Estimating linear models

ECON306 – Slides 2
Studenmund Ch. 1–3

Bruno Salcedo

The Pennsylvania State University

June 2013

[0]

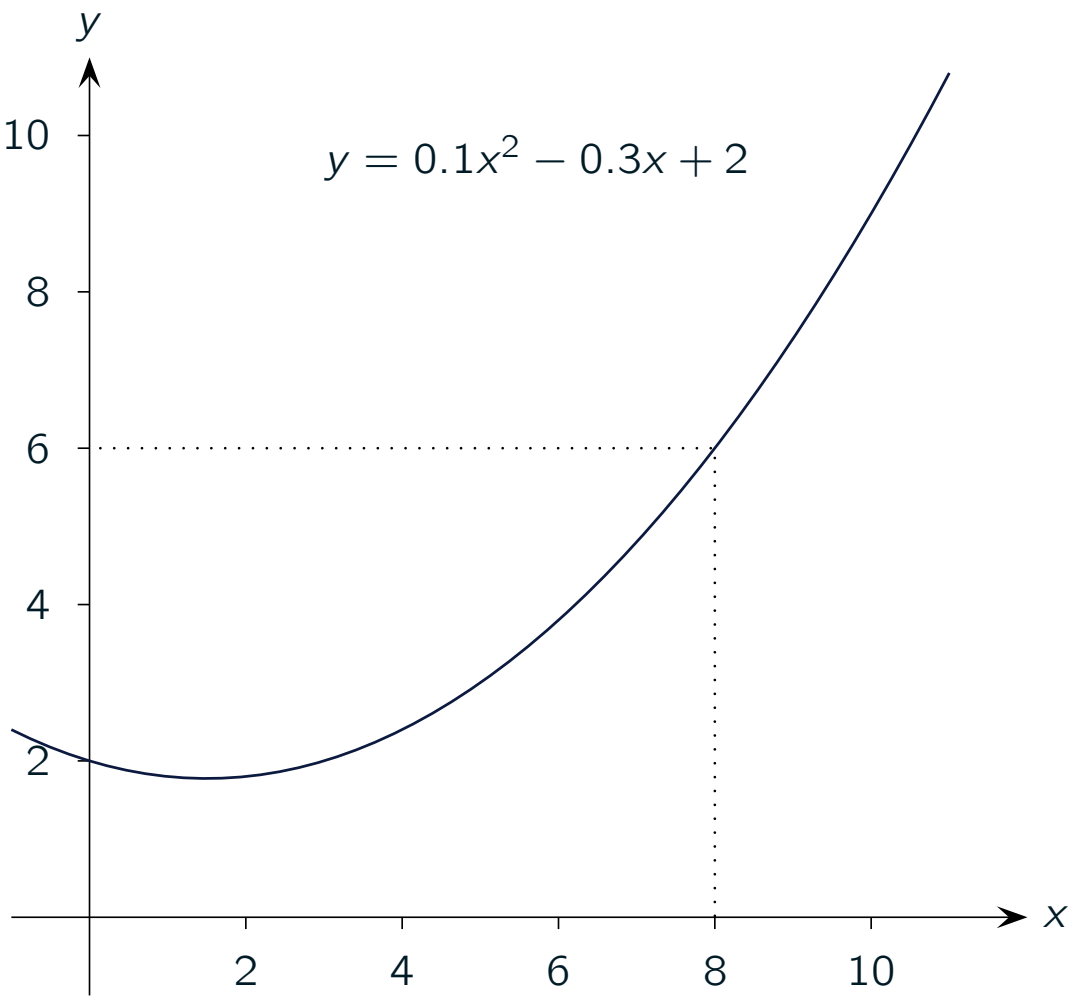
Functional relations

- Quantitative characteristics of the world are usually entangled in functional relations
- A regression or model specifies an explained variable as a function of an explanatory variable

$$y = f(x)$$

- y – regressand, response variable, explained variable, dependant variable, outcome
- x – regressor, predictor variable, explanatory variable, independent variable, control

Example: Quadratic regression



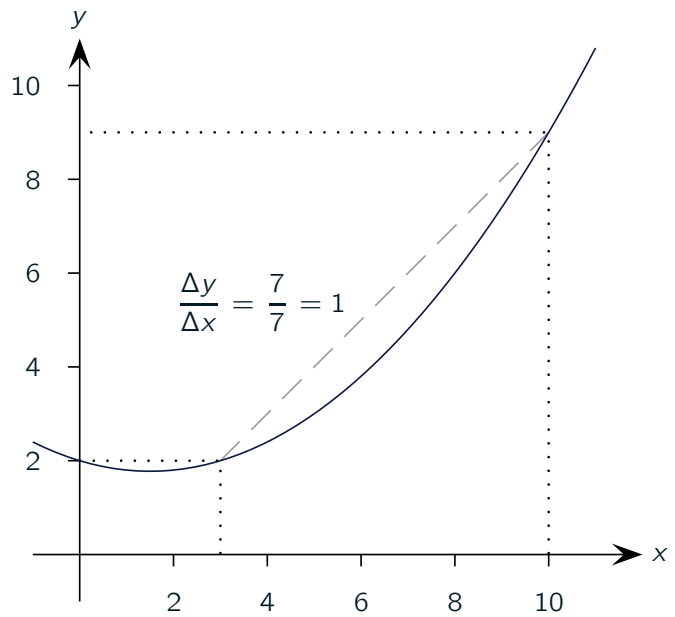
Rate of change

$$\Delta x = x_1 - x_0 \qquad \Delta y = y_1 - y_0 = f(x_1) - f(x_0)$$

- The rate of change measures how y responds to changes in x

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

- It depends both on the initial point and the magnitude of the change



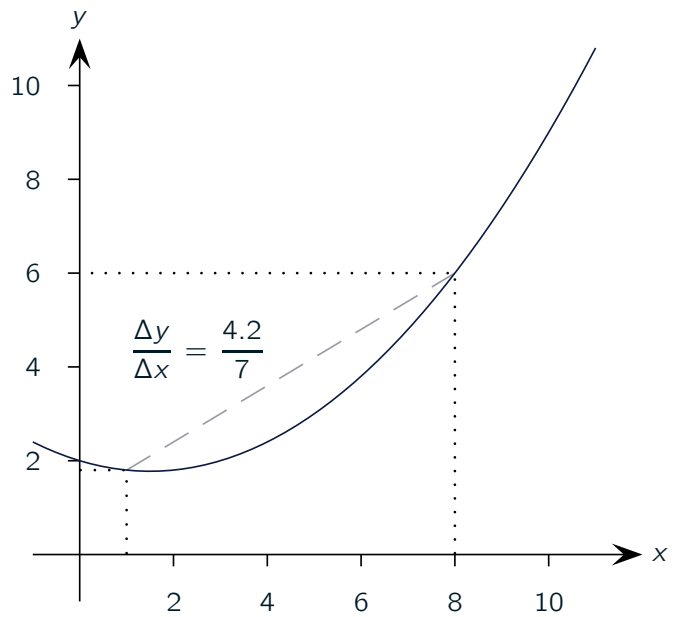
Rate of change

$$\Delta x = x_1 - x_0 \qquad \Delta y = y_1 - y_0 = f(x_1) - f(x_0)$$

- The rate of change measures how y responds to changes in x

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

- It depends both on the initial point and the magnitude of the change

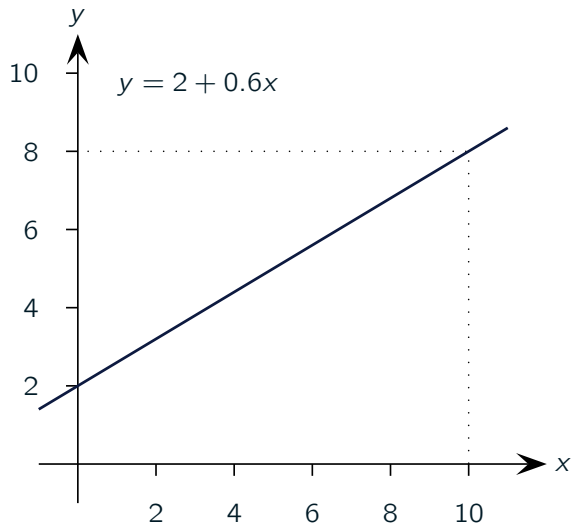


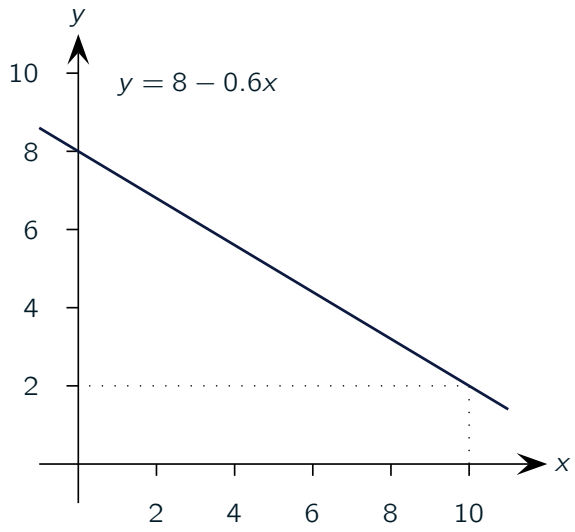
Linear models

- A model is linear if it can be written as:

