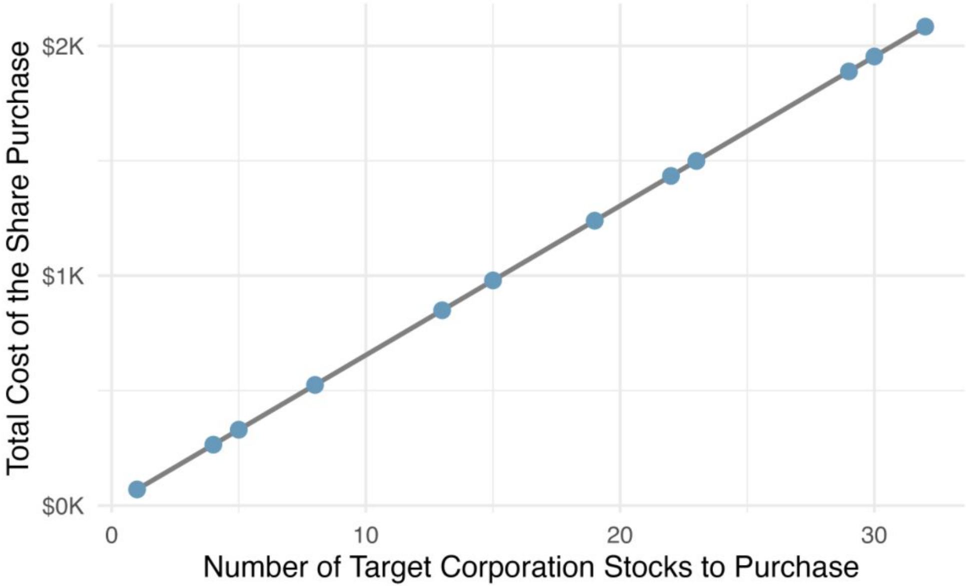
Simple linear regression

With a single predictor

Prof. Dr. Jan Kirenz
HdM Stuttgart

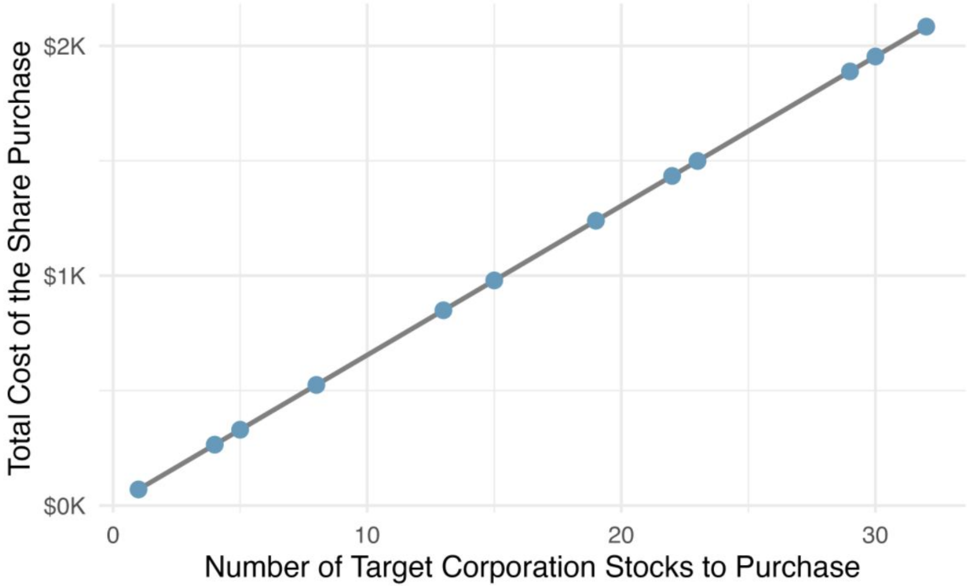
Fitting a line, residuals, and correlation

Fitting a line to data



12 stock purchases at a trading company
The total cost of the shares were reported.

