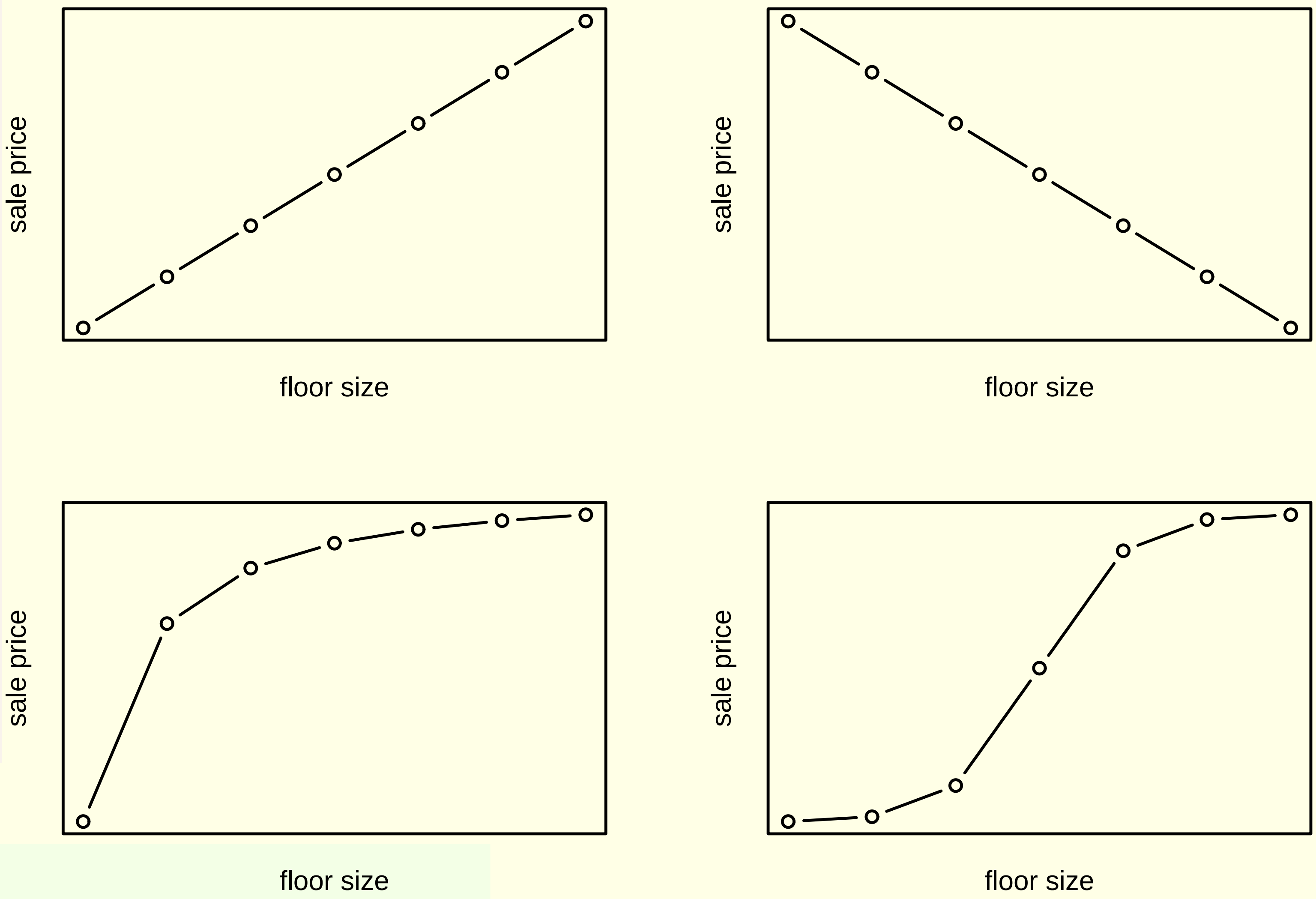
Applied Regression Modeling: A Business Approach
Chapter 2: Simple Linear Regression
Sections 2.1–2.3

by Iain Pardoe

Simple linear regression model for X and Y

- Y is a quantitative *response* variable (a.k.a. dependent, outcome, or output variable).
- X is a quantitative *predictor* variable (a.k.a. independent or input variable, or covariate).
- Two variables play different roles, so important to identify which is which and define carefully, e.g.:
  - Y is sale price, in $ thousands;
  - X is floor size, in thousands of square feet.
- How much do we expect Y to change by when we change the value of X?
- What do we expect the value of Y to be when we set the value of X at 2?
- Note: *association* (observational data) not *causation* (experimental data).

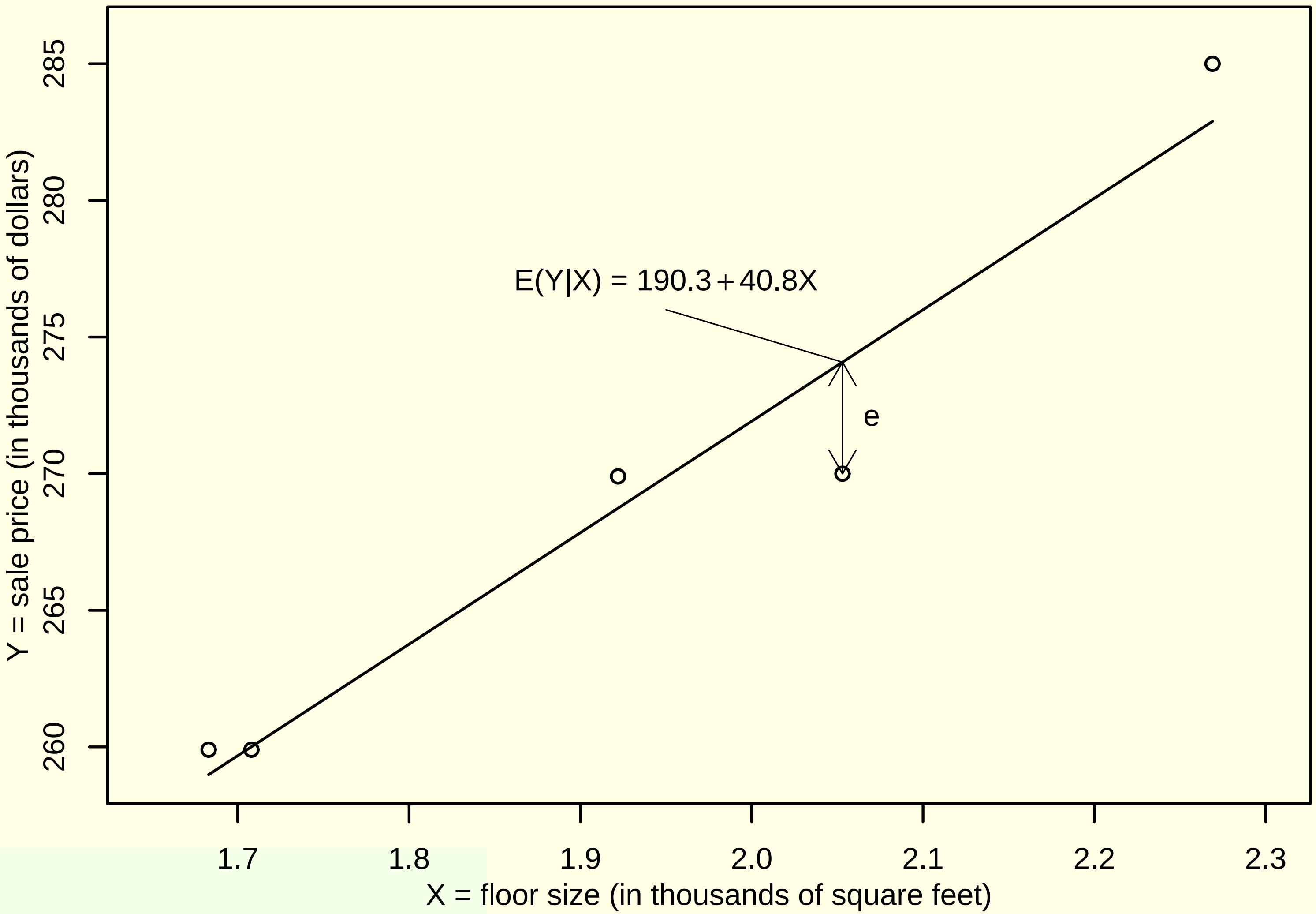
Which factors might lead to the different relationships?



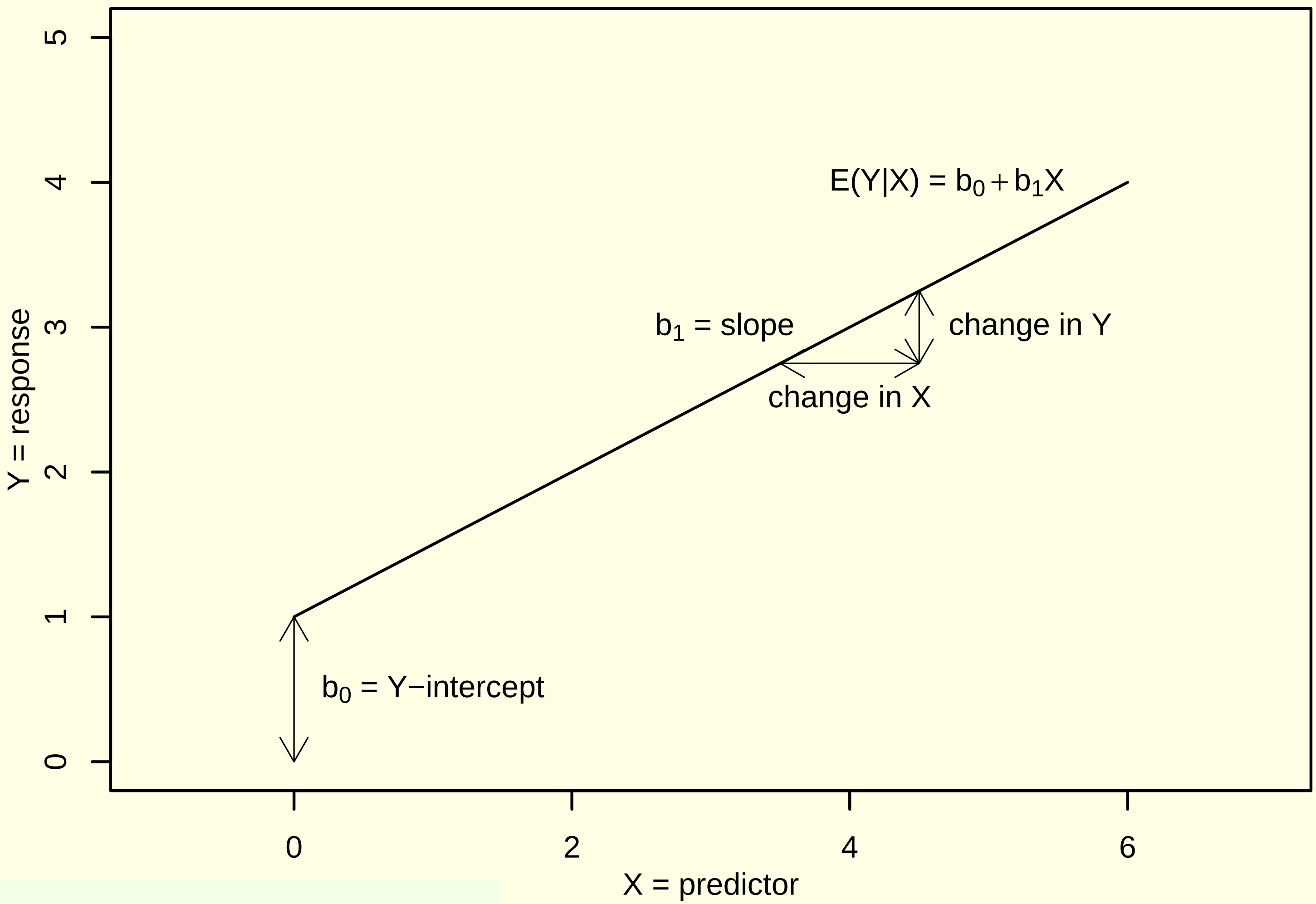
- Simple linear regression models straight-line relationships (like upper two plots on last slide).
- Suppose sale price is (on average) $\$190{,}300$ plus 40.8 times floor size.
  - $\mathrm{E}(Y \mid X_i) = 190.3 + 40.8X_i$, where $\mathrm{E}(Y \mid X_i)$ means "the expected value of Y given that X is equal to X_i".
- Individual sale prices can deviate from this expected value by an amount e_i (called a "random error").
  - $Y_i \mid X_i = 190.3 + 40.8X_i + e_i \quad (i = 1, \ldots, n)$.
  - $Y_i \mid X_i =$ deterministic part $+$ random error.
- Error, e_i, represents variation in Y due to factors other than X which we haven't measured, e.g., lot size, # beds/baths, age, garage, schools.

