MEDSTA 2: Regression models in medical research

16 January 2014

Øystein Ariansen Haaland, PHD

Department of Global Public Health and Primary Care, University of Bergen

Correlation

- Pearson’s **r**
- Measure of linear association between two variables (X and Y)
- Correlation coefficient, **r**, is between -1 and +1.
- If r=0, there is no linear association between the variables

Correlation

- Formula
- $r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$
- r>0: If x_i and y_i are small (or large) at the same time.
- r<0: If x_i is small when y_i is large (and vice versa).

Different values of r

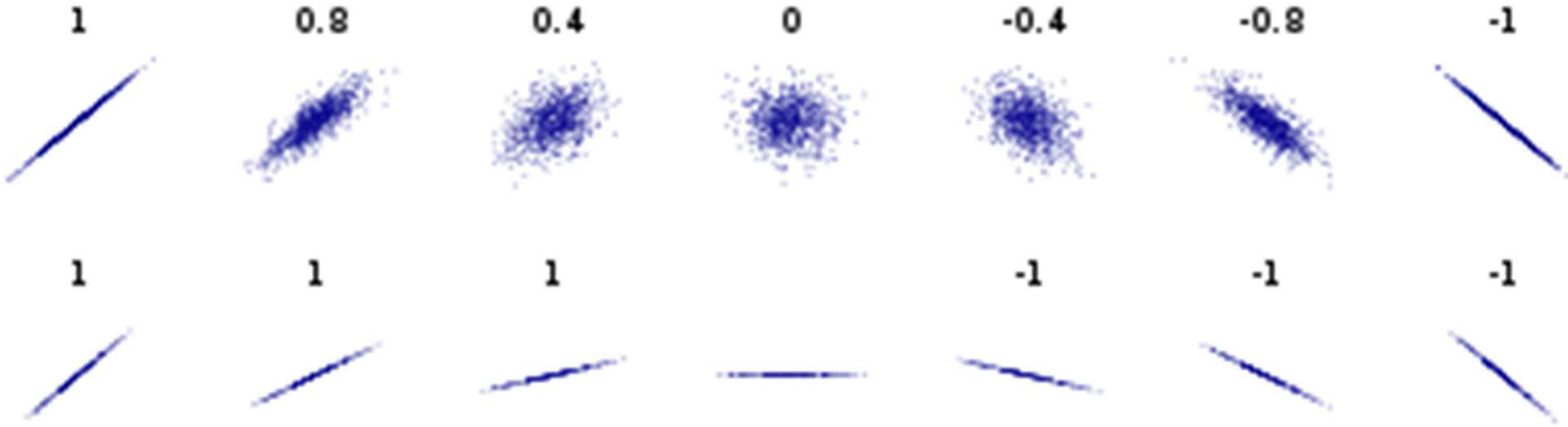
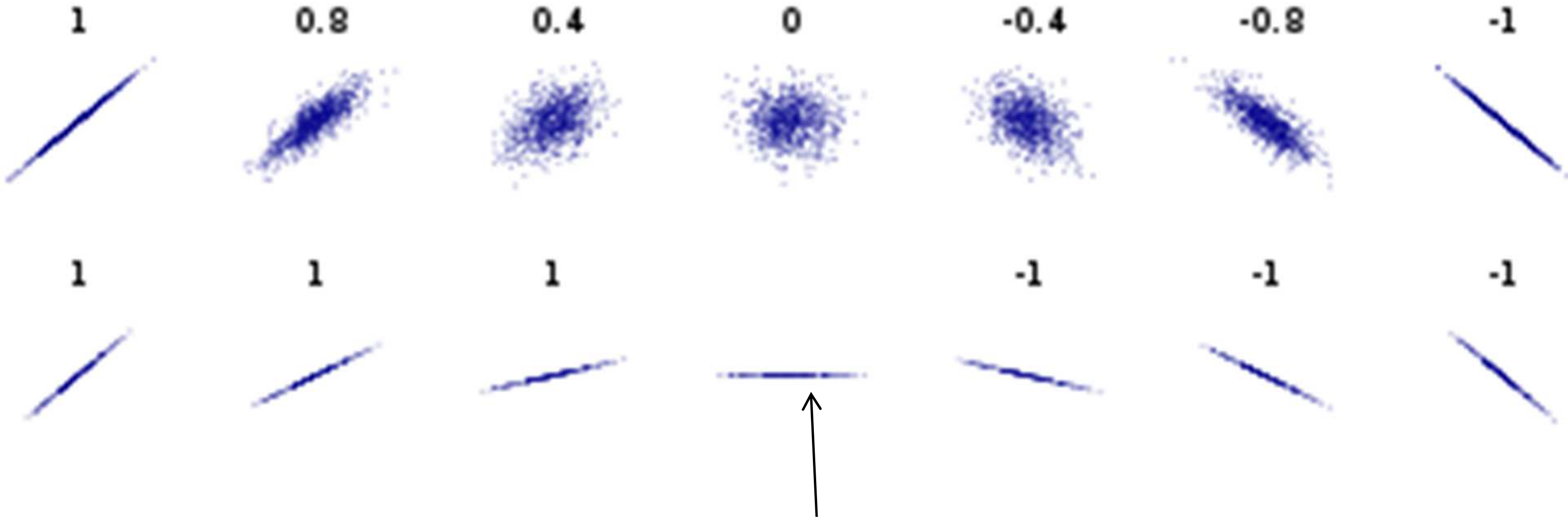


Figure from Wikipedia

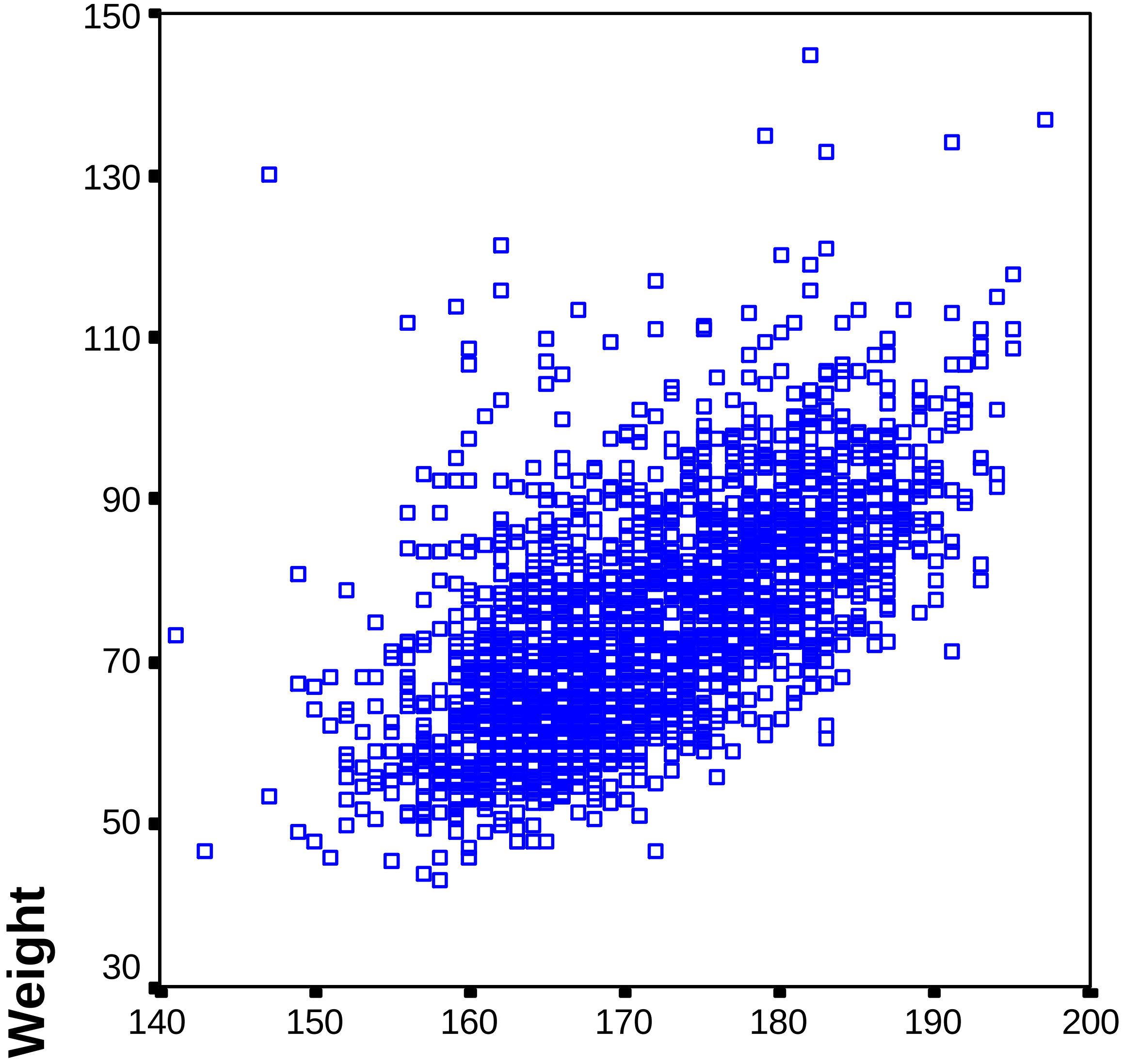
Different values of r



Var(Y)=0 (Y is constant)

Figure from Wikipedia

Real data



r = 0.62

Correlation

- If r is significantly different from 0, we take it as evidence of an association between X and Y.
- $r=0$ does not mean that there is no association between X and Y.
- Why?