Fitting a line to data



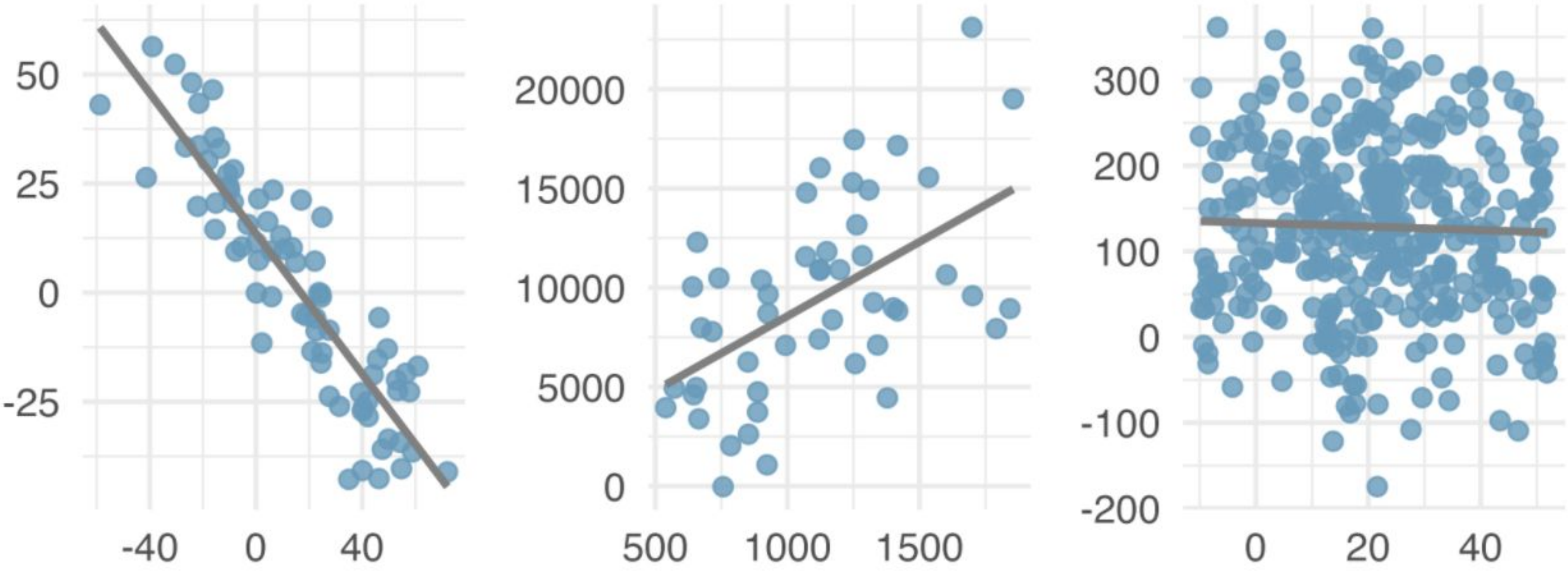
y: outcome x: predictor

y = 5 + 64.96x

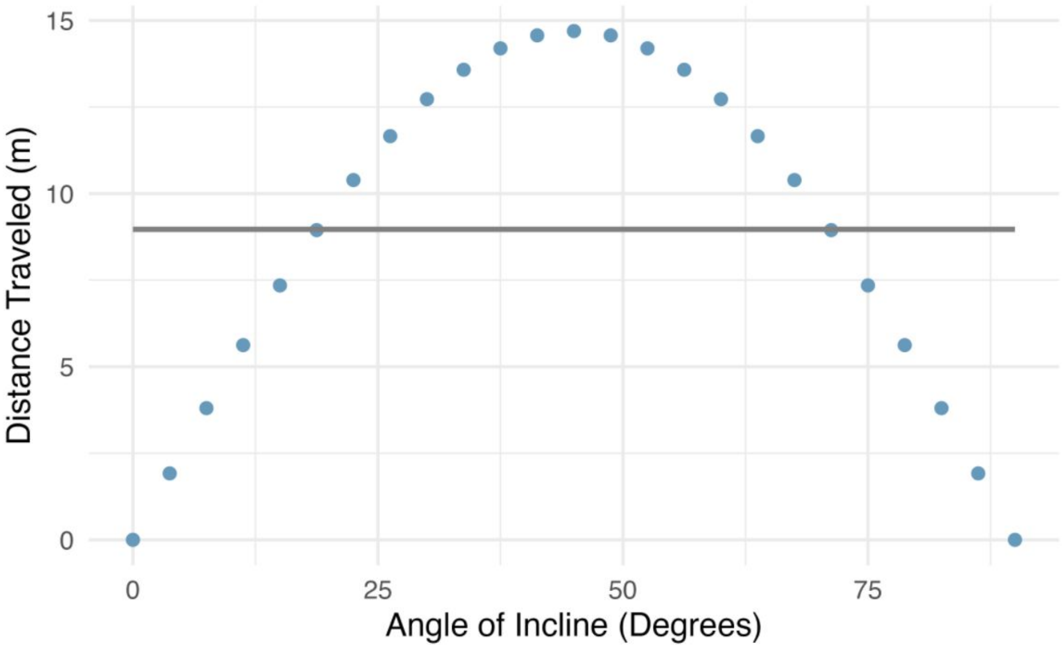
$$y = b_0 + b_1\, x + e$$

12 stock purchases at a trading company
The total cost of the shares were reported.

Three datasets where a linear model may be useful



A simple linear model is not useful



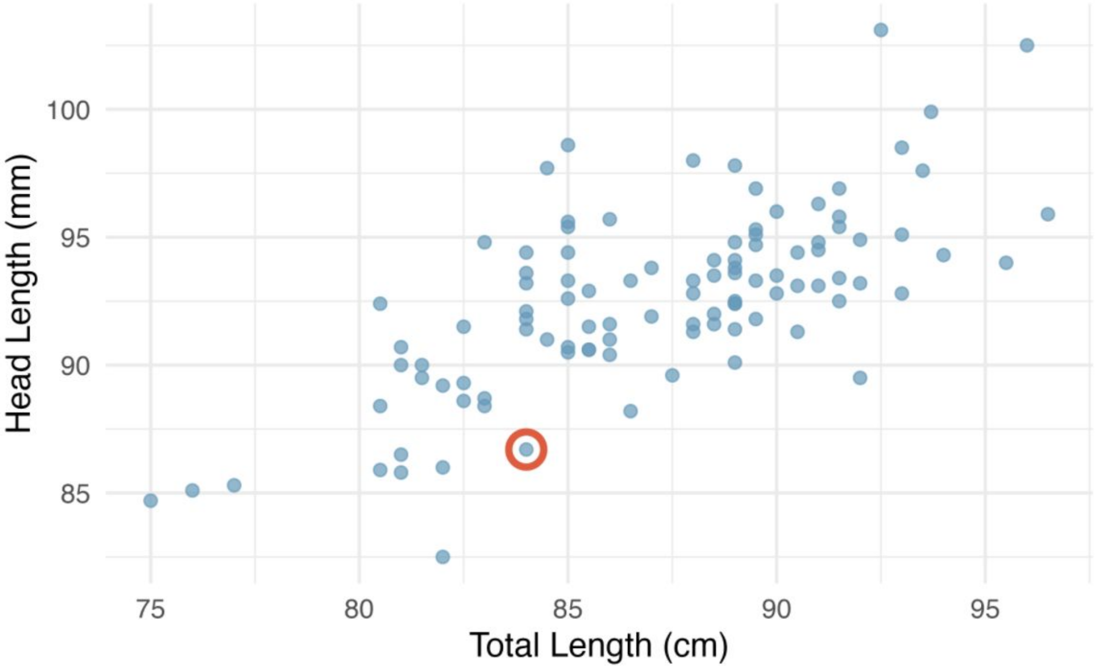
Using linear regression to predict possum head lengths

We consider two measurements:

- the **total length** of each possum, from head to tail, and
- the length of each possum's **head**.

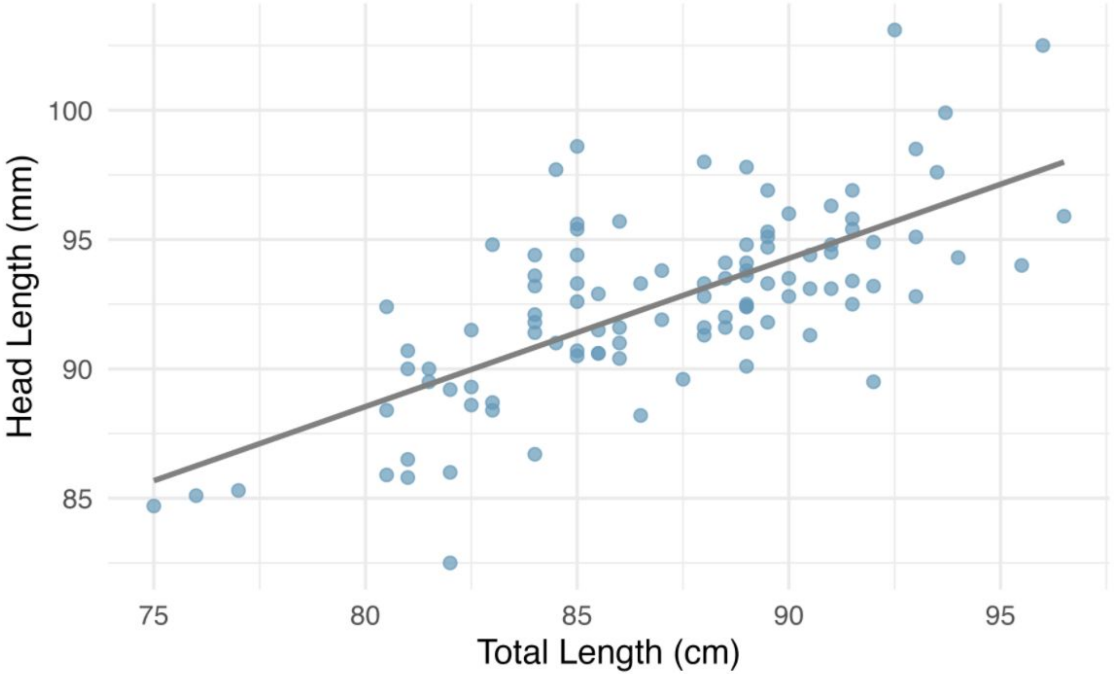
hoto by Greg Schecter, flic.kr/p/9BAFbR, CC BY 2.0 license.

A scatterplot showing head length against total length for 104 brushtail possums



A point representing a possum with head length 86.7 mm and total length 84 cm is highlighted.