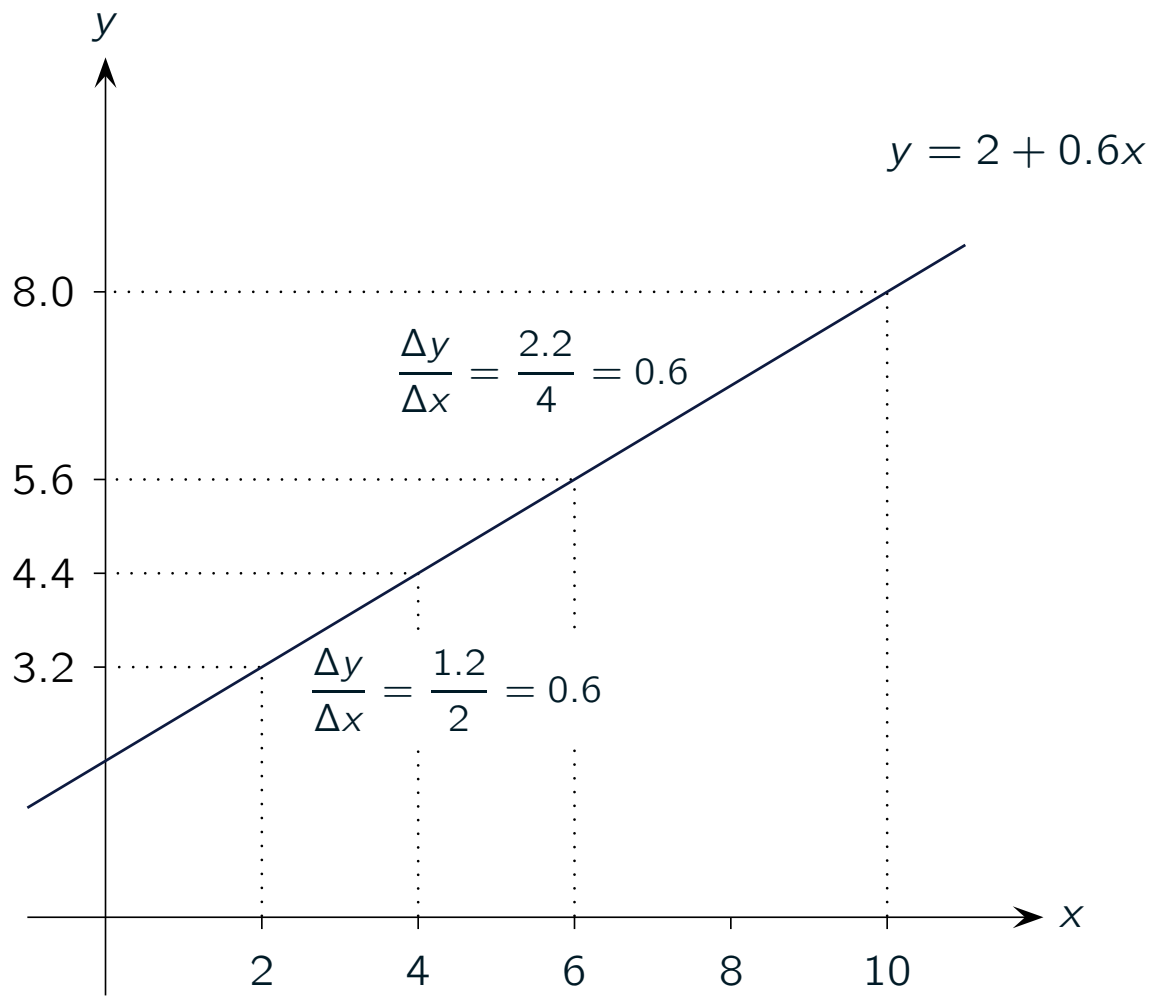
$$y = \beta_0 + \beta_1 x$$

- Which means that the graph of the regression is a (straight) line





Slope coefficient



Slope coefficient

- The slope of a linear model equals β_0 independently of x_0 and Δx

$$
\begin{aligned}
\frac{\Delta y}{\Delta x} &= \frac{y_1 - y_0}{x_1 - x_0} \\
&= \frac{(\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 x_0)}{x_1 - x_0} \\
&= \frac{(\beta_0 - \beta_0) + (\beta_1 x_1 - \beta_1 x_0)}{x_1 - x_0} \\
&= \frac{0 + \beta_1 (x_1 - x_0)}{x_1 - x_0} \\
&= \beta_1 \frac{x_1 - x_0}{x_1 - x_0} = \beta_1
\end{aligned}
$$

The linearity assumption

- The linearity assumption is less restrictive than it appears
- The following model is clearly nonlinear

$$y = \log(\gamma_0 x^{\gamma_1})$$

- However after some relabelling:

$$\beta_0 = \log(\gamma_0)$$
$$\beta_1 = \gamma_1$$
$$z = \log(x)$$

- We obtain a linear model

$$y = \log(\gamma_0 x^{\gamma_1}) = \log(\gamma_0) + \gamma_1 \log(x) = \beta_0 + \beta_1 z$$

Approximating non-linear models

- Supose that the true relationship between x and y is given by

$$y = f(x)$$

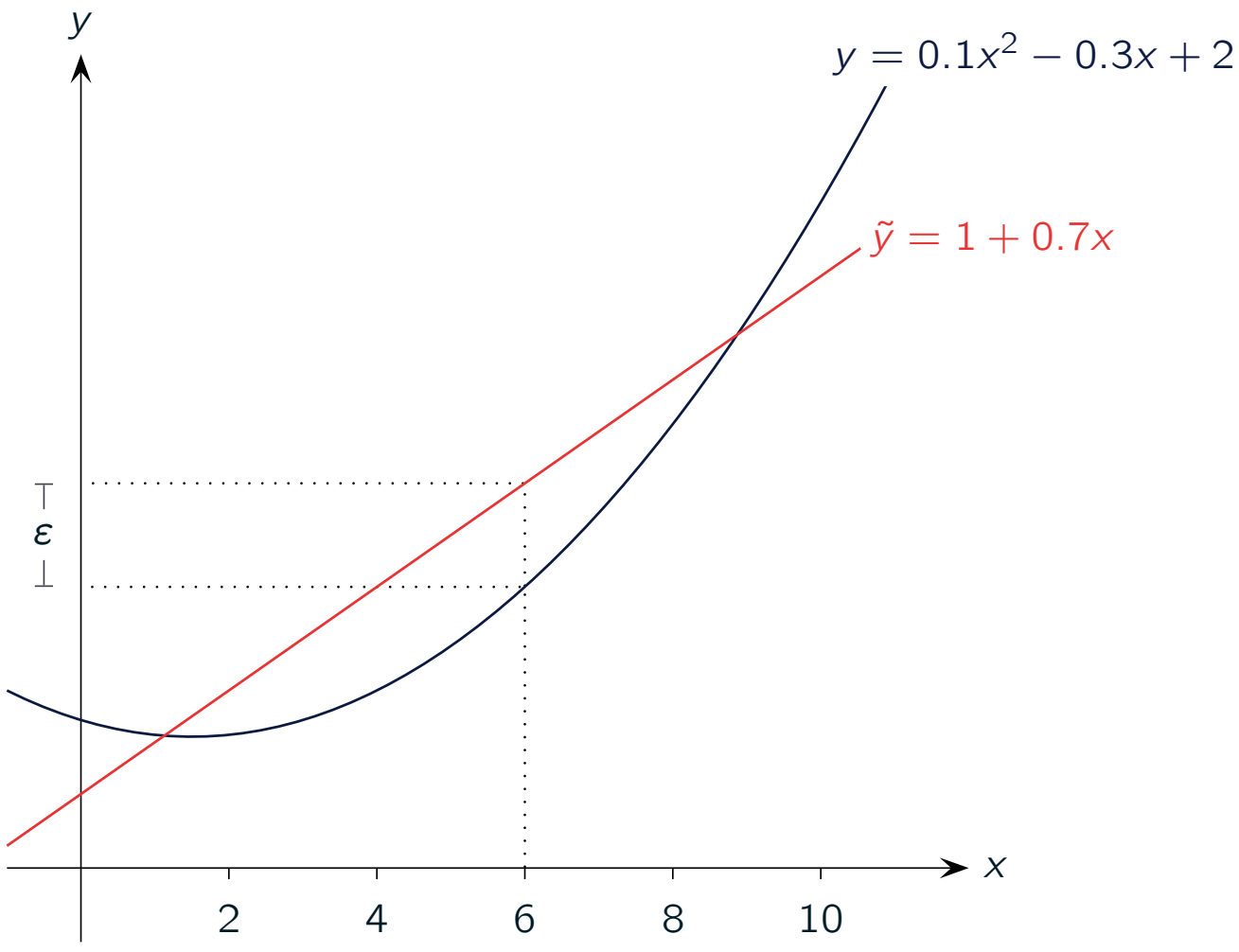
- We can always abstract from non potential linearities and use a linear model

$$\tilde{y} = \beta_0 + \beta_1 x \approx f(x) = y$$

- If f is not linear, then the approximation will be inexact and there will be approximation errors

$$\varepsilon = y - \tilde{y}$$

Approximating non-linear models



Multivariate regressions

- The value of the response variable may be a function of many regressors

$$y = f(x_1, x_2, \ldots, x_k)$$

- We can still have linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- In this case, each coefficient β_i is still a measure of change holding every other variable constant

$$\frac{\Delta y}{\Delta x_i} = \beta_i$$

- For multivariate regressions linearity assumes separability

Unobserved variables

- We may not know or observe all the variables which affect y

$$y = \beta_0 + \beta_1 x_1 + \underbrace{\beta_2 x_2 + \ldots + \beta_k x_k}_{\text{unobserved}}$$

- We can still approximate y with the variables that we do observe

$$\tilde{y} = \beta_0 + \beta_1 x_1 \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k = y$$

- As before, this approximation is inexact and has an approximation error

$$\varepsilon = y - \tilde{y} = \beta_2 x_2 + \ldots + \beta_k x_k$$

Stochastic regression

- Most of the time there is uncertainty because (at least)
  - We are not certain about the linearity of a regression
  - We cannot list all the relevant regressors
- Uncertainty is captured by a stochastic error term ε

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- $\beta_0 + \beta_1 x$ is called the deterministic component of the model

Stochastic regression

- Assume that the error has zero mean conditional on x
- Then the deterministic component corresponds to the mean of y conditional on x

$$\begin{aligned}\mathbb{E}[y|x] &= \mathbb{E}[\beta_0 + \beta_1 x + \varepsilon|x] \\ &= \mathbb{E}[\beta_0|x] + \mathbb{E}[\beta_1 x|x] + \mathbb{E}[\varepsilon|x] = \beta_0 + \beta_1 x\end{aligned}$$