Correlation

r=0

No association?

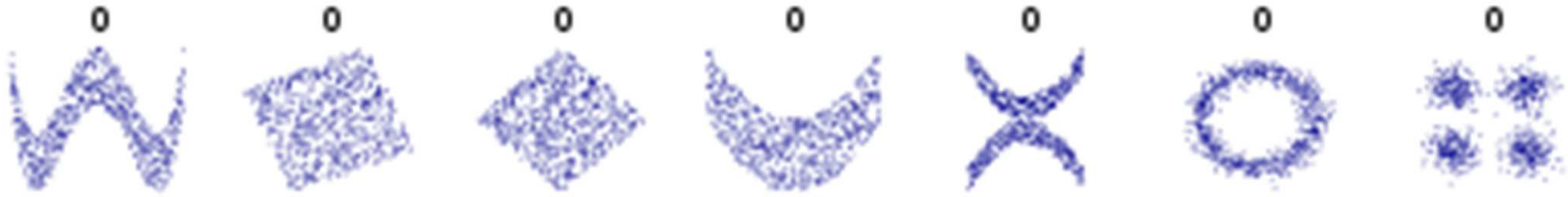


Figure from Wikipedia

Correlation

- Test if r=0
- $H_0\colon r = 0$, $H_1\colon r \neq 0$
- Will not give formula
- Stata
  - pwcorr [var1] [var2], sig
  - pwcorr [var1]…[varK], sig
  - correlate [var1] [var2] does not give p-value

Linear regression

- Study association between dependent variable and independent variable(s)
- Dependent variable =outcome variable =Y-variable =response variable
- Independent variable =predictor variable =covariate =X-variable
- Study variable
- Adjustment variables

Linear regression

- Simple
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - y_i is the observed value of subject i
  - x_i is the independent variable of subject i
  - β's are regression coefficients
  - ϵ_i is error due to chance of subject i

Linear regression

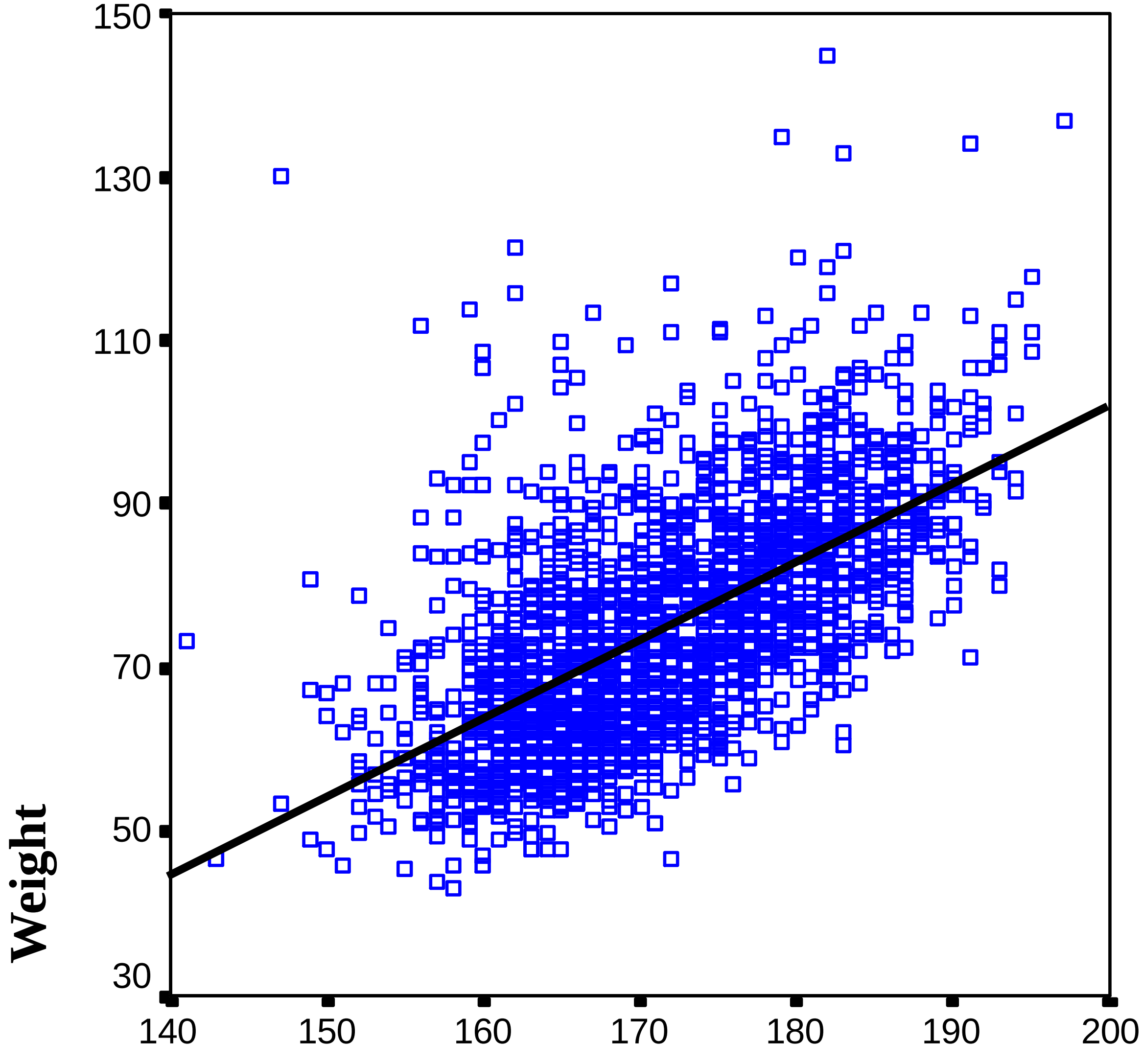
- Assumptions
  - ϵ_i
    - $\epsilon_i \sim N(0, \sigma^2)$
    - σ^2 constant for all i (homoscedasticity)
    - ϵ_i independent of ϵ_j if $i \neq j$ (between subjects)
  - y_i
    - Relationship between y_i and the x_i is linear
    - $y_i \sim N(\mu_i, \sigma^2)$ (because of ϵ_i)
    - $\mu_i = \beta_0 + \beta_1 x_i$

Linear regression

- Assumptions
  - x_i
    - Treated as a constant (number)
      - No measurment error
    - Can have any distribution
  - β's
    - β_0 is where the regression line intersects the y-axis (when $x_i = 0$)
    - If x_i changes 1 unit, y_i changes β_1 units