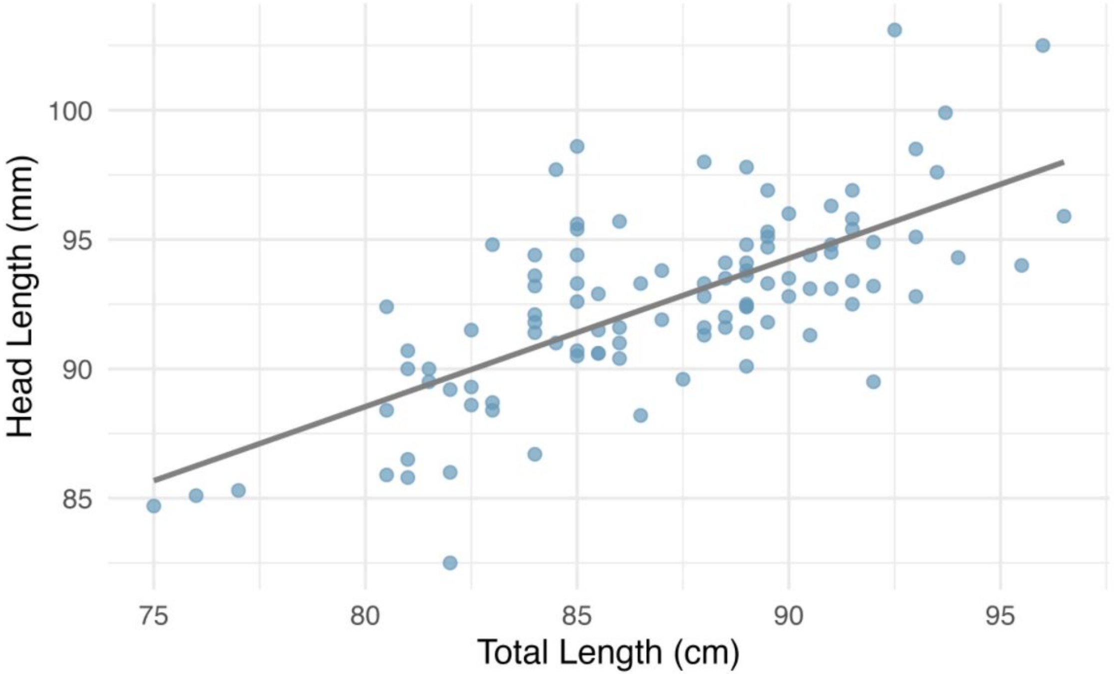
A scatterplot showing head length against total length for 104 brushtail possums



$$\hat{y} = 41 + 0.59x$$

A reasonable linear model was fit to represent the relationship between head length and total length.

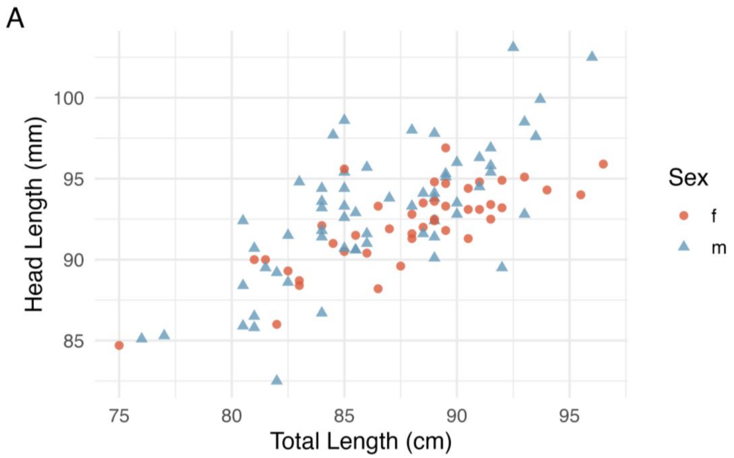
A scatterplot showing head length against total length for 104 brushtail possums

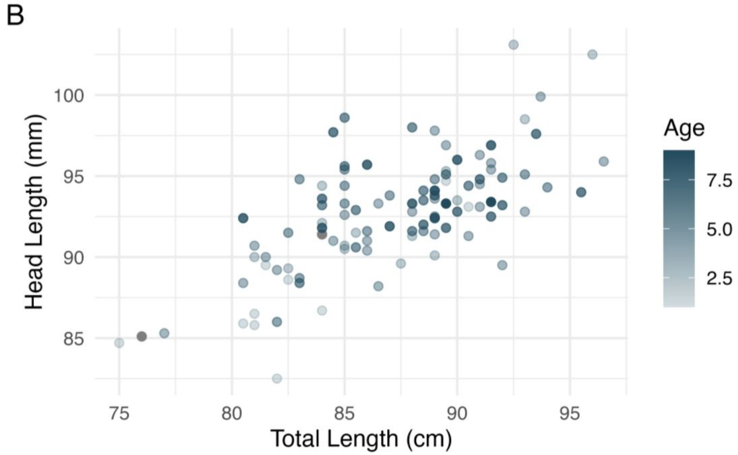


$$\hat{y} = 41 + 0.59x$$

This means a possum with a total length of 80 cm will have a head length of ____ mm on average

A reasonable linear model was fit to represent the relationship between head length and total length.

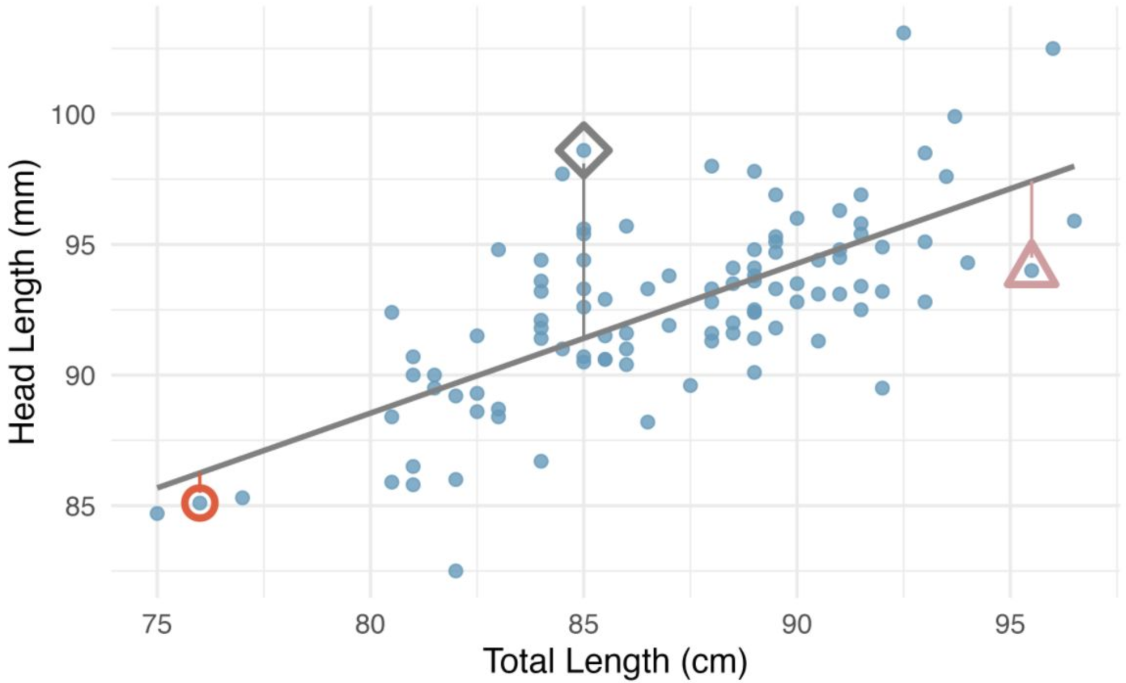




Relationship between total length and head length of brushtail possums, taking into consideration their sex (Plot A) or age (Plot B).

