Data Analysis:
Statistical Modeling and Computation in Applications

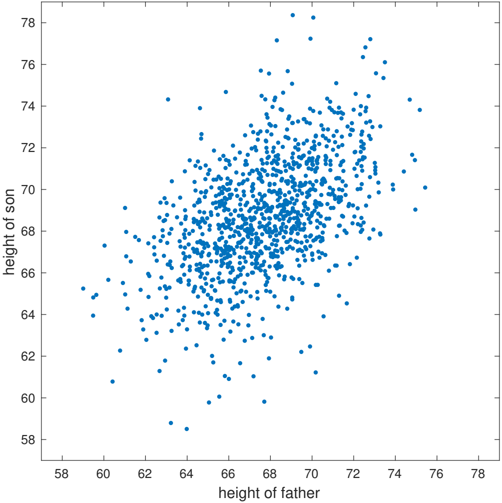
Correlation and Least Squares Regression

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

Scatter diagram: height of 1078 fathers and their sons

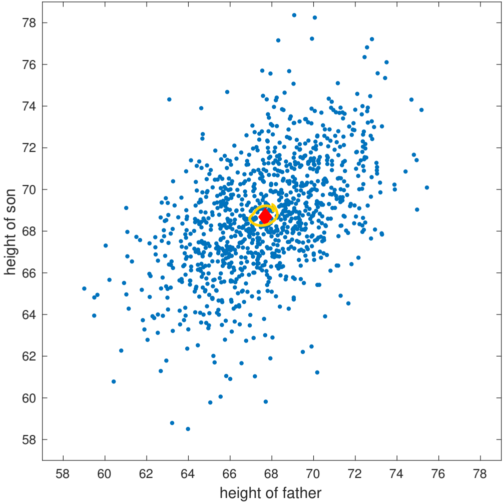
Is there an association?
What kind?



Data: Pearson K and Lee A. (1903). On the laws of inheritance in man. Biometrika, 2:357-462.
Downloaded from `https://myweb.uiowa.edu/pbreheny/data/pearson.html`

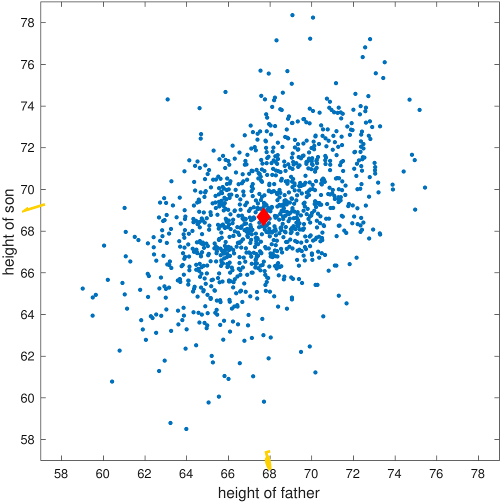
Summarizing the Plot

- average $\bar{x}$, $\bar{y}$



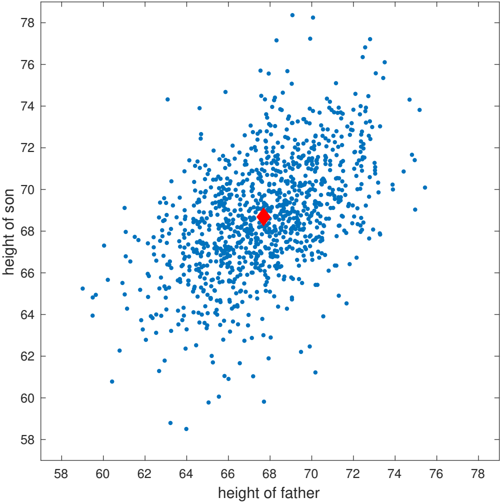
Summarizing the Plot

- average $\bar{x}$, $\bar{y}$
  fathers: $\bar{x} \approx 68$,
  sons: $\bar{y} \approx 69$



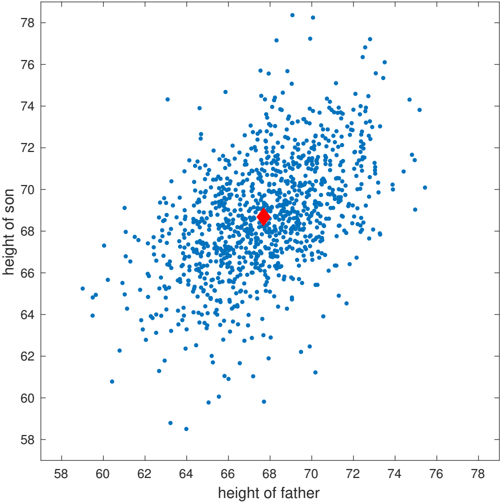
Summarizing the Plot

- average $\bar{x}$, $\bar{y}$
  fathers: $\bar{x} \approx 68$,
  sons: $\bar{y} \approx 69$
- standard deviation
  $s_x = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  here: $s_x \approx s_y \approx 2.7$



Summarizing the Plot

- average $\bar{x}$, $\bar{y}$
  fathers: $\bar{x} \approx 68$,
  sons: $\bar{y} \approx 69$
- standard deviation
  $s_x = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  here: $s_x \approx s_y \approx 2.7$
- correlation coefficient
  $r \approx 0.5$



Correlation Coefficient

$$r = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

(convert to standard units and take average product)

Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

(convert to standard units and take average product)

1. symmetric

Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

(convert to standard units and take average product)

1. symmetric
2. Why standard units?

Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

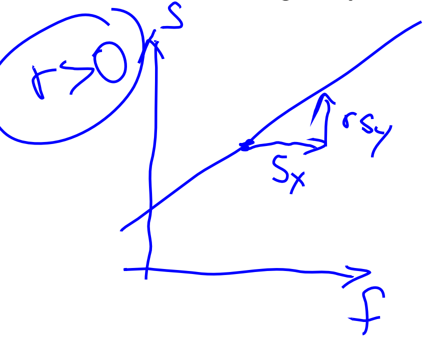
(convert to standard units and take average product)

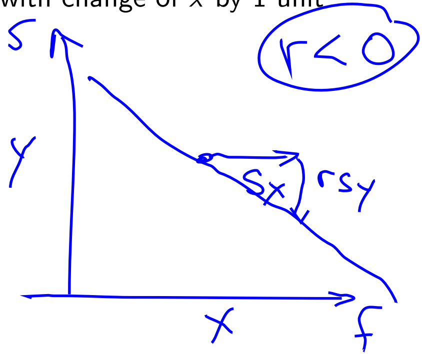
1. symmetric
2. Why standard units?
   adding or multiplying constants to all x_i or y_i does not change r
3. What does $r \approx 0.5$ mean?

What does the Correlation coefficient mean? (1)

$$r = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- measures *linear* association between variables:
  how much change of y is associated with change of x by 1 unit

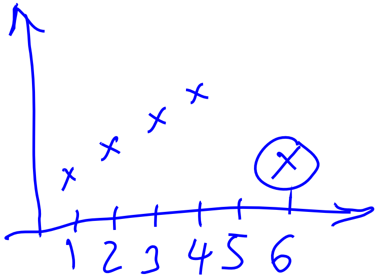




What does the Correlation coefficient mean? (1)

$$r = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

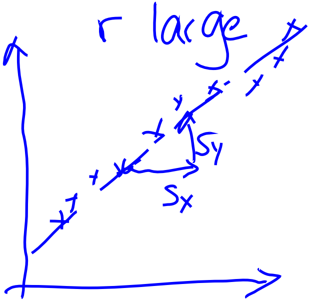
- measures *linear* association between variables:
  how much change of y is associated with change of x by 1 unit

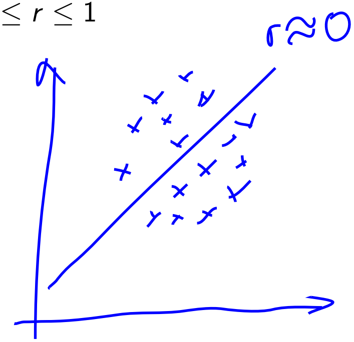


What does the Correlation coefficient mean? (2)

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- measures *clusteredness* along a line: $-1 \le r \le 1$
  sign?



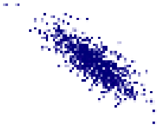


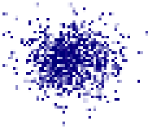
Examples

1.

2.

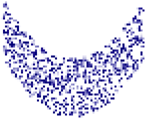
3.





4.

5.



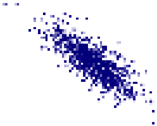
6.

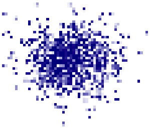
Examples

1. $r = 1$

2.

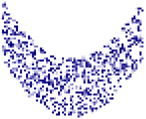
3.





4.

5.



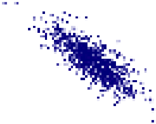
6.

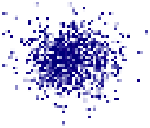
Examples

1. $r = 1$

2. $r = -1$

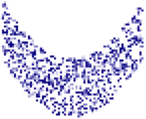
3.





4.

5.



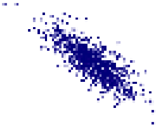
6.

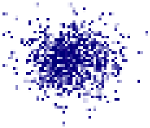
Examples

1. $r = 1$

2. $r = -1$

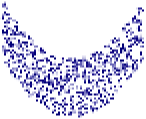
3.





4. $r = 0$

5.



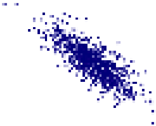
6.

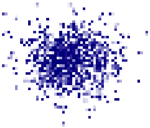
Examples

1. $r = 1$

2. $r = -1$

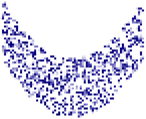
3. $r = -0.8$





4. $r = 0$

5.



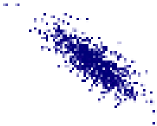
6.

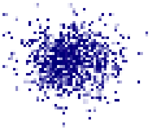
Examples

1. $r = 1$

2. $r = -1$

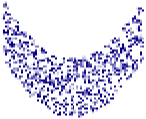
3. $r = -0.8$





4. $r = 0$

5. $r = 0$



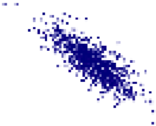
6.

Examples

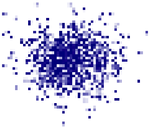
1. $r = 1$

2. $r = -1$

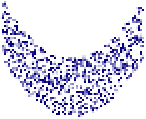
3. $r = -0.8$



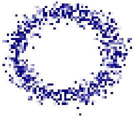
4. $r = 0$



5. $r = 0$



6. $r = 0$



Careful with nonlinearities and outliers!

Correlation coefficient: summary

$$r = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- measures *linear* association between variables:
- measures clusteredness along a line
- symmetric (swapping x and y)
- between -1 and 1, and invariant to
  - adding a constant to all x_i or all y_i
  - multiplying to all x_i (all y_i) by a positive constant

Data Analysis:
Statistical Modeling and Computation in Applications

Correlation and Least Squares Regression
Part 2

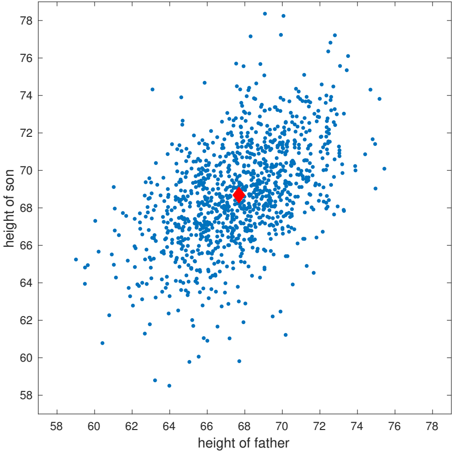
Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

Predicting a son's height from the father's height

- fathers: $\bar{x} \approx 68\text{in}$, $s_x = 2.7\text{in}$
- sons: $\bar{y} \approx 69\text{in}$, $s_y = 2.7\text{in}$
- $r \approx 0.5$

Suggestion: The sons' average is 1 inch more than the fathers' average. So, if the father's height is 64 inches we expect the son's height to be 65 inches.

