Data Analytics with R

Sumit Mishra

Institute for Financial Management and Research, Sri City

Linear Regression

23 December 2020

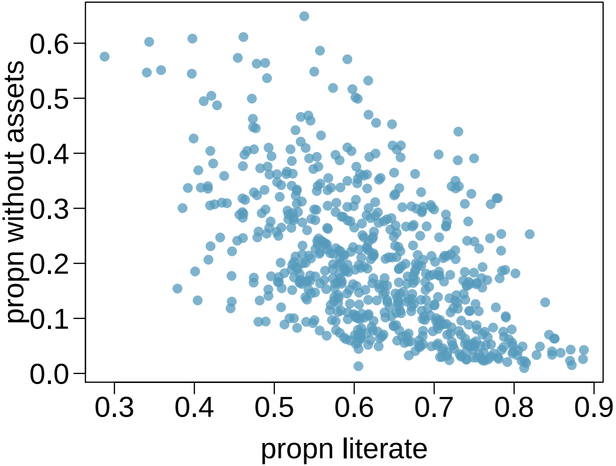
Line fitting, residuals, and correlation

Modeling numerical variables

In this unit we will learn to quantify the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable.

Asset Poverty vs. Literacy Rate

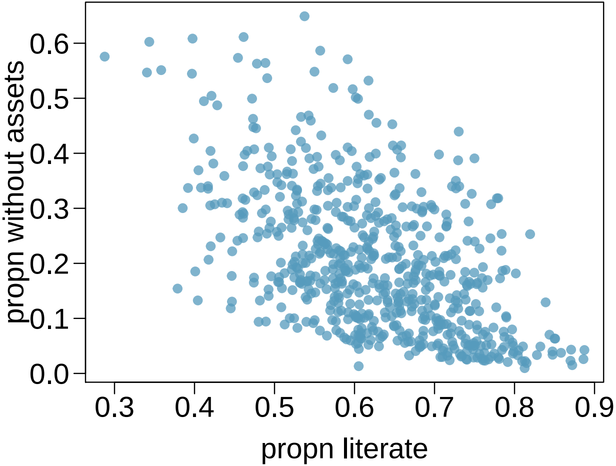
The *scatterplot* below shows the relationship between literacy rate in all 640 Indian districts and the % of households who don't own any assets.



Outcome variable?

Asset Poverty vs. Literacy Rate

The *scatterplot* below shows the relationship between literacy rate in all 640 Indian districts and the % of households who don't own any assets.

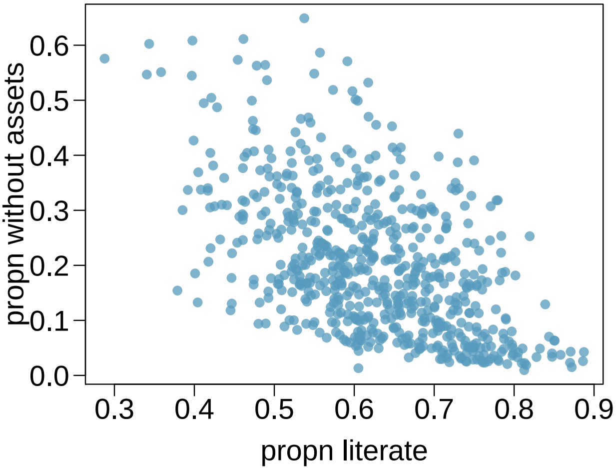


Outcome variable?

proportion of households without any assets

Asset Poverty vs. Literacy Rate

The *scatterplot* below shows the relationship between literacy rate in all 640 Indian districts and the % of households who don't own any assets.



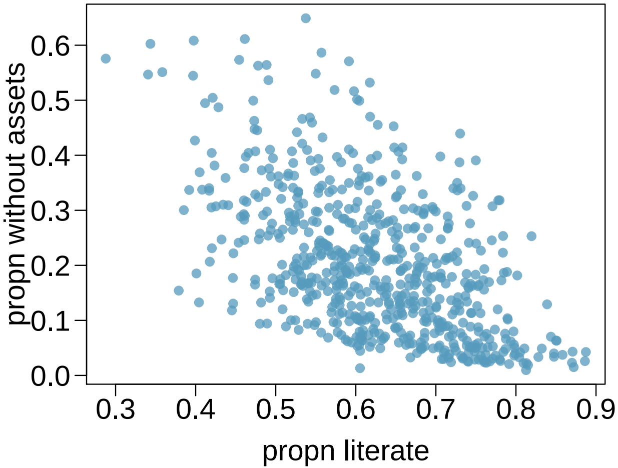
Outcome variable?

proportion of households without any assets

Predictor variable?