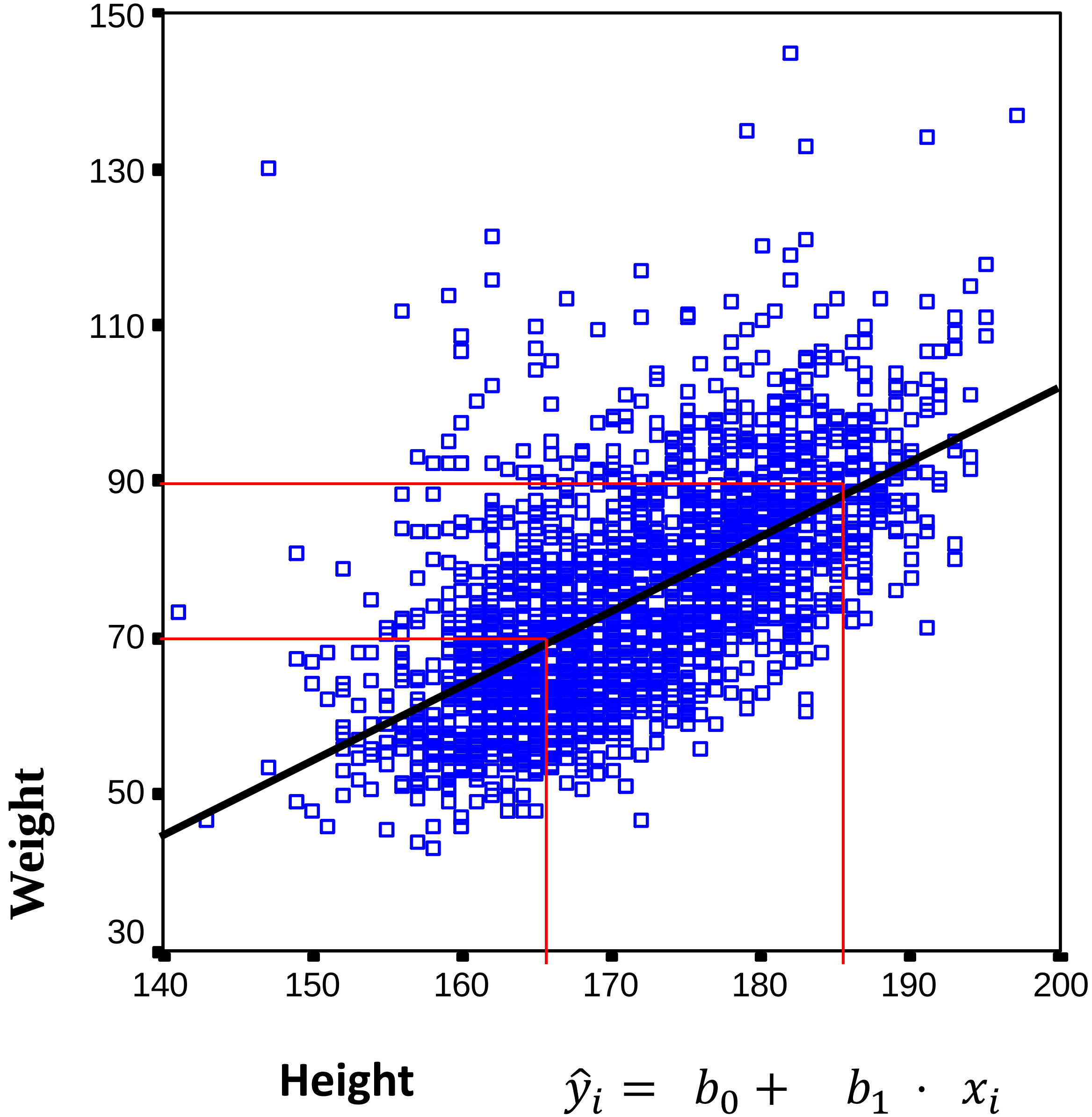
Linear regression

- β's are unknown
- ϵ's are unknown
- Estimated line
  - $\hat{y}_i = b_0 + b_1 x_i$ or
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$



$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

r = 0.62 **WEIGHT = -97 + 1.00 · HEIGHT**



$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

r = 0.62 **WEIGHT = -97 + 1.00 · HEIGHT**

Linear regression

- β's are unknown
- ϵ's are unknown
- Estimated line
  - $\hat{y}_i = b_0 + b_1 x_i$
- Estimated error term
  - $\hat{\epsilon}_i = y_i - \hat{y}_i$

Linear regression

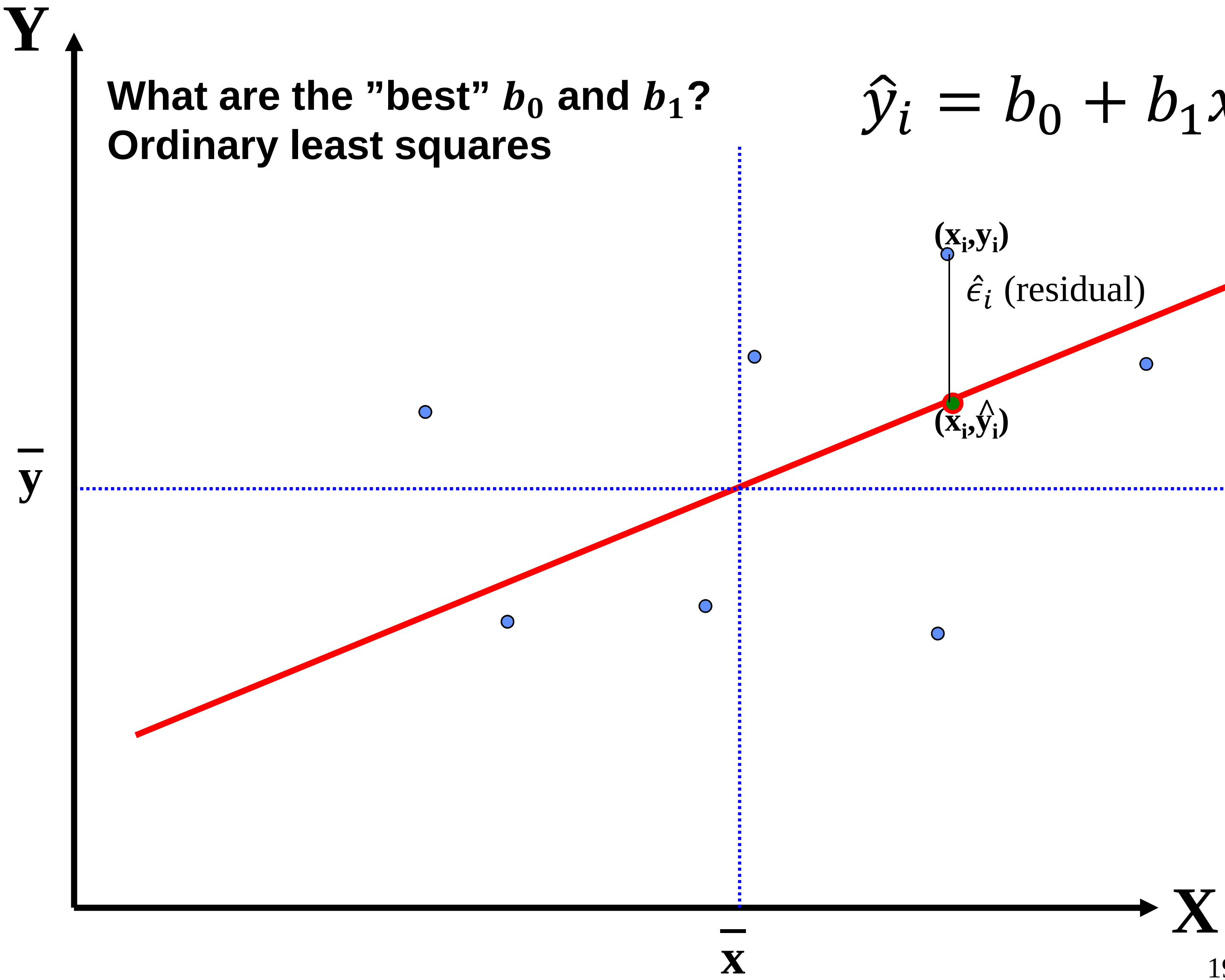
- Ordinary least squares (OLS)
- Minimize the square sum of $\hat{\epsilon}_i$
  - $\hat{\epsilon}_i = y_i - \hat{y}_i$

    $= y_i - (b_0 + b_1 x_i)$
- SSE=$\sum_i \hat{\epsilon}_i^2 = \sum_i (y_i - \hat{y}_i)^2$