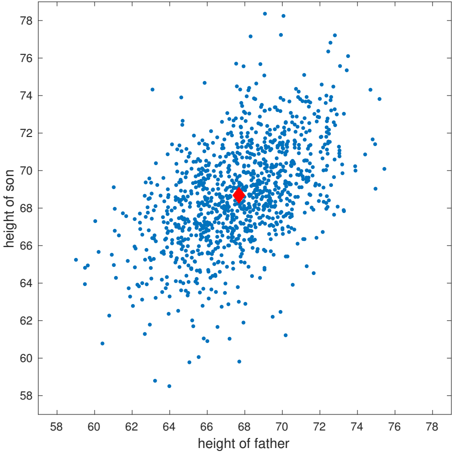


Predicting a son's height from the father's height

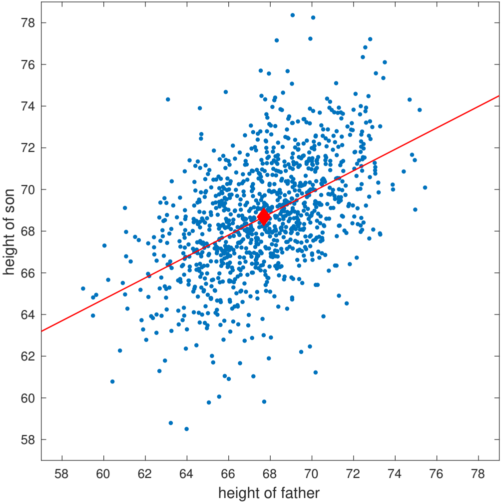
- fathers: $\bar{x} \approx 68\text{in}$, $s_x = 2.7\text{in}$
- sons: $\bar{y} \approx 69\text{in}$, $s_y = 2.7\text{in}$
- $r \approx 0.5$

Suggestion: The sons' average is 1 inch more than the fathers' average. So, if the father's height is 64 inches we expect the son's height to be 65 inches.

No! Correlation Coefficient...

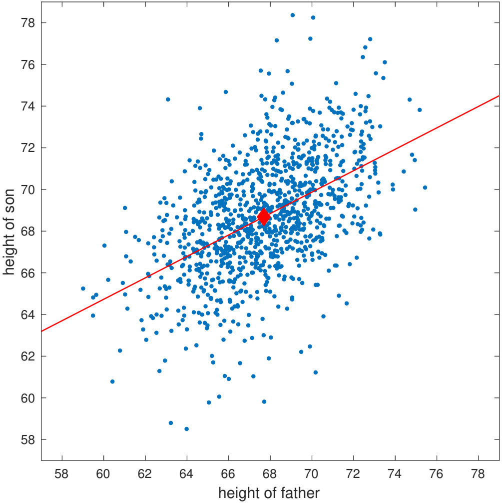


Regression line: what does it mean?



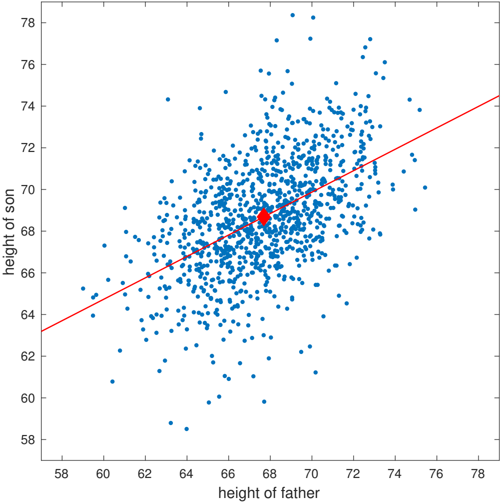
Regression line: what does it mean?

1. Increase of 1 std dev in x associated with increase of r std dev in y.



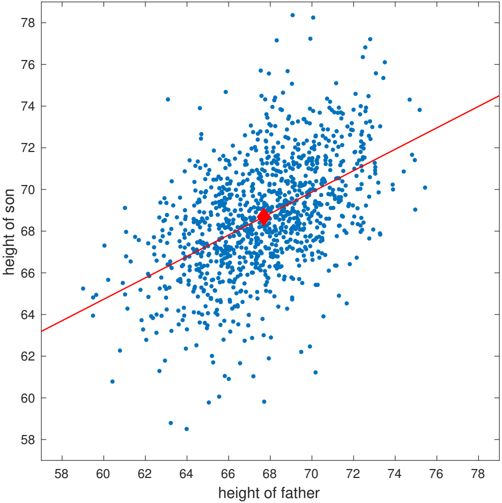
Regression line: what does it mean?

1. Increase of 1 std dev in x associated with increase of r std dev in y.
2. Interpolating conditional averages of y given x



Regression line: what does it mean?

