Multiple Linear Regression

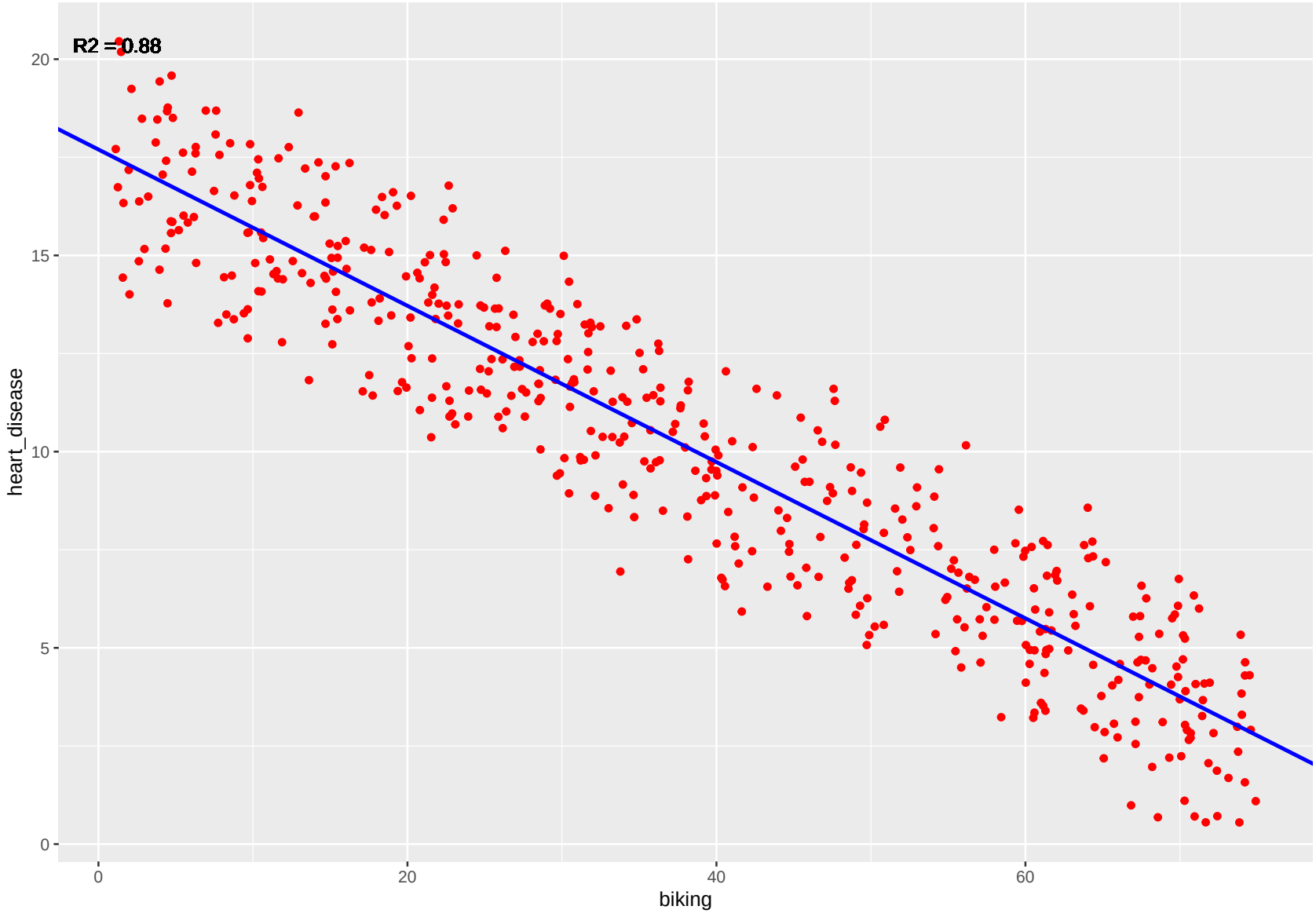
Univariate Case: Simple Linear Regression

- Is there a linear relation between biking and heart disease?

biking	heart_disease
30.801246	11.769423
65.129215	2.854081
1.959664	17.177803
44.800196	6.816647
69.428454	4.062223
54.403626	9.550046
49.056162	7.624507
4.784604	15.854654
65.730788	3.067462
35.257449	12.098484

Simple Linear Regression

- Regress heart_disease on biking
- Response/Dependent Variable: heart_disease
- Predictor variable: biking