Residuals

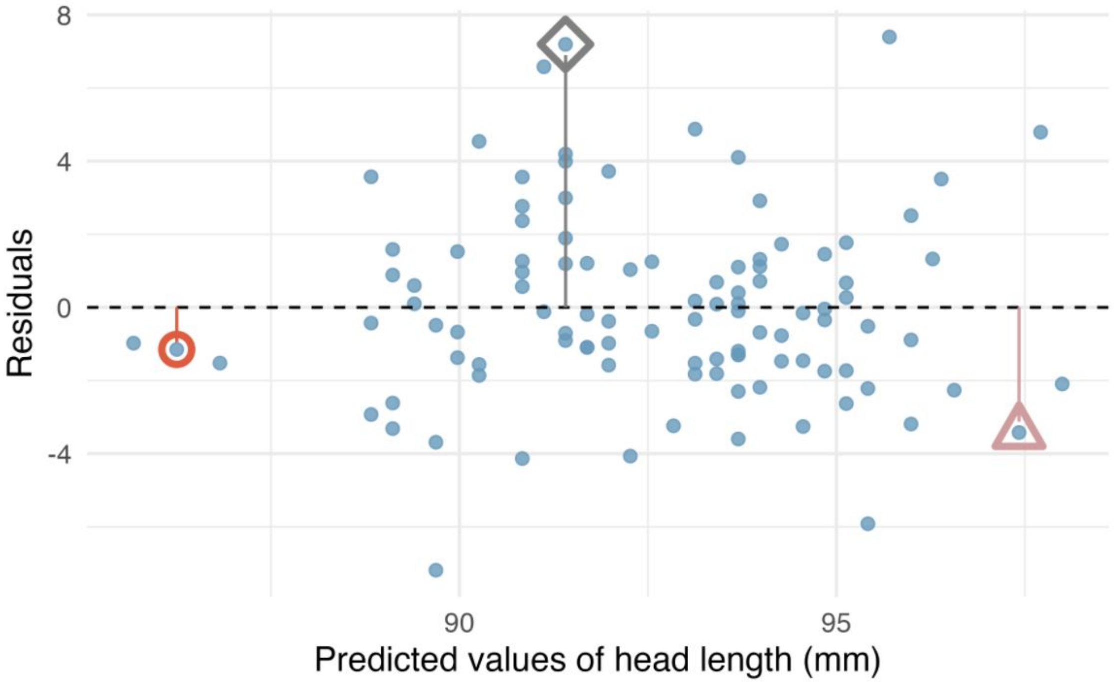
Data = Fit + Residual



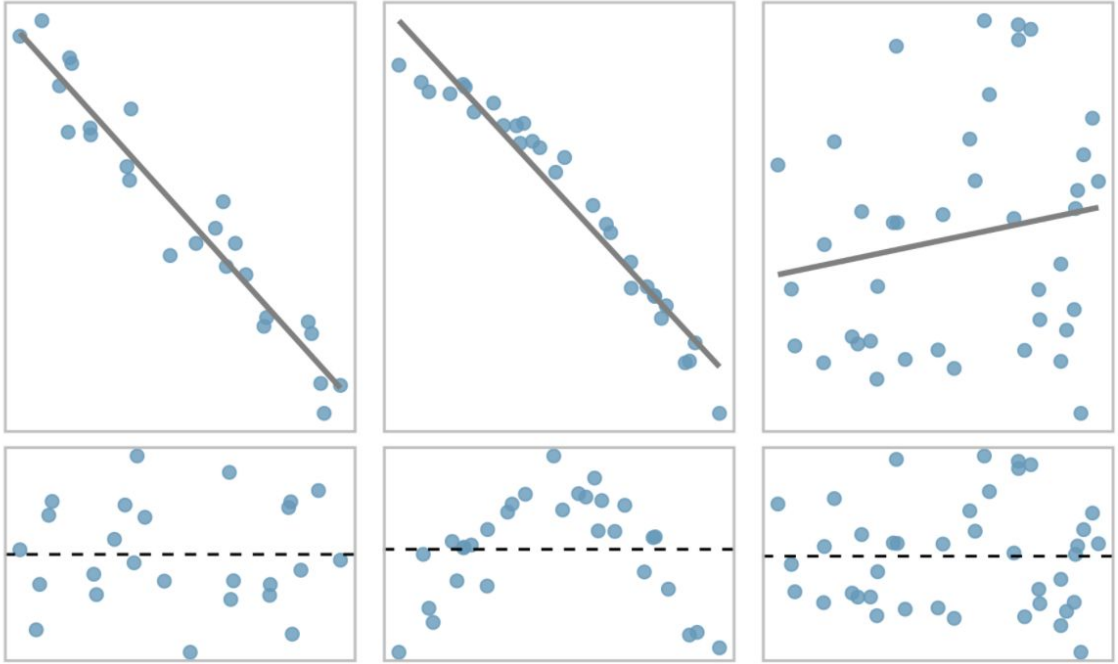
$$e_i = y_i - \hat{y}_i$$

A reasonable linear model was fit to represent the relationship between head length and total length, with three points highlighted.

Residual plot for the model



Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

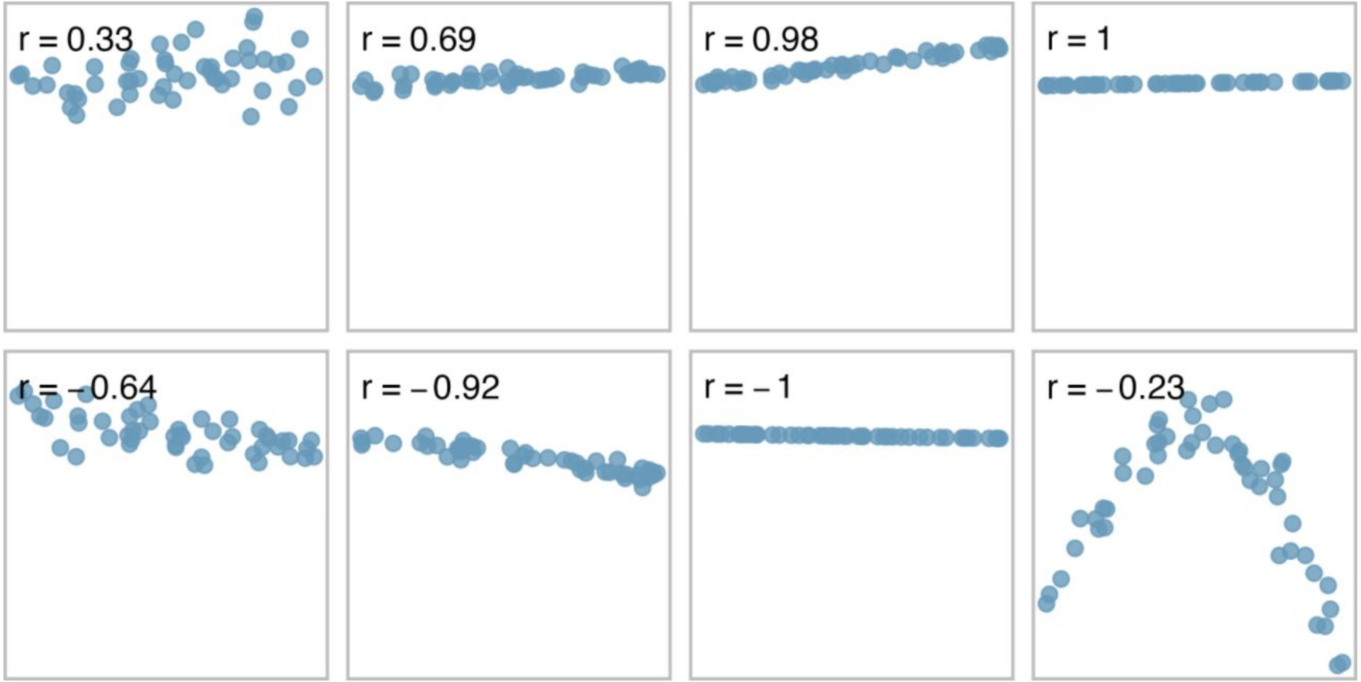


Describing linear relationships with correlation

Correlation: strength of a **linear relationship**.