Asset Poverty vs. Literacy Rate

The *scatterplot* below shows the relationship between literacy rate in all 640 Indian districts and the % of households who don't own any assets.



Outcome variable?

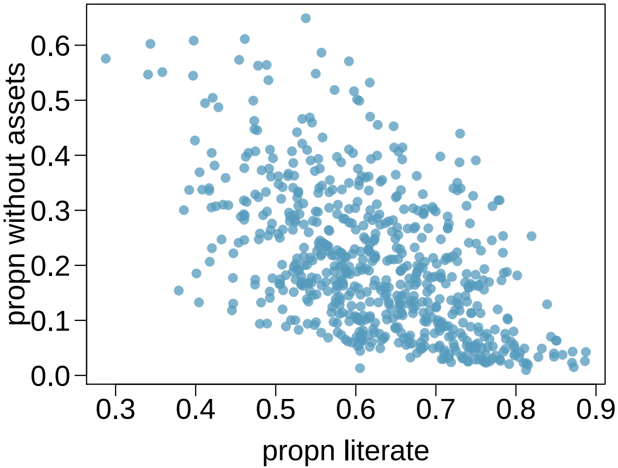
proportion of households without any assets

Predictor variable?

literacy proportion

Asset Poverty vs. Literacy Rate

The *scatterplot* below shows the relationship between literacy rate in all 640 Indian districts and the % of households who don't own any assets.



Outcome variable?

proportion of households without any assets

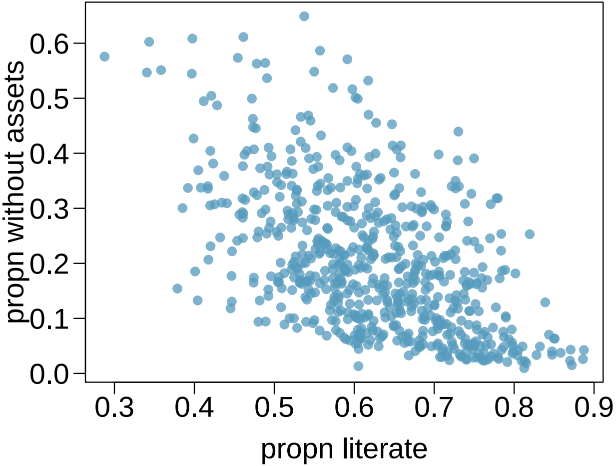
Predictor variable?

literacy proportion

Relationship?

Asset Poverty vs. Literacy Rate

The *scatterplot* below shows the relationship between literacy rate in all 640 Indian districts and the % of households who don't own any assets.



Outcome variable?

proportion of households without any assets

Predictor variable?

literacy proportion

Relationship?

linear, negative, moderately strong

The linear model for predicting asset-poverty from literacy rate in India is

$$\widehat{\text{poverty}} = 0.618 - 0.66 * \mathit{prop}_{Lit}$$

The "hat" is used to signify that this is an estimate.

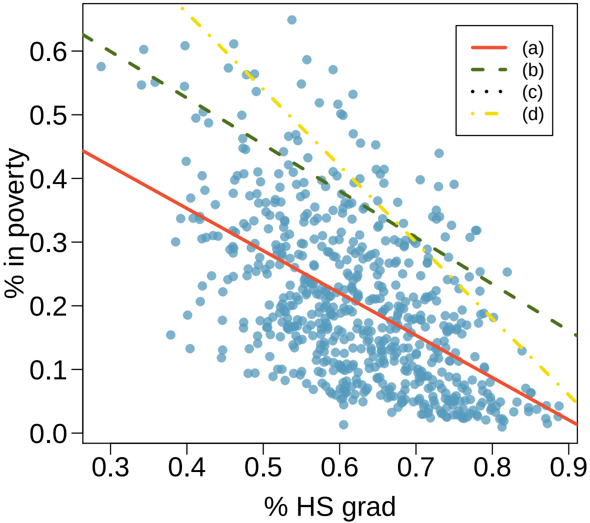
The literacy rate in Akola is 77.8%. What asset-poverty level does the model predict for this district?

