Tutorial 9

Biological Data Analysis

Spring 2023

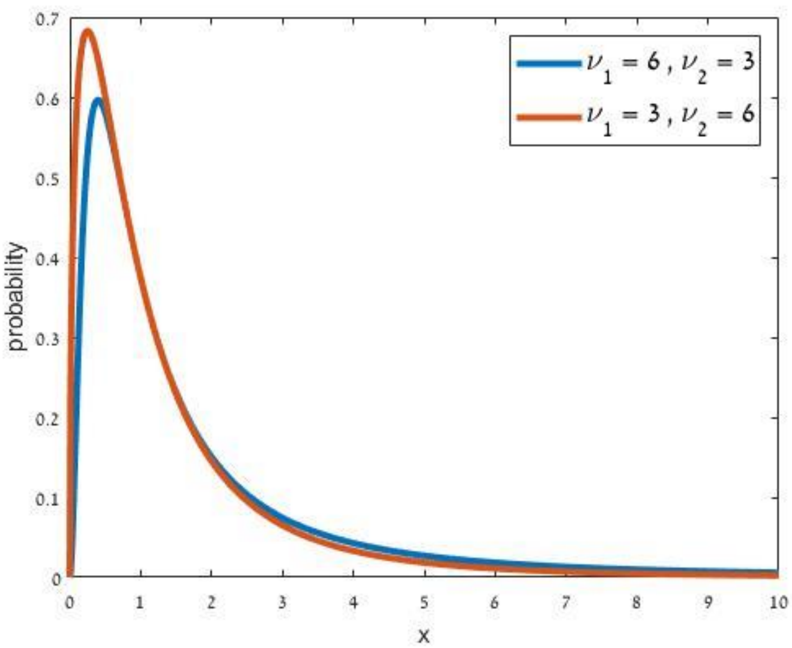
Outline

- F test
- Linear Regression

F Distribution

Ratio of two estimator of the same variance distributes F:

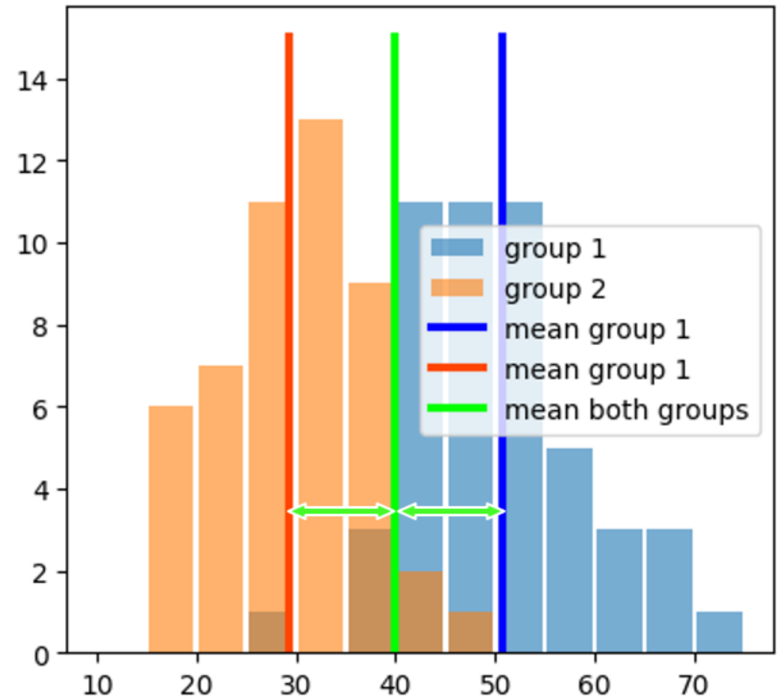
Defined by two parameters: degree of freedom of the nominator and degree of freedom of the denominator



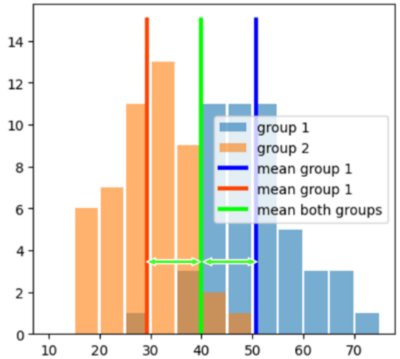
$$F_{\nu_1,\nu_2} = \frac{\sigma^2 \dfrac{\chi^2_{\nu_1}}{\nu_1}}{\sigma^2 \dfrac{\chi^2_{\nu_2}}{\nu_2}} = \frac{\dfrac{\chi^2_{\nu_1}}{\nu_1}}{\dfrac{\chi^2_{\nu_2}}{\nu_2}}$$