  $= \sum_i (y_i - b_0 - b_1 x_i)^2$

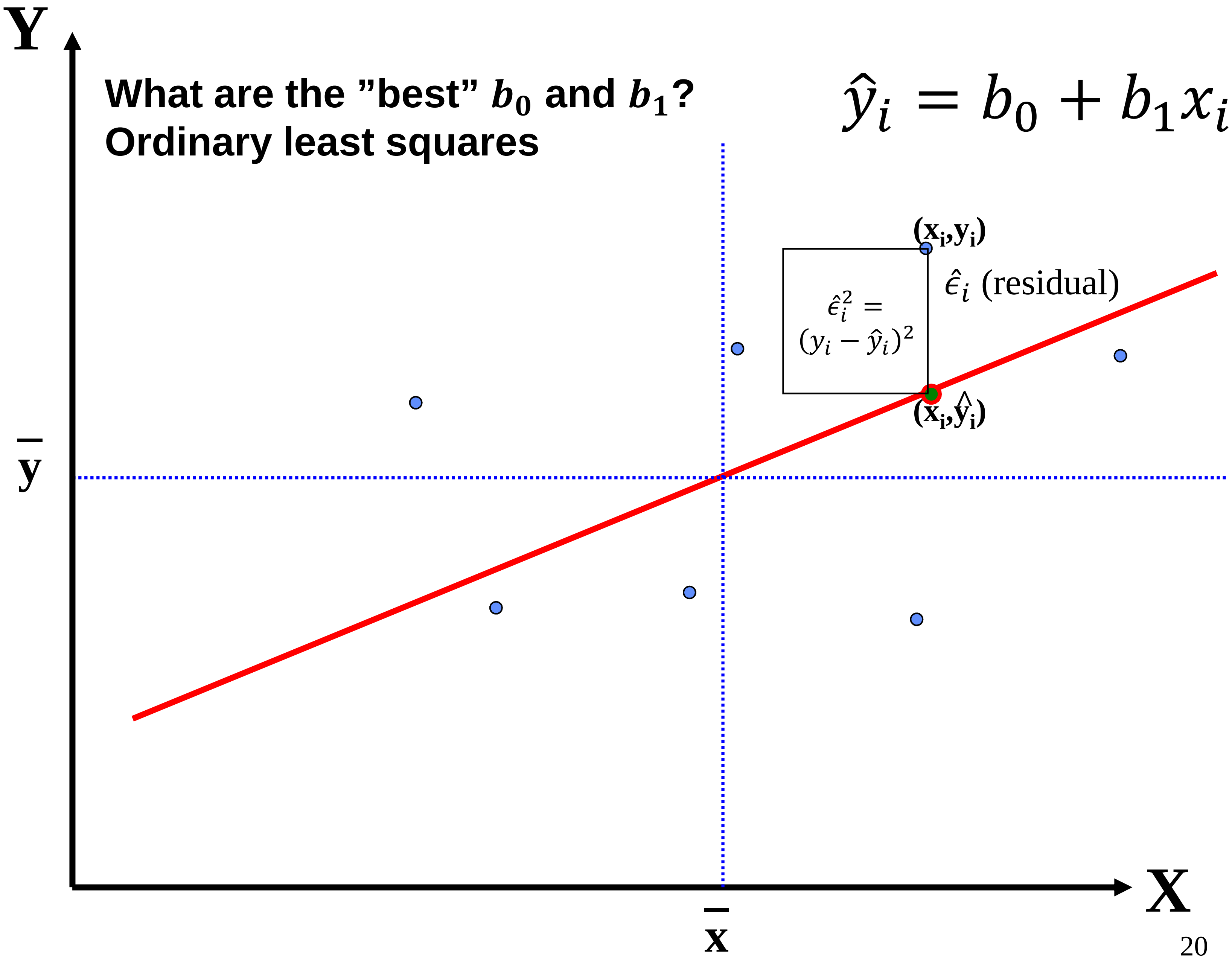
What are the "best" b_0 and b_1?
Ordinary least squares

$$\hat{y}_i = b_0 + b_1 x_i$$



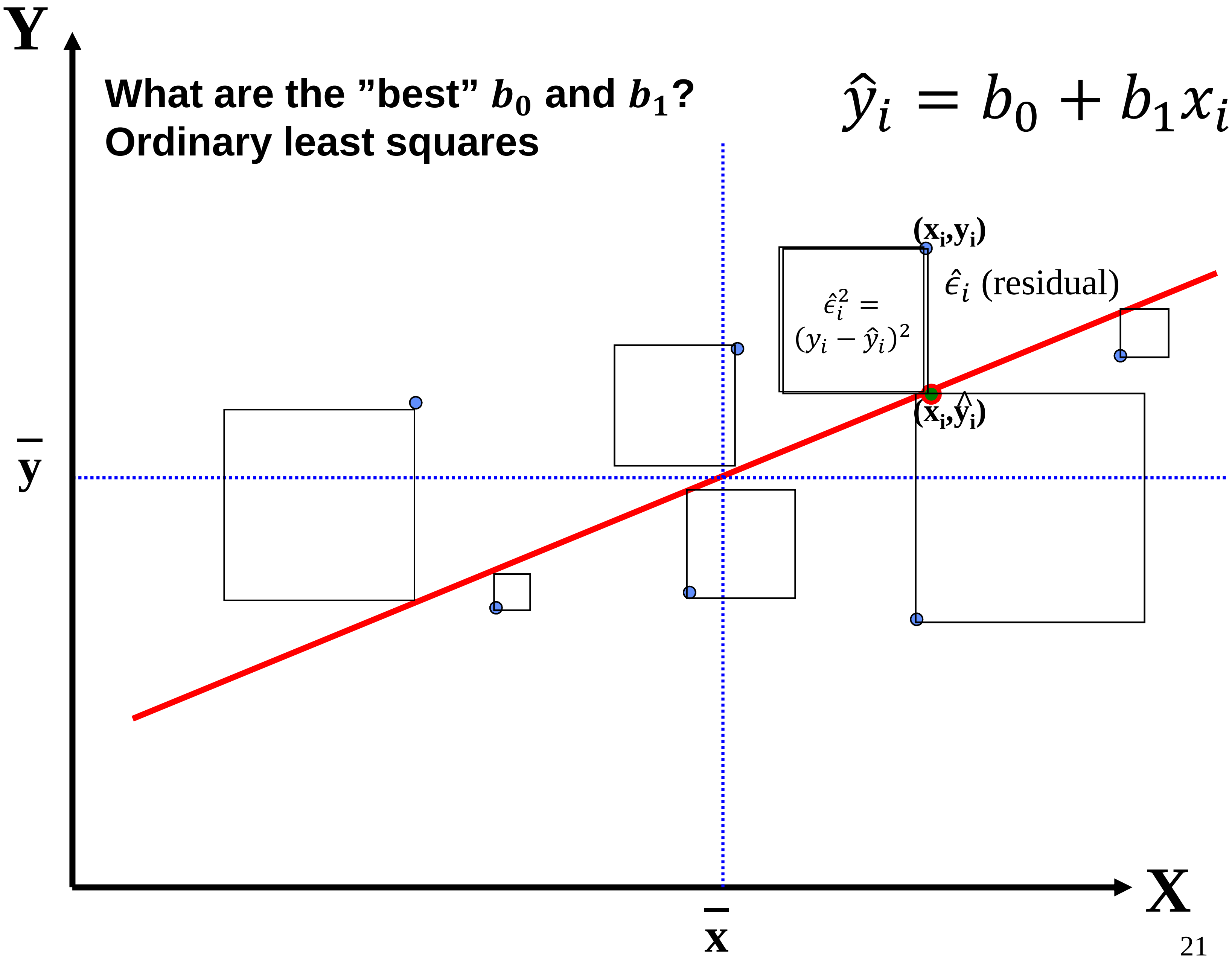
What are the ”best” b_0 and b_1?
Ordinary least squares

