Multiple Linear Regression

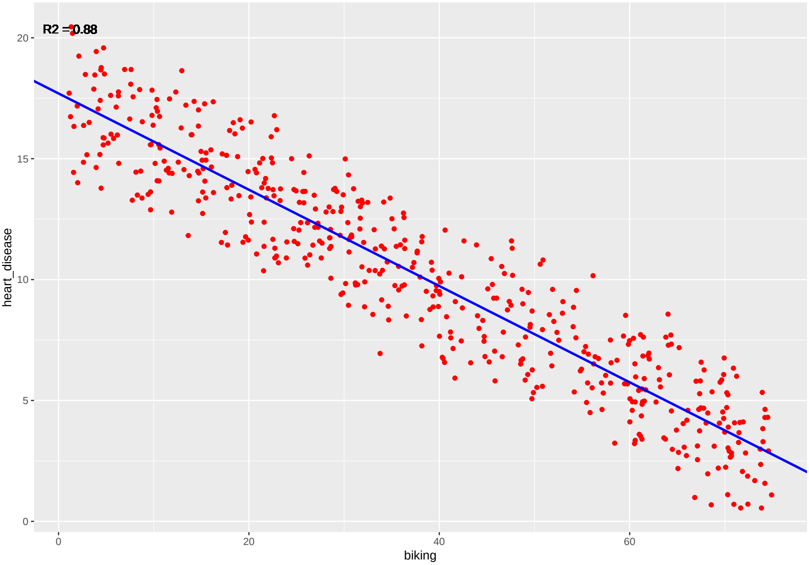
Univariate Case: Simple Linear Regression

- 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

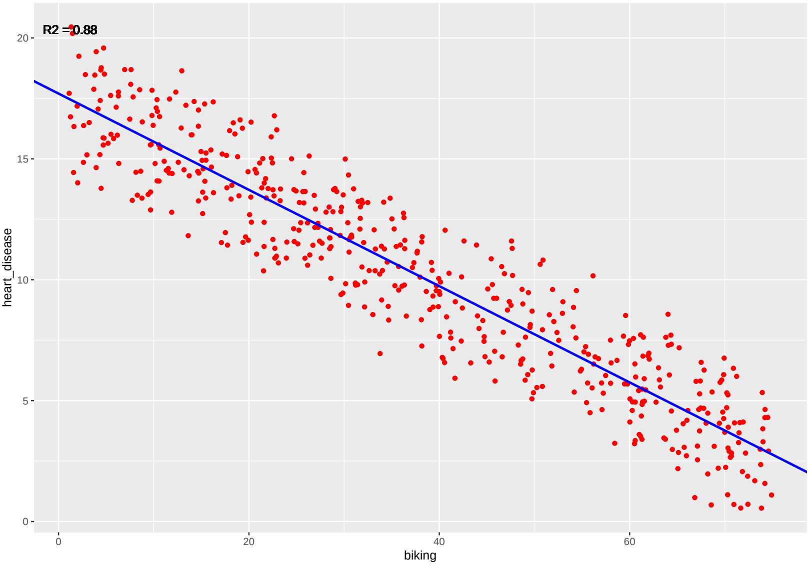
```
Call:
lm(formula = heart_disease ~ biking, data = d1)

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:  0.874
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

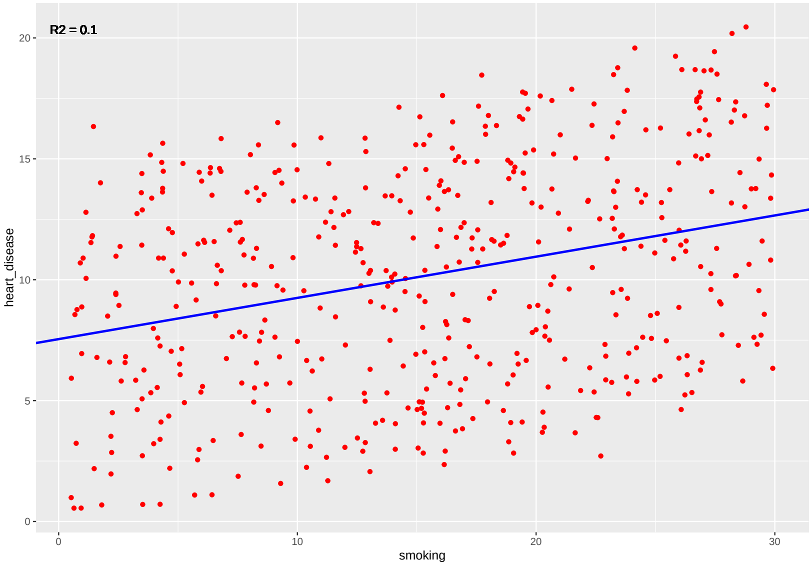
- Regress heart_disease on biking



