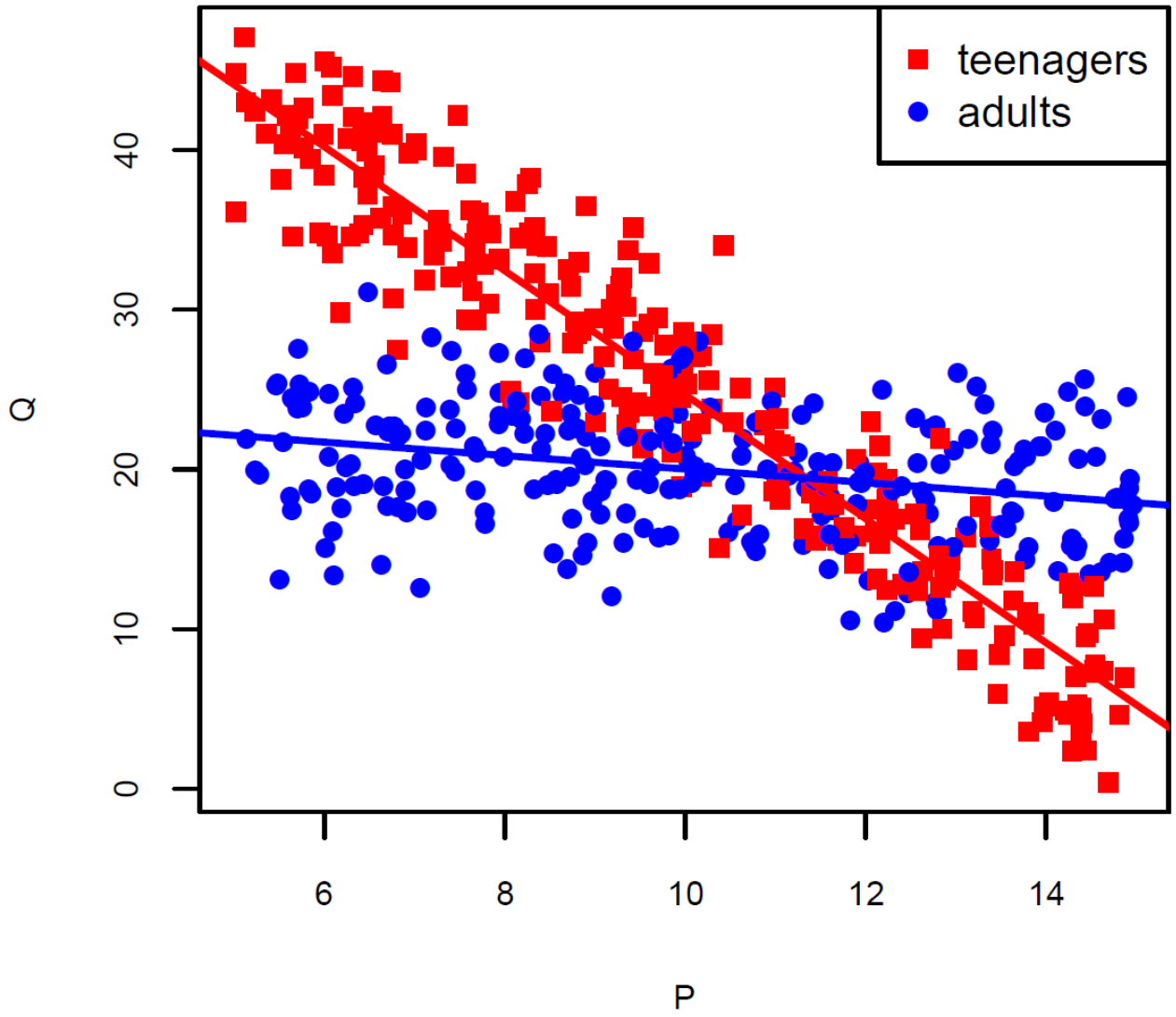
and include the new variable in the regression:

```
summary(lm(Q ~ P + adult + adult_P))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.48944    0.85166   74.55   <2e-16 ***
P            -3.88168    0.08339  -46.55   <2e-16 ***
adult       -39.25222    1.21030  -32.43   <2e-16 ***
adult_P       3.45993    0.11695   29.58   <2e-16 ***
```

The estimated value of 3.46 (on the `adult_P` dummy-continuous interaction term) means that the decrease in consumption due to an increase in price of \$1 is 3.46 grams/month less for adults than it is for teenagers. That is, the effect of price on quantity is -3.88 for teenagers, and (-3.88 + 3.46 = -0.42) for adults. The demand curve is much steeper for teenagers.

Figure 8.5: Two separate regression lines for the two different groups.



8.4.4 Hypothesis tests involving dummy interactions

An important use of dummy interaction terms is to test whether there is a different effect between two groups. In the marijuana example, the interaction term measures the difference in the slope of the demand curve between the two groups. To test the hypothesis that the sensitivity of marijuana consumption to changes in price is the same for teenagers as it is for adults, we could test the hypothesis:

$$H_0 : \beta_3 = 0$$
$$H_A : \beta_3 \neq 0$$

in the model:

$$Q = \beta_0 + \beta_1 P + \beta_2 adult + \beta_3 (adult \times P) + \epsilon$$

Differences-in-differences (DiD) [Not yet in textbook!]

Dummy-dummy interactions can be used for something called "Differences-in-differences" (DiD) estimation.

Example: increasing the minimum wage (image by Stable Diffusion)

- In 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour.
- Card and Krueger (1994) surveyed 410 fast-food restaurants before and after the increase, and asked about things like the number of employees.

Download Card and Krueger data:

```
did <- read.csv("https://rtgodwin.com/data/card.csv")
```

Some variables to look at for now:

EMP – number of full-time employees

TIME – a dummy equal to 0 for before the wage increase, 1 for after the increase

STATE – a dummy equal to 0 for Pennsylvania, equal to 1 for New Jersey

Difference in the number of employees before and after the wage increase:

```
mean(did$EMP[did$STATE == 1 & did$TIME == 1]) -
      mean(did$EMP[did$STATE == 1 & did$TIME == 0])
[1] 0.4666667
```

The difference is not significant:

```
dids <- subset(did, STATE==1)
summary(lm(EMP ~ TIME, data=dids))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.4306     0.5289  38.627   <2e-16 ***
TIME          0.4667     0.7480   0.624    0.533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 616 degrees of freedom
Multiple R-squared:  0.0006315,   Adjusted R-squared:  -0.0009909
F-statistic: 0.3892 on 1 and 616 DF,  p-value: 0.5329
```

So, the causal effect of the increase in minimum wage on employment is estimated to be an increase of 0.47 workers on average, but this increase is not statistically significant.

What is the problem with calling this a “causal effect”?

Next: “The Fundamental Problem of Causal Inference”