R2 = 0.88
heart_disease
biking
0
5
10
15
20
0
20
40
60

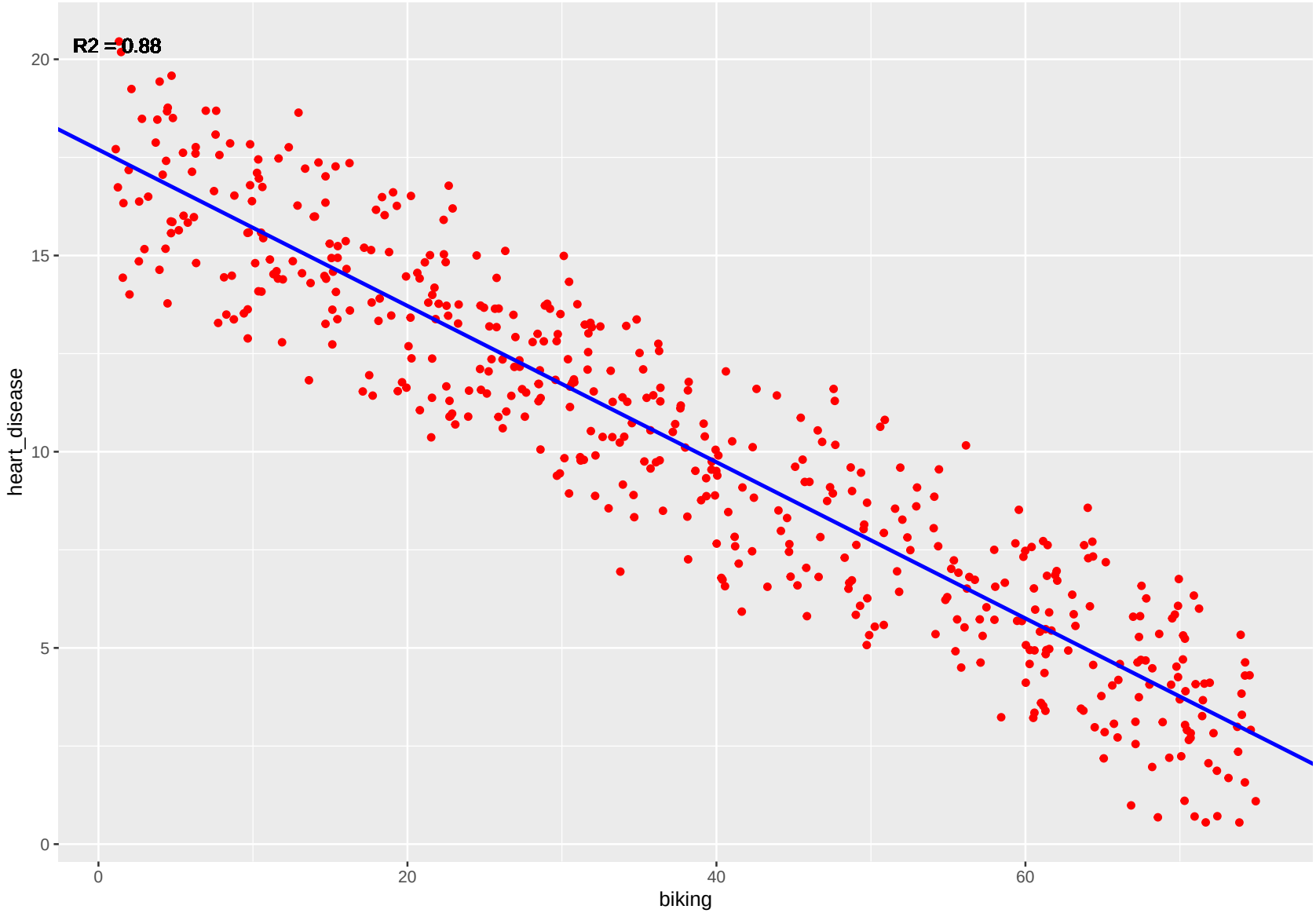
```
Call:
lm(formula = heart_disease ~ biking, data = d1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.028 -1.206 -0.004   1.151   3.643

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.697884   0.146780  120.57   <2e-16 ***
biking      -0.199091   0.003378  -58.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.618 on 496 degrees of freedom
Multiple R-squared:  0.8751,    Adjusted R-squared:  0.8748
F-statistic:  3474 on 1 and 496 DF,  p-value: < 2.2e-16
```