F test for difference between two groups

- Null Model – both samples are from the same population/distribution, so the difference between the means of the groups is due to the variance inside the population.
- Variance of sample means is from the same distribution as variance of the samples
- The ratio of two variances Comes from F distribution



F test for difference between two groups



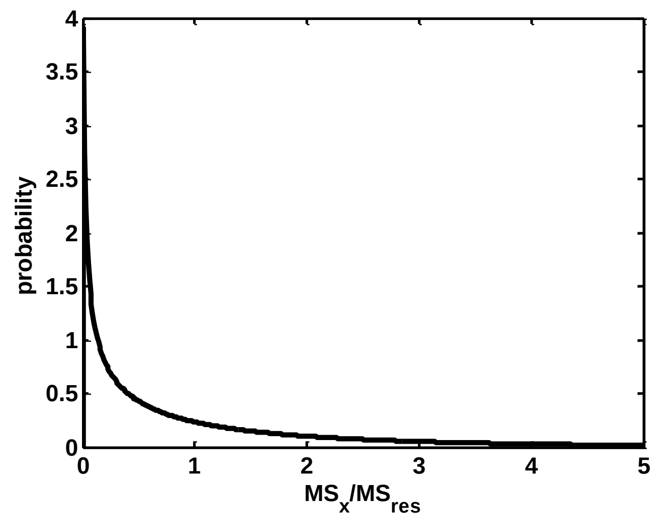
$$SS_{\bar{x}} = N\sum\left(\bar{x}_i - \bar{x}\right)^2 \sim \sigma^2 \chi^2_{2-1}$$

$$MS_{\bar{x}} = \frac{SS_{\bar{x}}}{K-1}$$

$$SS_{\text{res}} = \sum\left(x_{1,i} - \bar{x}_1\right)^2 + \sum\left(x_{2,i} - \bar{x}_2\right)^2 \sim \sigma^2 \chi^2_{2N-2}$$

$$MS_{res} = \frac{SS_{res}}{2N-K} = \frac{\sum_{j=1}^{K}\sum_{i=1}^{M_j}(x_{ij} - \bar{x}_j)^2}{\sum_{j=1}^{K} M_j - K}$$

$$\frac{MS_{\bar{x}}}{MS_{\text{res}}} \sim \frac{\sigma^2 \frac{\chi^2_{2-1}}{2-1}}{\sigma^2 \frac{\chi^2_{2N-2}}{2N-2}} = \frac{\frac{\chi^2_{2-1}}{2-1}}{\frac{\chi^2_{2N-2}}{2N-2}} = F_{2-1,2N-2}$$