- Then slope coefficient measures the average per-unit effect of changes in x over the average value of y conditional on x

$$\begin{aligned}\mathbb{E}[y|x_1] - \mathbb{E}[y|x_0] &= (\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 x_0) \\ &= \beta_1(x_1 - x_0)\end{aligned}$$

Random samples

- We are usually interested in different observations coming from
  - **Cross-sectional** – different sources
  - **Time series** – a single source at different times
  - **Panel data** – different time series from different sources
- We assume that the data comes from a random sample $\{x_i, y_i, \varepsilon_i\}$
- x_i and y_i are observed but ε_i is not
- In fact we have a collection of equations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- In case of a multivariate regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Predictions and residuals

- Suppose that we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- The estimated model is then:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Given an estimated model, for each realization of x_i the predicted value of y_i is:

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

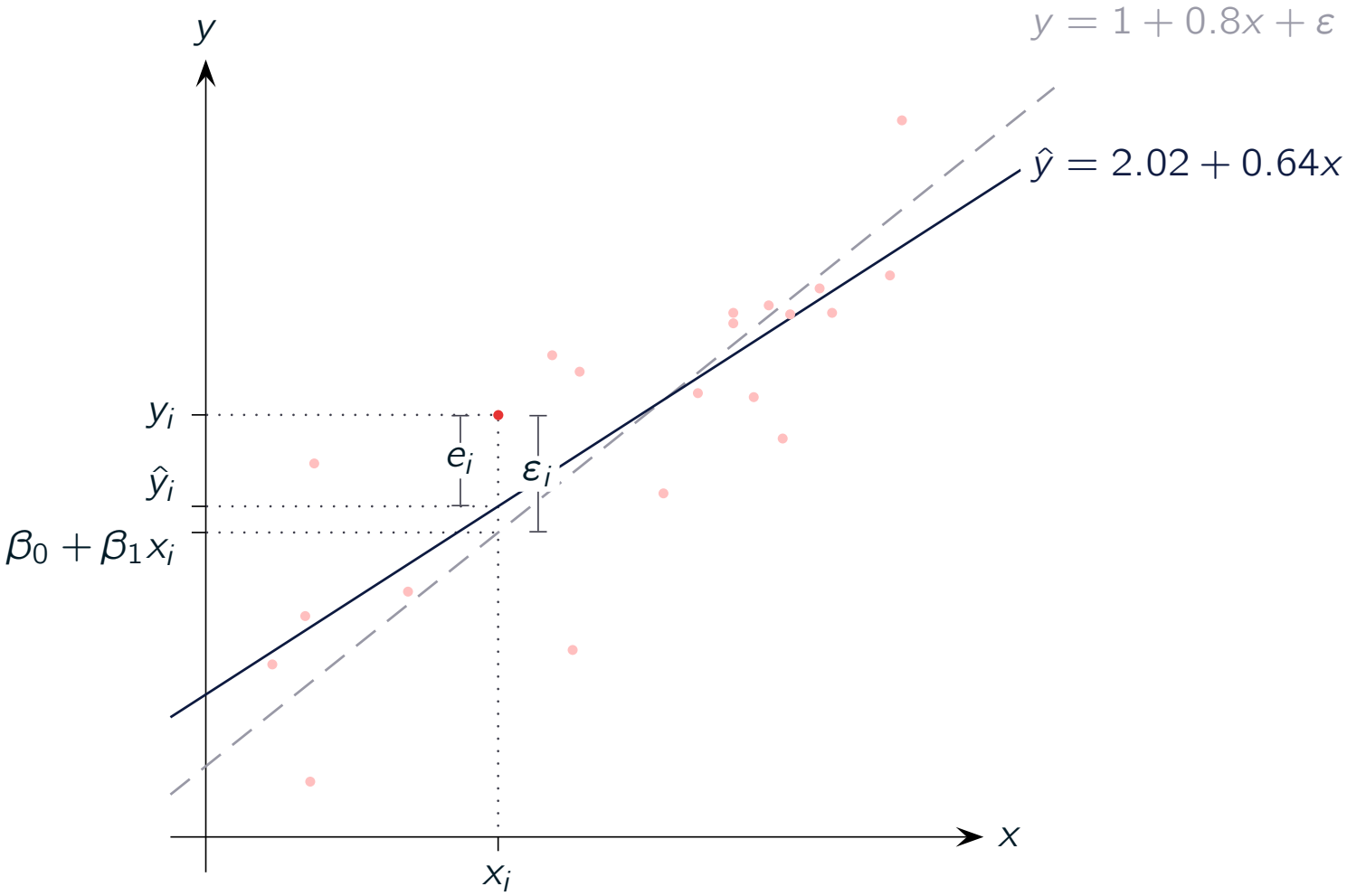
- The corresponding residual is:

$$e_i = y_i - \hat{y}_i$$

- Notice we cannot guarantee that $e_i = \varepsilon_i$ unless we know β_0 and β_1

Example: a linear regression

Errors vs. Residuals



Example: height and weight

Model

- Contest game:
  - If you guess the weight of a participant within 10lb of the actual weight, you get paid \$2
  - Otherwise you have to pay him/her \$3
- You could use height (observable) to estimate the weight

$$\text{WEIGHT}_i = \beta_0 + \beta_1 \text{HEIGHT}_i + \varepsilon_i$$

- Given estimated coefficients $\hat{\beta}_0 = 103.4$ and $\hat{\beta}_1 = 6.38$
- You can make predictions

$$\widehat{\text{WEIGHT}}_i = 103.4 + 6.38\,\text{HEIGHT}_i$$

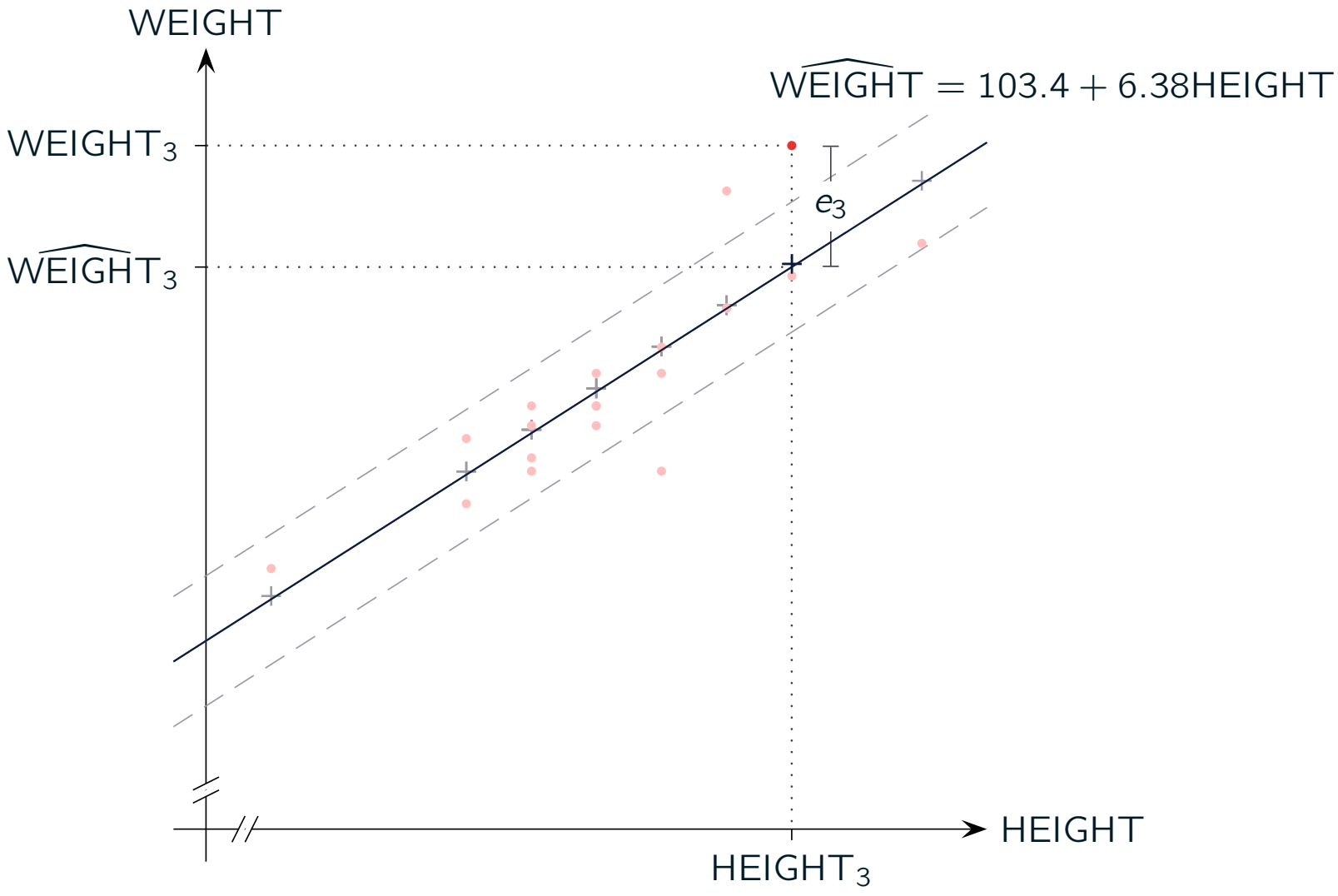
Example: height and weight

Predictions · observations · residuals

	HEIGHT_i	WEIGHT_i	$\widehat{\text{WEIGHT}}_i$	e_i	Profit
1	5	140	135.3	4.7	2
2	9	157	160.8	-3.8	2
3	13	205	186.3	18.7	-3
4	12	198	180.0	18.0	-3
5	10	162	167.2	-5.2	2
6	11	174	173.6	0.4	2
7	8	150	154.4	-4.4	2
8	9	165	160.8	4.2	2
9	10	170	167.2	2.8	2
10	12	180	180.0	0.0	2
11	11	170	173.6	-3.6	2
12	9	162	160.8	1.2	2
13	10	165	167.2	-2.2	2
14	12	180	180.0	0.0	2
15	8	160	154.4	5.6	2
16	9	155	160.8	-5.8	2
17	10	165	167.2	-2.2	2
18	15	190	199.1	-9.1	2
19	13	185	186.3	-1.3	2
20	11	155	173.6	-18.6	-3