$$\hat{y}_i = b_0 + b_1 x_i$$



What are the "best" b_0 and b_1?
Ordinary least squares

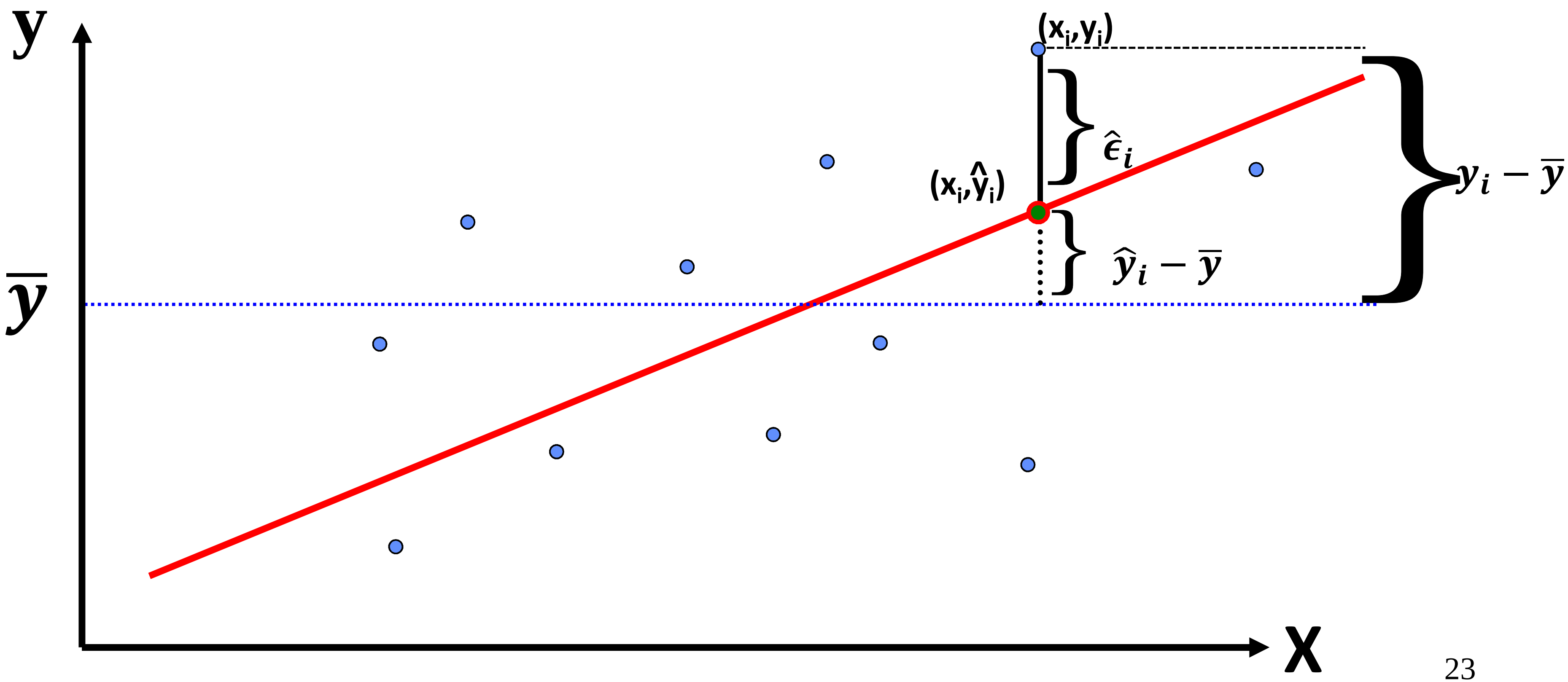
$$\hat{y}_i = b_0 + b_1 x_i$$



Linear regression

- Ordinary least squares (OLS)
- How to calculate coefficients
  - $b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{SS_{XY}}{SS_{XX}}$
  - $b_0 = \bar{y} - b_1 \bar{x}$
  - $\bar{x}$=mean x
  - $\bar{y}$=mean y
- Estimators are generally unbiased

How well is the regression line describing the data?



Linear regression

Total variation

$$SS_{YY} = \sum_i (y_i - \bar{y})^2$$

$$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

$$= SSE + SSR$$

$$SS_{Total} = SS_{Error} + SS_{Regression}$$

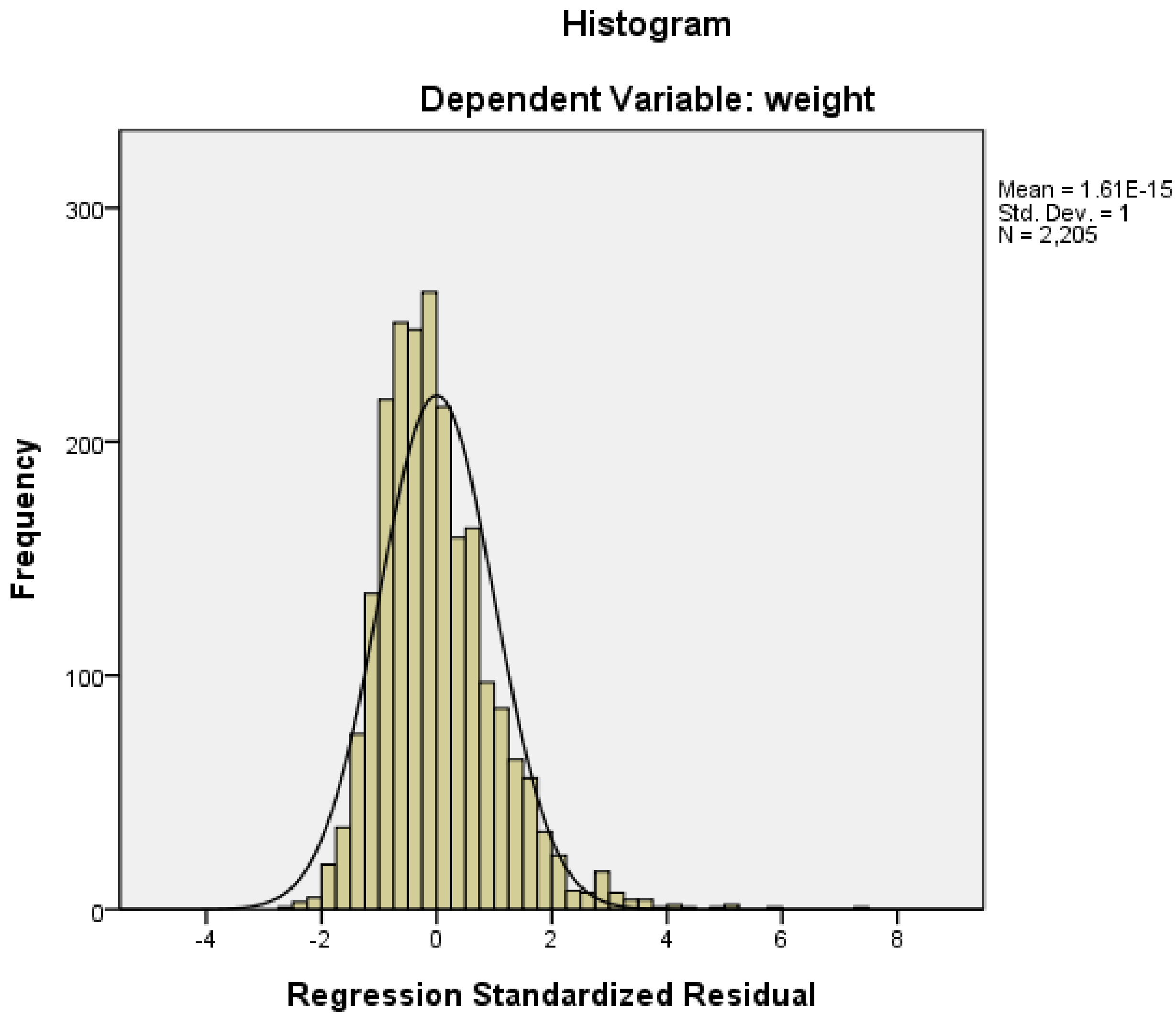
Linear regression

- Percentage of total variation explained by regression line
- $R^2 = \frac{SSR}{SS_{YY}}$
- R is correlation coefficient between x and y