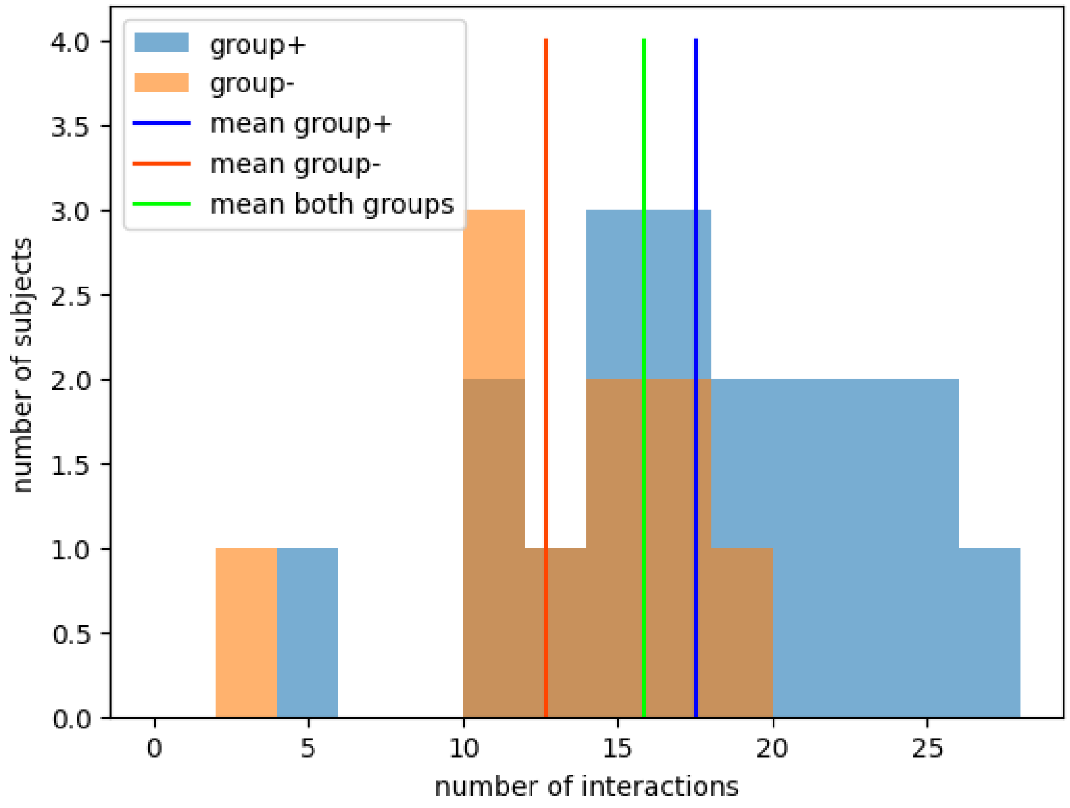


Exercise 1

Serotonin is a chemical that influences mood balance.
How does it affect mice behavior?
Scientists genetically altered mice by "knocking out" the expression of a gene, tryptophan hydroxylase 2 (Tph2), that regulates serotonin production.
With careful breeding, the scientists produced two types of mice that we label as “Minus” for Tph2-/-, “Plus” for Tph2+/+.
The variable ‘genotype’ records Minus/Plus.
The variable ‘interactions’ is the number of social contacts that a mouse had with other mice during an experiment.

interactions	genotype
23	Plus
15	Plus
15	Plus
19	Plus
20	Plus
25	Plus
16	Plus
26	Plus
17	Plus
22	Plus
17	Plus
21	Plus
5	Plus
12	Plus
11	Plus
11	Plus
19	Plus
15	Plus
24	Plus
2	Minus
15	Minus
12	Minus
16	Minus
16	Minus
11	Minus
11	Minus
15	Minus
11	Minus
18	Minus

Exercise 1


group+
group-
mean group+
mean group-
mean both groups
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
number of subjects
0
5
10
15
20
25
number of interactions

Example 1

$$SS_{\bar{x}} = N_1\left(\bar{x}_1 - \bar{x}_{1,2}\right)^2 + N_2\left(\bar{x}_2 - \bar{x}_{1,2}\right)^2$$

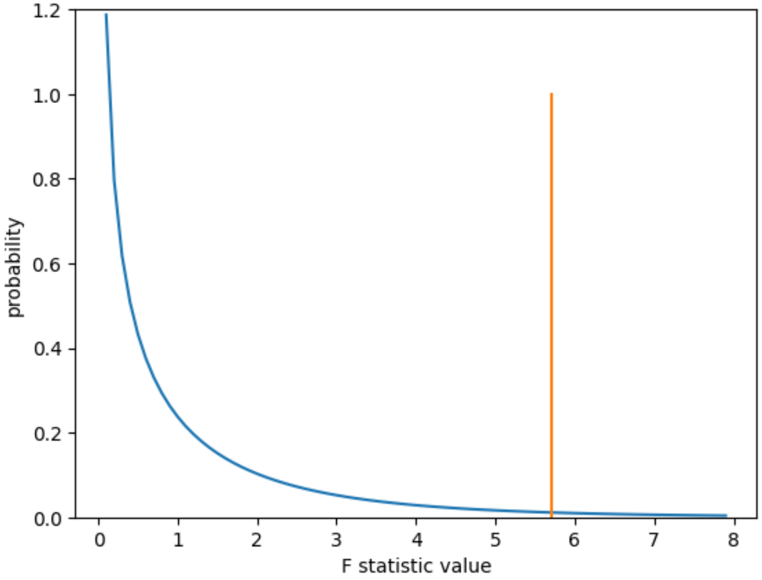
$$MS_{\bar{x}} = \frac{SS_{\bar{x}}}{2-1}$$

$$SS_{res} = \Sigma\left(x_{1i} - \bar{x}_1\right)^2 + \Sigma\left(x_{2i} - \bar{x}_2\right)^2$$