Example: height and weight

Predictions · observations · residuals



[0]

Estimating linear models

- Begin from dataset coming from a random sample $\{x_1, y_i\}$
- We assume that x and y are related by a model:

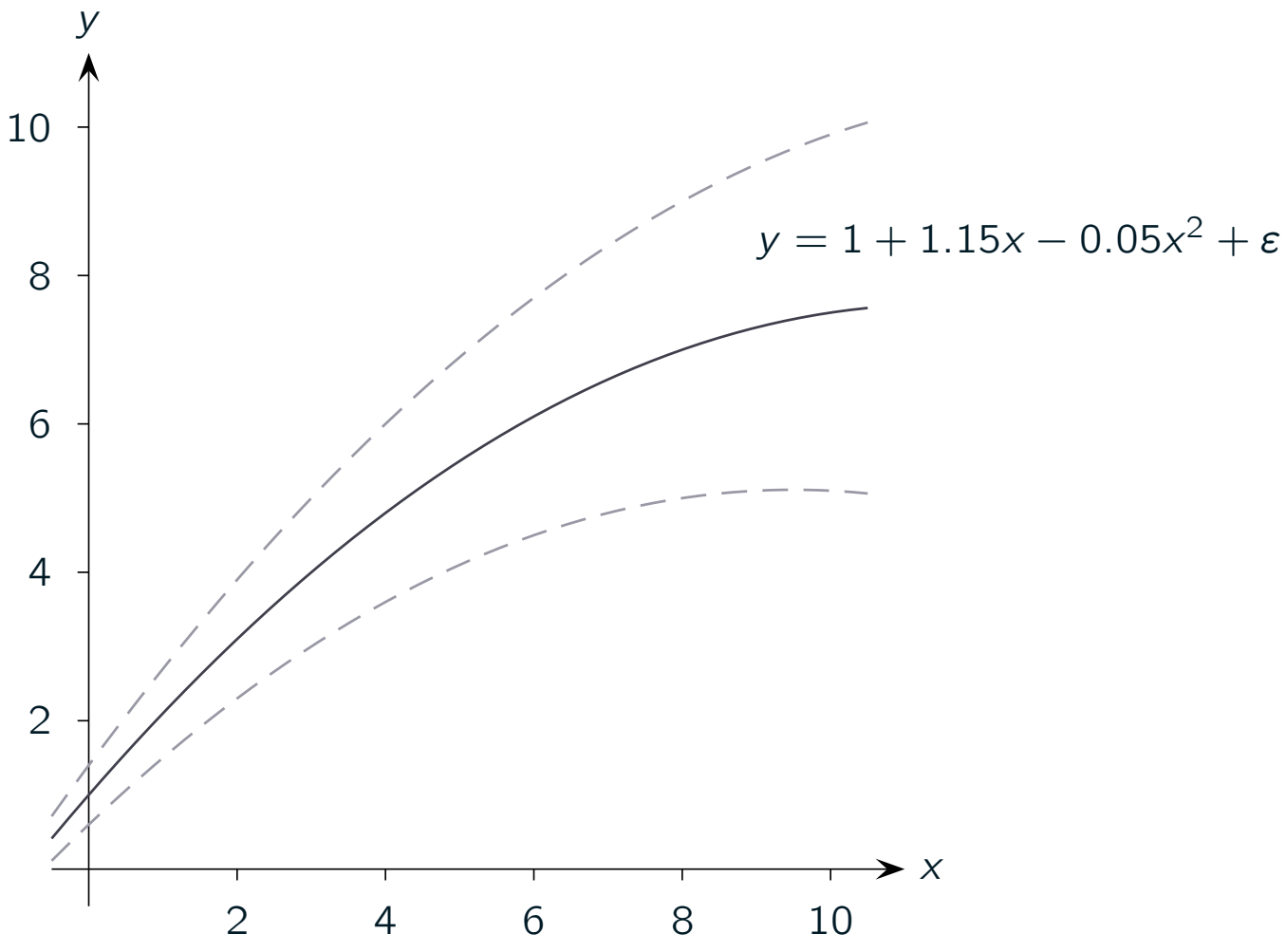
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- We do not observe ε_i
- We do not know the true coefficients β_0 and β_1
- Our objective now is to generate estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of these coefficients to obtain an estimated model