- Correlation which always takes values between -1 and 1, describes the strength and direction of the linear relationship between two variables. We denote the correlation by r
- The correlation value has no units and will not be affected by a linear change in the units (e.g., going from inches to centimeters).

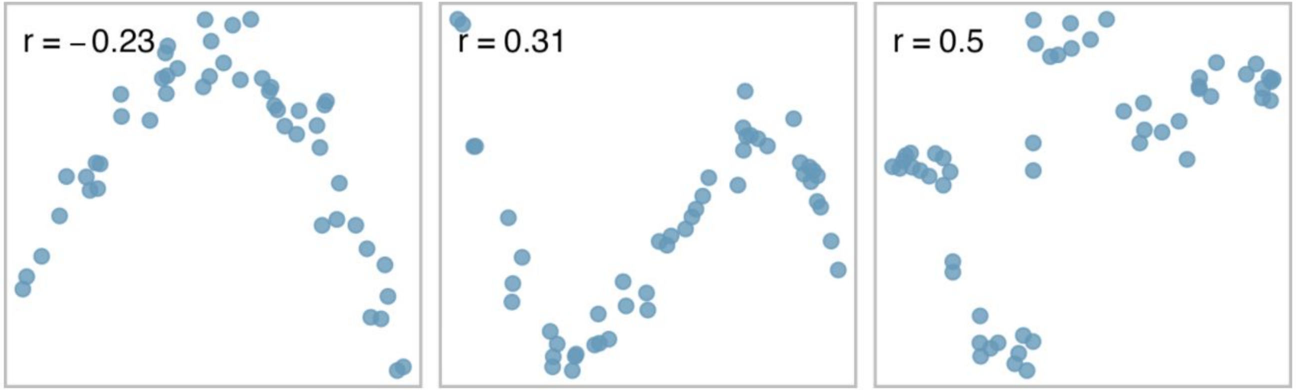
Sample scatterplots and their correlations



The first row shows variables with a positive relationship, represented by the trend up and to the right.

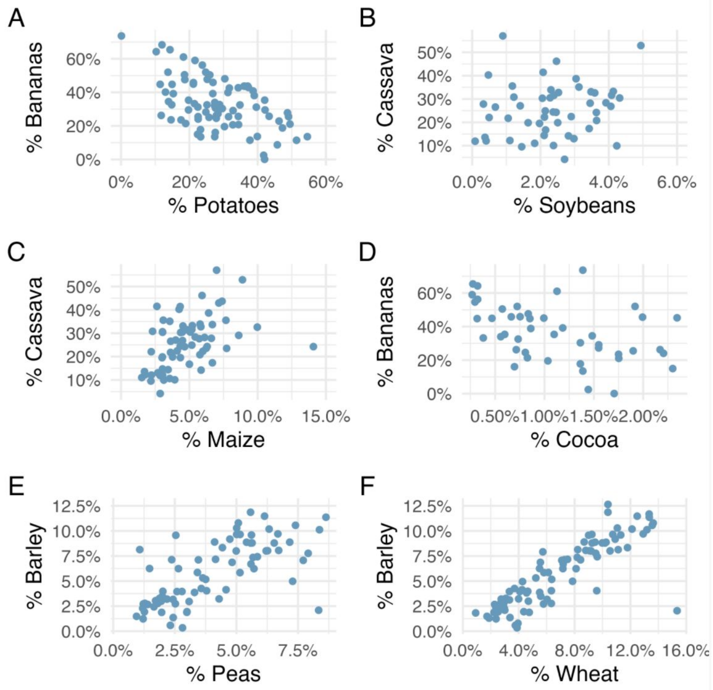
The second row shows variables with a negative trend, where a large value in one variable is associated with a lower value in the other.

Sample scatterplots and their correlations



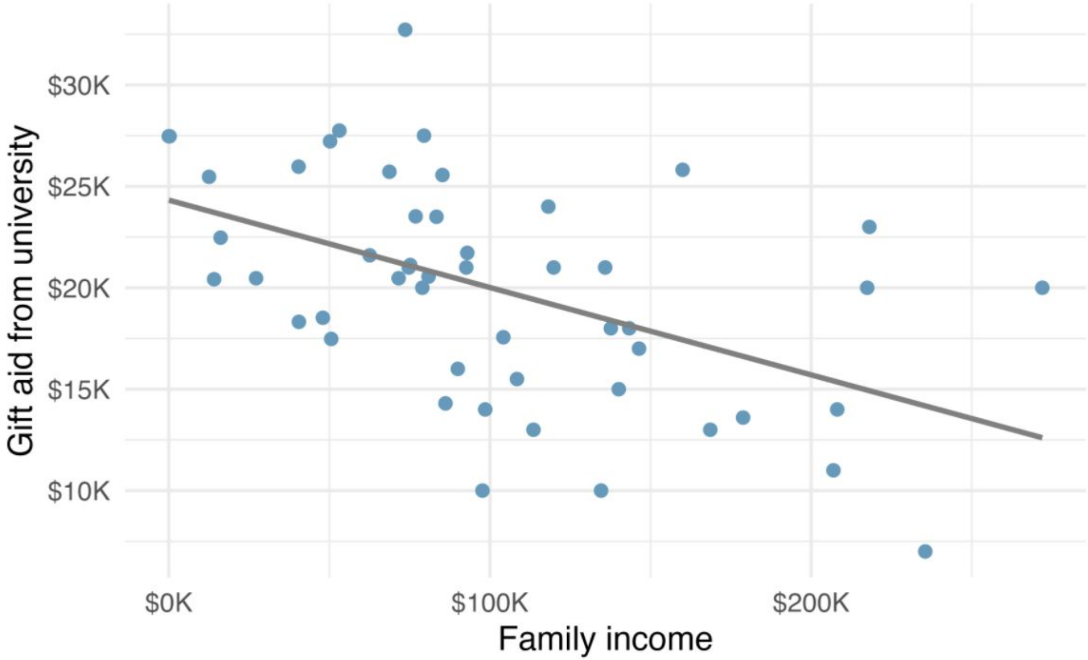
In each case, there is a strong relationship between the variables. However, because the relationship is not linear, the correlation is relatively weak.

Order the six scatterplots from strongest negative to strongest positive linear relationship.