HOMES2 data

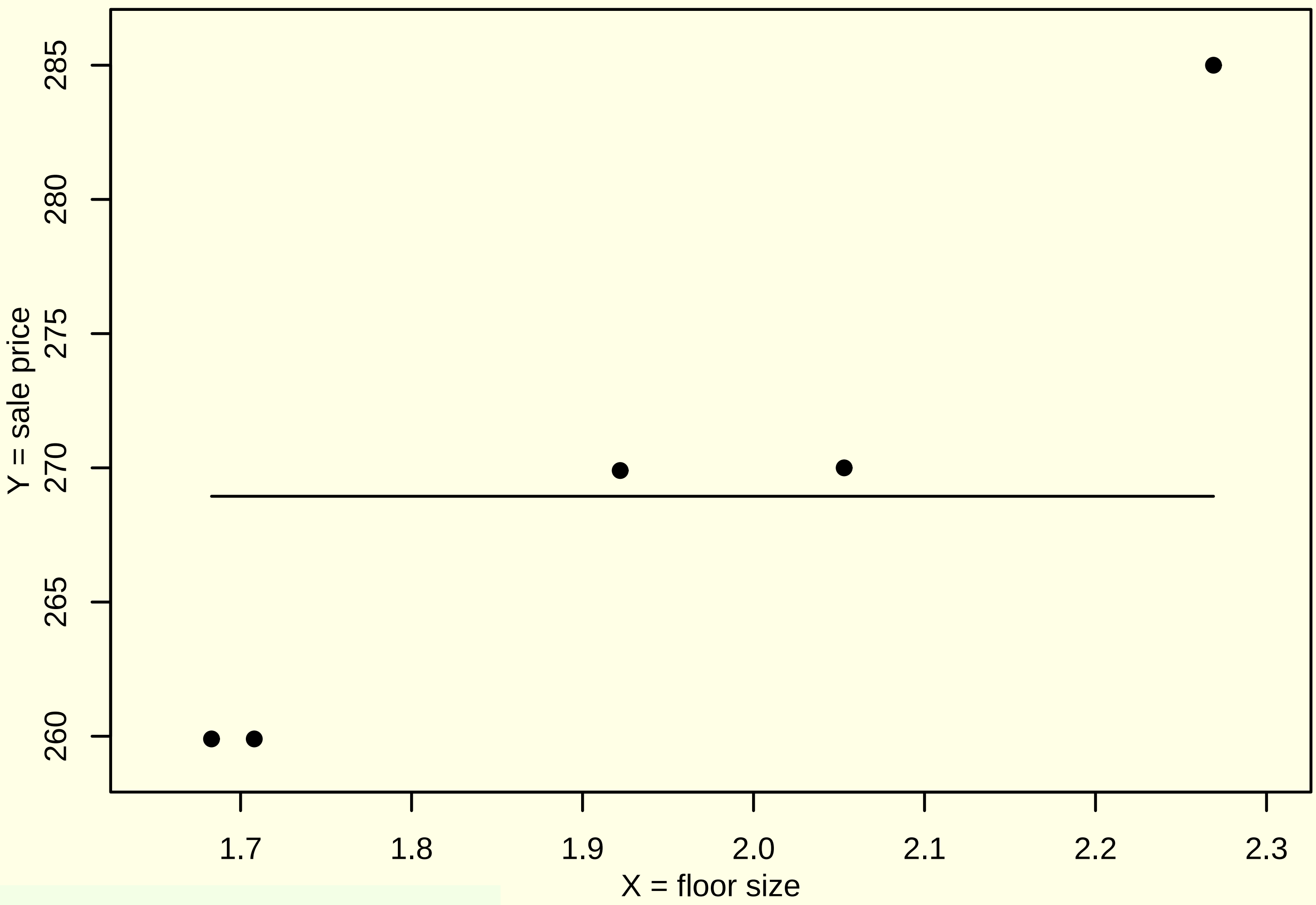
Y	259.9	259.9	269.9	270.0	285.0
X	1.683	1.708	1.922	2.053	2.269



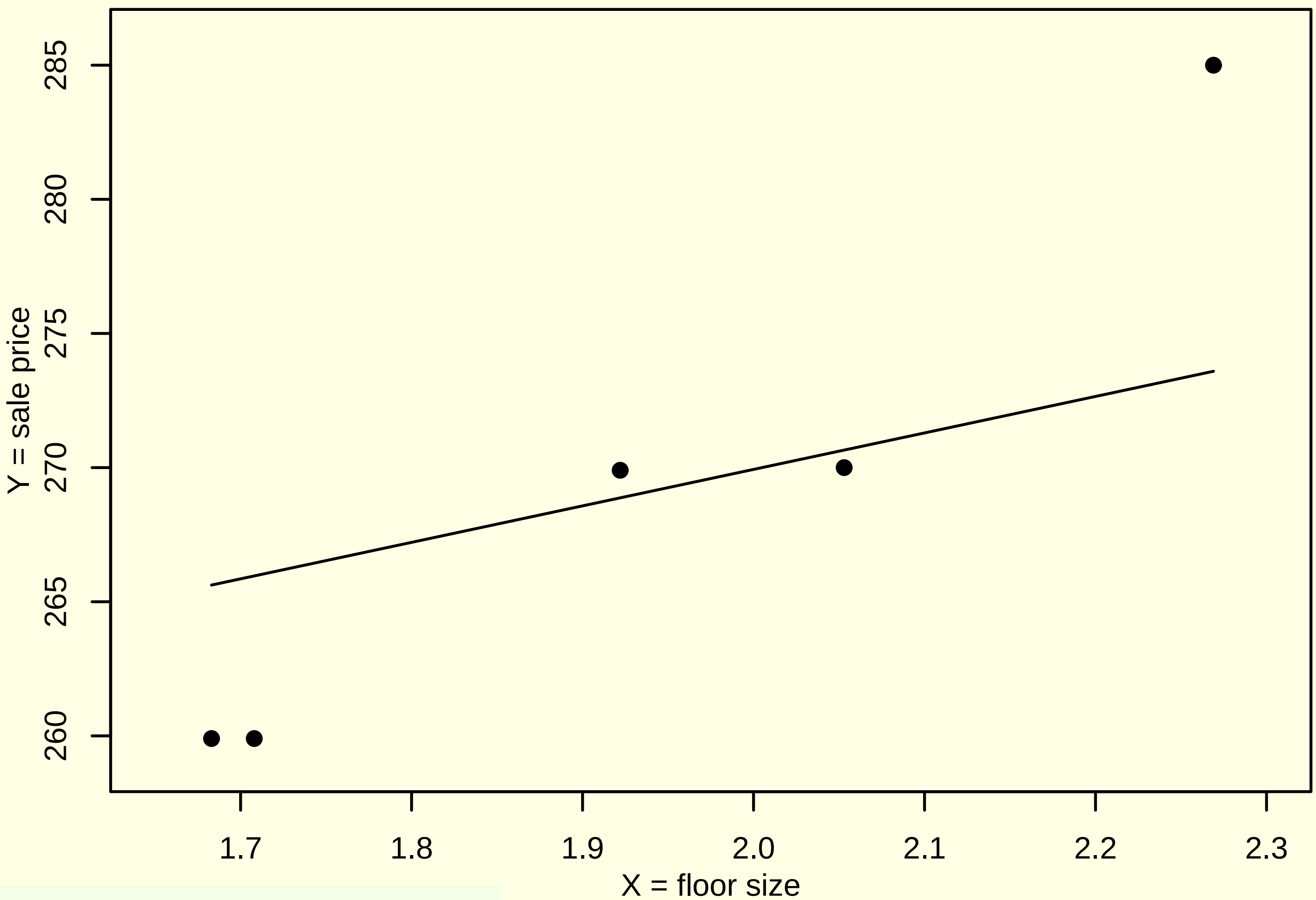
Population: $\mathrm{E}(Y\,|\,X) = b_0 + b_1 X.$



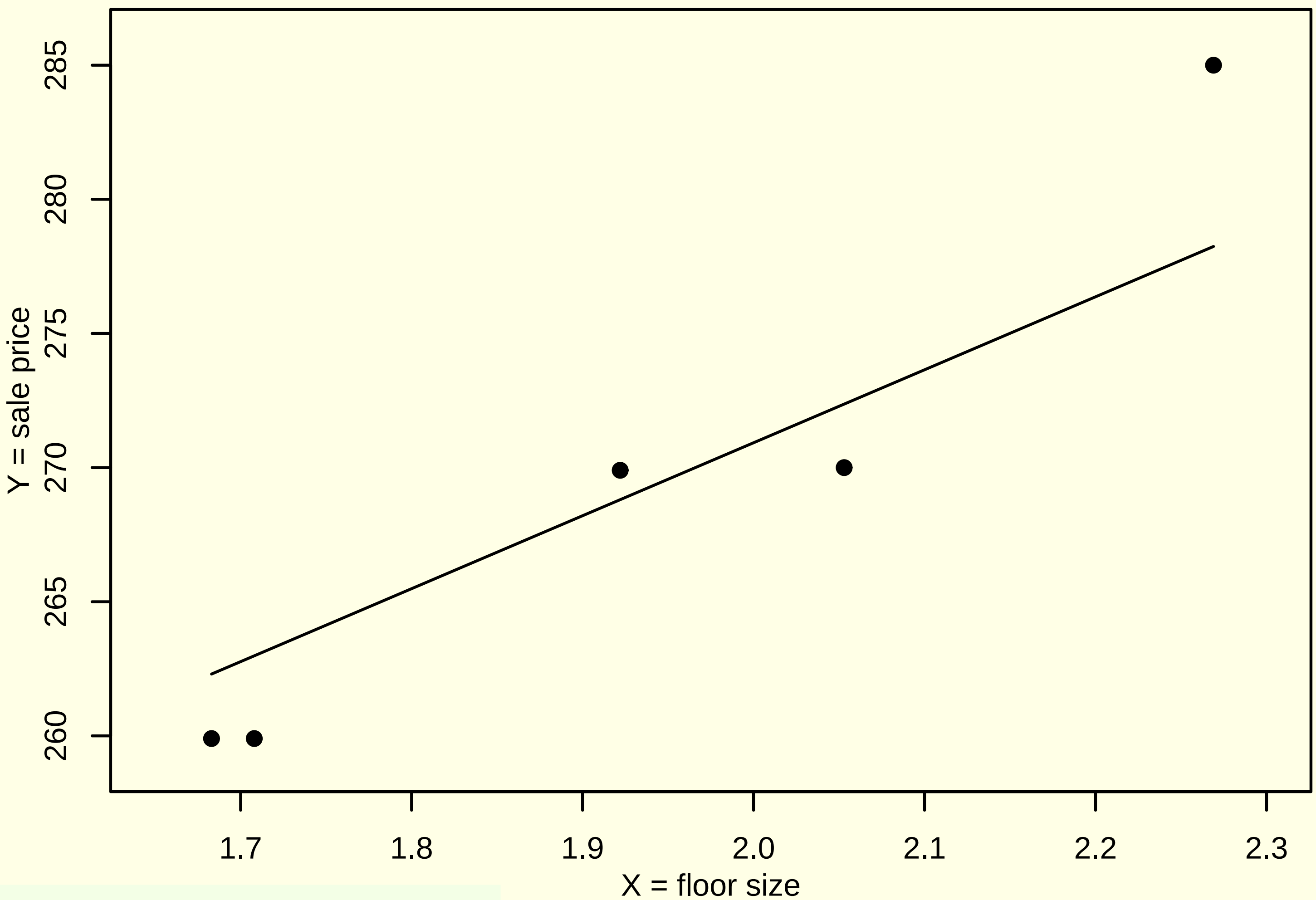
Which line fits the data best?