$$MS_{res} = \frac{SS_{res}}{N_1 + N_2 - 2}$$

$$\frac{MS_{\bar{x}}}{MS_{res}} \sim F_{2-1,N-2}$$

F = 5.7163
p-value = 0.024

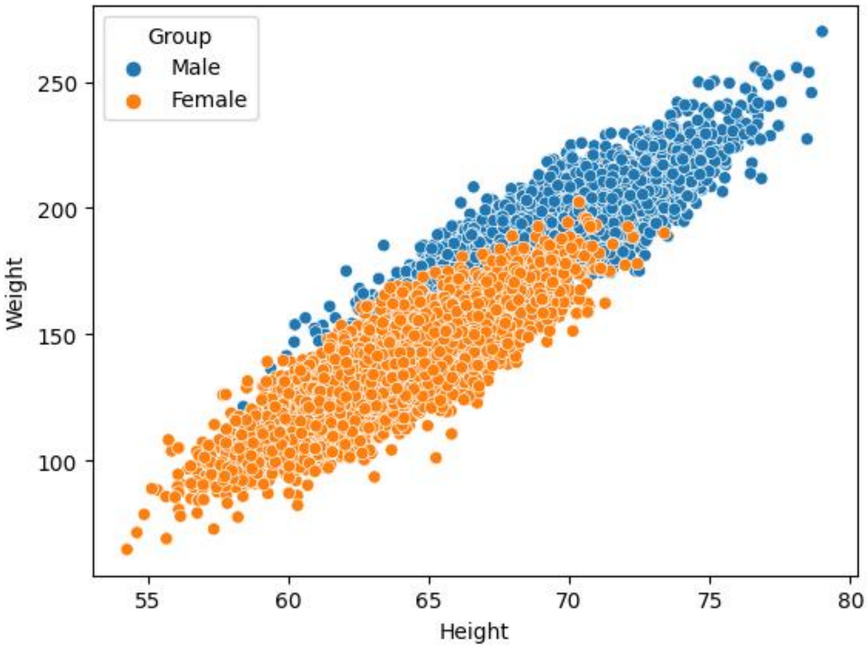


Example 2
Linear Regression

We have data height and weight of 5000 men and 5000 women

Is there a connection between height and weight?

Is this connection different in men and women



Example 2
Linear Regression

$$\hat{y}(x) = b_0 + b_1 x + r$$

$$r \sim N(0, S_{y|x}^2)$$

$$b_1 = \frac{\sum\left(x_n - \overline{x}\right)\left(y_n - \overline{y}\right)}{\sum\left(x_n - \overline{x}\right)^2}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

$$s_{y|x}^2 = \frac{SS_{res}}{N-2} = \frac{\sum_n (y_n - \hat{y}(x_n))^2}{N-2}$$

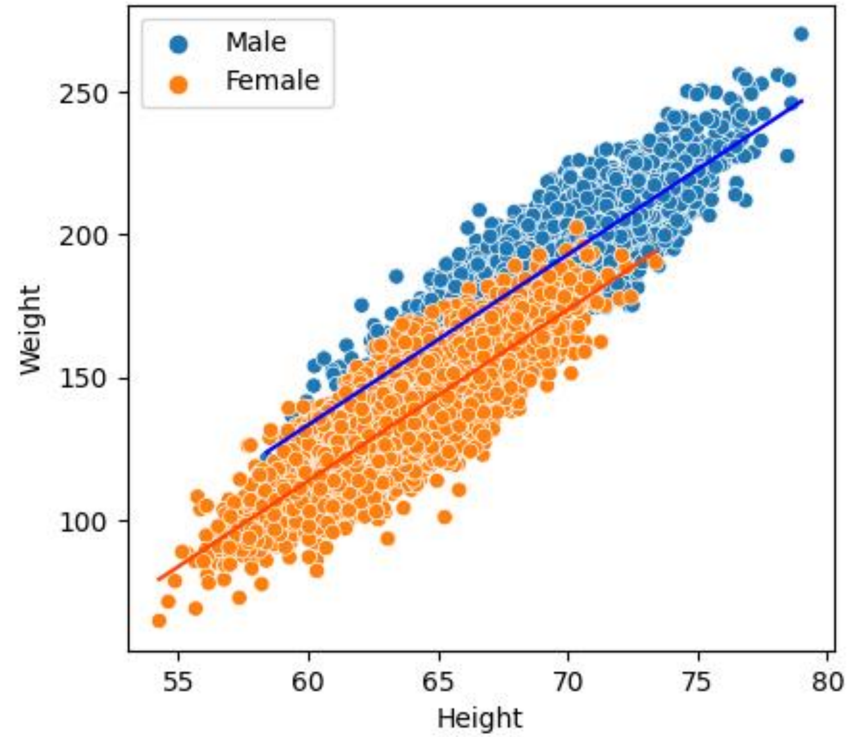
Example 2
Linear Regression

$$\hat{y}(x) = b_0 + b_1 x + r$$

$$r \sim N(0, S_{y|x}^2)$$

```
weight(height) for males model:
b0 = -224.4988 , b1 = 5.9618 ,
Syx2 = 99.9046
```

```
weight(height) for females model:
b0 = -246.0133 , b1 = 5.994 ,
Syx2 = 100.6754
```



Example 2
Linear Regression

Significance of coefficients:
p- value using t-test

$$se_{b_1} = \frac{s_{y|x}}{\sqrt{(N-1)s_x^2}} \qquad \frac{b_1}{se_{b_1}} \to t_{N-2}$$

```
b1 males: p-value = 0.0 ,
R2 = 0.7447 , f2 = 2.9175
```

```
b1 females: p-value = 0.0 ,
R2 = 0.7218 , f2 = 2.595
```