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

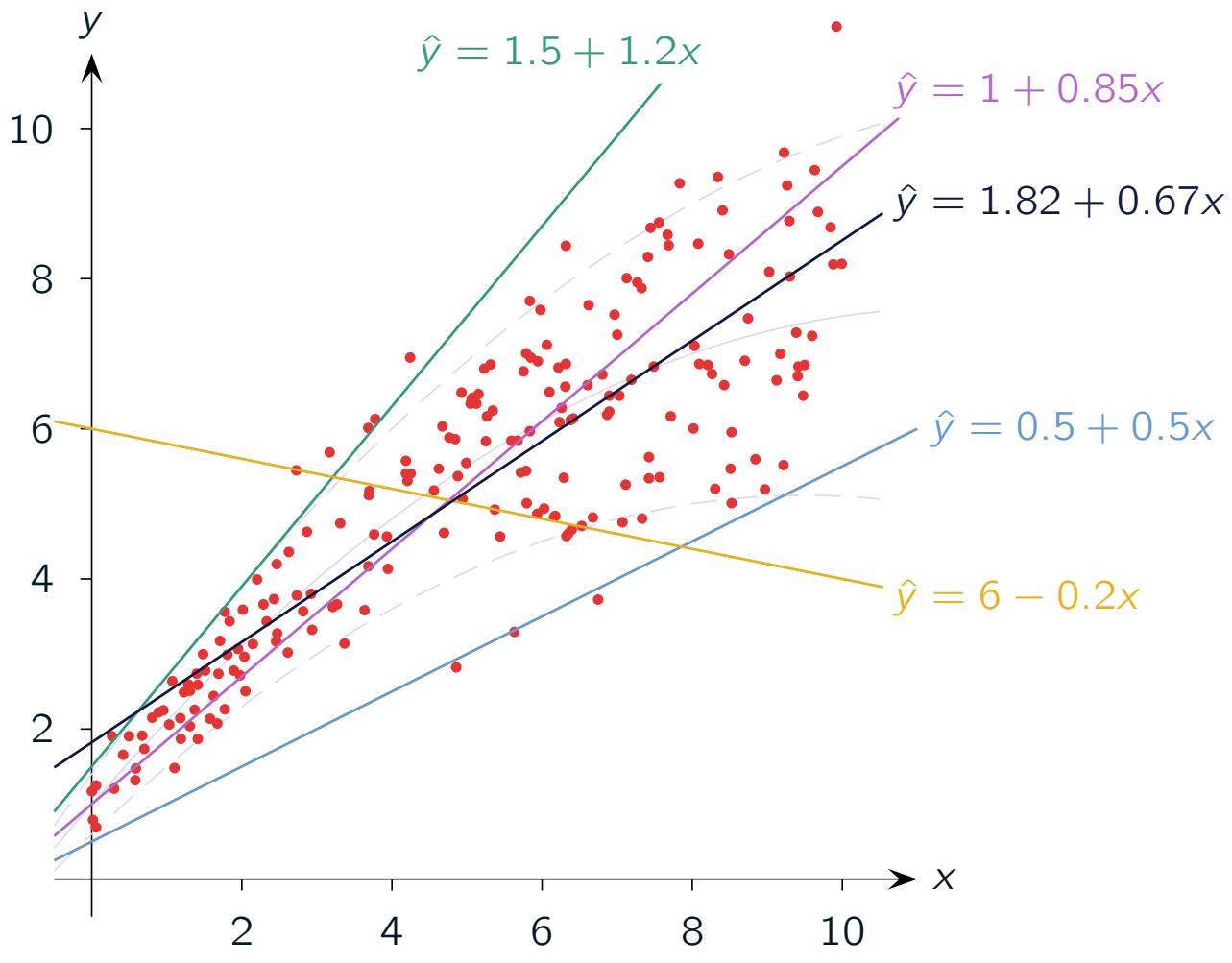
Example: linear regresion

Data generating process



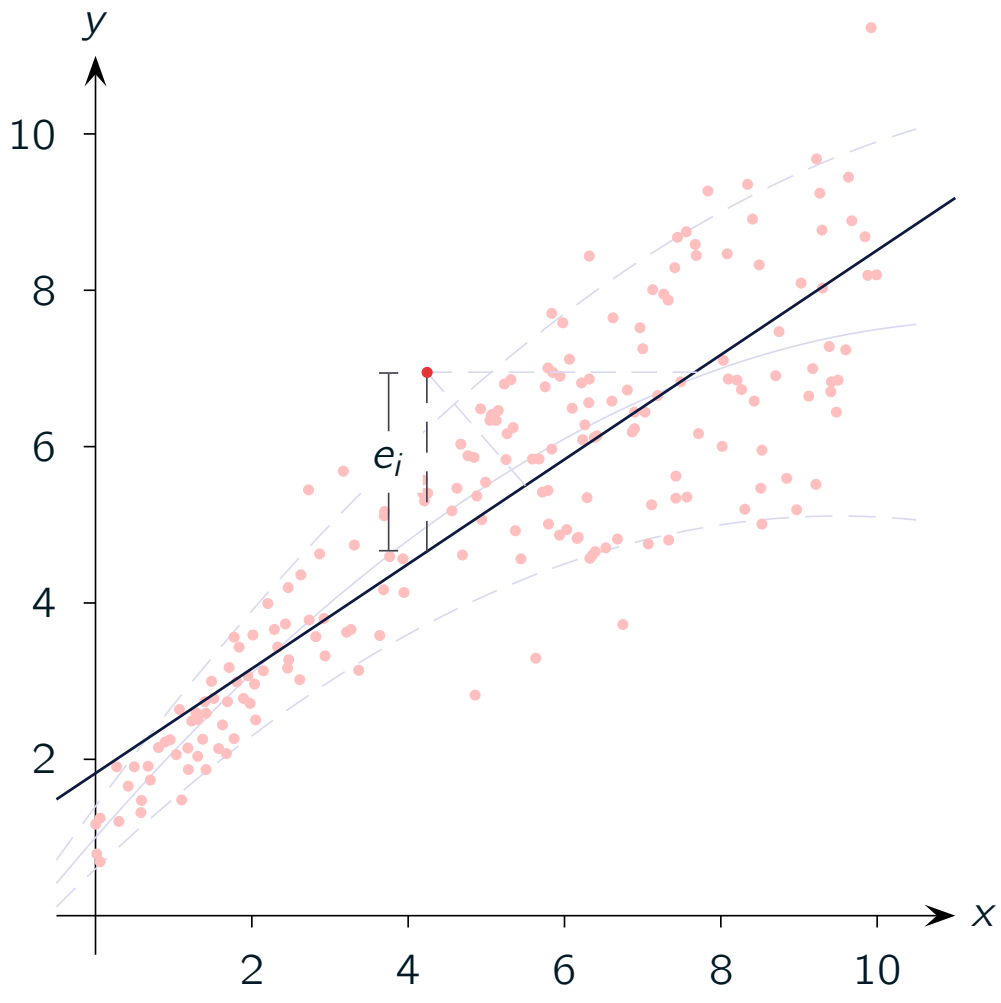
Example: linear regresion

The 'best' linear model



Example: linear regresion

The ‘closest’ linear model



The 'best' linear model

- Two uses for the estimated model:
  - **Prediction** – Given x_i, y_i should be around $\hat{y}_i$
  - **Policy** – Controlling x_i, y_i react on average according to:

$$\Delta y_i = \beta_1 \Delta x_i \approx \hat{\beta}_1 \Delta x_i$$

- Policy implications only make sense if we establish causality
- Better policy implications when $\hat{\beta}_1 \approx \beta_1$ and $e \approx 0$

- Better predictions when $y_i \approx \hat{y}_i$, i.e. when the residuals are small

Residual variance

- We wish to have small residuals:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Small means in magnitude not sign

- Minimize the total sum of of squared residuals:

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Ordinary least squares

Definition

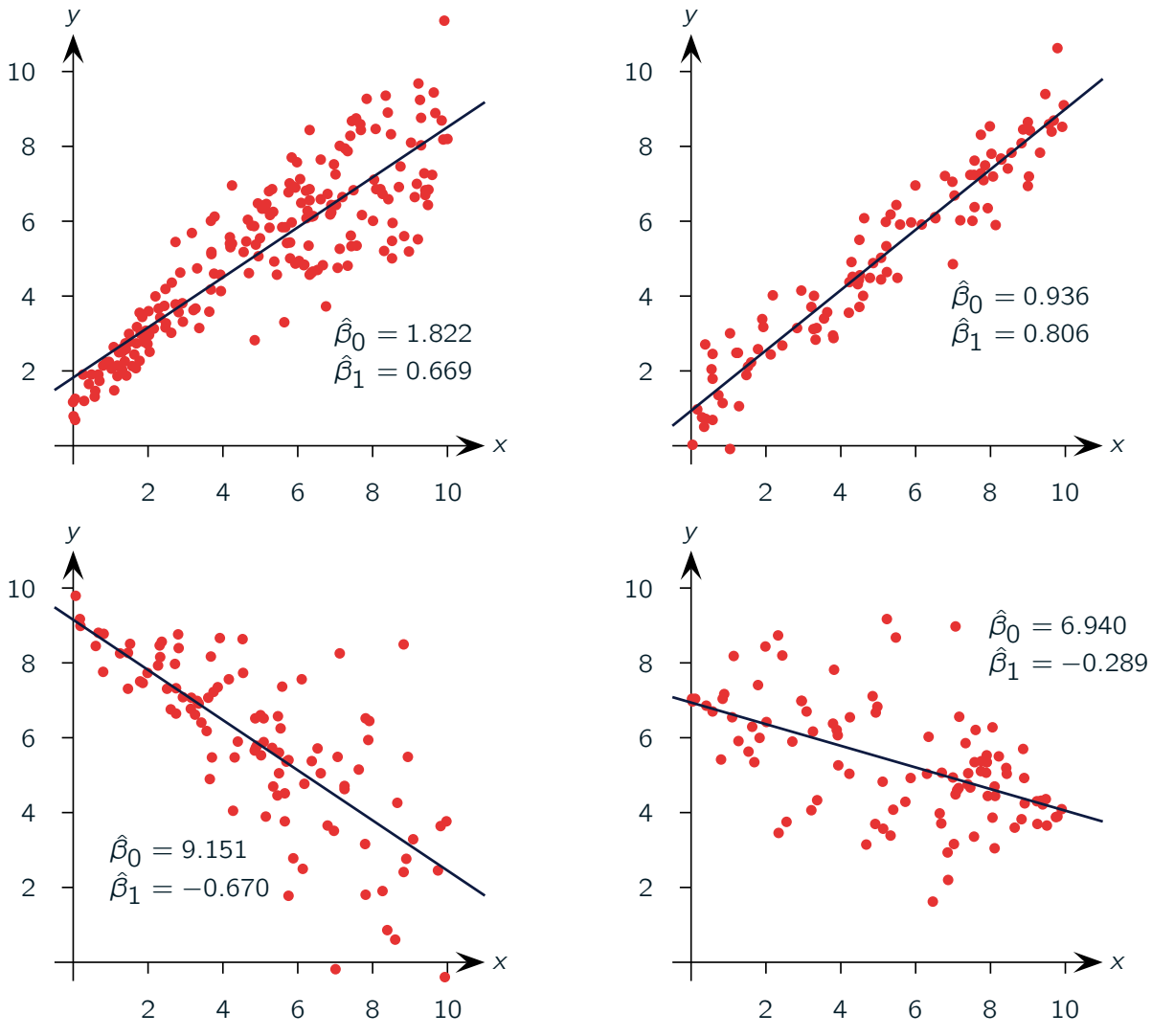
Given a data set, the ordinary least squares (OLS) estimates of β_0 and β_1, are the umbers $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of squared residuals:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

The OLS estimated model is:

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

Examples: OLS



Computing OLS

When $\beta_1 = 0$

- Suppose that we know that $\beta_1 = 0$, i.e.

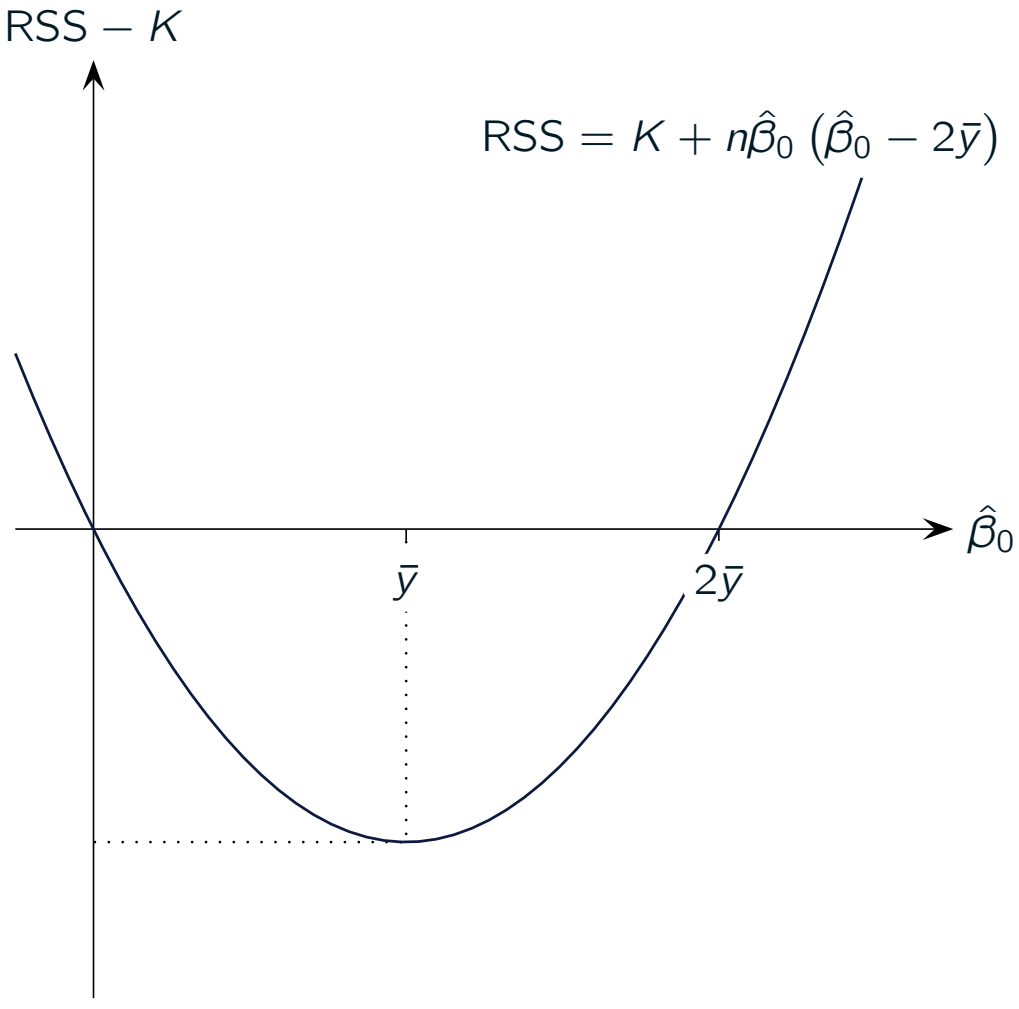
$$y_i = \beta_0 + \varepsilon_i$$

- The sum of residuals is then:

$$\begin{aligned} \text{RSS} = \sum \left(y_i - \hat{\beta}_0 \right)^2 &= \sum y_i^2 - 2\hat{\beta}_0 \sum y_i + n\hat{\beta}_0^2 \\ &= K - 2n\hat{\beta}_0 \bar{y} + n\hat{\beta}_0^2 = K + n\hat{\beta}_0 \left(\hat{\beta}_0 - 2\bar{y} \right) \end{aligned}$$