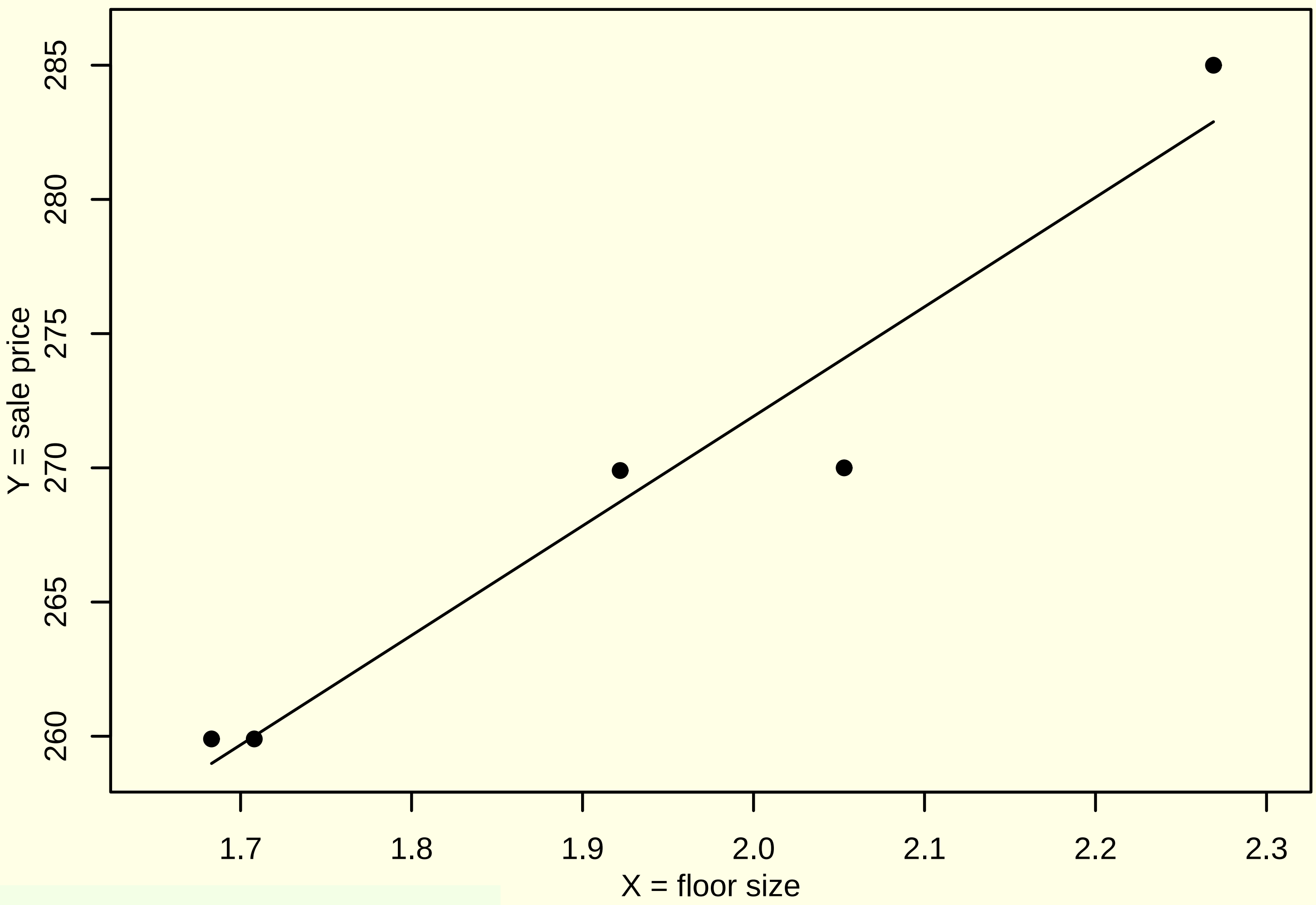
Which line fits the data best?



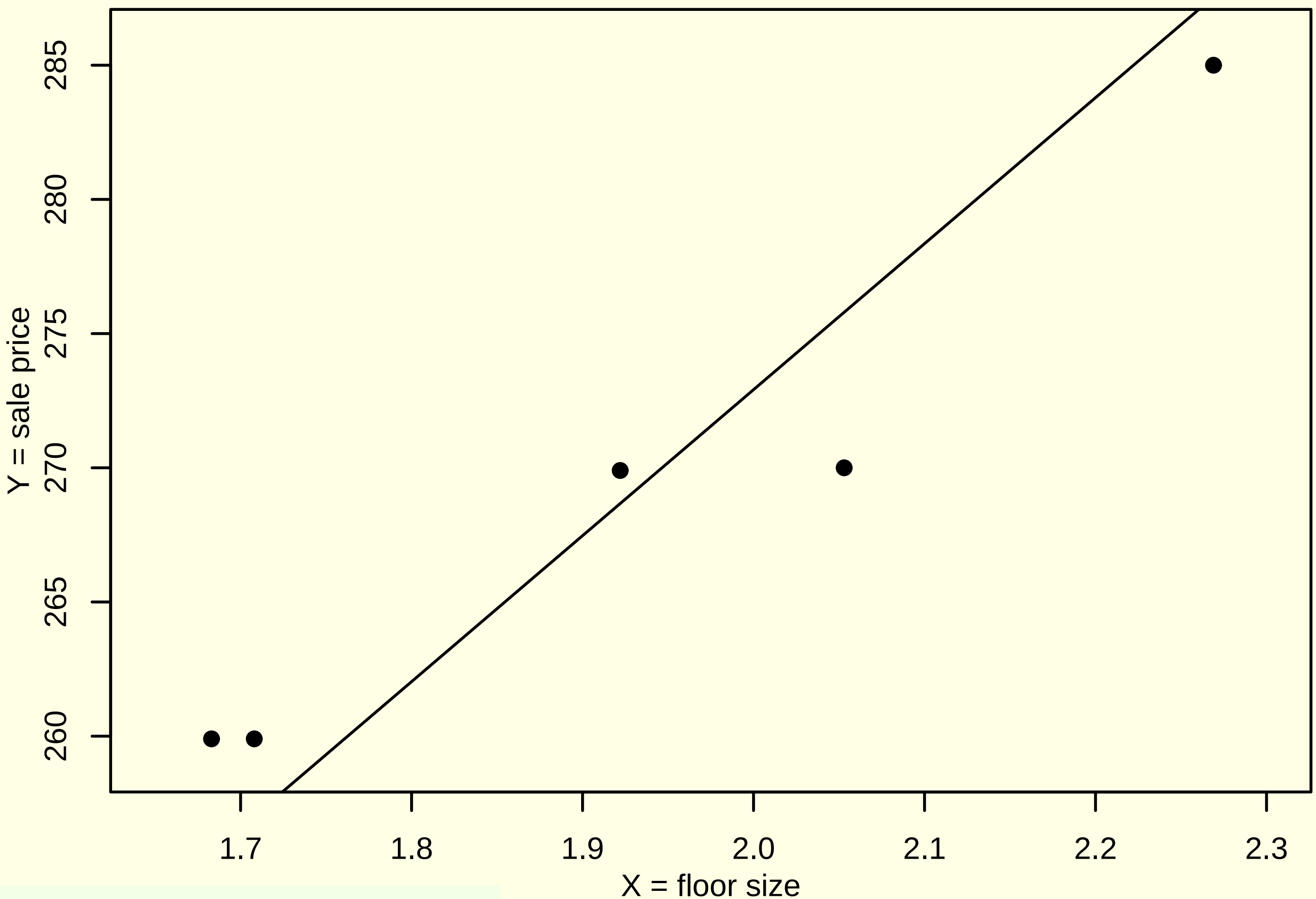
Which line fits the data best?



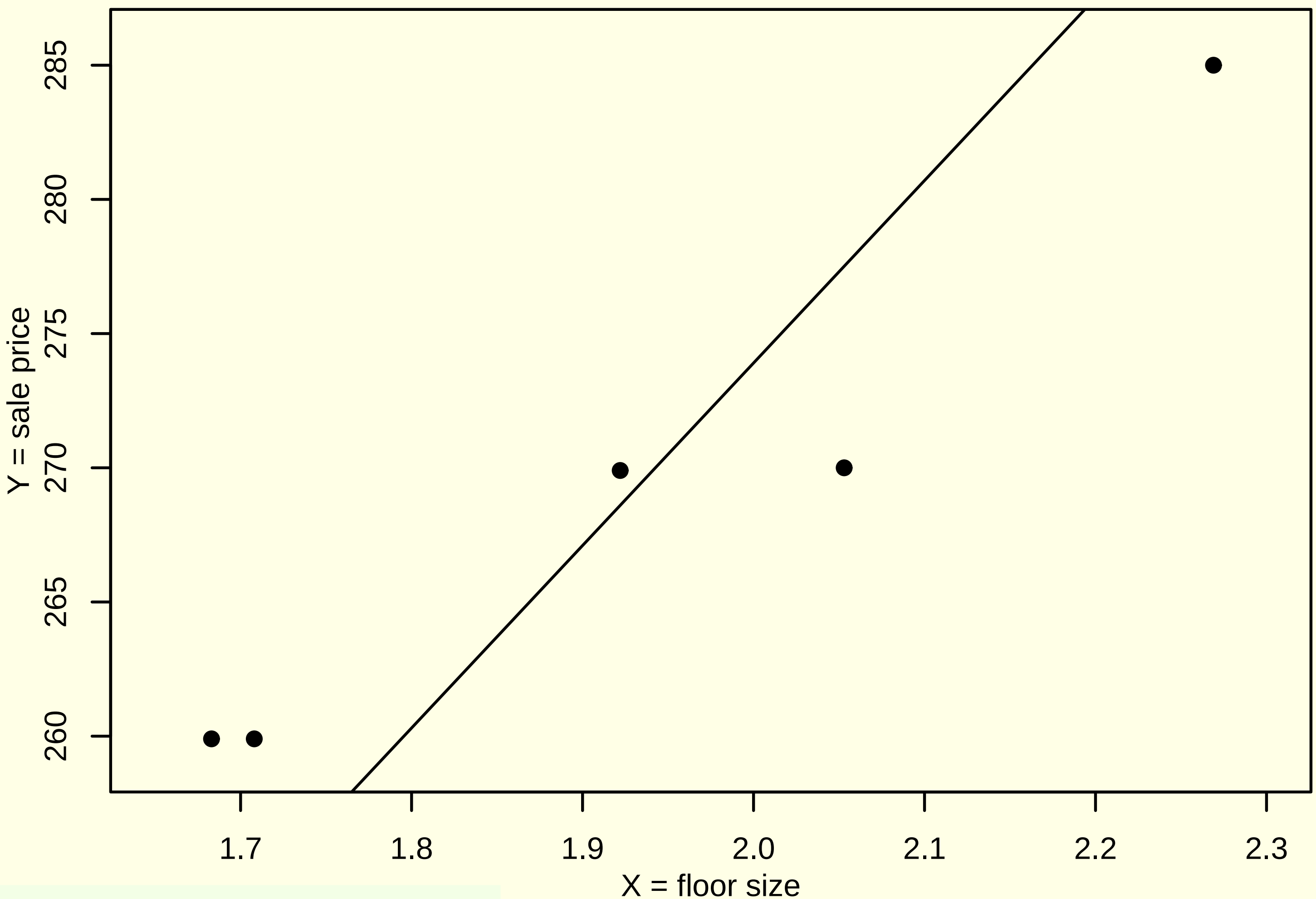
Which line fits the data best?