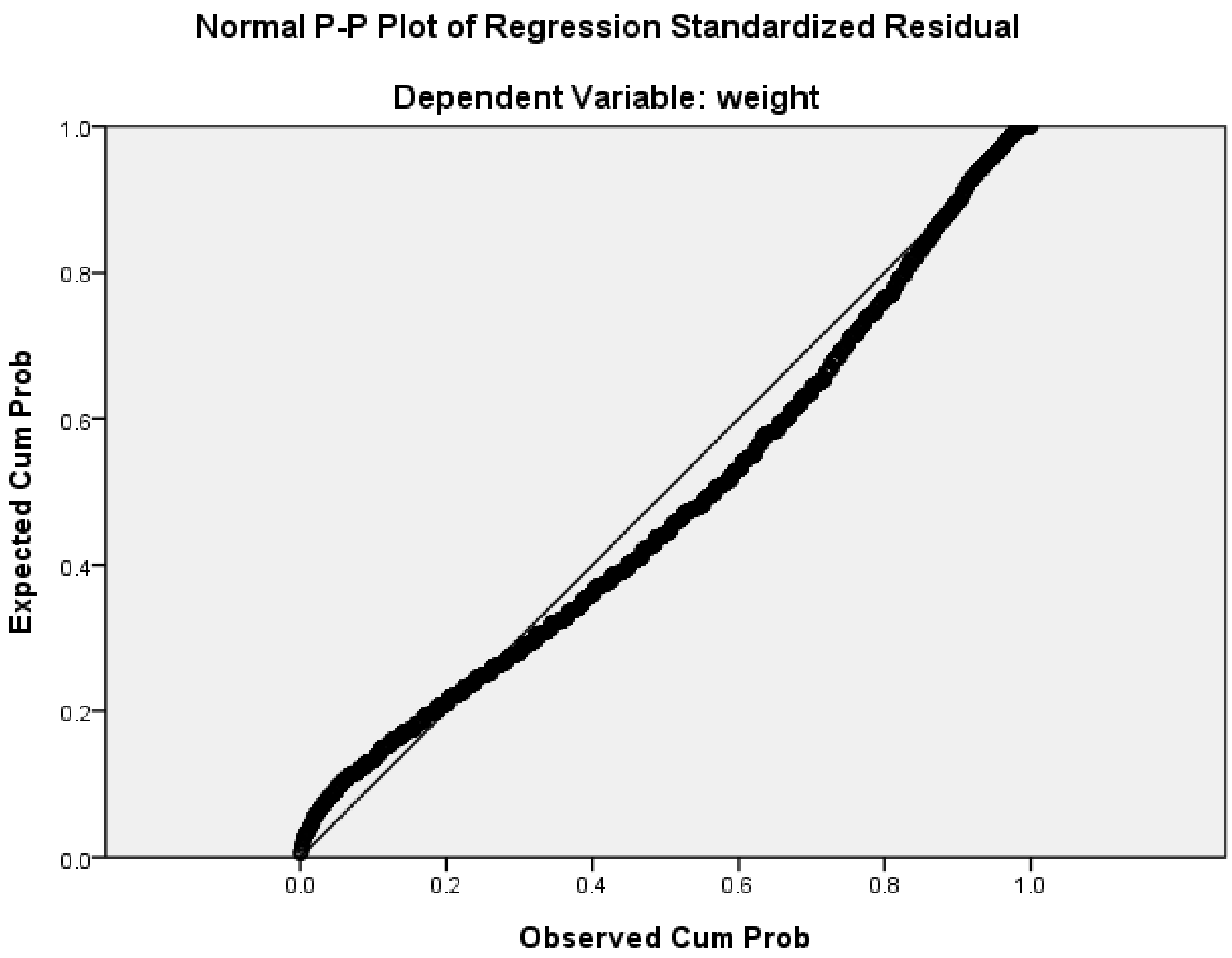
Example of STATA output

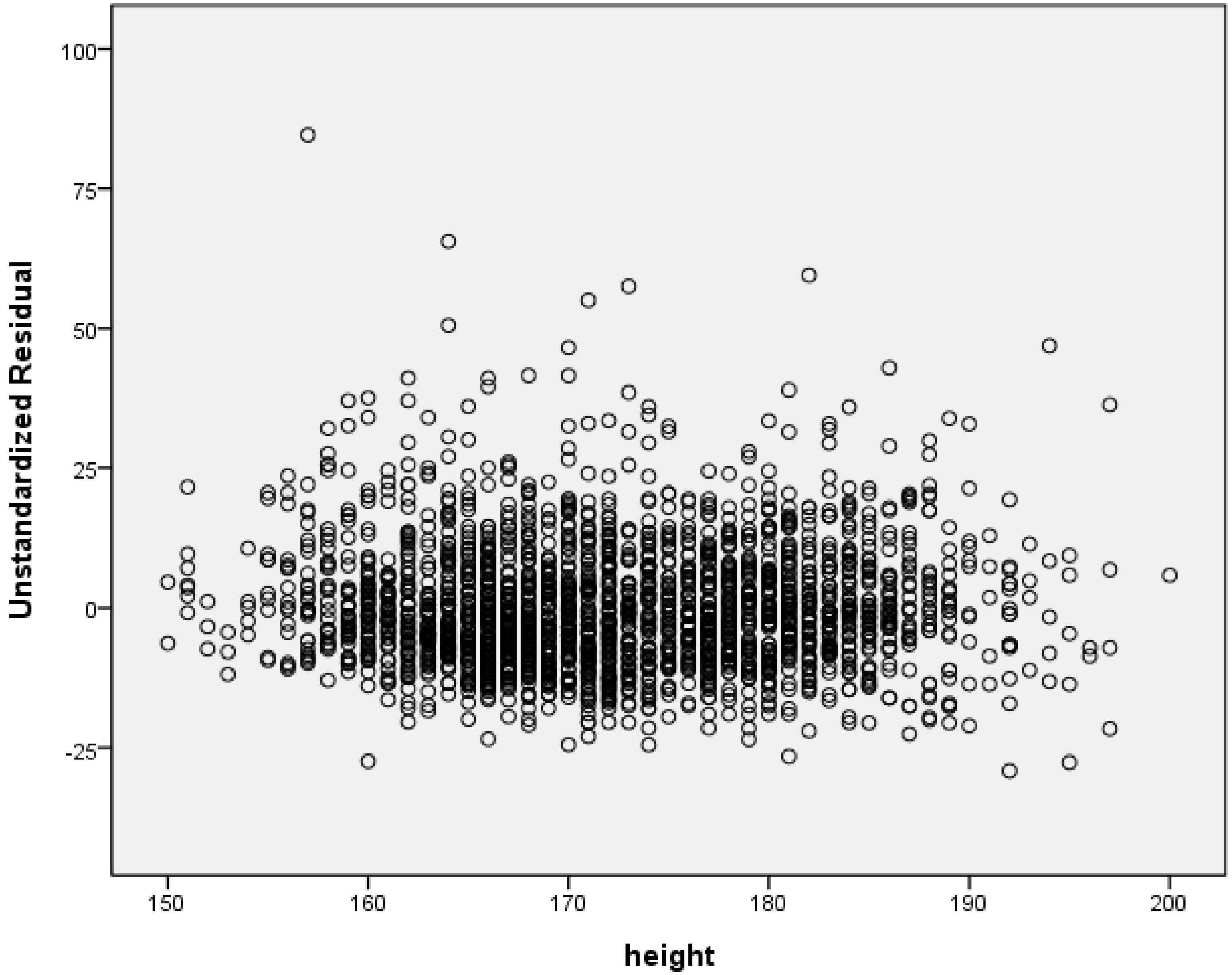
SS R SS E SS YY

```
. regress weight height, beta

      Source |       SS       df       MS              Number of obs =   22176
-------------+------------------------------           F(  1, 22174) =14427.86
       Model |  1802798.66     1  1802798.66           Prob > F      =  0.0000
    Residual |  2770699.56 22174  124.952627           R-squared     =  0.3942
-------------+------------------------------           Adj R-squared =  0.3942
       Total |  4573498.21 22175  206.245692           Root MSE      =  11.178

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      height |   1.006281   .0083776   120.12   0.000                 .6278405
       _cons |    -97.609   1.443956   -67.60   0.000                        .
------------------------------------------------------------------------------
```



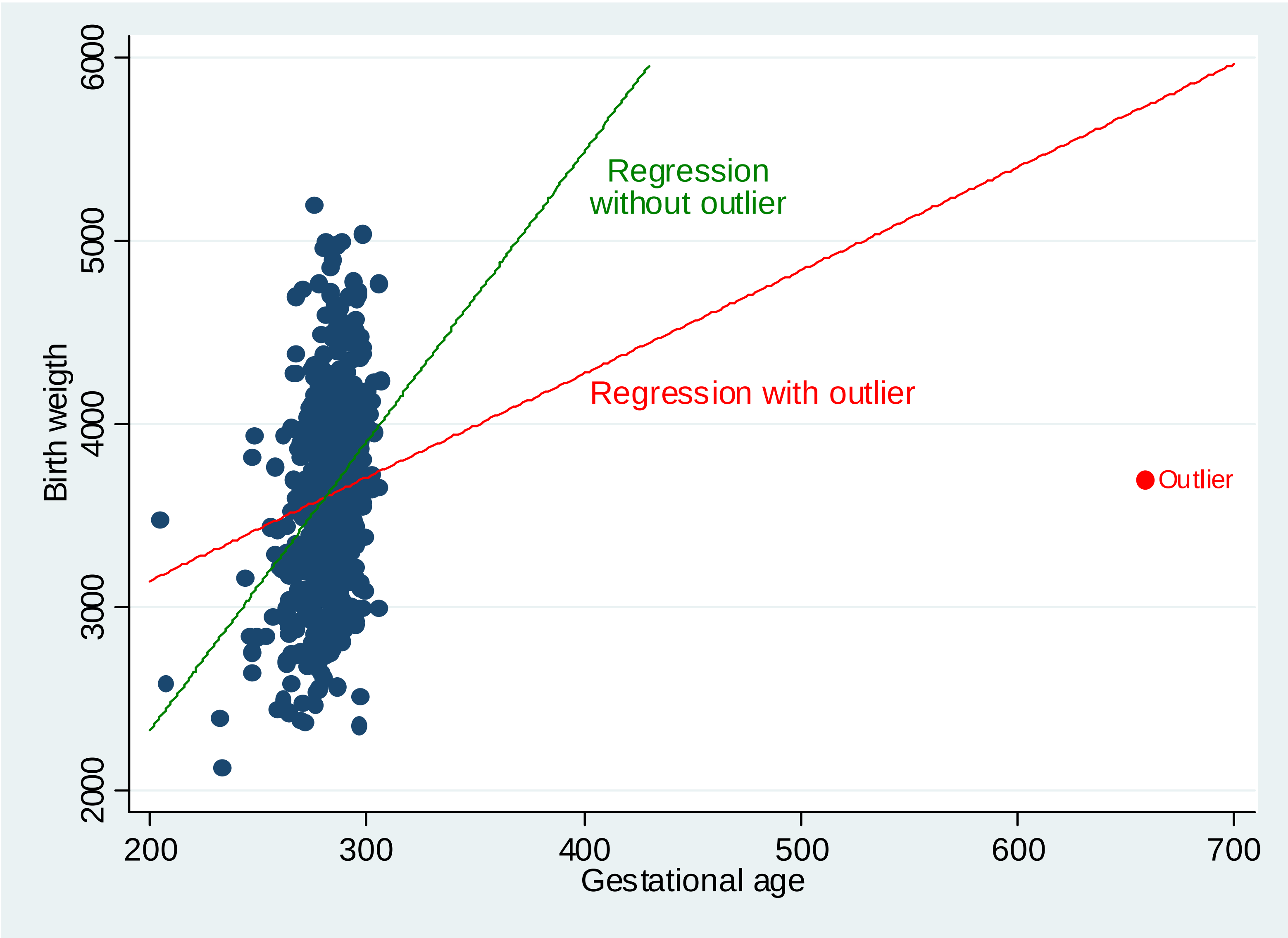
Histogram
Dependent Variable: weight
Mean = 1.61E-15
Std. Dev. = 1
N = 2,205
Frequency
300
200
100
0
-4
-2
0
2
4
6
8
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: weight
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0