1. Increase of 1 std dev in x associated with increase of r std dev in y.
2. Interpolating conditional averages of y given x
3. Solution to least squares



Regression Line for y on x

model:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

- fit to *minimize RMS error* (Gauss)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i - y_i)^2}$$

Regression Line for y on x

model:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0 \quad \text{with } \hat{\beta}_1 = r\frac{s_y}{s_x}$$

- fit to *minimize RMS error* (Gauss)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i - y_i)^2}$$

- RMS error is $\sqrt{1-r^2}s_y$

Regression Line for y on x

model:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0 \quad \text{with } \hat{\beta}_1 = r\frac{s_y}{s_x}$$

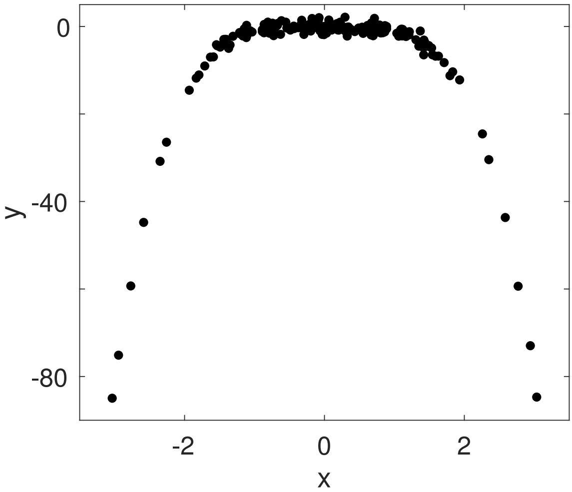
- fit to *minimize RMS error* (Gauss)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i - y_i)^2}$$

- RMS error is $\sqrt{1-r^2}s_y$
- not the same as the regression line of x on y

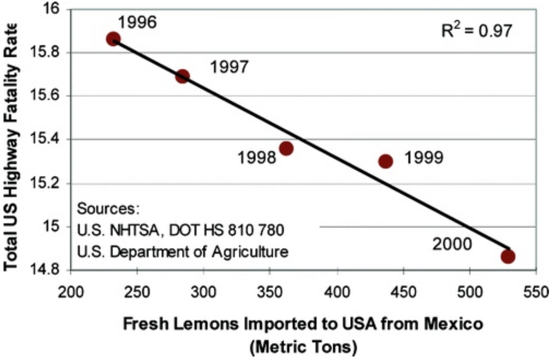
3 words of caution (1)

- Only measures a linear relationship.



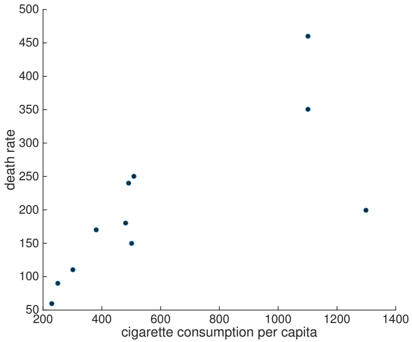
3 words of caution (2)

- Correlation is not equal to causation.



3 words of caution (3)

Country	*Cigarette consumption*	*Deaths per million*
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1,100	350
Great Britain	1,100	460
Iceland	230	60
Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
U.S.	1,300	200

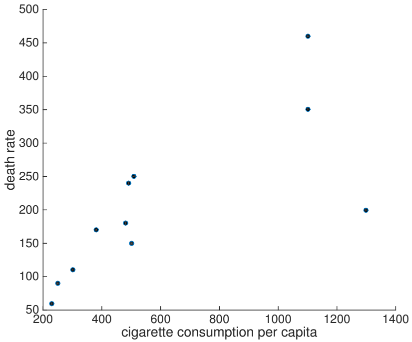


$r \approx 0.74$

(Source: Freedman, Pisani, Purves. Statistics)

3 words of caution (3)

Country	*Cigarette consumption*	*Deaths per million*
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1,100	350
Great Britain	1,100	460
Iceland	230	60
Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
U.S.	1,300	200



$r \approx 0.74$

Ecological correlations tend to overstate the strength of an association for individuals.

(Source: Freedman, Pisani, Purves. Statistics)

Ecological Correlation

Summary: Regression line

- interpolates conditional averages of y given x
- solves least squares problem
- slope: rs_y/s_x
- caution: linear relationship, and not impling causality
- caution: ecological correlations

Data Analysis: Statistical Modeling and Computation in Applications

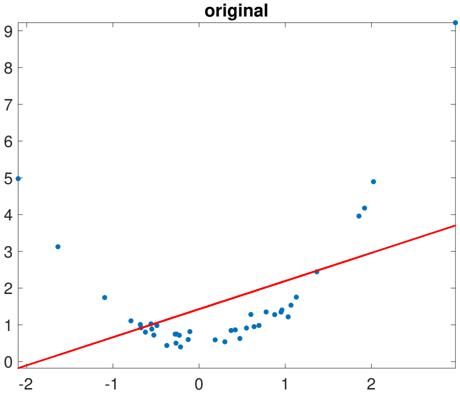
Correlation and Least Squares Regression
Part 3

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

How do we evaluate our regression line?

- We fit a model. Does it make sense?



Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

- should show no pattern (e.g. due to nonlinear association)
- points regularly scattered around 0

Does the model make sense?

Assumptions:

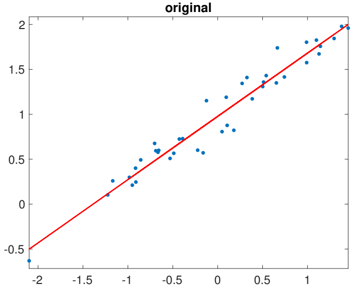
- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

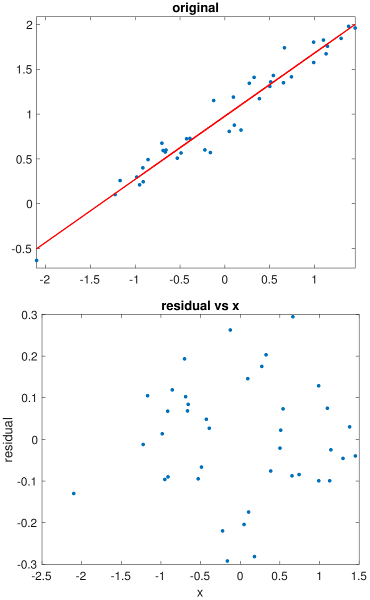
- should show no pattern (e.g. due to nonlinear association)
- points regularly scattered around 0