Model Effect size: R2 and f2

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \qquad R^2 = \frac{SS_{reg}}{SS_{tot}}$$

$$SS_{reg} = \sum_i (\hat{y}_i - \bar{y}_i)^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y}_i)^2 \qquad f^2 = \frac{SS_{reg}}{SS_{res}}$$

Example 2
Linear Regression

Significance of the model: p- value using F test

$$SS_{reg} = \sum_i \left(\hat{y}_i - \bar{y}_i \right)^2$$

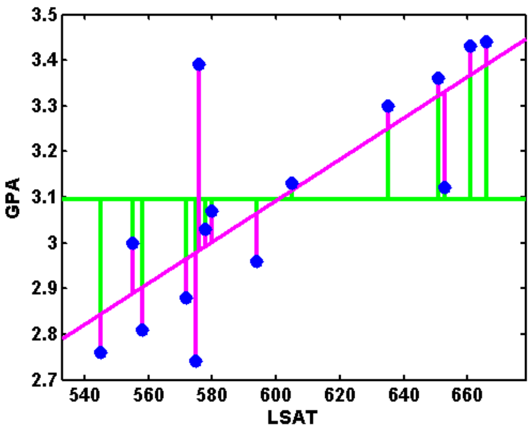
$$MS_{reg} = \frac{SS_{reg}}{K-1}$$

$$SS_{res} = \sum_i \left(y_i - \hat{y}_i \right)^2$$

$$MS_{res} = \frac{SS_{res}}{N-K}$$

K - number of independent variables (number of coefficients in a model)
N – number of data points

$$\frac{MS_{reg}}{MS_{res}} \rightarrow F_{K-1,N-K}$$




```
males model significance: F stat = 14581.4602 p-value = 0.0

females model significance: F stat = 12969.7374  p-value = 0.0
```

Comparing two regression lines

$$t = \frac{b_1 - b_2}{s_b} = \frac{b_1 - b_2}{SSres\sqrt{\frac{1}{(N_1 - 1)s_{x1}^2} + \frac{1}{(N_2 - 1)s_{x2}^2}}}$$

$$SSres^2 = \frac{(\mathrm{N}_1 - 2)SSres_{x1} + (\mathrm{N}_2 - 2)SSres_{x2}}{\mathrm{N}_1 + \mathrm{N}_2 - 4}$$

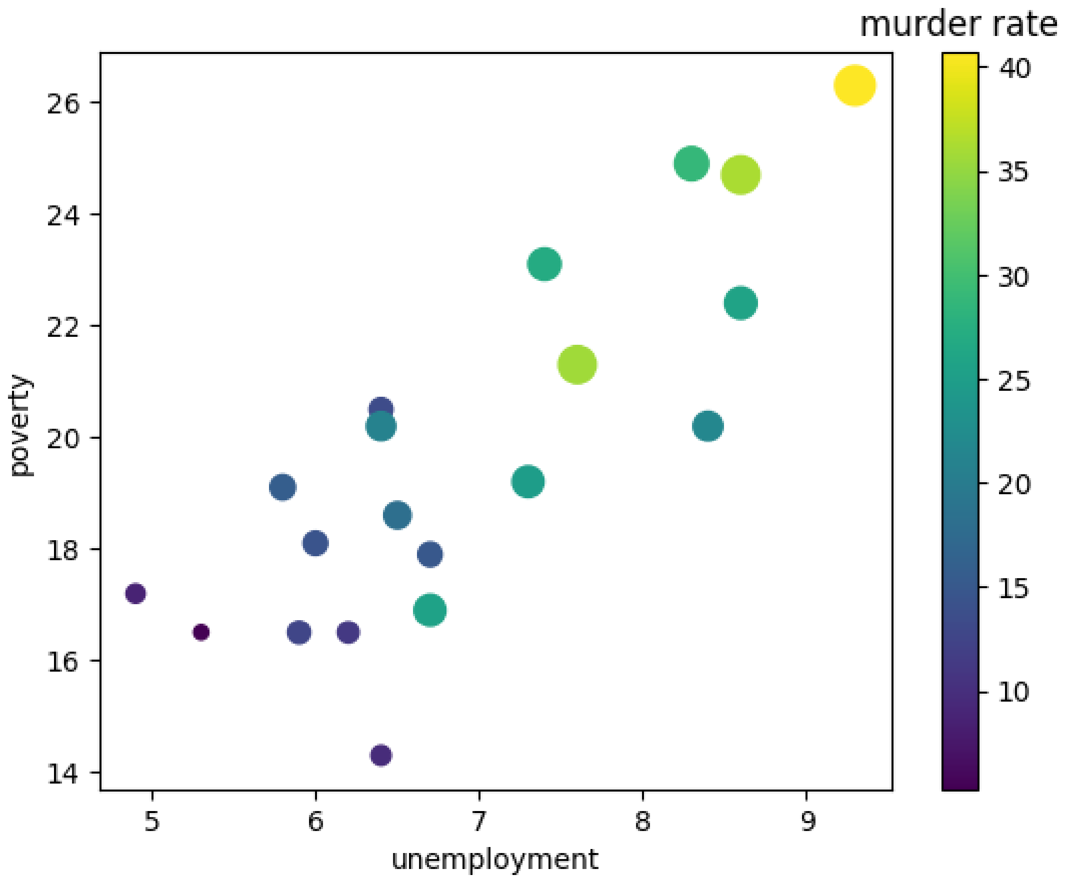
```
p-value for coefficient comparison = 0.4975
```