- Which is minimized when $\hat{\beta}_0 = \bar{y}$ (see next slide)
- $\bar{y}$ is indeed a natural estimator given $\beta_0 = \mathbb{E}[\, y \,]$

Minimizing quadratic functions



Computing OLS

When $\beta_0 = 0$

- Now suppose that we know that $\beta_0 = 0$, i.e.

$$y_i = \beta_1 x_i + \varepsilon_i$$

- The sum of residuals is then:

$$\begin{aligned} \text{RSS} &= \sum \left(y_i - \hat{\beta}_1 x_i \right)^2 = \sum y_i^2 - 2\hat{\beta}_1 \sum x_i y_i + \hat{\beta}_1^2 \sum x_i^2 \\ &= K + \hat{\beta}_0 \left(\hat{\beta}_0 \sum x_i y_i - 2 \sum x_i^2 \right) \end{aligned}$$

- In this case we obtain:

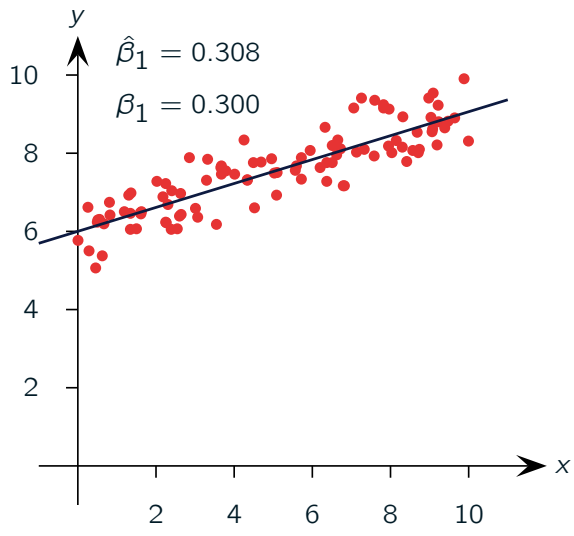
$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

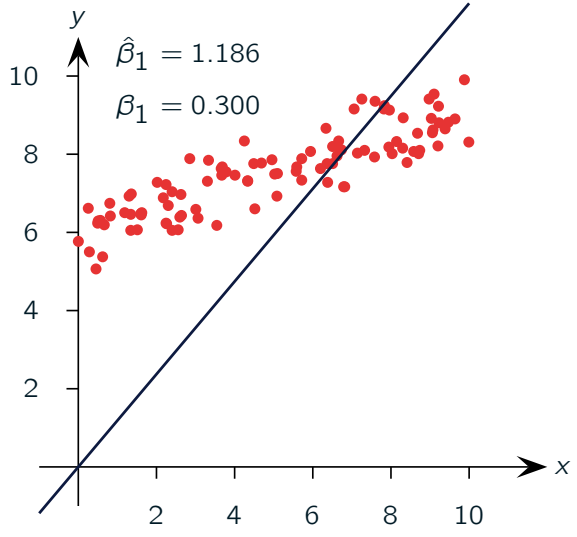
The need for an intercept

- Most of the time we will be interested in β_1 rather than β_0
- One could simply estimate

$$y_i = \beta_1 x_i + \varepsilon_i$$

- But if $\hat{\beta}_0 \neq 0$ we may get bad estimates





Computing OLS

OLS formulas

In the general case, the OLS estimates are given by:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Notice that $\hat{\beta}_1$ looks like a sample analogue of $\sigma_y^2 \cdot \rho_{xy}$
- The OLS estimates guarantee that $\sum e_i = 0$

Example: height and weight

Computing OLS

	x_i	y_i	$x_i - \bar{x}$	$y_1 - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	5	140	-5.35	-29.40	28.62	157.29
2	9	157	-1.35	-12.40	1.82	16.74
3	13	205	2.65	35.60	7.02	94.34
4	12	198	1.65	28.60	2.72	47.19
5	10	162	-0.35	-7.40	0.12	2.59
6	11	174	0.65	4.60	0.42	2.99
7	8	150	-2.35	-19.40	5.52	45.59
8	9	165	-1.35	-4.40	1.82	5.94
9	10	170	-0.35	0.60	0.12	-0.21
10	12	180	1.65	10.60	2.72	17.49
11	11	170	0.65	0.60	0.42	0.39
12	9	162	-1.35	-7.40	1.82	9.99
13	10	165	-0.35	-4.40	0.12	1.54
14	12	180	1.65	10.60	2.72	17.49
15	8	160	-2.35	-9.40	5.52	22.09
16	9	155	-1.35	-14.40	1.82	19.44
17	10	165	-0.35	-4.40	0.12	1.54
18	15	190	4.65	20.60	21.62	95.79
19	13	185	2.65	15.60	7.02	41.34
20	11	155	0.65	-14.40	0.42	-9.36
mean	10	169	0.00	0.00	4.63	29.51
sum	207	3388	0.00	0.00	92.55	590.20

Example: height and weight

Computing OLS

	x_i	y_i	$x_i - \bar{x}$	$y_1 - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	5	140	-5.35	-29.40	28.62	157.29
2	9	157	-1.35	-12.40	1.82	16.74
3	13	205	2.65	35.60	7.02	94.34
4	12	198	1.65	28.60	2.72	47.19
5	10	162	-0.35	-7.40	0.12	2.59
6	11	174	0.65	4.60	0.42	2.99
7	8	150	-2.35	-19.40	5.52	45.59
8	9	165	-1.35	-4.40	1.82	5.94
9	10	170	-0.35	0.60	0.12	-0.21
10	12	180	1.65	10.60	2.72	17.49
11	11	170	0.65	0.60	0.42	0.39
12	9	162	-1.35	-7.40	1.82	9.99
13	10	165	-0.35	-4.40	0.12	1.54
14	12	180	1.65	10.60	2.72	17.49
15	8	160	-2.35	-9.40	5.52	22.09
16	9	155	-1.35	-14.40	1.82	19.44
17	10	165	-0.35	-4.40	0.12	1.54
18	15	190	4.65	20.60	21.62	95.79
19	13	185	2.65	15.60	7.02	41.34
20	11	155	0.65	-14.40	0.42	-9.36
mean	10	169	0.00	0.00	4.63	29.51
sum	207	3388	0.00	0.00	92.55	590.20

Example: height and weight

Computing OLS

$\bar{x}$	$\bar{y}$	$\sum(x_i - \bar{x})^2$	$\sum(x_i - \bar{x})(y_i - \bar{y})$
10	169	92.55	590.2

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{590.2}{92.55} \approx 6.38$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 169 - 6.38 \cdot 10 \approx 105.22$$

$$\hat{y}_i = 105.22 + 6.38 x_i$$

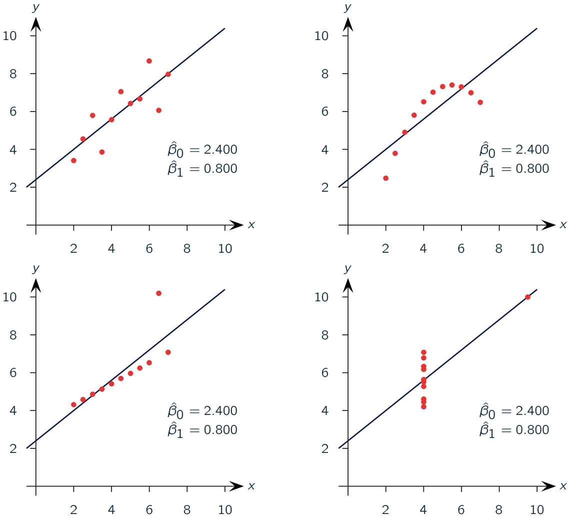
Example: Anscombe's quartet

Data

	(a)		(b)		(c)		(d)	
i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
1	5.0	6.4	5.0	7.3	5.0	6.0	4.0	5.3
2	4.0	5.6	4.0	6.5	4.0	5.4	4.0	4.6
3	6.5	6.1	6.5	7.0	6.5	10.2	4.0	6.2
4	4.5	7.0	4.5	7.0	4.5	5.7	4.0	7.1
5	5.5	6.7	5.5	7.4	5.5	6.2	4.0	6.8
6	7.0	8.0	7.0	6.5	7.0	7.1	4.0	5.6
7	3.0	5.8	3.0	4.9	3.0	4.9	4.0	4.2
8	2.0	3.4	2.0	2.5	2.0	4.3	9.5	10.0
9	6.0	8.7	6.0	7.3	6.0	6.5	4.0	4.4
10	3.5	3.9	3.5	5.8	3.5	5.1	4.0	6.3
11	2.5	4.5	2.5	3.8	2.5	4.6	4.0	5.5
μ	4.5	6.0	4.5	6.0	4.5	6.0	4.5	6.0
σ^2	2.8	2.6	2.8	2.6	2.8	2.6	2.8	2.6

Example: Anscombe's quartet

Estimated models



Multivariate regressions

- The analysis extends to multivariate models

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

- The interpretation is slightly different: $\hat{\beta}_k$ indicates the response to changes in x_k holding other regressors constant

- OLS is defined in the same way: minimizing RSS
- The formulas require linear algebra
- OLS is never done by hand: we use computers

Example: financial aid

Variables

- Response variable:
  - FINAID_i – grant per year to applicant i
- Regressors:
  - PARENT_i – feasible contributions from parents
  - HSRANK_i – GPA rank in high school
  - GENDER_i – gender dummy (1 if male and 0 if female)