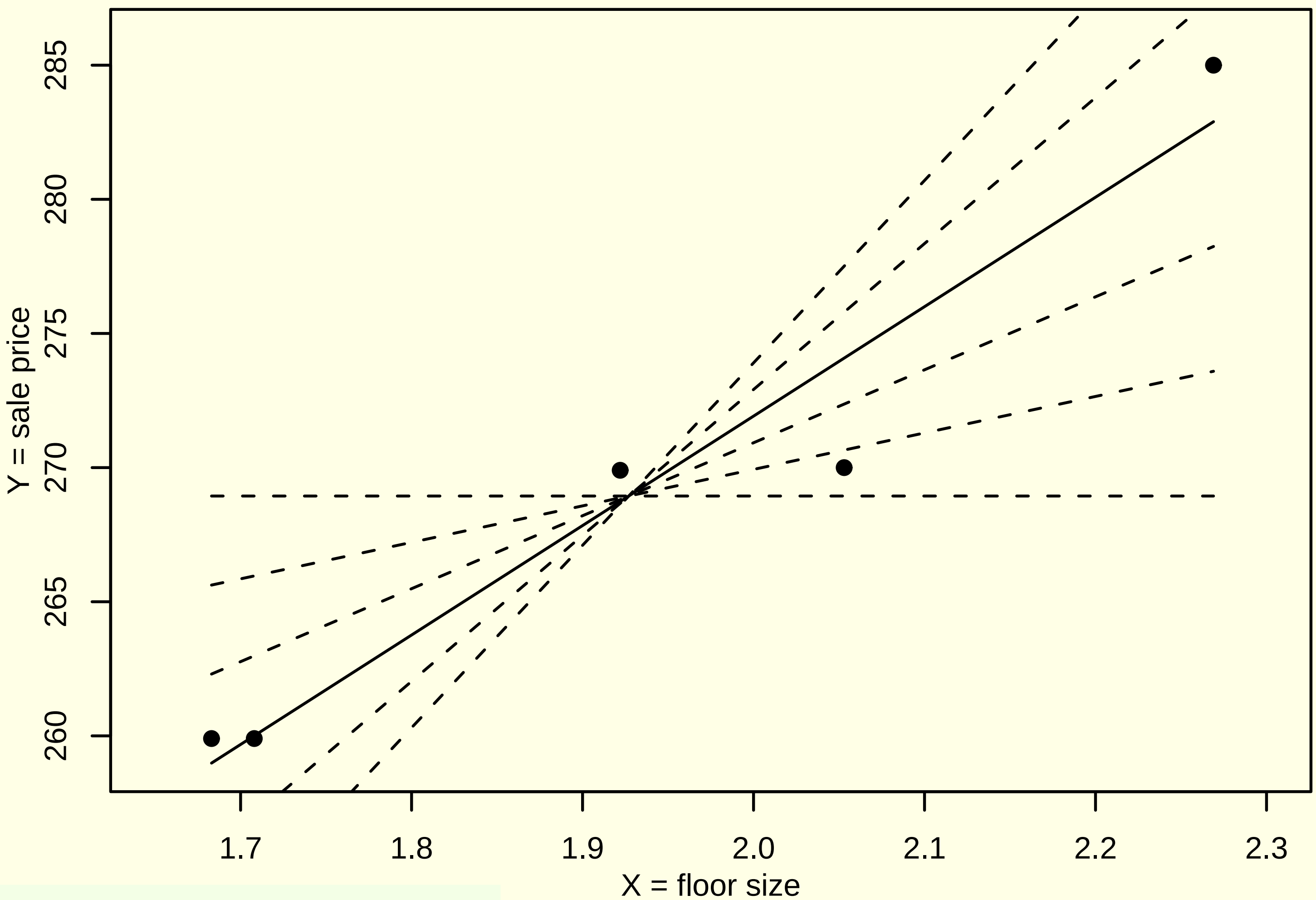
Which line fits the data best?



Which line fits the data best?



Which line fits the data best?



Estimating the model

- Population: $\mathrm{E}(Y \mid X) = b_0 + b_1 X$.
- Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ (estimated model).
- Obtain $\hat{b}_0$ and $\hat{b}_1$ by finding best fit line (least squares line).
- Mathematically, minimize sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{n} \hat{e}_i^2$$

Estimating the model

- Population: $\mathrm{E}(Y \mid X) = b_0 + b_1 X$.
- Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ (estimated model).
- Obtain $\hat{b}_0$ and $\hat{b}_1$ by finding best fit line (least squares line).
- Mathematically, minimize sum of squared errors:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} \hat{e}_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \end{aligned}$$

Estimating the model

- Population: $\mathrm{E}(Y \mid X) = b_0 + b_1 X$.
- Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ (estimated model).
- Obtain $\hat{b}_0$ and $\hat{b}_1$ by finding best fit line (least squares line).
- Mathematically, minimize sum of squared errors:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} \hat{e}_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2. \end{aligned}$$

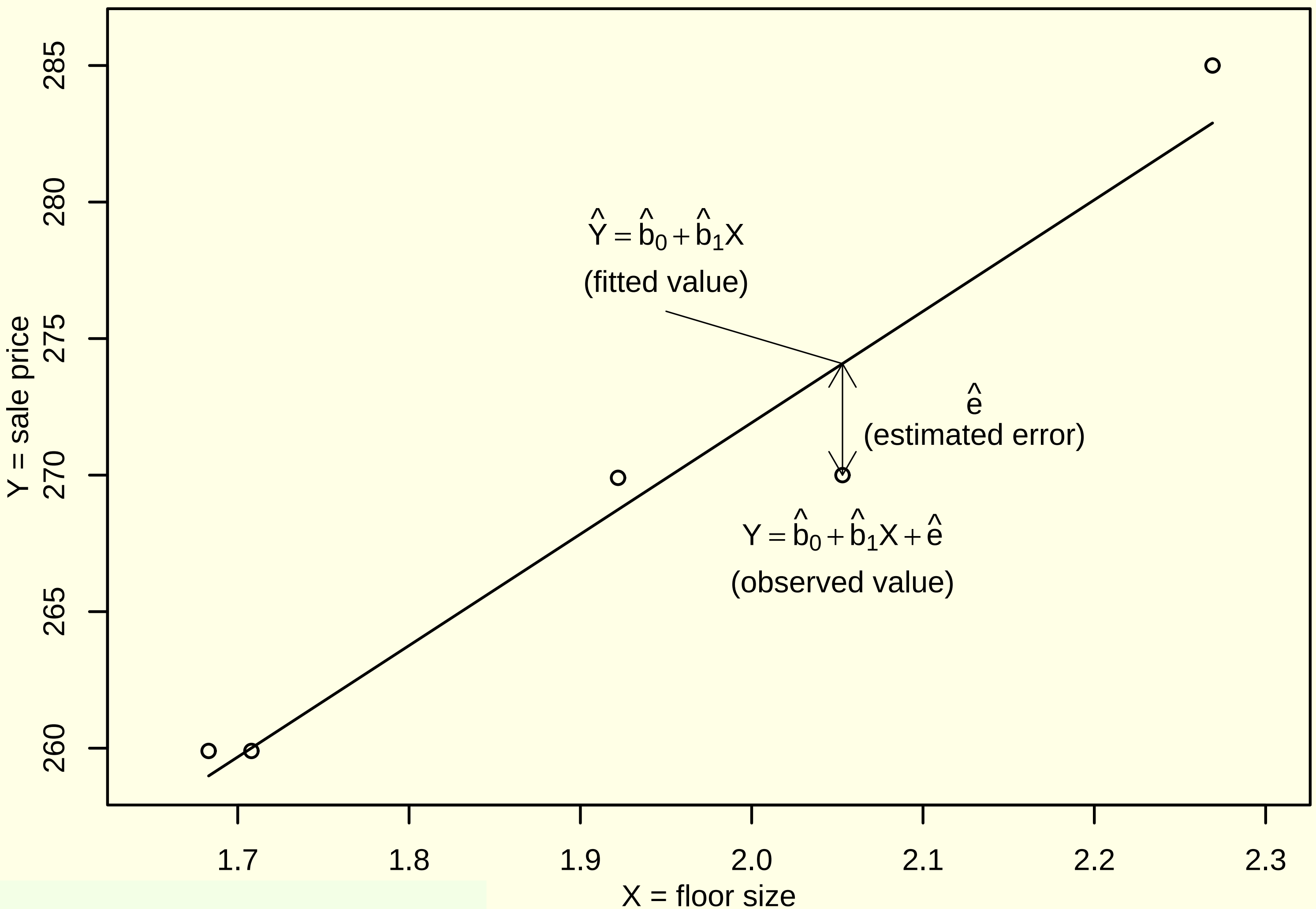
Estimating the model

- Population: $\mathrm{E}(Y \mid X) = b_0 + b_1 X$.
- Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ (estimated model).
- Obtain $\hat{b}_0$ and $\hat{b}_1$ by finding best fit line (least squares line).
- Mathematically, minimize sum of squared errors:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} \hat{e}_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2. \end{aligned}$$

- Can use calculus (partial derivatives), but we'll use computer software to find $\hat{b}_0$ and $\hat{b}_1$.

Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$.



Computer output

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(> |t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

[a] Response variable: Y.

- Estimated equation:
  $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.

Computer output

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

[a] Response variable: Y.

- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- We expect $Y = \hat{b}_0$ when $X = 0$, but *only* if this makes sense and we have data close to $X = 0$

Computer output

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

[a] Response variable: Y.

- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- We expect $Y = \hat{b}_0$ when $X = 0$, but *only* if this makes sense and we have data close to $X = 0$ (not the case here).

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(> |t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

[a] Response variable: Y.

- Estimated equation:
  $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- We expect $Y = \hat{b}_0$ when $X = 0$, but *only* if this makes sense and we have data close to $X = 0$ (not the case here).
- We expect Y to change by $\hat{b}_1$ when X increases by one unit, i.e., we expect sale price to increase by $\$40{,}800$ when floor size increases by 1000 sq. feet.