Example Data

biking	smoking	heart_disease
30.801246	10.896608	11.769423
65.129215	2.219563	2.854081
1.959664	17.588331	17.177803
44.800196	2.802559	6.816647
69.428454	15.974505	4.062223
54.403626	29.333175	9.550046
49.056162	9.060846	7.624507
4.784604	12.835021	15.854654
65.730788	11.991297	3.067462
35.257449	23.277683	12.098484
51.825567	14.435118	6.430248
52.936197	25.074869	8.608272
48.767479	11.023271	6.722524
26.166801	6.645749	10.597807
10.553075	5.990506	14.079478

Univeriate approach: Simple Linear Model

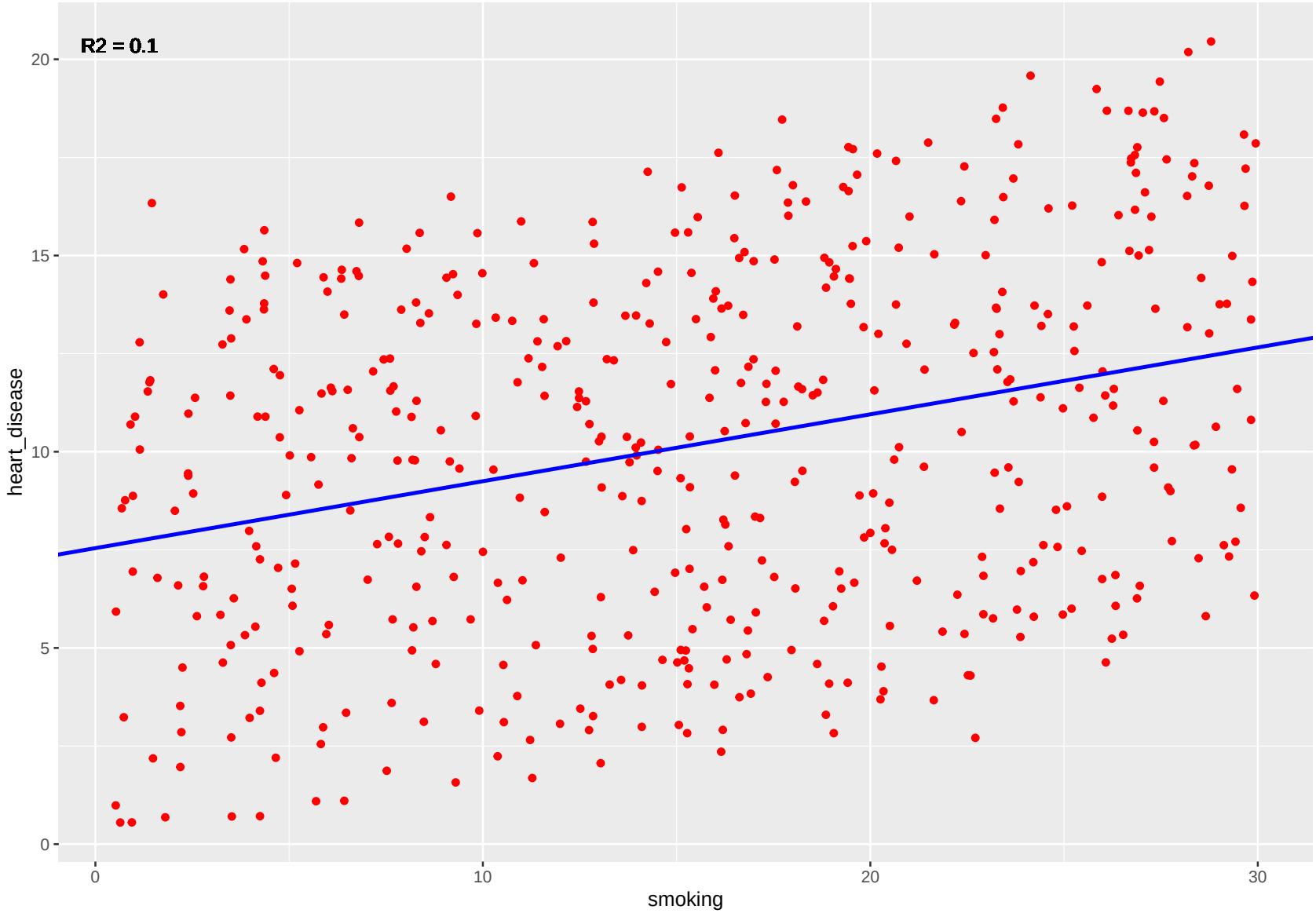
- Regress heart_disease on biking



```
Call:
lm(formula = heart_disease ~ biking, data = d1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.028 -1.206 -0.004   1.151   3.643

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.697884   0.146780  120.57   <2e-16 ***
biking      -0.199091   0.003378  -58.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.618 on 496 degrees of freedom
Multiple R-squared:  0.8751,    Adjusted R-squared:  0.8748
F-statistic:  3474 on 1 and 496 DF,  p-value: < 2.2e-16
```