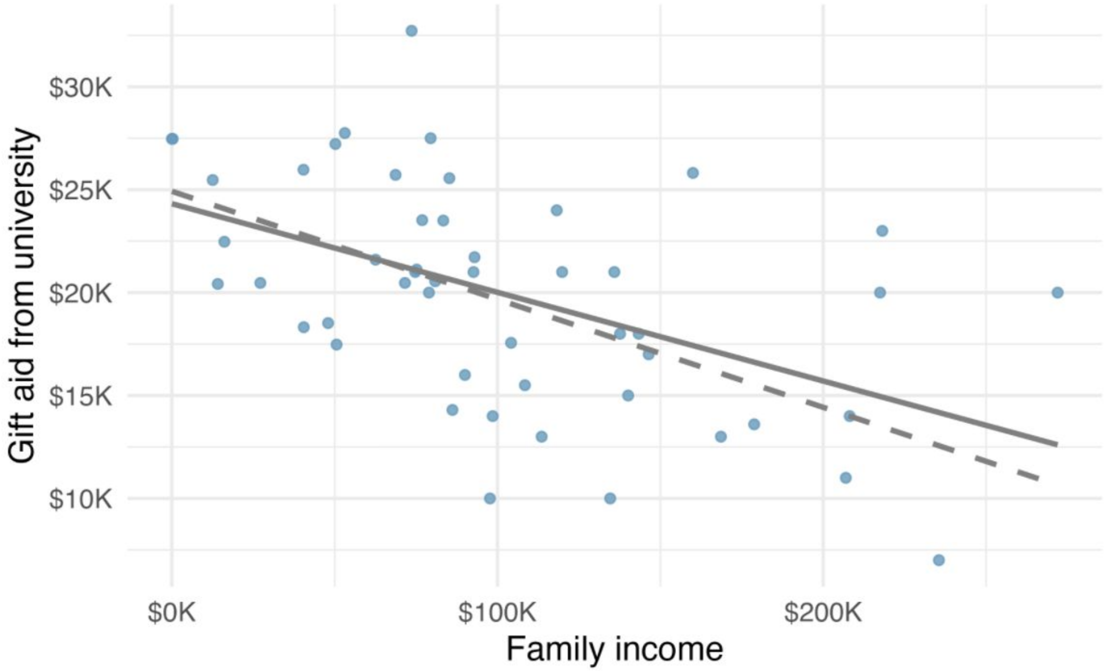
Least squares regression

Gift aid and family income for a random sample of 50 freshman students from Elmhurst College.



Is the correlation positive or negative?

An objective measure for finding the best line



____The solid line represents the line that minimizes the **sum of squared residuals, i.e., the least squares line.**

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

---- The dashed line represents the line that minimizes the **sum of the absolute value of residuals.**

$$|e_1| + |e_2| + \cdots + |e_n|$$

Finding and interpreting the least squares line

Data

Data is in \$1,000s 21.7 = \$21,700

family_income	gift_aid
92.92	21.7
0.25	27.5
53.09	27.8
50.20	27.2
137.61	18.0

Finding and interpreting the least squares line

Data

Data is in \$1,000s   21.7 = \$21,700

family_income	gift_aid
92.92	21.7
0.25	27.5
53.09	27.8
50.20	27.2
137.61	18.0

Equation of least squares regression

Prediction   Intercept   Slope   Predictor

$$\widehat{\texttt{aid}} = \boxed{\beta_0 + \beta_1} \times \texttt{family\_income}$$

Population parameters

Finding and interpreting the least squares line

Data

Data is in \$1,000s   21.7 = \$21,700

family_income	gift_aid
92.92	21.7
0.25	27.5
53.09	27.8
50.20	27.2
137.61	18.0

Equation of least squares regression

Prediction   Intercept   Slope   Predictor

$$\widehat{\texttt{aid}} = \boxed{\beta_0 + \beta_1} \times \texttt{family\_income}$$

Population parameters

Statistical model

Sample parameters

$$y = \boxed{b_0 + b_1} x + e$$

Finding and interpreting the least squares line

Data

Data is in $1,000s   21.7 = $21,700

family_income	gift_aid
92.92	21.7
0.25	27.5
53.09	27.8
50.20	27.2
137.61	18.0

Equation of least squares regression

Prediction   Intercept   Slope   Predictor

$$\widehat{\texttt{aid}} = \boxed{\beta_0 + \beta_1} \times \texttt{family\_income}$$

Population parameters

Statistical model

Sample parameters

$$y = \boxed{b_0 + b_1} x + e$$

Results of statistical model

	term	estimate	std.error	statistic	p.value
b_0	`(Intercept)`	24.32	1.29	18.83	<0.0001
b_1	`family_income`	-0.04	0.01	-3.98	2e-04

Finding and interpreting the least squares line

Data

Data is in \$1,000s 21.7 = \$21,700

family_income	gift_aid
92.92	21.7
0.25	27.5
53.09	27.8
50.20	27.2
137.61	18.0

Equation of least squares regression

Prediction Intercept Slope Predictor

$$\widehat{\texttt{aid}} = \boxed{\beta_0 + \beta_1} \times \texttt{family\_income}$$

Population parameters

Statistical model

Sample parameters

$$y = \boxed{b_0 + b_1} x + e$$

Results of statistical model

	term	estimate	std.error	statistic	p.value
b_0	`(Intercept)`	24.32	1.29	18.83	<0.0001
b_1	`family_income`	-0.04	0.01	-3.98	2e-04

b_1: For each additional \$1,000 of family income, we would expect a student to receive a net difference of 1,000 × (-0.0431) = -\$43.10 in aid on average, i.e., \$43.10 less.

Finding and interpreting the least squares line

Equation of least squares regression

Prediction Intercept Slope Predictor

$$\widehat{\texttt{aid}} = \boxed{\beta_0 + \beta_1} \times \texttt{family\_income}$$

Population parameters

Data

Data is in $1,000s 21.7 = $21,700

family_income	gift_aid
92.92	21.7
0.25	27.5
53.09	27.8
50.20	27.2
137.61	18.0

Statistical model

Sample parameters

$$y = \boxed{b_0 + b_1} x + e$$

The estimated intercept **b_0** = 24.319 describes the average aid if a student's family had no income, $24,319. The meaning of the intercept is relevant to this application since the family income for some students is $0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero.

Results of statistical model

	term	estimate	std.error	statistic	p.value
b_0	(Intercept)	24.32	1.29	18.83	<0.0001
b_1	family_income	-0.04	0.01	-3.98	2e-04

b_1: For each additional $1,000 of family income, we would expect a student to receive a net difference of 1,000 × (-0.0431) = -$43.10 in aid on average, i.e., $43.10 less.

Interpreting parameters estimated by least squares.

The **slope (b_1)** describes the estimated difference in the predicted average outcome of y if the predictor variable x happened to be one unit larger.

The **intercept (b_0)** describes the average outcome of y if x=0 and the linear model is valid all the way to x = 0 (values of x = 0 are not observed or relevant in many applications).

The slope of the least squares line can be estimated

$$b_1 = \frac{s_y}{s_x} r$$

where

- r is the correlation between the two variables, and
- s_x and s_y are the sample standard deviations of the predictor and outcome, respectively.

Family income, x		Gift aid, y		
mean	sd	mean	sd	r
102	63.2	19.9	5.46	-0.499

$$b_1 = \frac{s_y}{s_x} r = \frac{5.46}{63.2}(-0.499) = -0.0431$$

Limitations

$$\widehat{\mathtt{aid}} = 24.3 - 0.0431 \times \mathtt{family\_income}$$

Use this model to estimate the aid of another freshman student whose family had income of $1 million.

Limitations

$$\widehat{\texttt{aid}} = 24.3 - 0.0431 \times \texttt{family\_income}$$

Use this model to estimate the aid of another freshman student whose family had income of $1 million.

Applying a model estimate to values outside of the realm of the original data is called extrapolation. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

Describing the strength of a fit

We explain the strength of a linear fit using R^2, called R-squared

- Describes the amount of variation in the outcome variable that is explained by the least squares line

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29800 - 22400}{29800} = \frac{7500}{29800} \approx 0.25,$$

s^2_{aid}: variance of the outcome variable, aid received

s^2_{RES}: variability in the residuals describes how much variation remains after using the model

There was a reduction of about 25%, of the outcome variable's variation by using information about family income for predicting aid using a linear model.

$$r = -0.499 \rightarrow R^2 = 0.25$$

R-squared

- Is also called the coefficient of determination.
- R2 will always be between 0 and 1.