Variable transformations can help, e.g. $\log(y)$, $\sqrt{y}$, $\sqrt{x}$, $\log(x)$, x^2

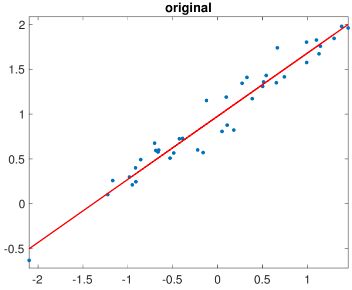
Example 1: assumptions hold

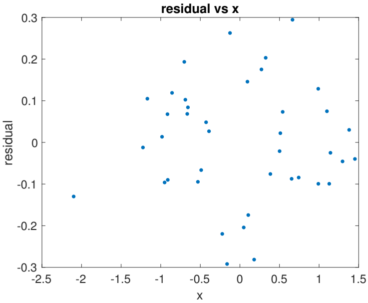


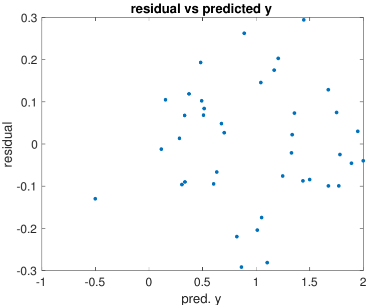
Example 1: assumptions hold



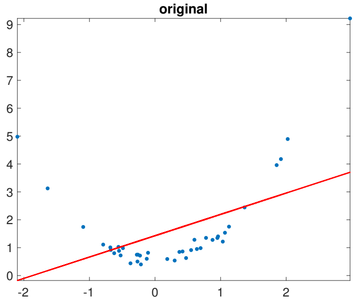
Example 1: assumptions hold



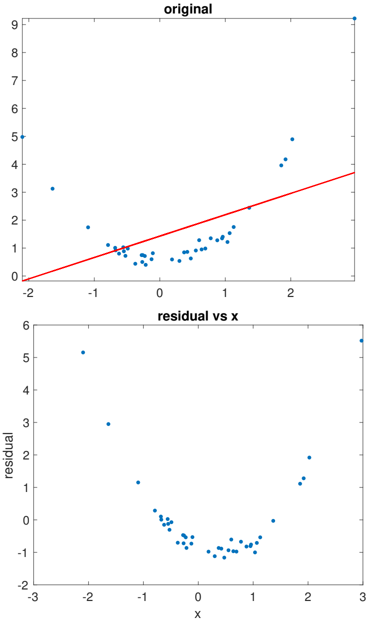




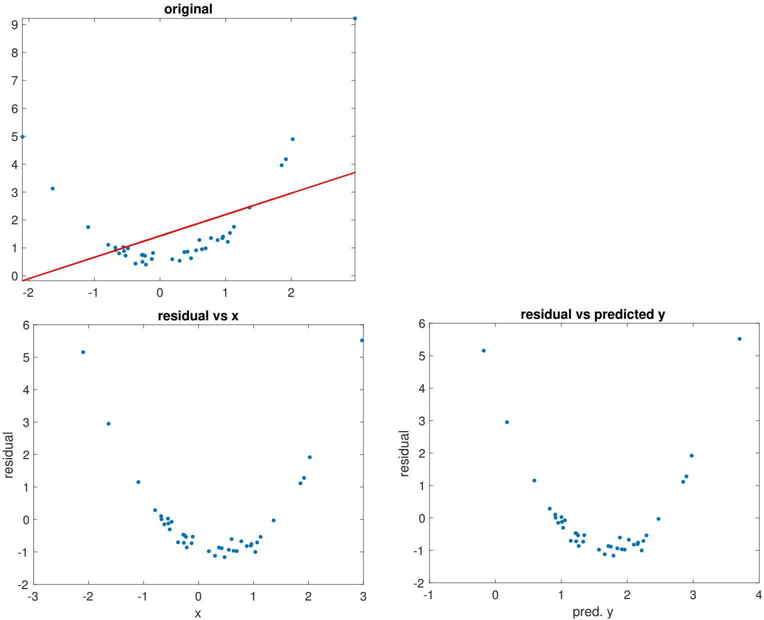
Example 2



Example 2

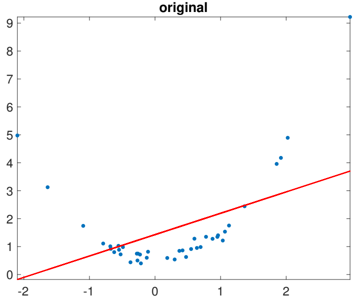

original
9
8
7
6
5
4
3
2
1
0
-2
-1
0
1
2
residual vs x
6
5
4
3
2
1
0
-1
-2
residual
-3
-2
-1
0
1
2
3
x

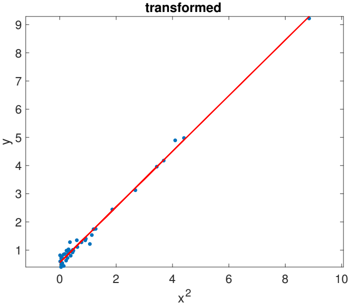
Example 2



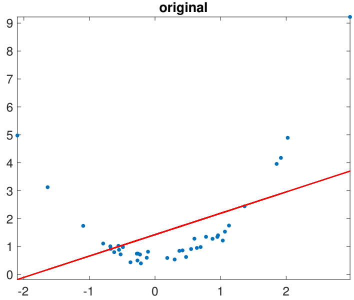
original
9
8
7
6
5
4
3
2
1
0
-2
-1
0
1
2
residual vs x
residual
6
5
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
x
residual vs predicted y
residual
6
5
4
3
2
1
0
-1
-2
-1
0
1
2
3
4
pred. y