Example: financial aid

Dataset

	FINAID_i	PARENT_i	HSRANK_i	GENDER_i
1	19640	0	92	0
2	8325	9147	44	1
3	12950	7063	89	0
4	700	33344	97	1
5	7000	20497	95	1
6	11325	10487	96	0
7	19165	519	98	1
8	7000	31758	70	0
9	7925	16358	49	0
10	11475	10495	80	0
11	18790	0	90	0
12	8890	18304	75	1
13	17590	2059	91	1
14	17765	0	81	0
15	14100	15602	98	0
16	18965	0	80	0
17	4500	22259	90	1
18	7950	5014	82	1
19	7000	34266	98	1
20	7275	11569	50	0
21	8000	30260	98	1
22	4290	19617	40	1
23	8175	12934	49	1
24	11350	8349	91	0
25	15325	5392	82	1

Example: financial aid

Dataset cont'd

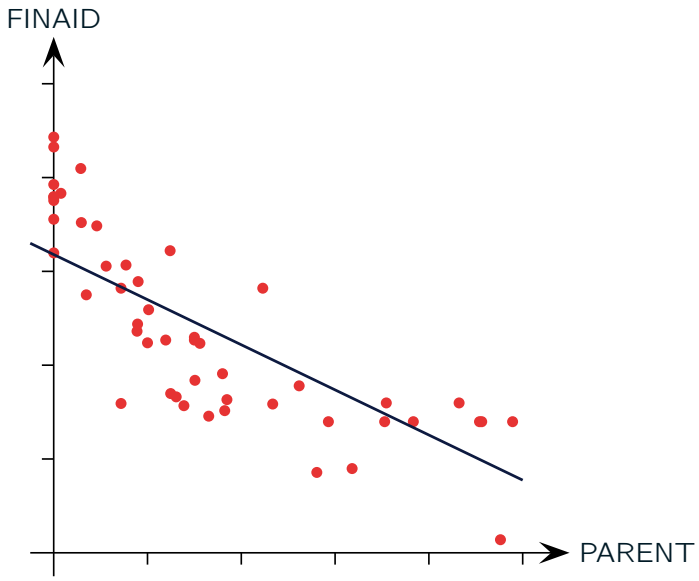
	FINAID_i	PARENT_i	HSRANK_i	GENDER_i
26	22148	0	98	0
27	17420	3207	99	0
28	18990	0	90	0
29	11175	10894	97	0
30	14100	5010	59	0
31	7000	24718	97	1
32	7850	9715	84	1
33	0	64305	84	0
34	7000	31947	98	1
35	16100	8683	95	1
36	8000	24817	99	0
37	8500	8720	20	1
38	7575	12750	89	1
39	13750	2417	41	1
40	7000	26846	92	1
41	11200	7013	86	1
42	14450	6300	87	0
43	15265	3909	84	0
44	20470	2027	99	1
45	9550	12592	89	0
46	15970	0	57	0
47	12190	6249	84	0
48	11800	6237	81	0
49	21640	0	99	0
50	9200	10535	68	0

Example: financial aid

OLS

- Estimated OLS model (ignoring GENDER and HSRANK):

$$\widehat{\text{FINAID}}_i = 15897 - 0.34\,\text{PARENT}_i$$

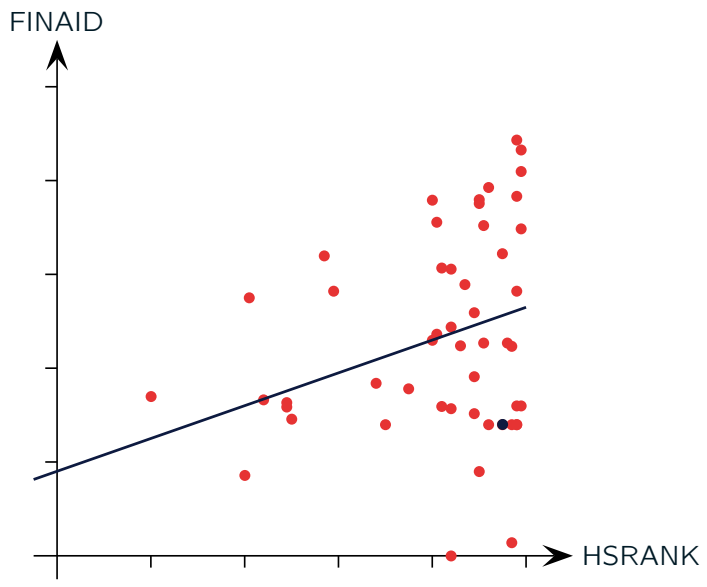


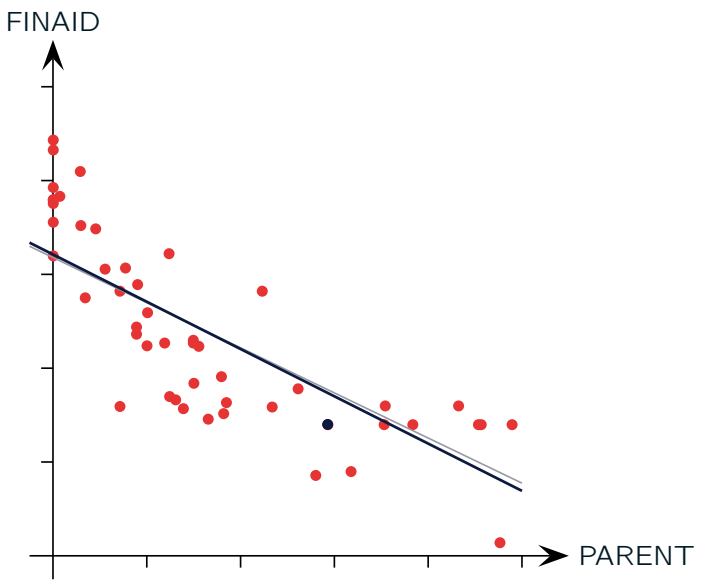
Example: financial aid

OLS

- Estimated OLS model (ignoring GENDER):

$$\widehat{\text{FINAID}}_i = 8927 - 0.36\,\text{PARENT}_i + 87.4\,\text{HSRANK}_i$$





[0]

Evaluating an estimated model

- Is the equation supported by sound theory/common sense?
- How well does the estimated model fit the data?
- Is the dataset reasonably large and accurate?
- Is OLS the best estimator to be used?
- Do estimated coefficients correspond to prior expectations?
- Are all the important variables included?

- In case we want to do policy: are the estimated parameters structural?

Explained variation

- Regressions are used to explain y
- In particular, we wish to explain why/when is y_i different from μ_y

- The variation in y can be decomposed as:

$$\begin{aligned} y_i - \mu_y =& \beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 \mu_x \\ =& \underbrace{\beta_1 (x_i - \mu_x)}_{\text{explained}} + \underbrace{\varepsilon_i}_{\text{unexplained}} \end{aligned}$$

- One way to evaluate models is to measure the proportion of the variance of y that we are able to explain

Explained variation

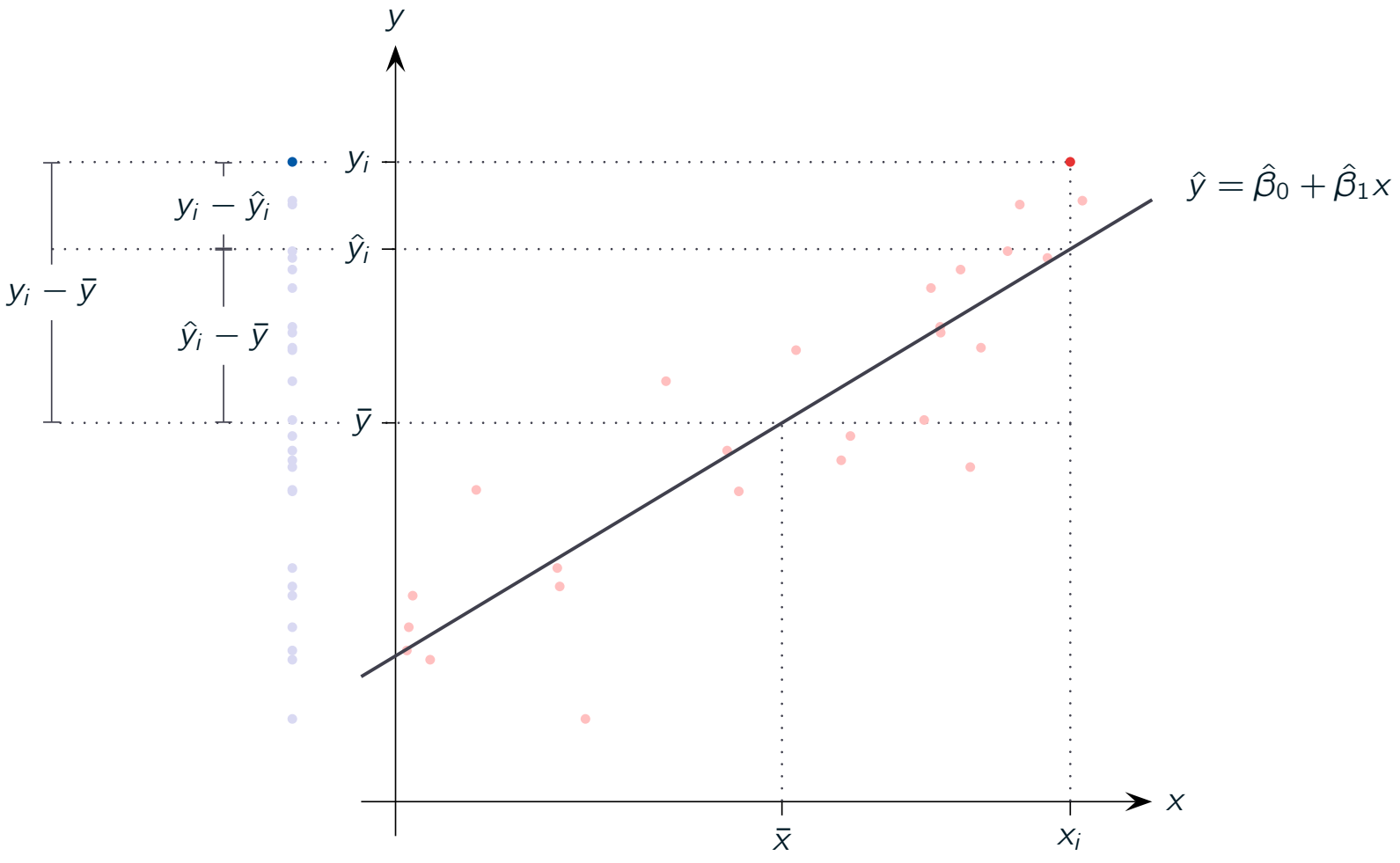
- Regressions are used to explain y
- In particular, we wish to explain why/when is y_i different from $\bar{y}$