Parameters [a]

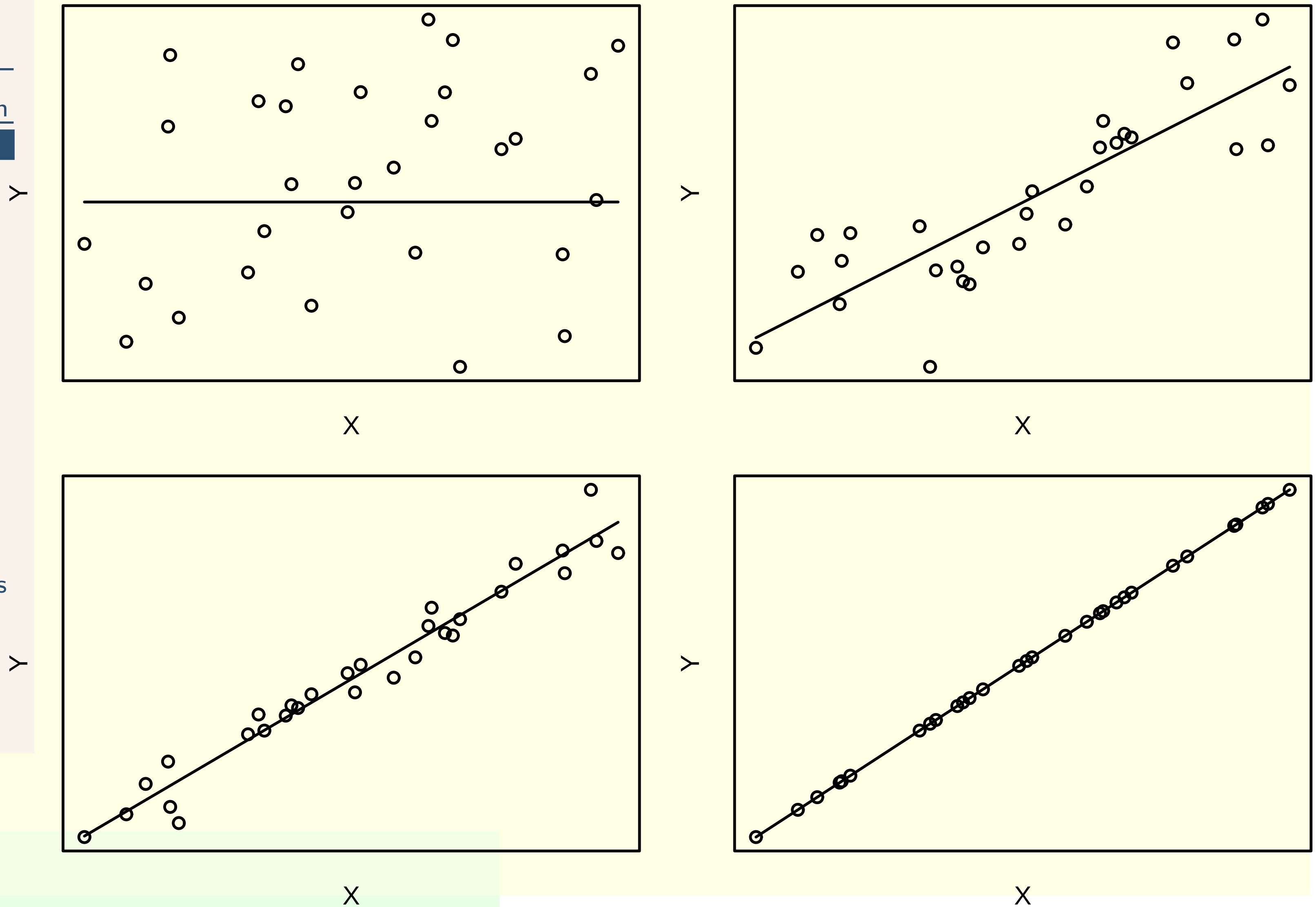
Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

[a] Response variable: Y.

- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- We expect $Y = \hat{b}_0$ when $X = 0$, but *only* if this makes sense and we have data close to $X = 0$ (not the case here).
- We expect Y to change by $\hat{b}_1$ when X increases by one unit, i.e., we expect sale price to increase by $\$40{,}800$ when floor size increases by 1000 sq. feet.
- For this example, more meaningful to say we expect sale price to increase by $\$4080$ when floor size increases by 100 sq. feet.

Model evaluation

How well does the model fit each dataset?



Three methods:

- How close are the actual observed Y-values to the model-based fitted values, $\hat{Y}$?
  - Calculate the *regression standard error*, s.

Evaluating fit numerically

Three methods:

- How close are the actual observed Y-values to the model-based fitted values, $\hat{Y}$?
  - Calculate the *regression standard error*, s.
- How much of the variability in Y have we been able to explain with our model?
  - Calculate the *coefficient of determination*, R^2.

Evaluating fit numerically

Three methods:

- How close are the actual observed Y-values to the model-based fitted values, $\hat{Y}$?
  - Calculate the *regression standard error*, s.
- How much of the variability in Y have we been able to explain with our model?
  - Calculate the *coefficient of determination*, R^2.
- How strong is the evidence of a straight-line relationship between Y and X?
  - Estimate and test the *slope parameter*, b_1.

Regression standard error, s

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972^a	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

- Regression standard error, s, estimates the std. dev. of the simple linear regression random errors:

$$s = \sqrt{\frac{\text{SSE}}{n-2}}.$$

- Unit of measurement for s is the same as unit of measurement for Y.
- Approximately 95% of the observed Y-values lie within plus or minus $2s$ of their fitted $\hat{Y}$-values.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 [a]	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

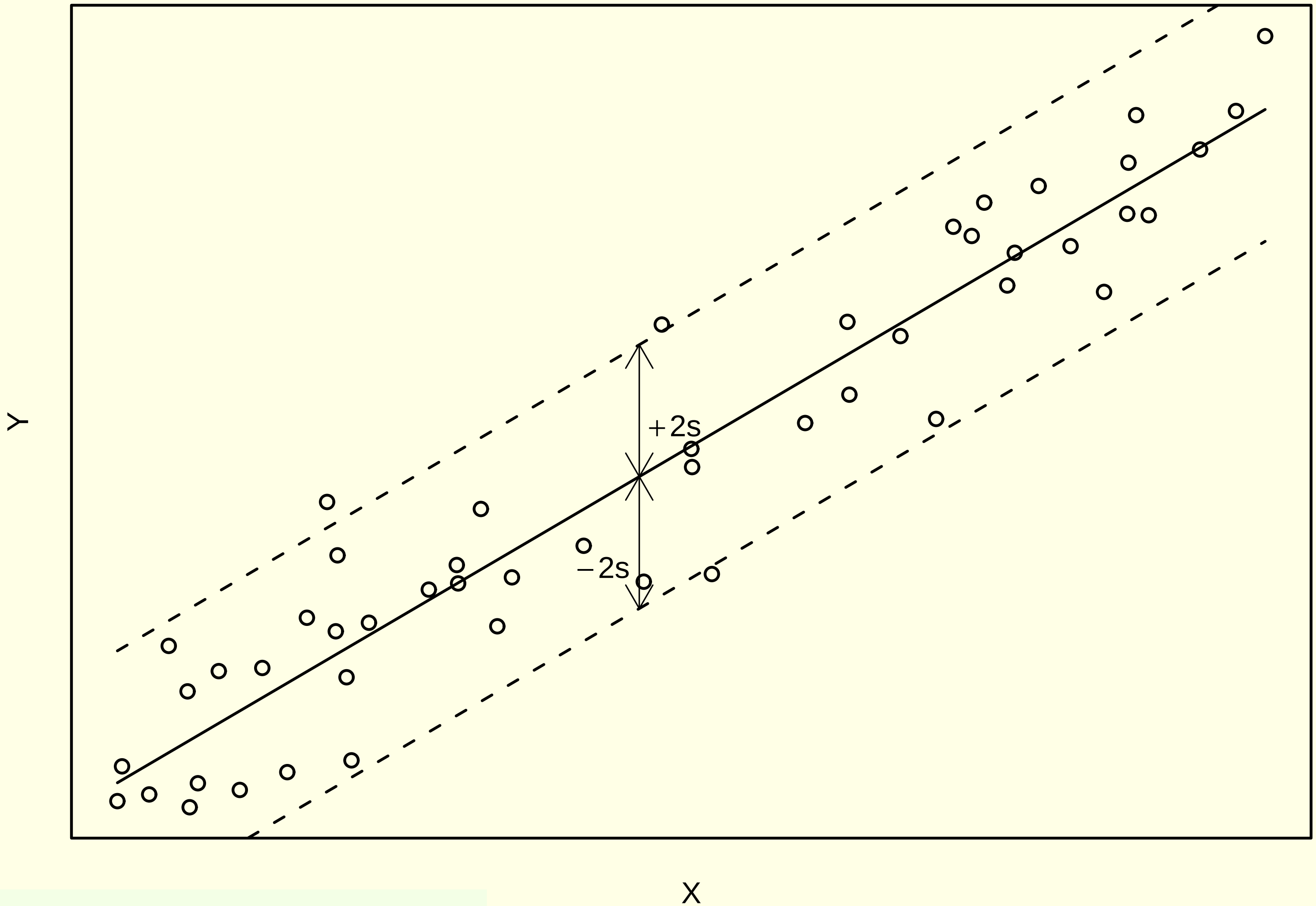
- Regression standard error, s, estimates the std. dev. of the simple linear regression random errors:

$$s = \sqrt{\frac{\text{SSE}}{n-2}}.$$

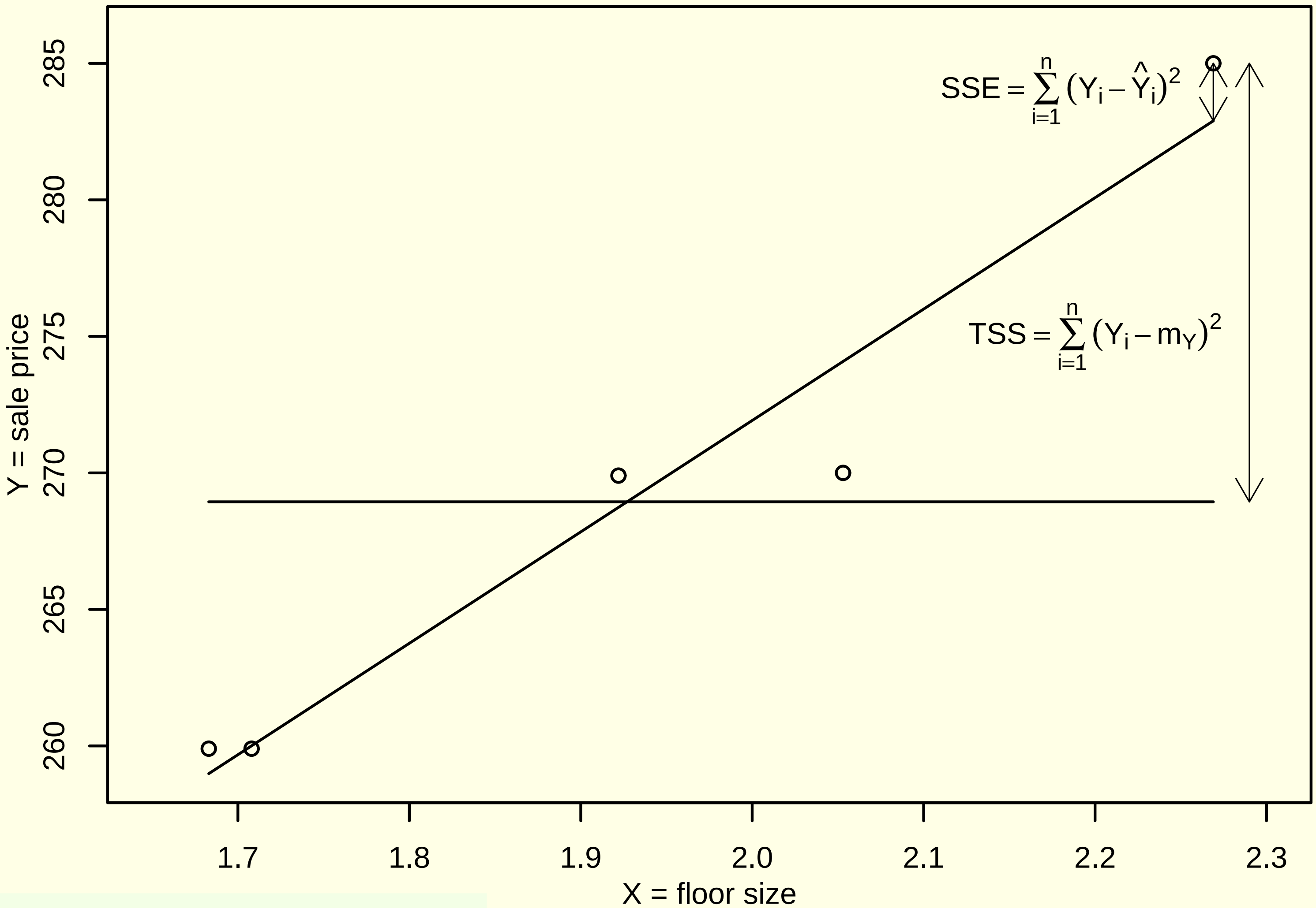
- Unit of measurement for s is the same as unit of measurement for Y.
- Approximately 95% of the observed Y-values lie within plus or minus $2s$ of their fitted $\hat{Y}$-values.
- Since $2s = 5.57$, we can expect to predict an unobserved sale price from a particular floor size to within approx. $\pm\$5570$ (at a 95% confidence level).