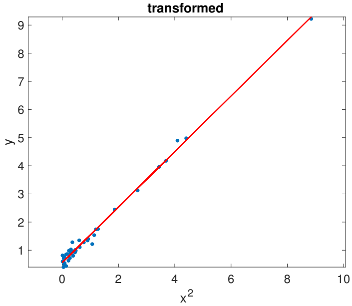
Example 2: transformation x^2

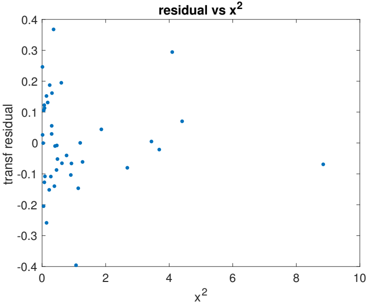




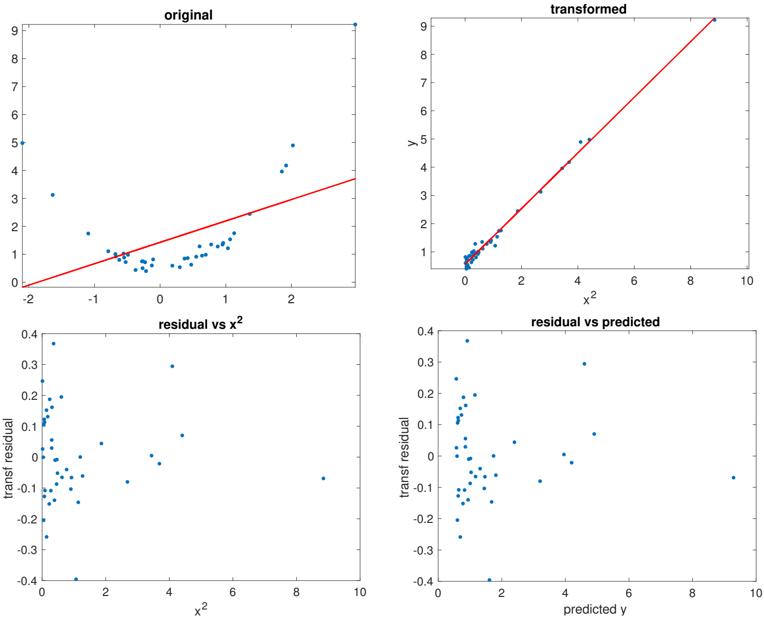
Example 2: transformation x^2



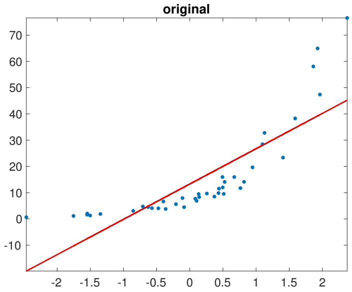




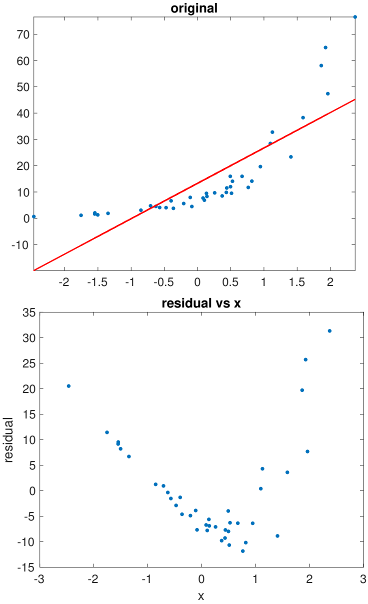
Example 2: transformation x^2



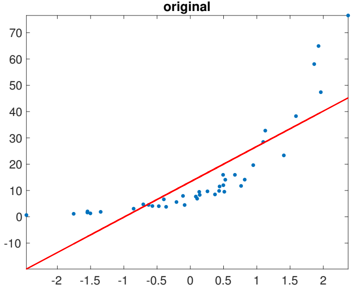
Example 3

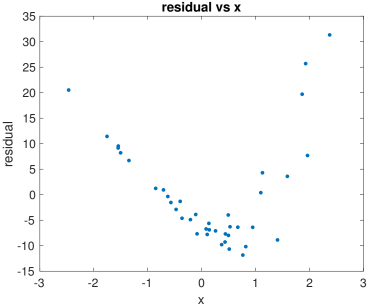


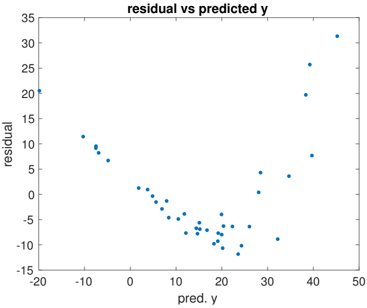
Example 3


original
70
60
50
40
30
20
10
0
-10
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
residual vs x
35
30
25
20
15
10
5
0
-5
-10
-15
residual
-3
-2
-1
0
1
2
3
x

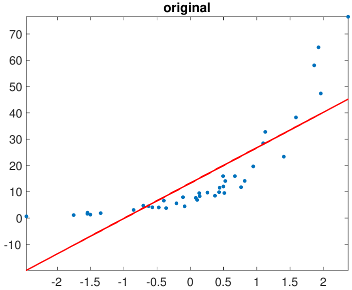
Example 3

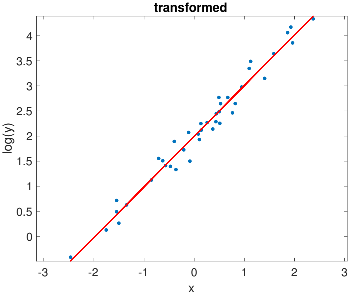

original
70
60
50
40
30
20
10
0
-10
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2