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

SST: total sum of squares,

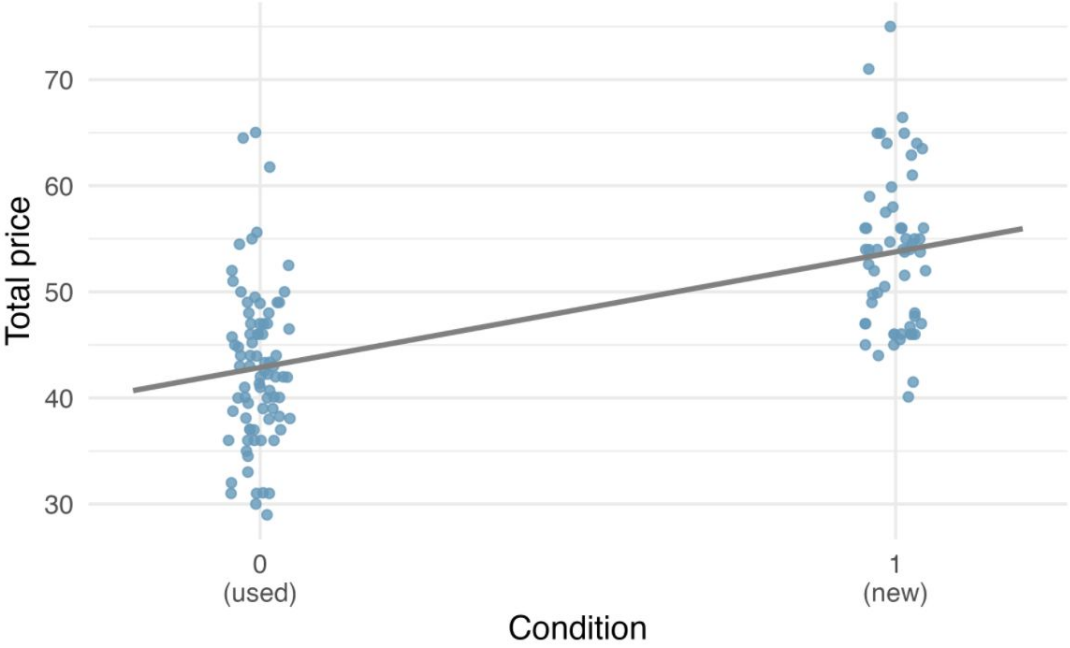
SSE: sum of squared errors

$$SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2.$$

$$\begin{aligned} SSE &= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \\ &= e_1^2 + e_2^2 + \cdots + e_n^2 \end{aligned}$$

Categorical predictors with two levels

Total auction prices for the video game Mario Kart, divided into used



Using categorical predictors

- **Convert** categories into a numerical form.
- We will do so using an **indicator variable** called:
  - **condnew**: **1** when the game is **new**
  - **condnew**: **0** when the game is **used**.

$$\widehat{\texttt{price}} = b_0 + b_1 \times \texttt{condnew}$$

Using categorical predictors

$$\widehat{\texttt{price}} = b_0 + b_1 \times \texttt{condnew}$$

$$\widehat{\texttt{price}} = 42.87 + 10.9 \times \texttt{condnew}$$

Interpret the two parameters estimated in the model for the price of Mario Kart in eBay auctions.

term	estimate	std.error	statistic	p.value
`(Intercept)`	42.9	0.81	52.67	<0.0001
`condnew`	10.9	1.26	8.66	<0.0001

Interpreting parameters

b_0

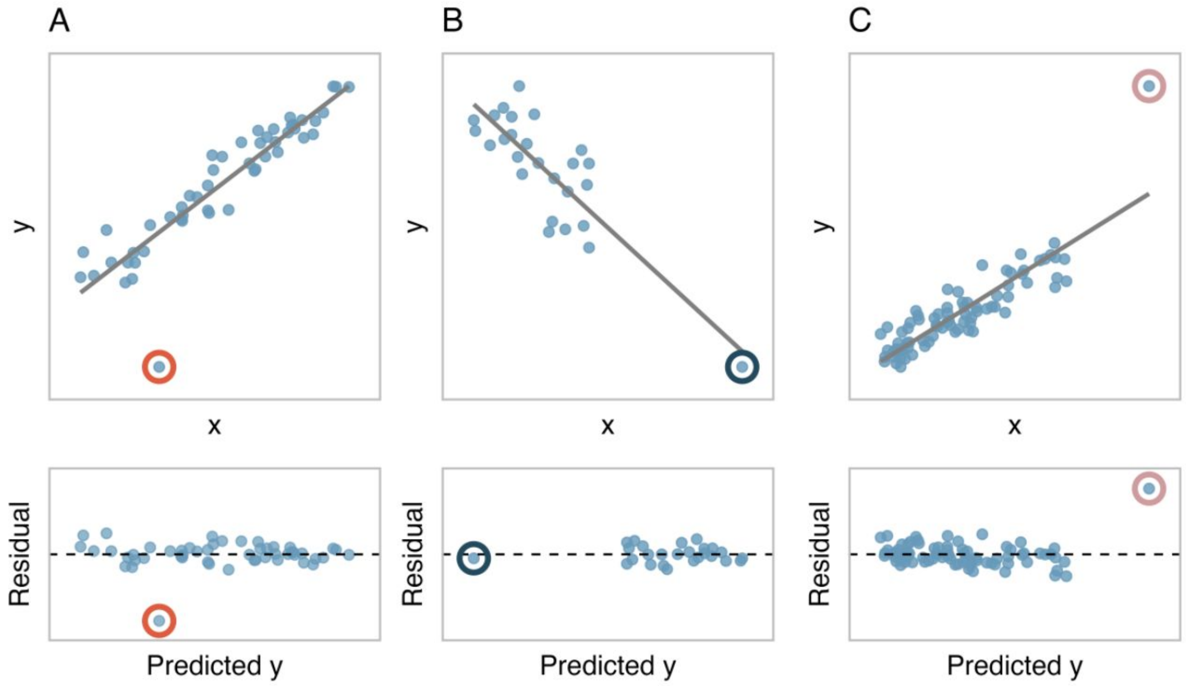
- **Intercept:** estimated price when condnew = 0 (**used condition**)
- Average selling price of a used version: $42.9.

b_1

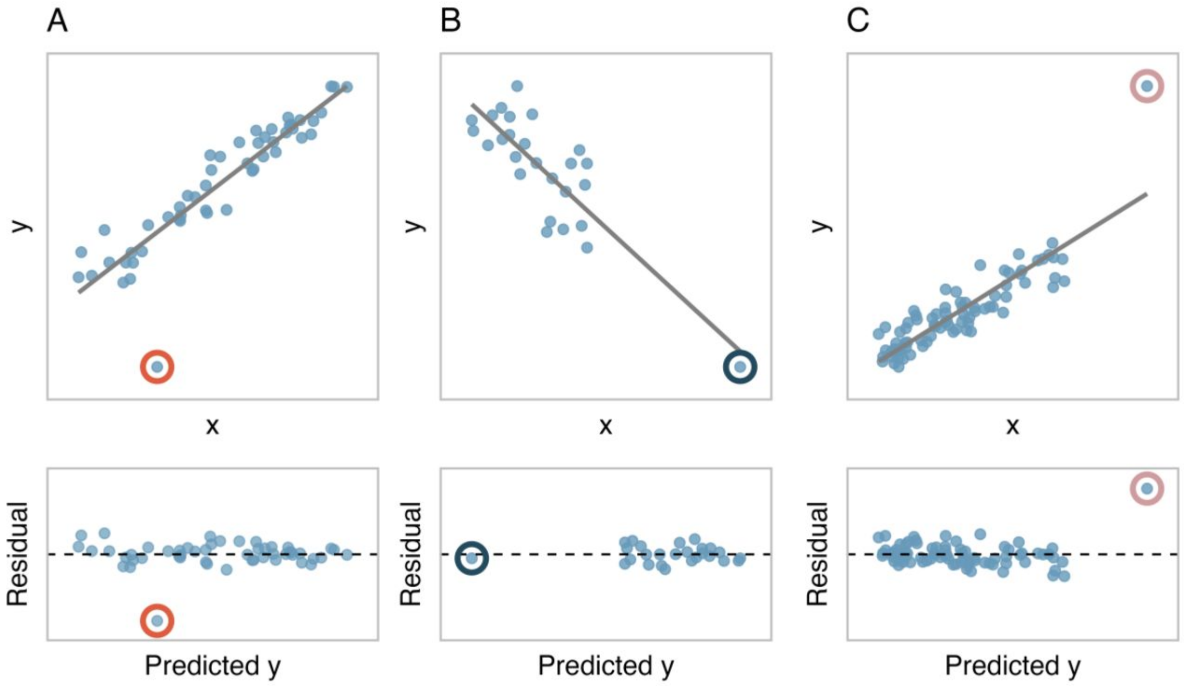
- **Slope:** on average, **new games** sell for about $10.9 more than used games (condnew = 1)
- Slope = the average change in the outcome variable between the two categories.