Unstandardized Residual
100
75
50
25
0
-25
150
160
170
180
190
200
height

Check for outliers



Linear regression

- Test of coefficients
- We want to test if $\beta_1 = 0$.
  - H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$
- Why $\beta_1 = 0$?

Linear regression

- Test of coefficients
- We want to test if $\beta_1 = 0$.
  - H_0: $\beta_1 = 0$ $\quad\quad\quad$ H_1: $\beta_1 \neq 0$
- Construct t-test
  - Find standard error of b_1
  - $SE(b_1) = \frac{s_{res}}{\sqrt{SS_{XX}}}$
  - $s_{res} = \sqrt{\frac{SSE}{n-2}}$

Linear regression

- Test of coefficients
- We want to test if $\beta_1 = 0$.
  - $H_0\colon \beta_1 = 0$ $\qquad\qquad$ $H_1\colon \beta_1 \neq 0$
- Construct t-test
  - $t = \frac{b_1}{SE(b_1)}$
  - t-distribution with n-2 degrees of freedom
  - $t \sim t(n-2)$

Linear regression

- Test of coefficients
- We want to test if $\beta_1 = 0$.
  - $H_0\colon \beta_1 = 0$ $\qquad$ $H_1\colon \beta_1 \neq 0$
- Construct t-test

	weight	Coef.	Std. Err.	t	P>|t|	Beta
β_1	height	1.006281	$\boxed{.0083776}$	$\boxed{120.12}$	$\boxed{0.000}$	.6278405
	_cons	-97.609	1.443956	-67.60	0.000	.

Linear regression

- Multivariate linear regression
- More than one independent variable
- Study variable
  - Drug, smoking, folate
- Adjustment variable
  - Sex, age, education

```
. summarize weight height  sex light_activity heavy_activity smoking
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	22177	75.59839	14.36101	34	163.5
height	22181	172.1248	8.960389	140	209
sex	22204	1.540308	.4983838	1	2
light_acti~y	21574	3.205664	.8449124	1	4
heavy_acti~y	21377	2.271507	1.032141	1	4
smoking	22137	.9050459	.7977201	0	2

```
. pwcorr weight height sex light_activity heavy_activity smoking, sig
```

	weight	height	sex	light_~y	heavy_~y	smoking
weight	1.0000					
height	0.6278	1.0000				
	0.0000					
sex	-0.5492	-0.7299	1.0000			
	0.0000	0.0000				
light_acti~y	-0.0807	-0.0196	0.0683	1.0000		
	0.0000	0.0040	0.0000			
heavy_acti~y	0.0168	0.1076	-0.0951	0.3881	1.0000	
	0.0143	0.0000	0.0000	0.0000		
smoking	0.0364	0.0167	-0.0145	-0.0205	-0.0031	1.0000
	0.0000	0.0132	0.0309	0.0027	0.6555	

Linear regression

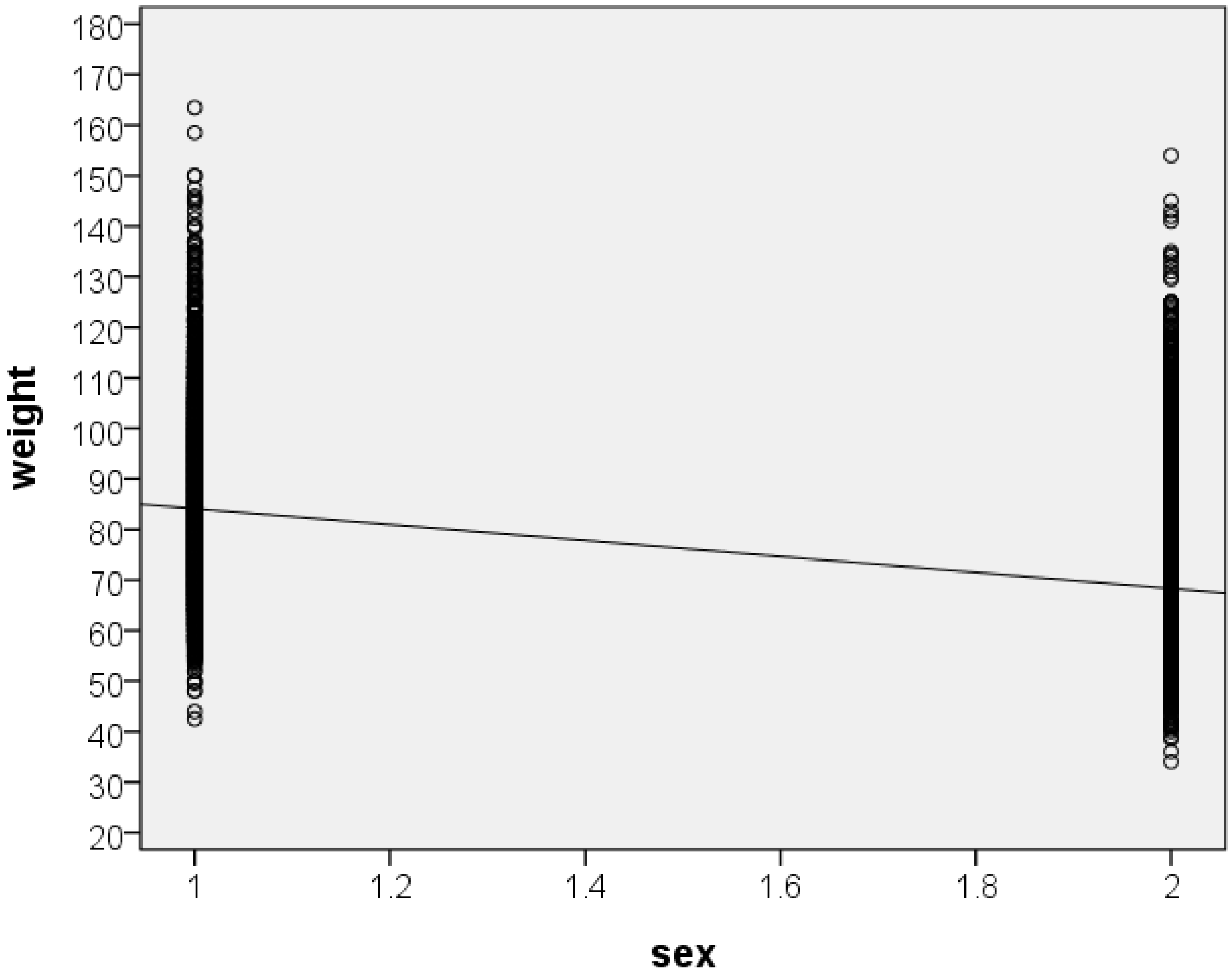
- Multivariate linear regression
  - $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \epsilon_i$
  - y_i is the observed value of subject i
  - x_{ki} is the k'th observation of subject i
    - There are K observations per subject
  - β's are regression coefficients
  - ϵ_i is error (as before)

Linear regression

- New assumptions
  - y_i
    - Relationship between y_i and ALL x_{ki} is linear
  - β's
    - If x_{ki} changes 1 unit, y_i changes β_k units
  - x_k
    - No multicollinearity
      - When x_k is highly correlated with x_l if k$\neq l$
      - E.g., birth weight and gestational age

Linear regression

- β's are unknown
- ϵ's are unknown
- Estimated line
  - $\hat{y}_i = b_0 + b_1 x_{1i} + \cdots + b_{Ki} x_K$
- Similar approach as with univariate linear regression


weight
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
1
1.2
1.4
1.6
1.8
2
sex