Example 3
Multivariate Linear Regression

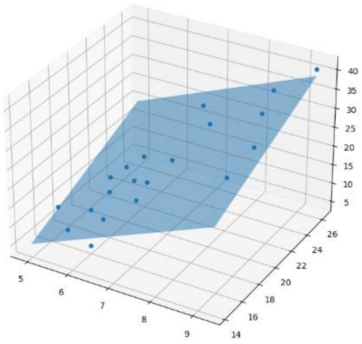
unemployment	poverty	murder_rate
6.2	16.5	11.2
6.4	20.5	13.4
9.3	26.3	40.7
5.3	16.5	5.3
7.3	19.2	24.8
5.9	16.5	12.7
6.4	20.2	20.9
7.6	21.3	35.7
4.9	17.2	8.7
6.4	14.3	9.6
6.0	18.1	14.5
7.4	23.1	26.9
5.8	19.1	15.7
8.6	24.7	36.2
6.5	18.6	18.1
8.3	24.9	28.9
6.7	17.9	14.9
8.6	22.4	25.8
8.4	20.2	21.7
6.7	16.9	25.7

In 1991, a group of researches gathered data on murder rate, poverty rate and unemployment in 20 different cities.
We will check the dependence of murder rate on poverty and unemployment

Example 3 presentation



Multivariate Linear Regression



The model $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}; \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{Xb}$$

$$SS_{\text{res}} = \sum \left(y_n - \hat{y}_n \right)^2 \qquad \hat{y}_n = b_0 + x_{n,1}b_1 + x_{n,2}b_2$$

$$\sum \left(y_n - \hat{y}_n \right)^2 = \left(\mathbf{y} - \mathbf{Xb} \right)^T \left(\mathbf{y} - \mathbf{Xb} \right)$$

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

$$s_{y|x}^2 = MS_{\text{res}} = \frac{SS_{\text{res}}}{\nu_{\text{res}}} = \frac{\left(\mathbf{y} - \mathbf{Xb} \right)^T \left(\mathbf{y} - \mathbf{Xb} \right)}{n - 3}$$

$$\mathbf{s}_{\mathbf{b}}^2 = s_{y|x}^2 \left(\mathbf{X}^T \mathbf{X} \right)^{-1}$$

Example 3

$$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i + r \quad ; \quad \mathrm{r \sim N(0, s^2_{y|X})}$$

Effect size for regression model $R^2 = \frac{SS_{reg}}{SS_{tot}}$ $f^2 = \frac{SS_{reg}}{SS_{res}}$

```
b0 = -34.0725 , b1 = 4.3989 , b2 = 1.2239
```

```
Syx2 = 21.6084   , R2 = 0.802 , f2 = 4.0503
```

Multivariate Linear Regression

- Using F test we check how good is our model in explaining the data
- Using t-test we check the significance of the coefficient estimators

Example 3

Significance of the regression model

$$\frac{MS_{reg}}{MS_{res}} \rightarrow F_{K-1,N-K}$$

```
model significance: F stat = 34.4278 , p-value = 1.05e-06
```

Significance of the model coefficients

$$\frac{b_i}{s_{bi}} \rightarrow t_{N-K} \qquad \mathbf{s}_{\mathbf{b}}^2 = s_{y|x}^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

```
model coefficient significance:
b1 p-value = 0.0052 , b2 p-value = 0.0229
```

Addition test

We check if the addition of a factor (independent variable) significantly improves the model, i.e. adds to how good our model explains the data.

We compare the full model that includes the variable we are checking to a partial model without this variable.

