Suppose we hypothesize that the variable *x* causes the variable *y*, and we want to estimate the marginal effect of *x* on *y*. So, we estimate the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i \ ,$$

and find:

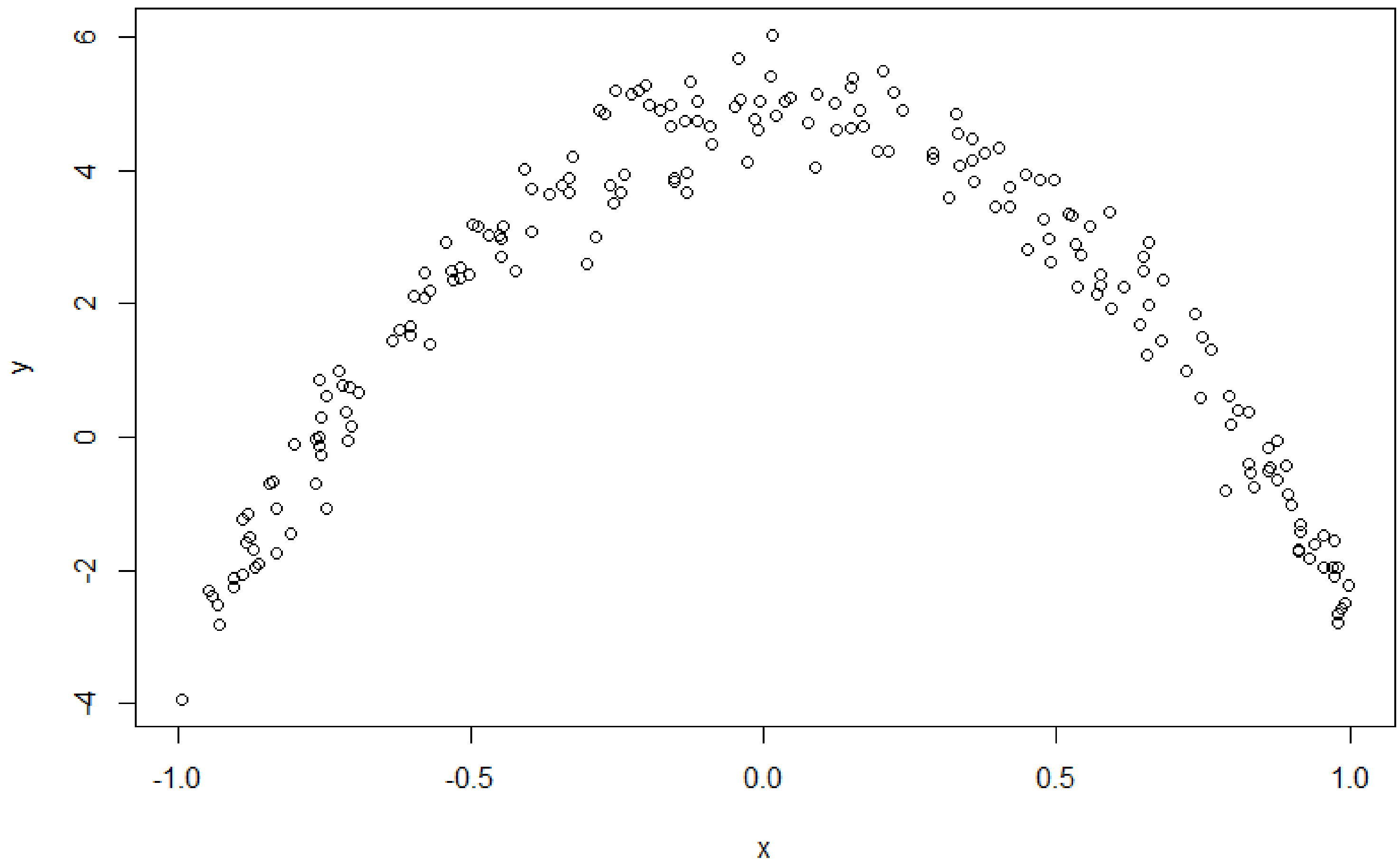
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.98142    0.17845   11.10   <2e-16 ***
x           -0.02331    0.29188   -0.08    0.936
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.521 on 198 degrees of freedom
Multiple R-squared:  3.22e-05,	Adjusted R-squared:  -0.005018
F-statistic: 0.006376 on 1 and 198 DF,  p-value: 0.9364
```

What do you conclude?

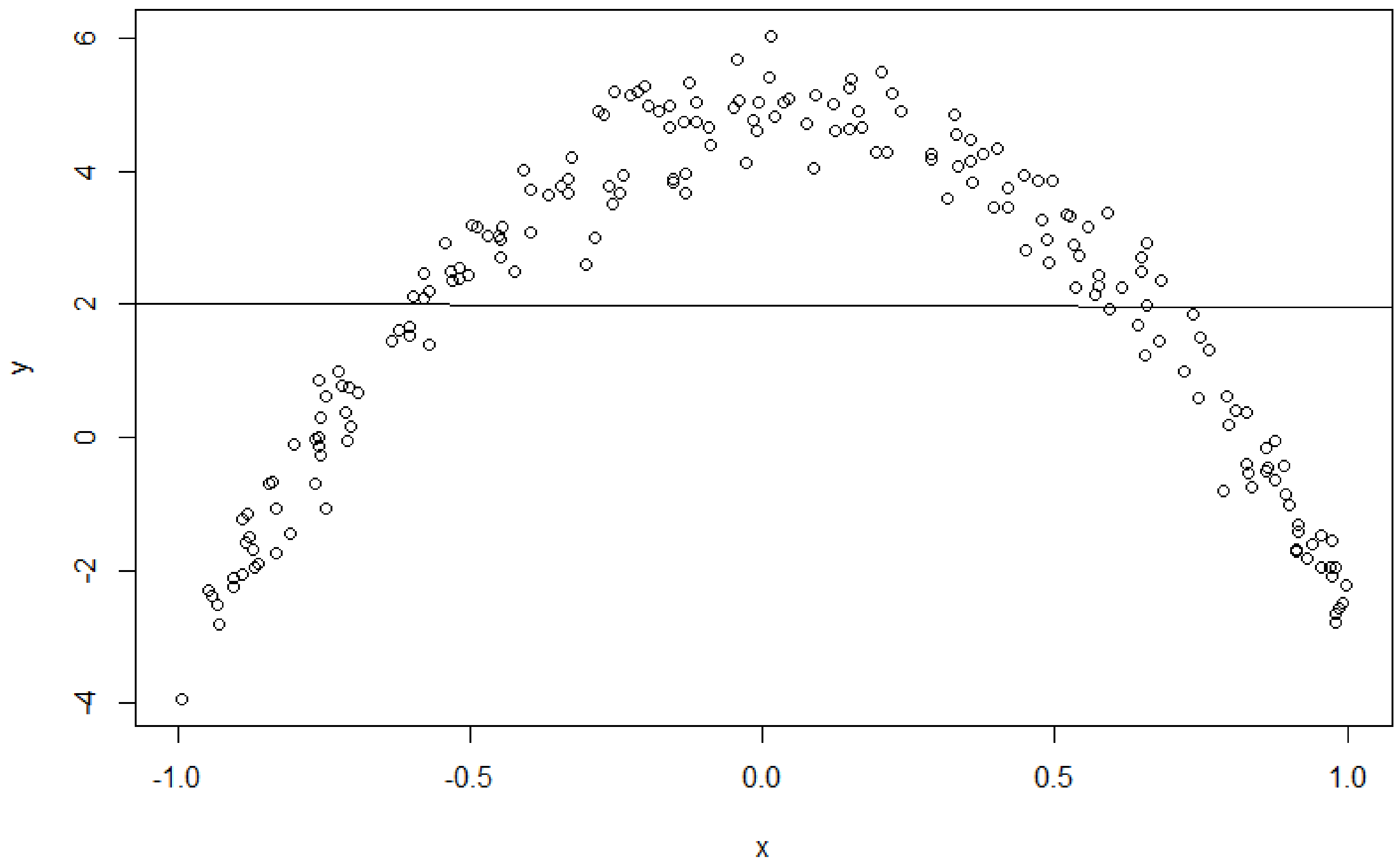
We are missing the possibility of a nonlinear relationship between *y* and *x*.

```
plot(x,y)
```



Plot the fitted line form the linear regression:

```
abline(lm(y ~ x))
```



The linear model is *misspecified* (a form of omitted variable bias). We can approximate the nonlinear relationship using a polynomial, and instead specify the population model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i^* \,.$$

Usually a quadratic form is enough, but we have included x_i^3 as well.

We create the new variables:

```
x2 <- x^2
x3 <- x^3
```

and run OLS:

```
summary(lm(y ~ x + x2 + x3))
```

```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.93434    0.05298  93.131  < 2e-16 ***
x              0.60236    0.15031   4.008 8.71e-05 ***
x2            -7.93666    0.11065 -71.730  < 2e-16 ***
x3            -0.10524    0.22175  -0.475    0.636    
```

We find that x_i^3 is insignificant, so we remove it from the estimated model:

```
summary(lm(y ~ x + x2))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.93537    0.05283   93.41   <2e-16 ***
x            0.53617    0.05591    9.59   <2e-16 ***
x2          -7.94480    0.10909  -72.83   <2e-16 ***
```

How to interpret the estimated model? Have to consider specific values for x_i.