```
. regress weight sex, beta

      Source |       SS       df       MS              Number of obs =   22177
-------------+------------------------------           F(  1, 22175) = 9577.59
       Model |  1379527.35     1  1379527.35           Prob > F      =  0.0000
    Residual |  3194021.26 22175  144.037035           R-squared     =  0.3016
-------------+------------------------------           Adj R-squared =  0.3016
       Total |   4573548.6 22176  206.238664           Root MSE      =  12.002

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         sex |  -15.82461   .1616981   -97.87   0.000                -.5492101
       _cons |   99.96717   .2617206   381.96   0.000                        .
------------------------------------------------------------------------------


. ttest weight, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |   10203    84.14256    .1208465    12.20669    83.90567    84.37944
       2 |   11974    68.31795    .1080544    11.82393    68.10614    68.52975
---------+--------------------------------------------------------------------
combined |   22177    75.59839    .0964348    14.36101    75.40938    75.78741
---------+--------------------------------------------------------------------
    diff |            15.82461    .1616981                15.50767    16.14155
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  97.8652
Ho: diff = 0                                     degrees of freedom =    22175

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Do men weigh more than women only because they are taller?

```
. regress weight height sex, beta

      Source |       SS       df       MS              Number of obs =   22176
-------------+------------------------------           F(  2, 22173) = 7764.52
       Model |  1883772.28     2  941886.138           Prob > F      =  0.0000
    Residual |  2689725.94 22173  121.306361           R-squared     =  0.4119
-------------+------------------------------           Adj R-squared =  0.4118
       Total |  4573498.21 22175  206.245692           Root MSE      =  11.014

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      height |   .7785727   .0120753    64.48   0.000                 .4857686
         sex |  -5.608661   .2170847   -25.84   0.000                 -.194652
       _cons |  -49.77759   2.334861   -21.32   0.000                        .
------------------------------------------------------------------------------
```

Do men weigh more than women only because they are taller?

```
. regress weight height sex, beta

      Source |       SS       df       MS              Number of obs =   22176
-------------+------------------------------           F(  2, 22173) = 7764.52
       Model |  1883772.28     2  941886.138           Prob > F      =  0.0000
    Residual |  2689725.94 22173  121.306361           R-squared     =  0.4119
-------------+------------------------------           Adj R-squared =  0.4118
       Total |  4573498.21 22175  206.245692           Root MSE      =  11.014

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      height |   .7785727   .0120753    64.48   0.000                 .4857686
         sex |  -5.608661   .2170847   -25.84   0.000                 -.194652
       _cons |  -49.77759   2.334861   -21.32   0.000                        .
------------------------------------------------------------------------------
```

Regression analyses including height, sex, and light and heavy physical leisure time activity as independent variables

```
. regress weight height sex  light_activity  heavy_activity, beta

      Source |       SS       df       MS              Number of obs =   20979
-------------+------------------------------           F(  4, 20974) = 3761.26
       Model |  1811930.39     4  452982.597           Prob > F      =  0.0000
    Residual |   2525980.3 20974  120.433885           R-squared     =  0.4177
-------------+------------------------------           Adj R-squared =  0.4176
       Total |  4337910.69 20978  206.783806           Root MSE      =  10.974

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      height |    .790329   .0124467    63.50   0.000                 .4923801
         sex |   -5.49161   .2242904   -24.48   0.000                -.1903996
light_acti~y |  -.7872808   .0980221    -8.03   0.000                -.0462277
heavy_acti~y |  -.5068424   .0808189    -6.27   0.000                -.0361951
       _cons |  -48.29099   2.408748   -20.05   0.000                        .
------------------------------------------------------------------------------
```

Linear regression

- Prediction
- Weight=-48.3+0.79*height-5.5*sex-0.79*light activity-0.51*heavy activity
- Example
  - Woman of 165cm (sex=2)
  - light physical activity 1-2 times a week (light activity=3)
  - heavy physical activity 1-2 times a week (heavy activity=3)
- What is her predicted weight?

Linear regression

- Prediction
- Weight=-48.3+0.78*height-5.5*sex-0.79*light activity-0.51*heavy activity
- Example
  - Woman of 165cm (sex=2)
  - light physical activity 1-2 times a week (light activity=3)
  - heavy physical activity 1-2 times a week (heavy activity=3)
- What is her predicted weight?
- Weight = -48.3+0.79*165-5.5*2-0.79*3-0.51*3 = 67.15kg

What about smoking?

```
. mean weight, over(smoking)

Mean estimation                    Number of obs    =    22110

    _subpop_1: smoking = Never-smoker
    _subpop_2: smoking = Current-smoker
    _subpop_3: smoking = Ex-smoker

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
weight       |
   _subpop_1 |   75.84302   .1605173      75.52839    76.15764
   _subpop_2 |   73.92477   .1581386      73.61481    74.23474
   _subpop_3 |   77.43938   .1837396      77.07924    77.79952
--------------------------------------------------------------
```

```
. regress weight  smoking, beta

      Source |       SS       df       MS              Number of obs =   22110
-------------+------------------------------           F(  1, 22108) =   29.27
       Model |  6024.52709     1  6024.52709           Prob > F      =  0.0000
    Residual |  4549692.24 22108  205.793932           R-squared     =  0.0013
-------------+------------------------------           Adj R-squared =  0.0013
       Total |  4555716.77 22109  206.057115           Root MSE      =  14.346

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     smoking |   .6544092   .1209495     5.41   0.000                  .036365
       _cons |   75.00955   .1459137   514.07   0.000                        .
------------------------------------------------------------------------------
```

Creating indicator variables

Smoking	Never	Current	Ex
	0	1	2
Index _var_1	0	1	0
Index _var_2	0	0	1

Creating indicator variables

```
. regress weight i.smoking, beta
```

Source	SS	df	MS
Model	43077.8065	2	21538.9033
Residual	4512638.96	22107	204.127152
Total	4555716.77	22109	206.057115

Number of obs	=	22110
F(2, 22107)	=	105.52
Prob > F	=	0.0000
R-squared	=	0.0095
Adj R-squared	=	0.0094
Root MSE	=	14.287

weight	Coef.	Std. Err.	t	P>|t|	Beta
smoking					
1	-1.918242	.2257697	-8.50	0.000	-.0639349
2	1.596366	.241861	6.60	0.000	.0496669
_cons	75.84302	.1579407	480.20	0.000	.

```
. regress weight height sex  light_activity  heavy_activity i.smoking, beta

      Source |       SS       df       MS              Number of obs =   20922
-------------+------------------------------           F(  6, 20915) = 2577.27
       Model |  1837625.09     6  306270.848           Prob > F      =  0.0000
    Residual |  2485442.81 20915   118.83542           R-squared     =  0.4251
-------------+------------------------------           Adj R-squared =  0.4249
       Total |   4323067.9 20921  206.637727           Root MSE      =  10.901

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      height |   .7821648   .0123927    63.11   0.000                 .4874925
         sex |  -5.592277   .2232835   -25.05   0.000                -.1939631
light_acti~y |  -.8499602   .0976794    -8.70   0.000                -.0499033
heavy_acti~y |  -.6278786   .0807802    -7.77   0.000                -.0448362
             |
     smoking |
          1  |  -2.065611   .1782763   -11.59   0.000                -.0687006
          2  |   1.103528   .1897556     5.82   0.000                 .0342332
             |
       _cons |  -45.82763   2.403266   -19.07   0.000                        .
------------------------------------------------------------------------------
```

Linear regression

- Model selection
  - backward selection:
    - Exclude exposure variable with largest p-value in multivariate analysis and re-estimate
    - repeat this until all terms are significant.
  - forward selection:
    - Include exposure variable with lowest p-value in univariate analysis and re-estimate
    - repeat this until all terms are no longer significant

Linear regression

- Model selection
  - "Change in estimate"
    - Adjusting for confounding
    - Starting model: One study variable
    - Enter independent variables if they change the coefficient (effect estimate) of the study variable
    - "Change" can be defined as, e.g., 10%

Linear regression

- Model selection
  - Akaike information criterion (AIC)
  - Model which is “closest” to the data
    - regress [dep] [ind1] … [indK]
    - estat ic
  - Low AIC is good.
    - Select among models with the lowest AIC
      - A difference of less than 2 is small
      - A difference of 4-7 is large
      - A difference of more than 10 is huge
    - Akaike weights (advanced!)

Linear regression

- Model selection
  - WARNING: NONE OF THESE METHODS ARE VERY GOOD!
  - Best to use other information than statistical selection methods.
    - What is known about the relationships between your variables?
  - Selection probably will affect effect estimates and p-values.
  - Always run different approaches and compare