If both models – full and partial – explain the data equally, MSadd/MSres ~ F(K_full-K_part , N-K_full)

If the probability of getting a F statistic or more extreme is lower than alpha error, we conclude that the addition of the variable significantly improves the model

Addition test

```
poverty variable contribution:
p-value = 0.0459
```

```
unemployment variable
contribution: p-value = 0.0103
```

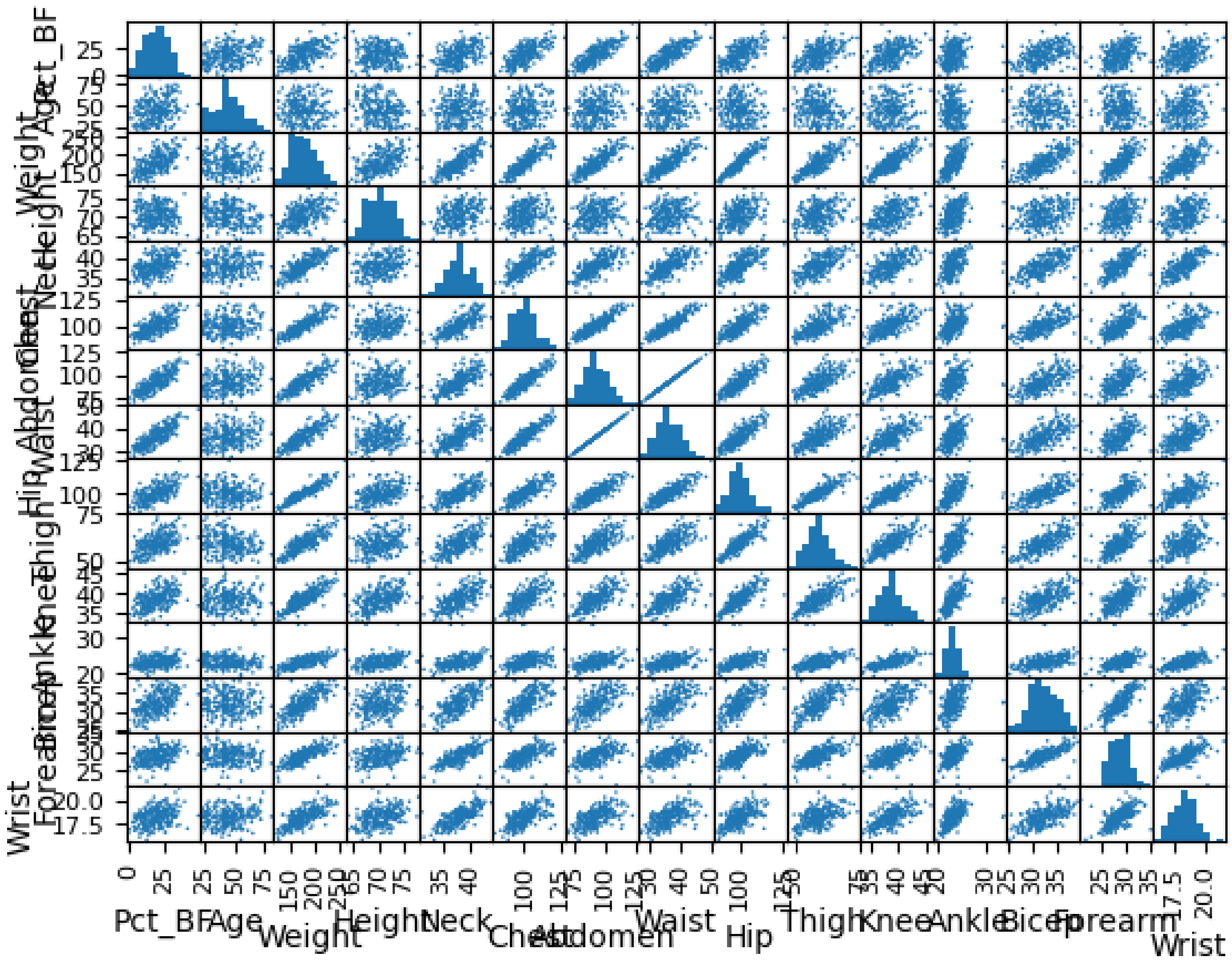
Both additions are significant, so we should use both variables in our model

Sometimes the variable coefficient is significant according to the t-test, but the addition test is not significant and we conclude that the variable is not essential for the model

Example 4
Multivariate Linear Regression

A group of researches gathered data on blood fat percentage in 250 adult males. They also did additional non-invasive measurements. Can we predict the blood fat given easily done measurements?