residual vs x
35
30
25
20
15
10
5
0
-5
-10
-15
residual
-3
-2
-1
0
1
2
3
x

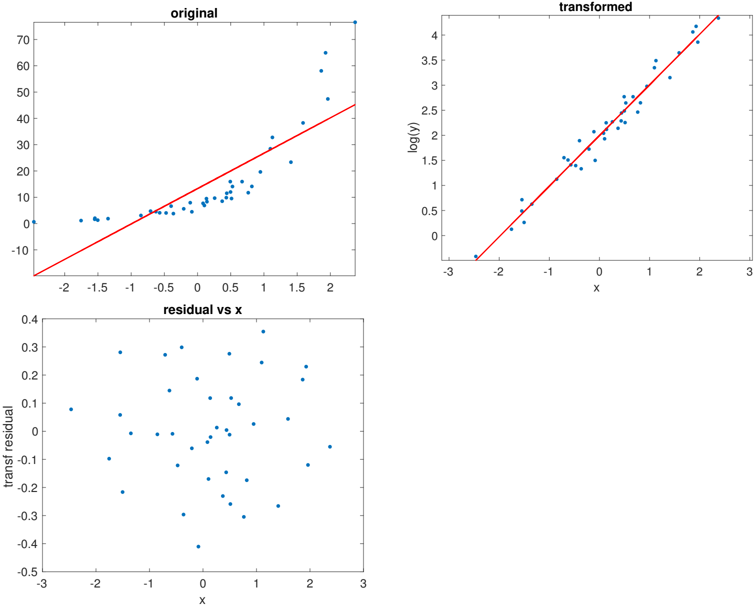

residual vs predicted y
35
30
25
20
15
10
5
0
-5
-10
-15
residual
-20
-10
0
10
20
30
40
50
pred. y

Example 3: transformation log(y)

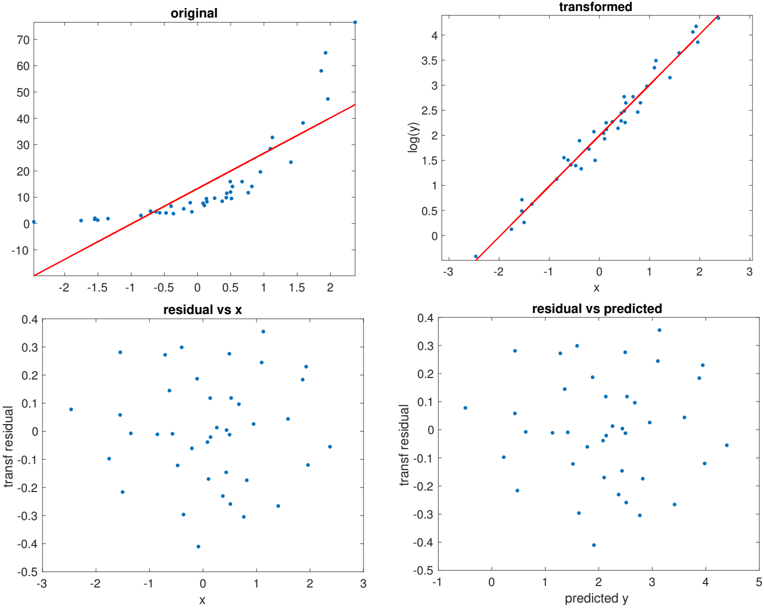




Example 3: transformation $\log(y)$



Example 3: transformation log(y)



Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

- should show no pattern (e.g. due to nonlinear association)
- points regularly scattered around 0

Variable transformations can help, e.g. $\log(y)$, $\sqrt{y}$, $\sqrt{x}$, $\log(x)$, x^2

Data Analysis:
Statistical Modeling and Computation in Applications

Correlation and Least Squares Regression
Part 4

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

Multiple regression

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

y_i x_{i1} x_{i2}

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$

$$41 = \beta_0 + 190\beta_1 + 67\beta_2 + \epsilon_1$$

$$\underbrace{\begin{pmatrix} \boxed{41} \\ 36 \\ 12 \\ 18 \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}}_{\epsilon}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$

$$N \underbrace{\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix}}_{Y} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} \overset{N}{=} \overbrace{\begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix}}^{p} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$
  - $\boldsymbol{\beta}$ parameters: $p \times 1$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \overbrace{\begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix}} \left\{ \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \right. + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$
  - $\boldsymbol{\beta}$ parameters: $p \times 1$
  - ϵ: random error / disturbances
    ϵ_i are iid, $\mathbb{E}[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Examples of multiple regression

- **Simple linear regression**:

$$p = 2,\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \ y_i = \beta_0 + \beta_1 x_1$$

Examples of multiple regression

- **Simple linear regression**:

$$p = 2,\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \ y_i = \beta_0 + \beta_1 x_1$$

- **Quadratic (polynomial) regression**:

$$p = 3,\ X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & X_N^2 \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix},\ \ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$$

Examples of multiple regression

- **Effect on groups**. Consider an example where we have data obtained on different days. The effect of the days can be modeled as
  $y_i = \underbrace{\beta_0}_{\text{day 1}} + \underbrace{\beta_1}_{\text{day 2}} + \underbrace{\beta_2}_{\text{day 3}} + \epsilon_i$

- **Effect on groups**. Consider an example where we have data obtained on different days. The effect of the days can be modeled as
$y_i = \underbrace{\beta_0}_{\text{day 1}} + \underbrace{\beta_1}_{\text{day 2}} + \underbrace{\beta_2}_{\text{day 3}} + \epsilon_i$

$$p = 3, \quad X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$
  - $\boldsymbol{\beta}$ parameters: $p \times 1$
  - ϵ: random error / disturbances
    ϵ_i are iid, $\mathbb{E}[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

$$Y = X \beta + \epsilon$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\|a\|^2 = a^T a = \sum_{j=1}^{N} a_j^2$$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \underbrace{\mathbf{x}_i\boldsymbol{\beta}}_{\hat{y}_i})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- setting derivative to zero gives *normal equations*

$$(X^\top X)^{-1} \quad \underbrace{\mathbf{X}^\top\mathbf{X}}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

- if $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^\top X)^{-1}X^\top}_{\text{hat matrix}} y$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

- if $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
- fitted values: $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\text{“hat matrix”}}\mathbf{y}$

Deriving the normal equations

- least squares objective:

$$f(\beta) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

Deriving the normal equations

- least squares objective:

$$f(\beta) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

- set gradient to zero. *Gradient* is the vector of partial derivatives:

Deriving the normal equations

- least squares objective:

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- set gradient to zero. *Gradient* is the vector of partial derivatives:

$$\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial f}{\partial \beta_0} \\ \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_{p-1}} \end{pmatrix} = 0$$

$$\frac{\partial f}{\partial \beta_0} = 0 \qquad \frac{\partial f}{\partial \beta_1} = 0 \qquad \vdots$$

If $\boldsymbol{\beta}$ is $p \times 1$, then $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ is $p \times 1$.