Outliers in linear regression

Each dataset has at least one outlier.



Each dataset has at least one outlier.

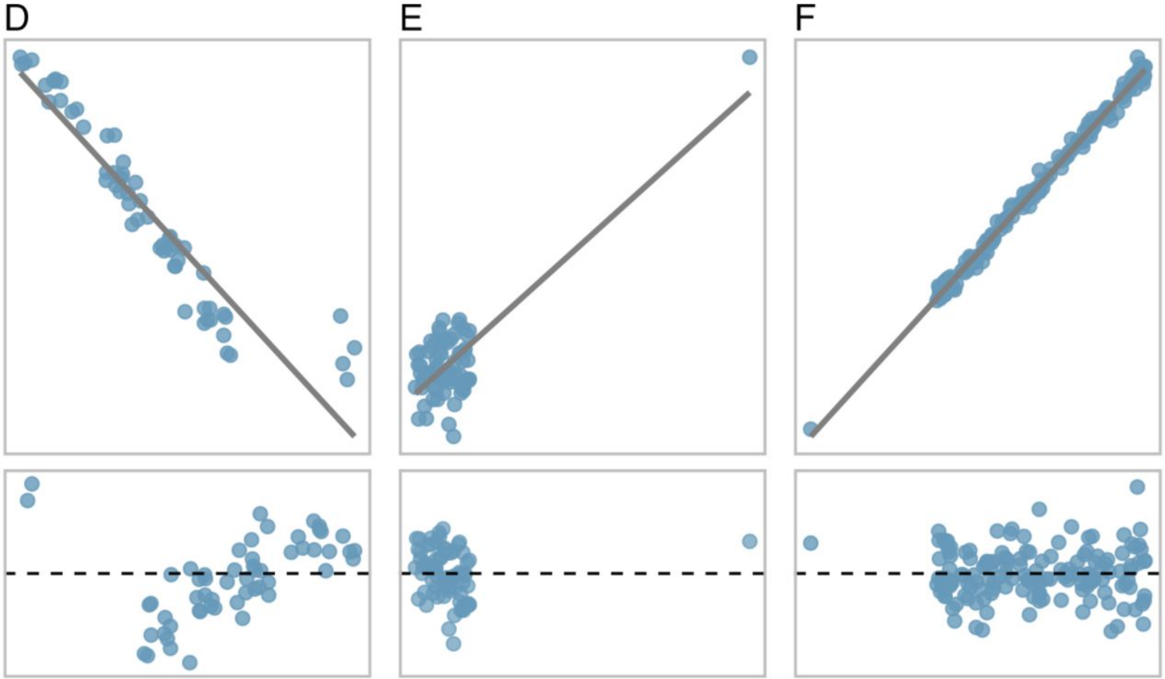


A: There is one outlier far from the other points, though it only appears to slightly influence the line.

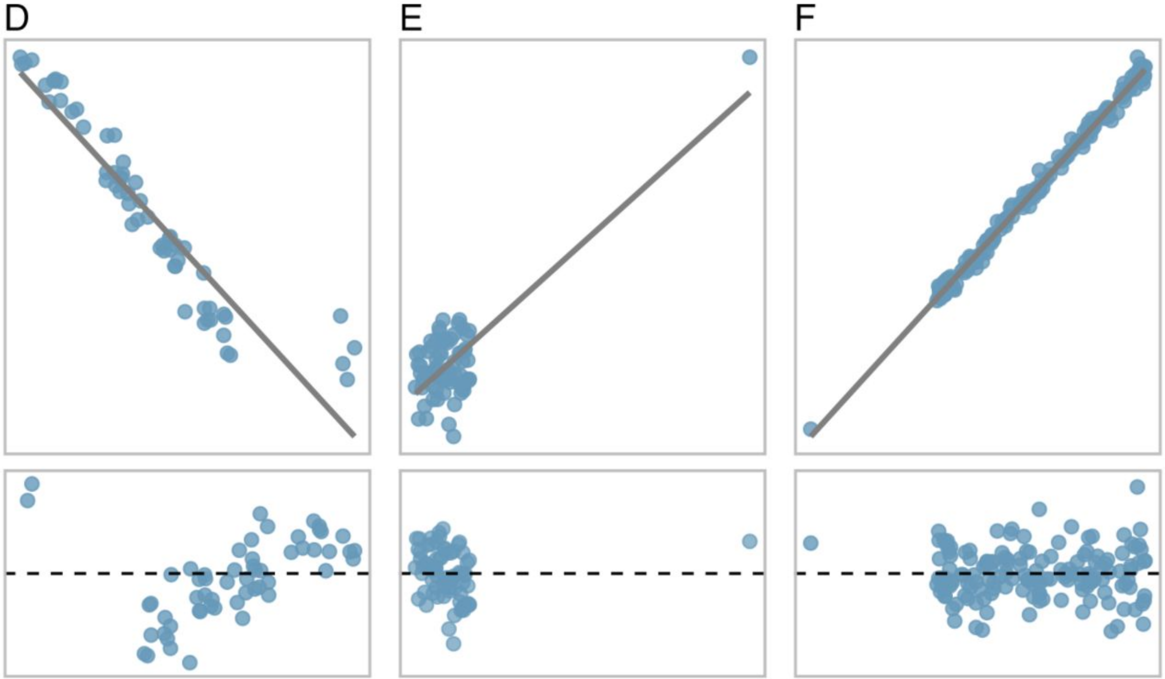
B: There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn’t very influential.

C: There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn’t appear to fit very well.

Each dataset has at least one outlier.



Each dataset has at least one outlier.



D: The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

E: There is no obvious trend in the main cloud of points and the outlier on the right appears to largely (and problematically) control the slope of the least squares line.

F: There is one outlier far from the cloud. However, it falls quite close to the least squares line and does not appear to be very influential

Types of outlier

Outlier:

A point (or a group of points) that stands out from the rest of the data

Don’t remove outliers without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly.

High leverage or leverage points (influential points):

Points that fall horizontally away from the center of the cloud tend to pull harder on the line

Terms you should know

coefficient of determination

correlation

extrapolation

high leverage

indicator variable

influential point

least squares line

leverage point

outcome

outlier

predictor

R-squared

residuals

sum of squared error

total sum of squares