Regression standard error interpretation

CLT: 95% of Y-values lie within $\pm 2s$ of regression line.



Measures of variation for simple linear regression.



Calculating R^2

- Without model, estimate Y with sample mean m_Y.
- With model, estimate Y using fitted $\hat{Y}$-value.
- How much do we reduce our error when we do this?

- Without model, estimate Y with sample mean m_Y.
- With model, estimate Y using fitted $\hat{Y}$-value.
- How much do we reduce our error when we do this?
- Total error without model:
  TSS $= \sum_{i=1}^{n}(Y_i - m_Y)^2$, variation in Y about m_Y.
- Remaining error with model:
  SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, unexplained variation.

- Without model, estimate Y with sample mean m_Y.
- With model, estimate Y using fitted $\hat{Y}$-value.
- How much do we reduce our error when we do this?
- Total error without model:
  $\text{TSS} = \sum_{i=1}^{n}(Y_i - m_Y)^2$, variation in Y about m_Y.
- Remaining error with model:
  $\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, unexplained variation.
- Proportional reduction in error: $R^2 = \frac{\text{TSS}-\text{SSE}}{\text{TSS}}$.

- Without model, estimate Y with sample mean m_Y.
- With model, estimate Y using fitted $\hat{Y}$-value.
- How much do we reduce our error when we do this?
- Total error without model:
  TSS $= \sum_{i=1}^{n}(Y_i - m_Y)^2$, variation in Y about m_Y.
- Remaining error with model:
  SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, unexplained variation.
- Proportional reduction in error: $\mathrm{R}^2 = \frac{\mathrm{TSS}-\mathrm{SSE}}{\mathrm{TSS}}$.
- Home prices example: $\mathrm{R}^2 = \frac{423.4-23.3}{423.4} = 0.945$.

- Without model, estimate Y with sample mean m_Y.
- With model, estimate Y using fitted $\hat{Y}$-value.
- How much do we reduce our error when we do this?
- Total error without model:
  TSS $= \sum_{i=1}^{n}(Y_i - m_Y)^2$, variation in Y about m_Y.
- Remaining error with model:
  SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, unexplained variation.
- Proportional reduction in error: $R^2 = \frac{\text{TSS}-\text{SSE}}{\text{TSS}}$.
- Home prices example: $R^2 = \frac{423.4-23.3}{423.4} = 0.945$.
- 94.5% of the variation in sale price (about its mean) can be explained by a straight-line relationship between sale price and floor size.

Interpreting R^2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 [a]	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

- R^2 measures the proportion of variation in Y (about its mean) that can be explained by a straight-line relationship between Y and X.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 [a]	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

- R^2 measures the proportion of variation in Y (about its mean) that can be explained by a straight-line relationship between Y and X.
- If TSS = SSE then $R^2 = 0$: using X to predict Y hasn't helped and we might as well use m_Y to predict Y regardless of the value of X.

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 [a]	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

- R^2 measures the proportion of variation in Y (about its mean) that can be explained by a straight-line relationship between Y and X.
- If TSS $=$ SSE then $R^2 = 0$: using X to predict Y hasn't helped and we might as well use m_Y to predict Y regardless of the value of X.
- If SSE $= 0$ then $R^2 = 1$: using X allows us to predict Y perfectly (with no random errors).

Interpreting R^2

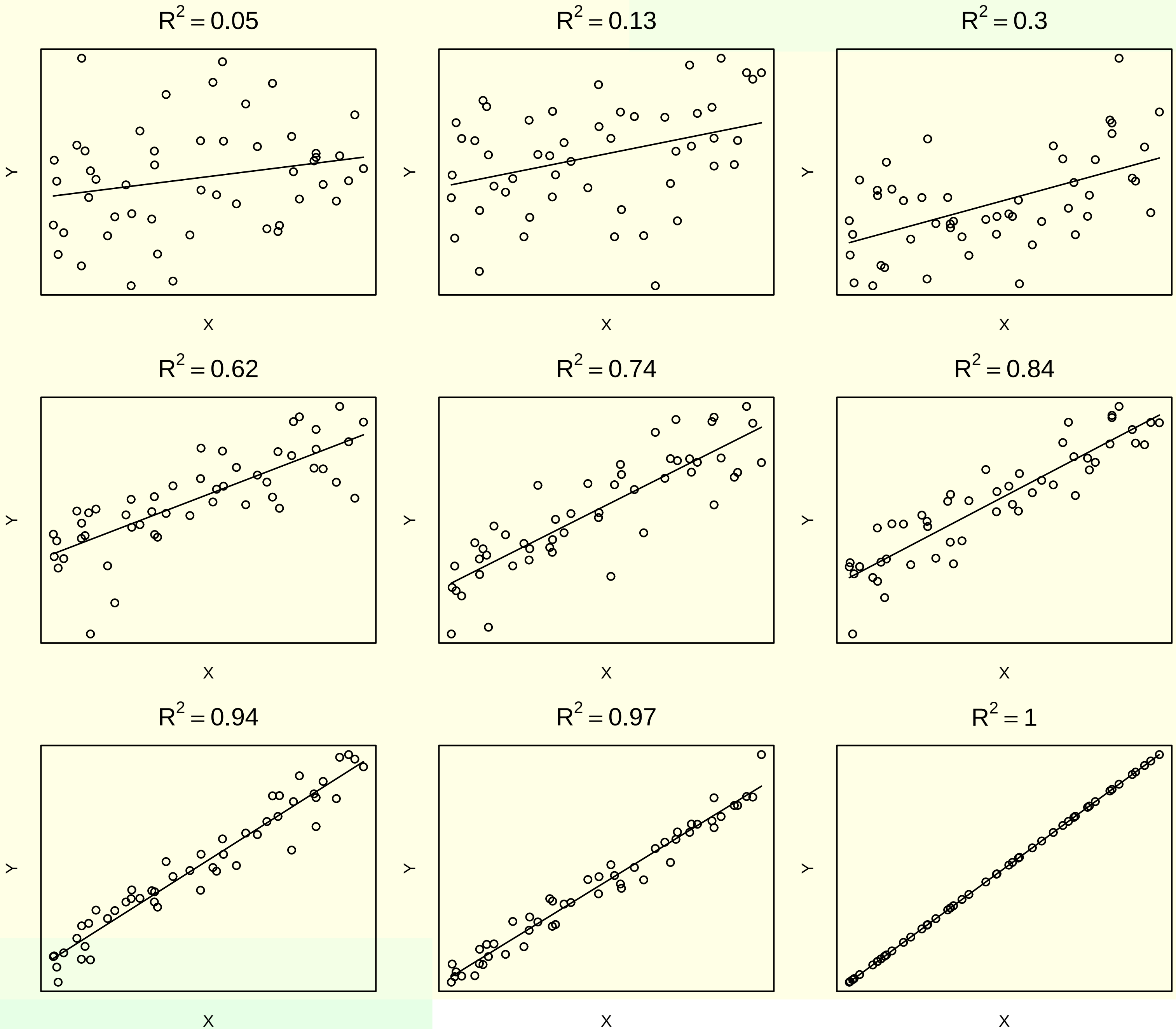
Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 [a]	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

- R^2 measures the proportion of variation in Y (about its mean) that can be explained by a straight-line relationship between Y and X.
- If TSS = SSE then $R^2 = 0$: using X to predict Y hasn't helped and we might as well use m_Y to predict Y regardless of the value of X.
- If SSE $= 0$ then $R^2 = 1$: using X allows us to predict Y perfectly (with no random errors).
- Such extremes rarely occur and usually R^2 lies between zero and one, with higher values of R^2 corresponding to better fitting models.