Partial derivative

- example: 1 data point, $p = 2$:

$$f(\boldsymbol{\beta}) = (\underbrace{y_1 - x_{11}\beta_1 - \beta_0}_{\hat{y}_1})^2$$

$$\frac{\partial f}{\partial \beta_1} = 0$$

- example: 1 data point, $p = 2$:

$$f(\boldsymbol{\beta}) = (y_1 - x_{11}\beta_1 - \beta_0)^2$$

- derivative:

$$\frac{\partial f}{\partial \beta_1} = -2x_{11}(y_1 - x_{11}\beta_1 - \beta_0) \stackrel{!}{=} 0$$

Partial derivative

- example: 1 data point, $p = 2$:

$$f(\boldsymbol{\beta}) = (y_1 - x_{11}\beta_1 - \beta_0)^2$$

- derivative:

$$\frac{\partial f}{\partial \beta_1} = -2x_{11}(y_1 - x_{11}\beta_1 - \beta_0)$$

- similarly:

$$\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \overset{!}{=} 0$$

$$X\beta = X^\top X y$$

Data Analysis:
Statistical Modeling and Computation in Applications

Correlation and Least Squares Regression
Part 5

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

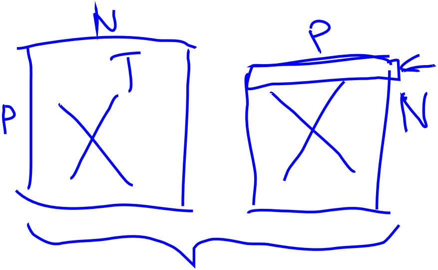
- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

- if $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

When is $\mathbf{X}^\top\mathbf{X}$ invertible?

- if $\mathbf{X}^\top\mathbf{X}$ has *full rank*:

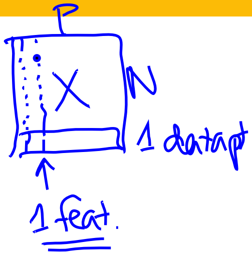


1 datapoint

pxp matrix

- $N \geq P$
- all cols linearly independent

When is $\mathbf{X}^\top \mathbf{X}$ invertible?



- if $\mathbf{X}^\top \mathbf{X}$ has *full rank*:
- $N \geq p$

$$\beta_0 + 2\beta_1 = 5$$

$N = 1$
$p = 2$

$\beta_1 \quad \beta_2$

When is $\mathbf{X}^\top \mathbf{X}$ invertible?

- if $\mathbf{X}^\top \mathbf{X}$ has *full rank*:
- $N \geq p$
- all columns of $\mathbf{X}$ linearly independent

If $p > N$...

Regularize!

If $p > N \ldots$

Regularize!

- ℓ_2 **penalty**: minimize

$$\underbrace{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{} + \lambda \underbrace{\underbrace{\|\boldsymbol{\beta}\|_2^2}_{\sum_{j=0}^{p-1} \beta_j^2}}_{}$$

penalizes large values of β_j
always unique $\hat{\boldsymbol{\beta}}$.

If $p > N$...

Regularize!

- ℓ_2 **penalty**: minimize

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_2^2}_{\sum_{j=0}^{p-1} \beta_j^2}$$

  penalizes large values of β_j
  always unique $\hat{\boldsymbol{\beta}}$.

- ℓ_1 **penalty (Lasso)**: minimize

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_1}_{\sum_{j=0}^{p-1} |\beta_j|}$$

  prefers *sparse* β (few nonzero coordinates)

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad\qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad\qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- **OLS is (conditionally) unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}} | \mathbf{X}] = \boldsymbol{\beta}$.

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- **OLS is (conditionally) unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}} | \mathbf{X}] = \boldsymbol{\beta}$.

- **Gaussianity:** If $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, model correct and $\mathbf{X}$ fixed, then $\hat{\boldsymbol{\beta}}$ is normal: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad\qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- **OLS is (conditionally) unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}} | \mathbf{X}] = \boldsymbol{\beta}$.

- **Gaussianity:** If $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, model correct and $\mathbf{X}$ fixed, then $\hat{\boldsymbol{\beta}}$ is normal: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$
- *t-test to test $\beta_j = 0$ vs. $\beta_j \neq 0$*: estimate σ^2 as $\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i-1}^{N} (y_i - \hat{y}_i)^2$, then $(N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p}$.

Backward Model Selection

Which variables should I include in my model?
$\beta_j = 0$ would mean I exclude variable j from the prediction.

Backward Model Selection

Which variables should I include in my model?
$\beta_j = 0$ would mean I exclude variable j from the prediction.

- Fit a model that uses all variables:
  $y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_{p-1} x_{i,p-1} + \epsilon_i$

Backward Model Selection

Which variables should I include in my model?
$\beta_j = 0$ would mean I exclude variable j from the prediction.

- Fit a model that uses all variables:
  $y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_{p-1} x_{i,p-1} + \epsilon_i$
- Use the t-test to determine variables that are not significant. Of those, remove the one with the largest p-value. Re-fit and repeat until all variables have significant p-values.

References

- D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007.
  Part III.

- D. Freedman. *Statistical Models – Theory and Practice*. 2009.
  Chapters 2–4.