```
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:  0.874
F-statistic:  3474 on 1 and 496 DF,  p-value: < 2.2e-16
```

Univeriate approach: Simple Linear Model

- Regress heart_disease on smoking



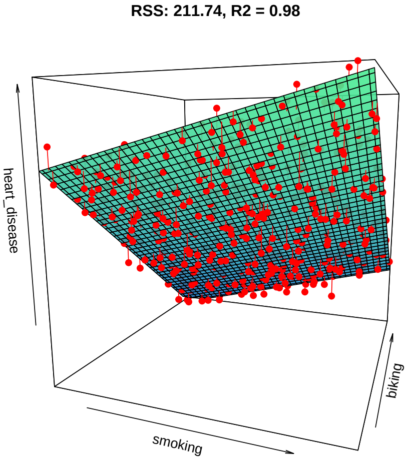
```
Call:
lm(formula = heart_disease ~ smoking, data = d1)

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:  0.093
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 $+$ -0.2 $\cdot$ biking $+$ 0.18 $\cdot$ 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:  0.979
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) $\epsilon \sim N(0, \sigma^2)$

Parameters Estimation

Data Presentation

Observation	Response Variable	Predictors			
	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Matrix Equation of MLR

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Goodness of Fit

Coefficient of Determination

- Similarly to the case of SLR, we have

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Reg SS}}$$

- And

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

F-test

- Full Model:

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

- Baseline Model or i.i.d model:

$$y = \beta_0 + \epsilon$$

- The baseline model is equivalent to

$$\beta_1 = \beta_2 = ... = \beta_p = 0$$

- We would like to test for the joint significant of all predictors, or if the full model is a significant improvement over the baseline model, or

$$\mathrm{H}_0: \quad \underbrace{\beta_1 = \beta_2 = \cdots = \beta_p = 0}_{\text{i.i.d. model}} \quad \text{vs.} \quad \mathrm{H}_a: \quad \underbrace{\text{at least one } \beta_j \text{ is non-zero}}_{\text{MLR model}}.$$

- Test Statistics

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS}_1)/p}{\mathrm{RSS}\ /(n-p-1)} = \frac{\mathrm{Reg\ SS}\ /p}{\mathrm{RSS}\ /(n-p-1)},$$

ANOVA Table

- The results of MLR are usually summarized in the ANOVA table

Source	Sum of Squares	df	Mean Square	F
Regression	Reg SS	p	Reg SS$/p$	$\dfrac{\text{Reg SS}/p}{\text{RSS}/[n-(p+1)]}$
Error	RSS	$n-(p+1)$	$s^2 = \text{RSS}/[n-(p+1)]$	
Total	TSS	$n-1$		

Example

An actuary uses multiple regression model with three predictors and 20 observations and has the following results.

	Sum of Squares
Regression	150
Total	200

He wants to test the following hypothesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

H_1 : At least one of β_1, β_2, and β_3 is zero

Calculate the F-statistics of the test.

From R2 to F-test

- The R^2 and the $F - statistics$ have the following relation

$$F = \frac{RegSS/p}{RSS/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

and

$$R^2 = \frac{Fp}{Fp + n - p - 1}$$

Example

Sarah performs a regression of the return on a mutual fund (y) on four predictors plus an intercept. She uses monthly returns over 105 months. Her software calculates the $R^2 = .8$ but then it quits working before it calculates the value of F. Calculates the F-statistics for Sarah.

Generalized F-test

- Full Model:

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

- Reduced Model:

$$y = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + ... + \beta_{p-q} x_{p-q} + \epsilon$$

	Reduced model		Full model
RSS	RSS_0	$\geq$	RSS_1
Reg SS	$(\mathrm{Reg\ SS})_0$	$\leq$	$(\mathrm{Reg\ SS})_1$
TSS	TSS	$=$	TSS

Model	RSS	$RegSS$	
Reduced	RSS_0	$RegSS_0$	TSS
Full	RSS_1	$RegSS_1$	TSS

- $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_{p-q} = 0$ or Reduced model is adequate
- Test Statistics

$$F = \frac{\text{Extra SS}/q}{\text{RSS}_1/(n-p-1)} = \frac{(\text{RSS}_0 - \text{RSS}_1)/q}{\text{RSS}_1/(n-p-1)}.$$

Example

- Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

The results of the regression are as follows:

Model Number	Residual Sum of Squares	Regression Sum of Squares
1	13.47	22.75
2	10.53	25.70

The null hypothesis is $\mathrm{H}_0 : \beta_3 = \beta_4 = 0$ with the alternative hypothesis that the two betas are not equal to zero.

Calculate the statistic used to test H_0.

Example

You wish to find a model to predict insurance sales, using 27 observations and 8 variables x_1, x_2,...,x_8. The analysis of variance (ANOVA) tables are below. Model A contains all 8 variables and Model B contains x_1 and x_2 only.

Calculate the F-statistics for testing
$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

Model A

Source	SS	df	MS
Regression	115,175	8	14,397
Error	76,893	18	4,272
Total	192,068	26	

Model B

Source	SS	df	MS
Regression	65,597	2	32,798
Error	126,471	24	5,270
Total	192,068	26	