R^2 examples



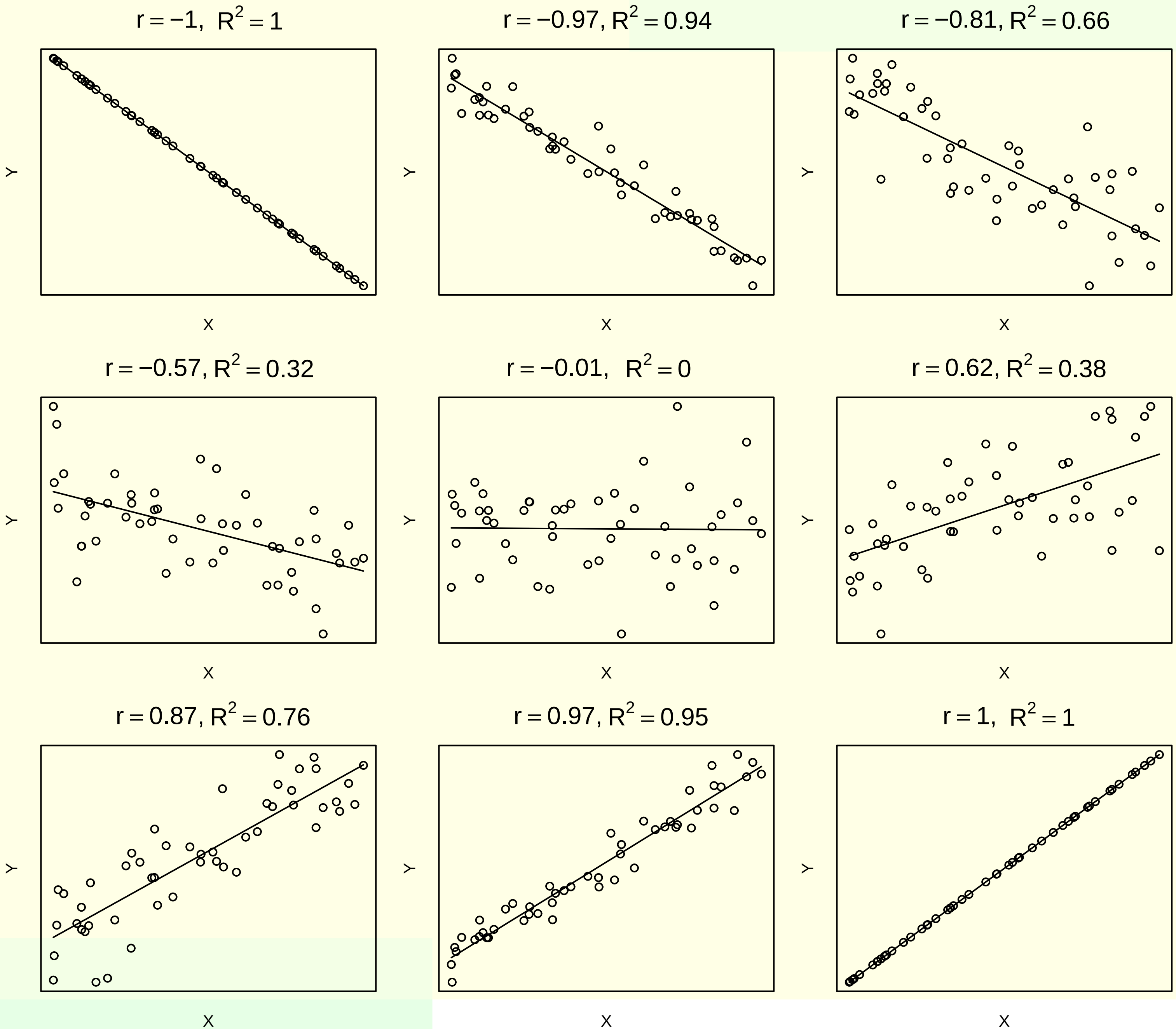
Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972^{a}	0.945	0.927	2.7865

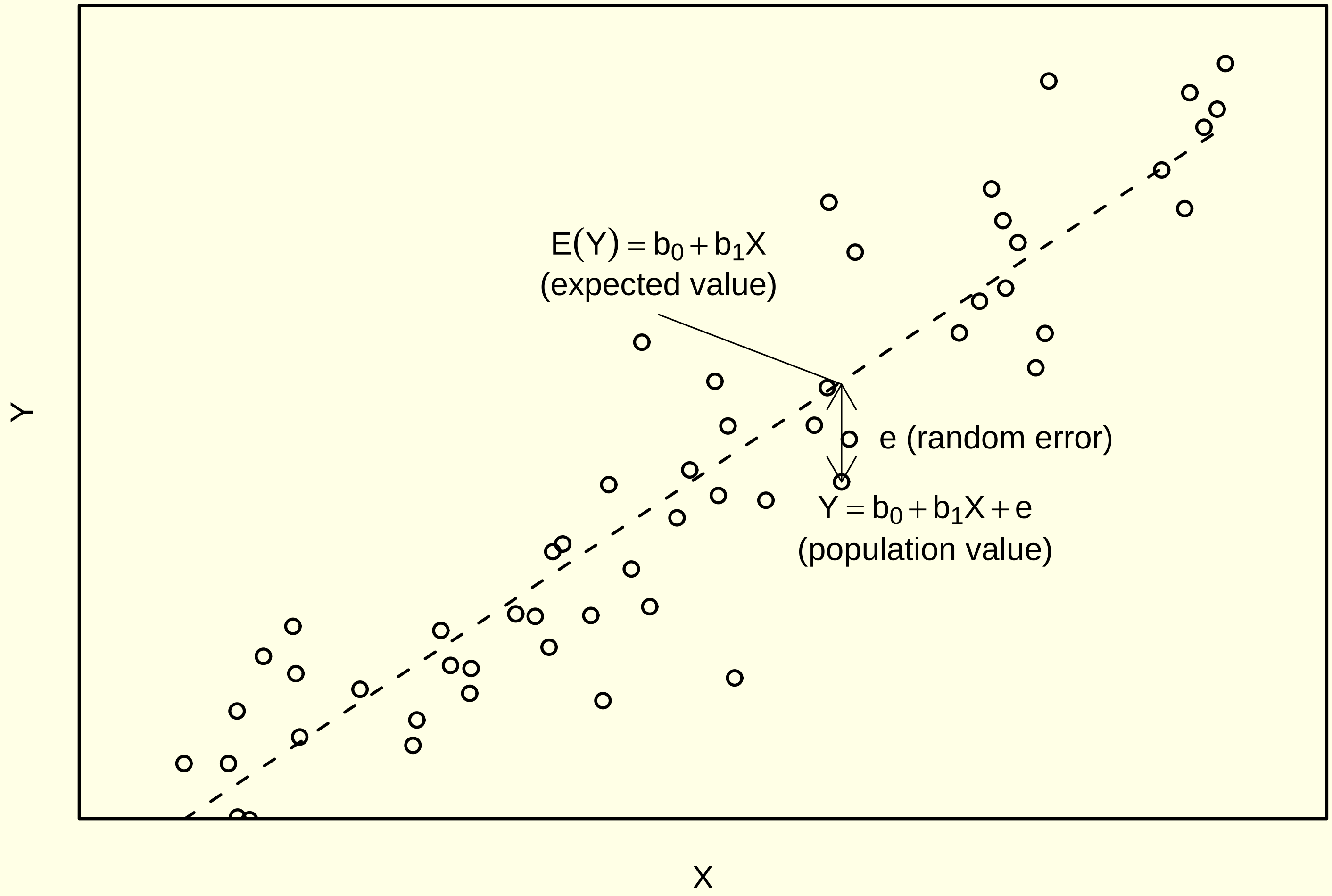
[a] Predictors: (Intercept), X.

- Correlation coefficient, r, measures the strength and direction of linear association between Y and X:
  - $r \approx -1$ indicates a negative linear relationship;
  - $r \approx +1$ indicates a positive linear relationship;
  - $r \approx 0$ indicates no *linear* relationship.
- Simple linear regression: $\sqrt{R^2}$ = absolute value of r ("multiple R" above).
- But, correlation is less useful than R^2 in *multiple* linear regression.