	Pct_BF	Age	Weight	Height	Neck	Chest	Abdomen	Waist	Hip	Thigh	Knee	Ankle	Bicep	Forearm	Wrist
0	12.3	23	154.25	67.75	36.2	93.1	85.2	33.543307	94.5	59.0	37.3	21.9	32.0	27.4	17.1
1	6.1	22	173.25	72.25	38.5	93.6	83.0	32.677165	98.7	58.7	37.3	23.4	30.5	28.9	18.2
2	25.3	22	154.00	66.25	34.0	95.8	87.9	34.606299	99.2	59.6	38.9	24.0	28.8	25.2	16.6
3	10.4	26	184.75	72.25	37.4	101.8	86.4	34.015748	101.2	60.1	37.3	22.8	32.4	29.4	18.2
4	28.7	24	184.25	71.25	34.4	97.3	100.0	39.370079	101.9	63.2	42.2	24.0	32.2	27.7	17.7
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
245	11.0	70	134.25	67.00	34.9	89.2	83.6	32.913386	88.8	49.6	34.8	21.5	25.6	25.7	18.5
246	33.6	72	201.00	69.75	40.9	108.5	105.0	41.338583	104.5	59.6	40.8	23.2	35.2	28.6	20.1
247	29.3	72	186.75	66.00	38.9	111.1	111.5	43.897638	101.7	60.3	37.3	21.5	31.3	27.2	18.0
248	26.0	72	190.75	70.50	38.9	108.3	101.3	39.881890	97.8	56.0	41.6	22.7	30.5	29.4	19.8
249	31.9	74	207.50	70.00	40.8	112.4	108.5	42.716535	107.1	59.3	42.2	24.6	33.7	30.0	20.9




```
['Pct_BF', 'Age', 'Weight', 'Height', 'Neck', 'Chest', 'Abdomen', 'Waist',
'Hip', 'Thigh', 'Knee', 'Ankle', 'Bicep', 'Forearm', 'Wrist']
```

Example 4

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Pct_BF   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.736
Method:                 Least Squares   F-statistic:                     50.50
Date:                Mon, 29 May 2023   Prob (F-statistic):           1.12e-62
Time:                        10:50:21   Log-Likelihood:                -709.57
No. Observations:                 250   AIC:                             1449.
Df Residuals:                     235   BIC:                             1502.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.7087     23.423      0.073      0.942     -44.438      47.855
Age            0.0717      0.032      2.224      0.027       0.008       0.135
Weight        -0.0173      0.067     -0.257      0.797      -0.150       0.115
Height        -0.2490      0.192     -1.296      0.196      -0.628       0.130
Neck          -0.3841      0.236     -1.627      0.105      -0.849       0.081
Chest         -0.1201      0.109     -1.105      0.270      -0.334       0.094
Abdomen    -5.851e+04   3.91e+05     -0.150      0.881   -8.29e+05    7.12e+05
Waist       1.486e+05   9.94e+05      0.150      0.881   -1.81e+06    2.11e+06
Hip           -0.1573      0.147     -1.074      0.284      -0.446       0.131
Thigh          0.1720      0.147      1.168      0.244      -0.118       0.462
Knee          -0.0432      0.247     -0.175      0.861      -0.529       0.443
Ankle          0.1839      0.220      0.834      0.405      -0.250       0.618
Bicep          0.1747      0.174      1.005      0.316      -0.168       0.517
Forearm        0.2797      0.209      1.340      0.182      -0.132       0.691
Wrist         -1.7976      0.535     -3.361      0.001      -2.851      -0.744
==============================================================================
```

Example 4

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Pct_BF   R-squared:                       0.750
Model:                            OLS   Adj. R-squared:                  0.737
Method:                 Least Squares   F-statistic:                     54.61
Date:                Mon, 29 May 2023   Prob (F-statistic):           1.50e-63
Time:                        10:55:31   Log-Likelihood:                -709.58
No. Observations:                 250   AIC:                             1447.
Df Residuals:                     236   BIC:                             1496.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.6852     23.374      0.072      0.943     -44.363      47.734
Age            0.0719      0.032      2.234      0.026       0.009       0.135
Weight        -0.0176      0.067     -0.263      0.793      -0.150       0.115
Height        -0.2468      0.191     -1.291      0.198      -0.623       0.130
Neck          -0.3868      0.235     -1.647      0.101      -0.850       0.076
Chest         -0.1192      0.108     -1.101      0.272      -0.332       0.094
Waist          2.2975      0.232      9.897      0.000       1.840       2.755
Hip           -0.1588      0.146     -1.089      0.277      -0.446       0.129
Thigh          0.1730      0.147      1.178      0.240      -0.116       0.462
Knee          -0.0458      0.246     -0.186      0.852      -0.530       0.438
Ankle          0.1850      0.220      0.842      0.401      -0.248       0.618
Bicep          0.1797      0.170      1.054      0.293      -0.156       0.515
Forearm        0.2761      0.207      1.334      0.183      -0.132       0.684
Wrist         -1.8016      0.533     -3.380      0.001      -2.852      -0.751
==============================================================================
```

Example 4

Addition test on ‘abdomen’ variable

```
abdomen variable contribution:
p-value = 0.8925
```

Variable addition is not significant, we can remove it from our model