- The variation in y can be decomposed as:

$$\begin{aligned} y_i - \bar{y} =& \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i - \beta_0 - \beta_1 \bar{x} \\ =& \underbrace{\beta_1 (x_i - \bar{x})}_{\text{explained}} + \underbrace{e_i}_{\text{unexplained}} \end{aligned}$$

- One way to evaluate estimated models is to measure the proportion of the variance of y that we are able to explain

Example: Variance decomposition



Variance decomposition

$$\begin{aligned}
\text{TSS} &= \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i + e_i - \bar{y})^2 \\
&= \sum (\hat{y}_i - \bar{y})^2 - 2 \sum e_i (\hat{y}_i - \bar{y}) + \sum e_i^2 \qquad \sum e_i (\hat{y}_i - \bar{y}) \to 0 \\
&= \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{explained}} + \underbrace{\sum e_i^2}_{\text{unexplained}}
\end{aligned}$$

Total sum of squares (TSS)	=	Explained sum of squares (ESS)	+	Residual sum of squares (RSS)

Goodness of fit (R^2)

- We have decomposed the total variation (TSS) into the explained variation (ESS) and the unexplained or residual variation (RSS)
- A measure of the explanatory power of the model is the proportion of explained variation

$$\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

- The higher the R^2 the closer the model is to the data
- Since $0 \leq \mathrm{RSS} \leq \mathrm{TSS}$ we know that

$$0 \leq \mathrm{R}^2 \leq 1$$

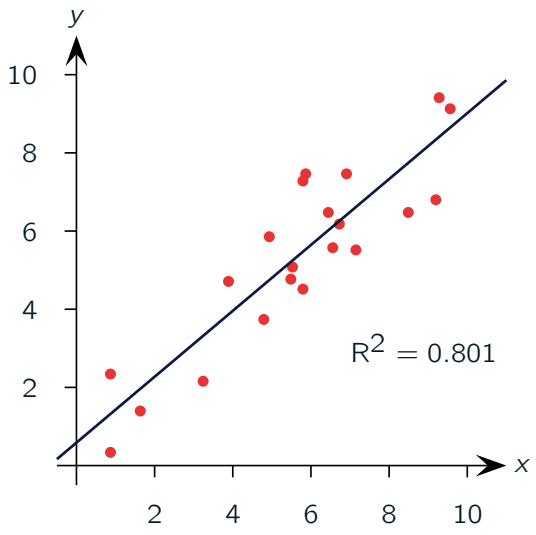
Interpreting the R^2

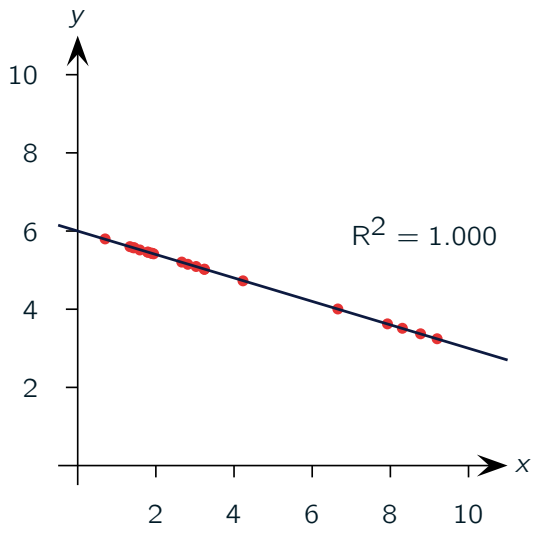
- The R^2 coefficient measures:

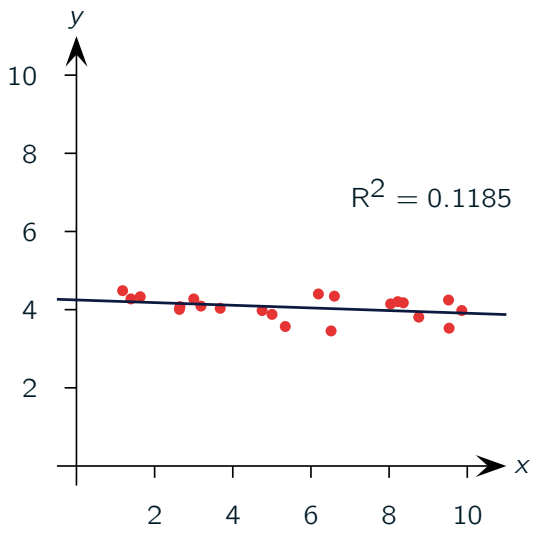
  How much of the variation of y can be explained by the variation of x according to the estimated model

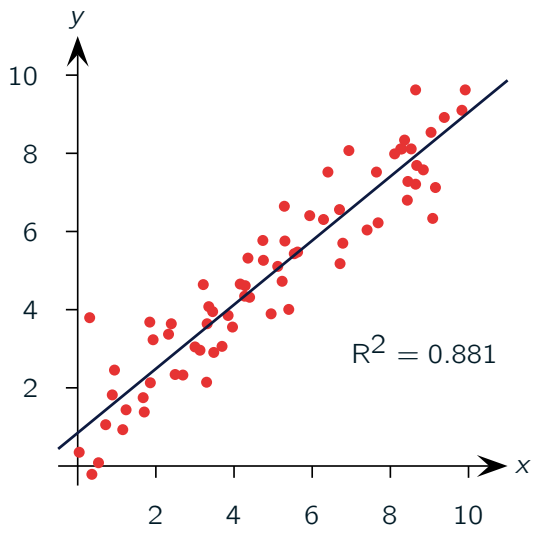
- It does NOT measure:
  - How linear/tight the relation between x ad y is (correlation)
  - The inclination of the estimated line (slope coefficient)
  - The strength of the causal relation between x and y

Examples: R^2


y
10
8
6
4
2
$R^2 = 0.801$
x
2
4
6
8
10


y
10
8
6
4
2
$R^2 = 1.000$
x
2
4
6
8
10


y
10
8
6
4
2
$R^2 = 0.1185$
x
2
4
6
8
10


y
10
8
6
4
2
$R^2 = 0.881$
x
2
4
6
8
10

Example: height and weight

Computing OLS

	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	e_i	e_i^2
1	5	140	135.3	-29.40	864.36	4.70	22.09
2	9	157	160.8	-12.40	153.76	-3.80	14.44
3	13	205	186.3	35.60	1267.36	18.70	349.69
4	12	198	179.9	28.60	817.96	18.00	324
5	10	162	167.2	-7.40	54.76	-5.20	27.04
6	11	174	179.9	4.60	21.16	0.40	0.16
7	8	150	173.6	-19.40	376.36	-4.40	19.36
8	9	165	160.8	-4.40	19.36	4.20	17.64
9	10	170	167.2	0.60	0.36	2.80	7.84
10	12	180	179.9	10.60	112.36	0.00	0
11	11	170	173.7	0.60	0.36	-3.60	12.96
12	9	162	160.8	-7.40	54.76	1.20	1.44
13	10	165	167.2	-4.40	19.36	-2.20	4.84
14	12	180	179.9	10.60	112.36	0.00	0
15	8	160	154.4	-9.40	88.36	5.60	31.36
16	9	155	160.8	-14.40	207.36	-5.80	33.64
17	10	165	167.2	-4.40	19.36	-2.20	4.84
18	15	190	199.1	20.60	424.36	-9.10	82.81
19	13	185	186.3	15.60	243.36	-1.30	1.69
20	11	155	173.6	-14.40	207.36	-18.60	345.96
mean	10	169	170.70	0.00	253.24	0.00	65.09
sum	207	3388	3413.90	0.00	5064.80	0.00	1301.8

$$R^2 = 1 - 1301.8/5064.80 \approx 0.743$$

Adding more regressors

- What would happen to our model if we add a new regressor x_2?
- Recall that OLS minimizes RSS:

$$\text{RSS} = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} \right)^2$$

- Adding a new regressor adds a degree of freedom (we can always set $\hat{\beta}_2 = 0$) and hence always decreases RSS
- This implies that adding a regressor always decreases the R^2 coefficient even if y is almost independent from it!

$\bar{R}^2$ – Adjusted R^2

- Having more data or more variables improves the R^2 because it increases the degrees of freedom
- The adjusted R^2 controls for this bias:

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-k-1)}{\sum (y_i - \bar{y})^2/(n-1)}$$

  where:

  - n – number of observations (sample size)
  - k – number of regressors

- $\bar{R}^2 = R^2$ when $k = 1$ and $\bar{R}^2 \approx R^2$ when n is very large

ANOVA

```
                Number of obs =      32     R-squared     =  0.9214
                Root MSE      = .877971     Adj R-squared =  0.8985

    Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
     Model |      217.00     7       31.00       40.22     0.0000
           |
        x1 |       3.125     1       3.125        4.05     0.0554
        x2 |      194.50     3  64.8333333       84.11     0.0000
        x3 |      19.375     3  6.45833333        8.38     0.0006
           |
  Residual |       18.50    24  .770833333
-----------+----------------------------------------------------
     Total |      235.50    31  7.59677419
```

Example: water supply

Variables

- Response variable:
  - WATER_i – water consumed in period i
- Regressors:
  - PRICE_i – price of water in period i
  - POP_i – population in period i
  - RAIN_i – rainfall during period i

$$\widehat{\text{WATER}}_i = 24000 + 0.62\text{POP} - 400\text{RAIN} \qquad \bar{R}^2 = 0.847$$

$$\widehat{\text{WATER}}_i = 24000 + 48000\text{PRICE} + 0.4\text{POP} - 370\text{RAIN} \qquad \bar{R}^2 = 0.859$$