Univeriate approach: Simple Linear Model

- Regress heart_disease on smoking



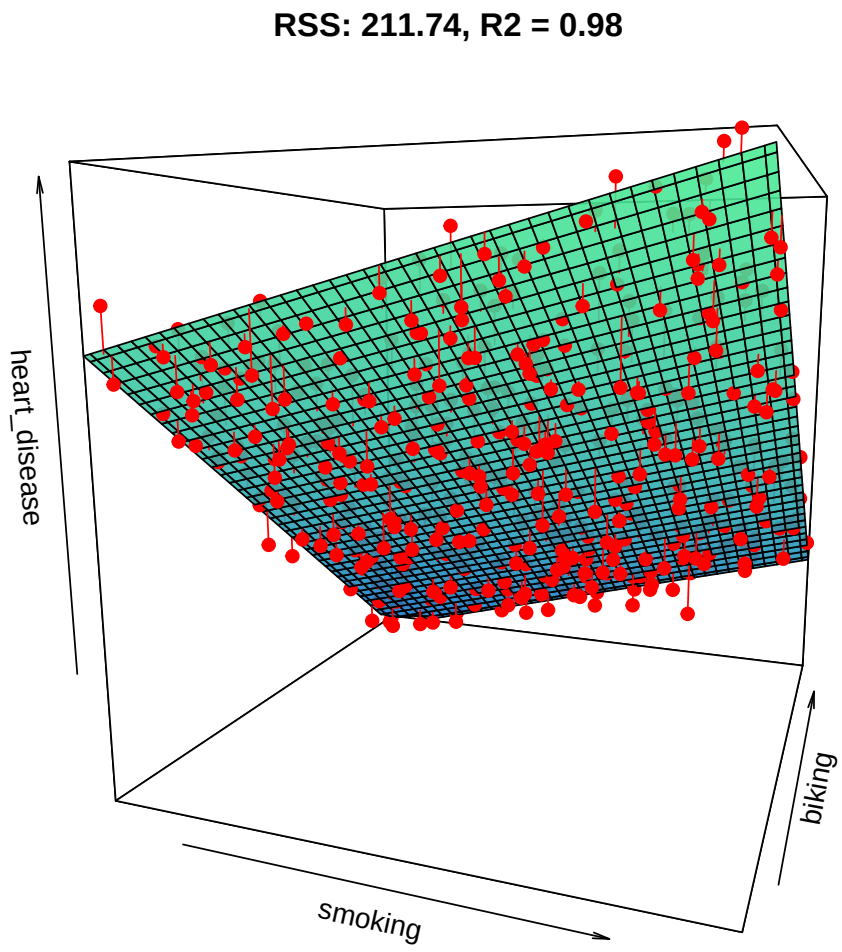
```
Call:
lm(formula = heart_disease ~ smoking, data = d1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7065 -3.7069  0.5007  3.6597  8.5434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.54311    0.41251  18.286  < 2e-16 ***
smoking      0.17048    0.02355   7.239 1.73e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 4.352 on 496 degrees of freedom
Multiple R-squared:  0.09556,   Adjusted R-squared:  0.0937
F-statistic: 52.41 on 1 and 496 DF,  p-value: 1.729e-12
```

- Is there a better way? better model?

Multivariate Approach: Multiple Regression Model

- heart_disease $= \beta_0 + \beta_1 \cdot$ biking $+ \beta_2 \cdot$ smoking $+ \epsilon$
- $\epsilon \sim N(0, \sigma^2)$

Graphing the solution



- heart_disease $= 14.98 + \text{-}0.2 \cdot \text{biking} + 0.18 \cdot \text{smoking}$

```
Call:
lm(formula = z ~ x + y)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1789 -0.4463  0.0362  0.4422  1.9331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.984658   0.080137  186.99   <2e-16 ***
x            0.178334   0.003539   50.39   <2e-16 ***
y           -1.400931   0.009561 -146.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.654 on 495 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9795
F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

Model Definition

$$y = \beta_0 + \beta_1 x_1 + + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

- Model Assumptions
  - (A1) The response variable y is a random variable and the predictor $x_1, x_2, ..., x_n$ is non-random
  - (A2) $\epsilon \sim N(0, \sigma^2)$

Assignment 1: Linear Models in R

Assignment 1