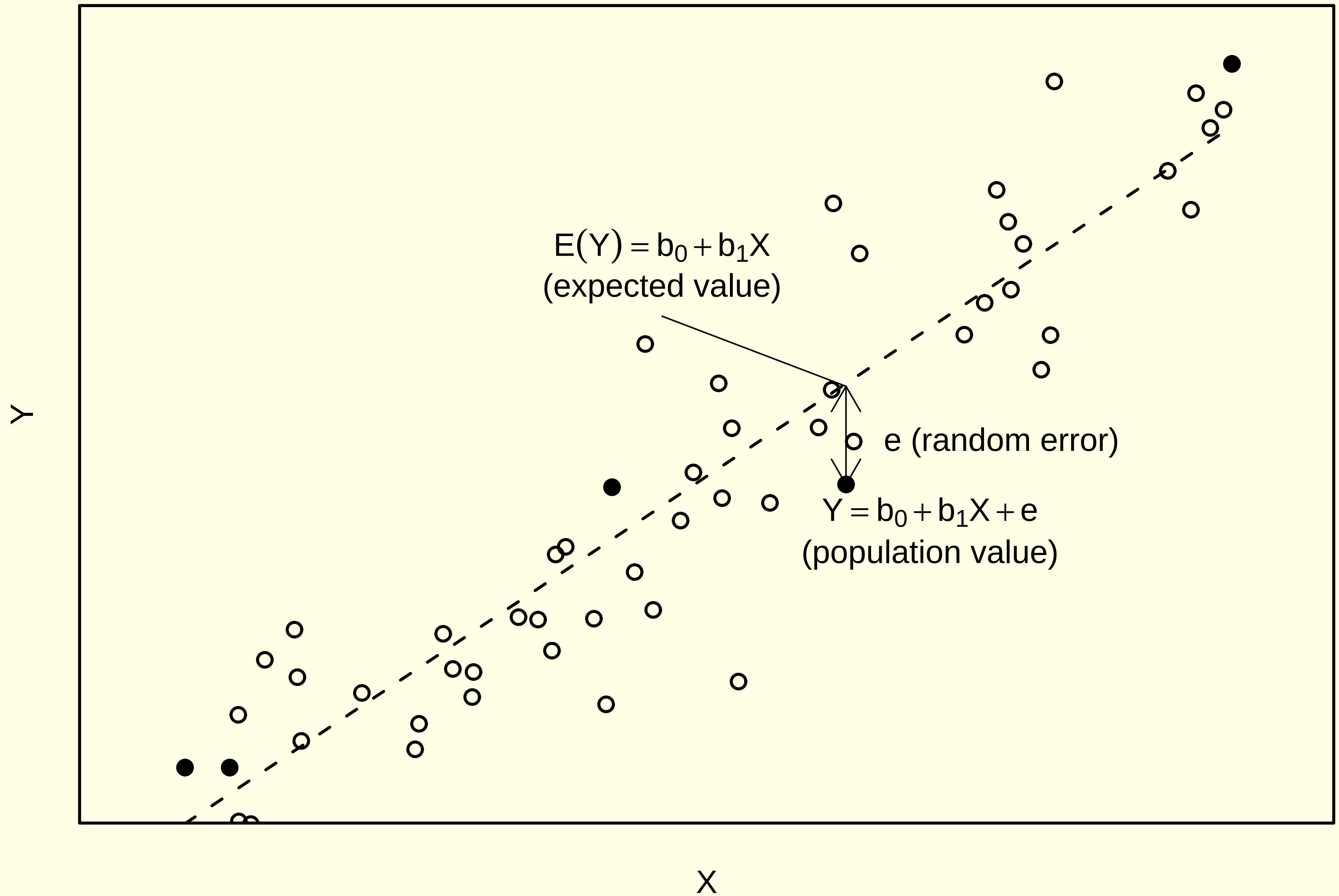
Correlation examples



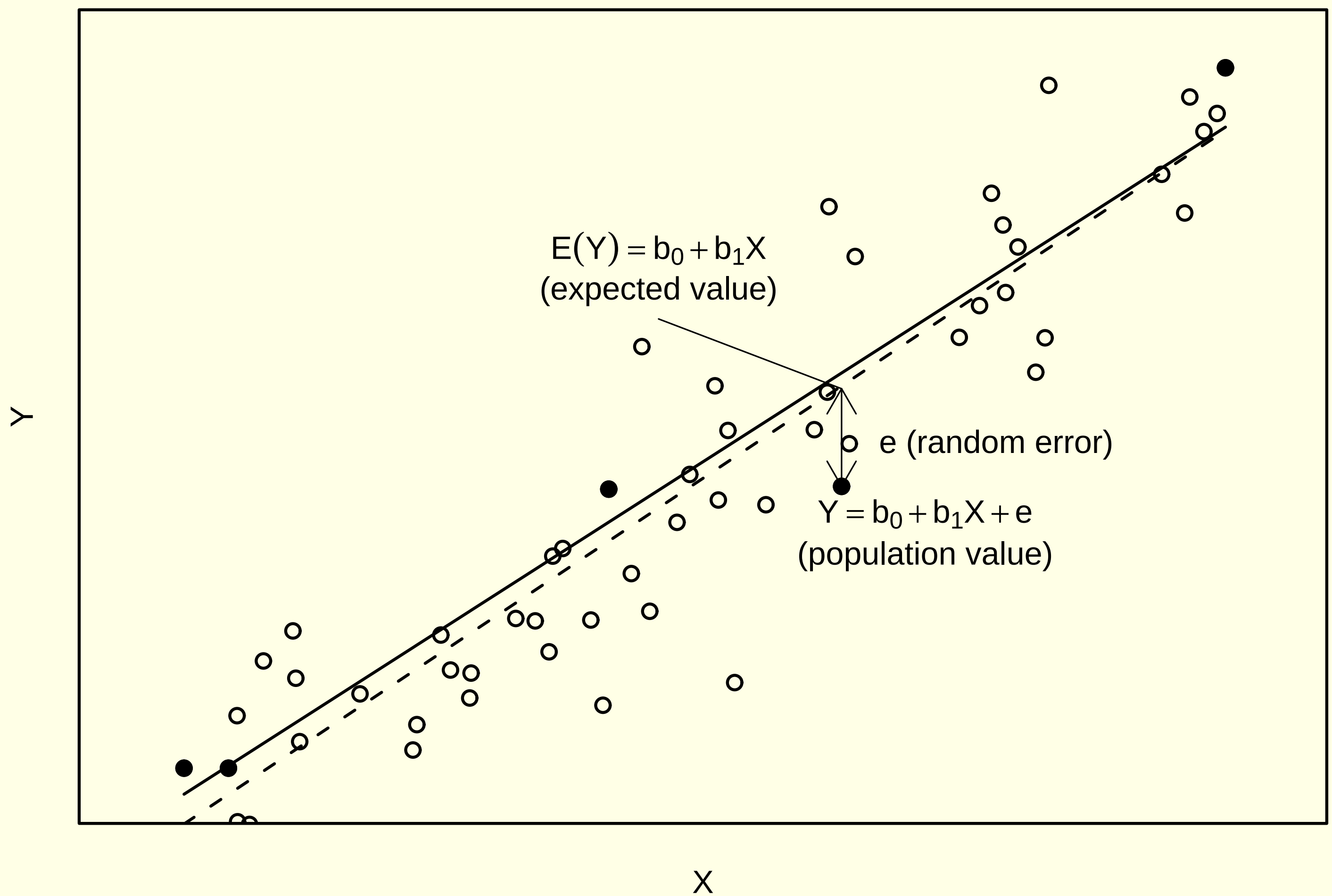
Infer from sample slope about the population slope.



Infer from sample slope about the population slope.



Infer from sample slope about the population slope.



Hypothesis test for b_1

- Recall univariate t-statistic $= \dfrac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.

Hypothesis test for b_1

- Recall univariate t-statistic $= \dfrac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.

Hypothesis test for b_1

- Recall univariate t-statistic $= \dfrac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \dfrac{40.8 - 0}{5.684} = 7.18$.

Hypothesis test for b_1

- Recall univariate t-statistic $= \dfrac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \dfrac{40.8 - 0}{5.684} = 7.18$.
- Significance level $= 5\%$.

Hypothesis test for b_1

- Recall univariate t-statistic $= \dfrac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \dfrac{40.8 - 0}{5.684} = 7.18$.
- Significance level $= 5\%$.
- Critical value is 3.182 (97.5^{th} percentile of t_3).

Hypothesis test for b_1

- Recall univariate t-statistic $= \frac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{40.8 - 0}{5.684} = 7.18$.
- Significance level $= 5\%$.
- Critical value is 3.182 (97.5^{th} percentile of t_3).
- Since t-statistic (7.18) is between 5.841 and 10.215, p-value is between 0.01 and 0.002.

- Recall univariate t-statistic $= \frac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{40.8 - 0}{5.684} = 7.18$.
- Significance level $= 5\%$.
- Critical value is 3.182 (97.5^{th} percentile of t_3).
- Since t-statistic (7.18) is between 5.841 and 10.215, p-value is between 0.01 and 0.002.
- Since t-statistic (7.18) $>$ critical value (3.182) and p-value $<$ signif. level, reject NH in favor of AH.

Hypothesis test for b_1

- Recall univariate t-statistic $= \frac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{40.8 - 0}{5.684} = 7.18$.
- Significance level $= 5\%$.
- Critical value is 3.182 (97.5^{th} percentile of t_3).
- Since t-statistic (7.18) is between 5.841 and 10.215, p-value is between 0.01 and 0.002.
- Since t-statistic $(7.18) >$ critical value (3.182) and p-value $<$ signif. level, reject NH in favor of AH.
- In other words, the sample data favor a nonzero slope (at a significance level of 5%).

Hypothesis test for b_1

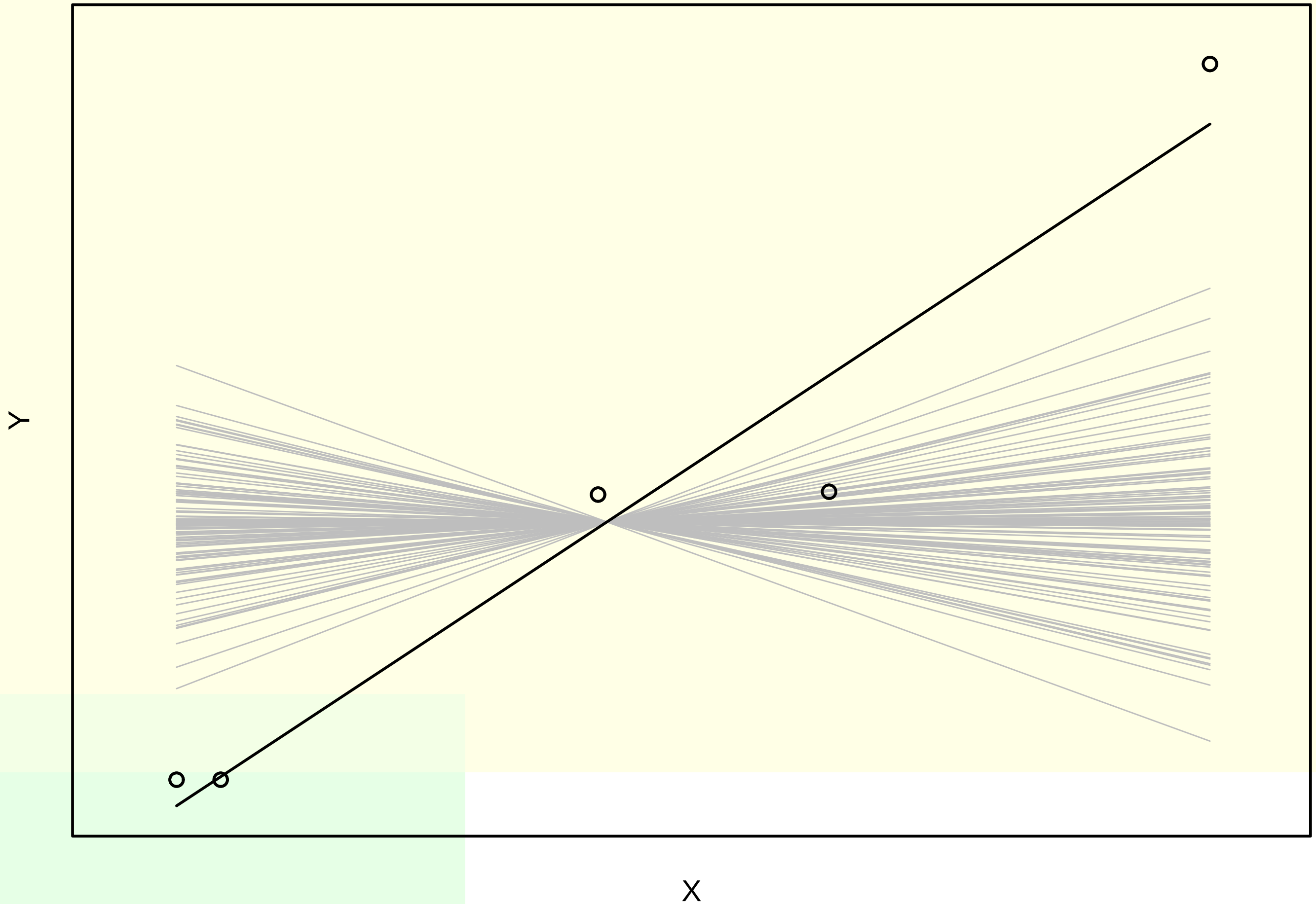
- Recall univariate t-statistic $= \dfrac{m_Y - \mathrm{E}(Y)}{s_Y/\sqrt{n}} \sim t_{n-1}$.
- Here, slope t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$.
- NH: $b_1 = 0$ versus AH: $b_1 \neq 0$.
- t-statistic $= \dfrac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \dfrac{40.8 - 0}{5.684} = 7.18$.
- Significance level $= 5\%$.
- Critical value is 3.182 (97.5^{th} percentile of t_3).
- Since t-statistic (7.18) is between 5.841 and 10.215, p-value is between 0.01 and 0.002.
- Since t-statistic $(7.18) >$ critical value (3.182) and p-value $<$ signif. level, reject NH in favor of AH.
- In other words, the sample data favor a nonzero slope (at a significance level of 5%).
- Exercise: do an upper tail test for this example.

Computer output and slope test illustration

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

[a] Response variable: Y.



Slope confidence interval

- Calculate a 95% confidence interval for b_1.

Slope confidence interval

- Calculate a 95% confidence interval for b_1.
- 97.5^{th} percentile of t_3 is 3.182.

Slope confidence interval

- Calculate a 95% confidence interval for b_1.
- 97.5^{th} percentile of t_3 is 3.182.
- $\hat{b}_1 \pm 97.5^{\text{th}}$ percentile $(s_{\hat{b}_1}) =$ $40.8 \pm 3.182 \times 5.684 = 40.8 \pm 18.1 = (22.7, 58.9)$.

Slope confidence interval

- Calculate a 95% confidence interval for b_1.
- 97.5^{th} percentile of t_3 is 3.182.
- $\hat{b}_1 \pm 97.5^{\text{th}}\text{ percentile}\,(s_{\hat{b}_1}) =$ $40.8 \pm 3.182 \times 5.684 = 40.8 \pm 18.1 = (22.7, 58.9)$.
- Loosely speaking: based on this dataset, we are 95% confident that the population slope, b_1, is between 22.7 and 58.9.

Slope confidence interval

- Calculate a 95% confidence interval for b_1.
- 97.5^{th} percentile of t_3 is 3.182.
- $\hat{b}_1 \pm 97.5^{\text{th}}$ percentile $(s_{\hat{b}_1}) =$ $40.8 \pm 3.182 \times 5.684 = 40.8 \pm 18.1 = (22.7, 58.9)$.
- Loosely speaking: based on this dataset, we are 95% confident that the population slope, b_1, is between 22.7 and 58.9.
- More precisely: if we were to take a large number of random samples of size 5 from our population of homes and calculate a 95% confidence interval for each, then 95% of those confidence intervals would contain the (unknown) population slope.

Slope confidence interval

- Calculate a 95% confidence interval for b_1.
- 97.5^{th} percentile of t_3 is 3.182.
- $\hat{b}_1 \pm 97.5^{\text{th}}$ percentile $(s_{\hat{b}_1}) =$ $40.8 \pm 3.182 \times 5.684 = 40.8 \pm 18.1 = (22.7, 58.9)$.
- Loosely speaking: based on this dataset, we are 95% confident that the population slope, b_1, is between 22.7 and 58.9.
- More precisely: if we were to take a large number of random samples of size 5 from our population of homes and calculate a 95% confidence interval for each, then 95% of those confidence intervals would contain the (unknown) population slope.
- Exercise: calculate a 90% confidence interval for b_1.