The literacy rate in Akola is 77.8%. What asset-poverty level does the model predict for this district?

$$0.618 - 0.66 * 0.778 = 0.102$$

Eyeballing the line

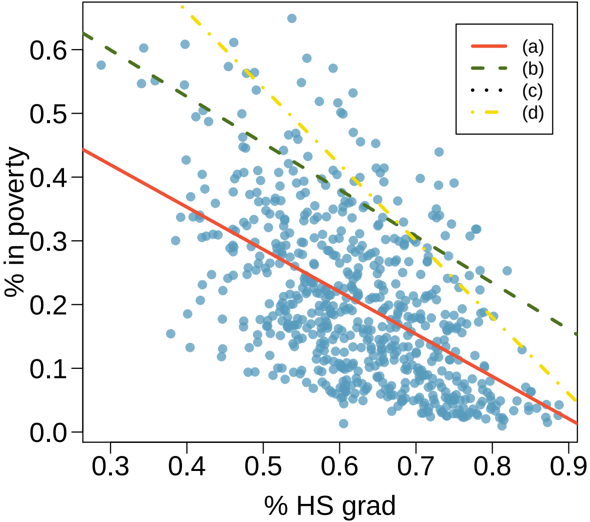
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % literate? Choose one.



Eyeballing the line

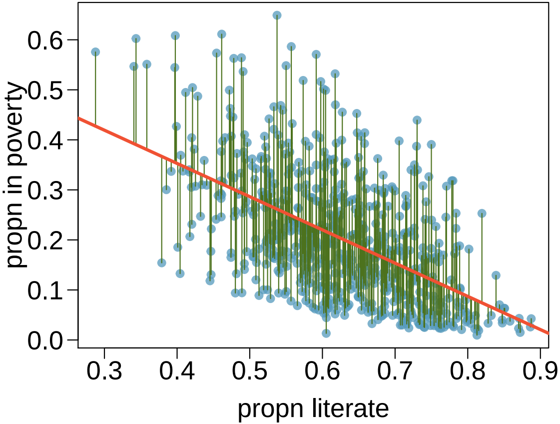
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % literate? Choose one.

(a)



Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

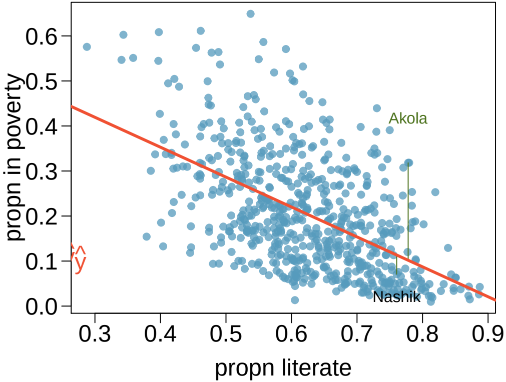


Residuals (cont.)

Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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

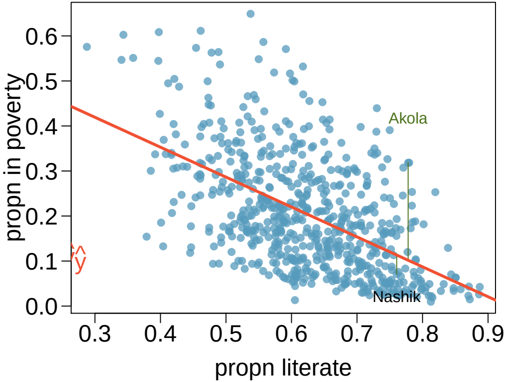


Residuals (cont.)

Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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



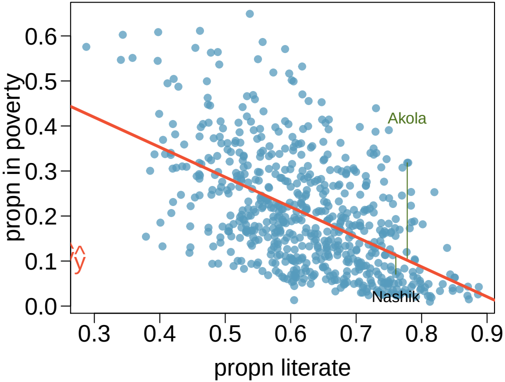
- % living in poverty in Akola is 21% more than predicted.

Residuals (cont.)

Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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



- % living in poverty in Akola is 21% more than predicted.
- % living in poverty in Nashik is 4% less than predicted.

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.

Guessing the correlation

Which of the following is the best guess for the correlation between propn in asset poverty and propn literate?

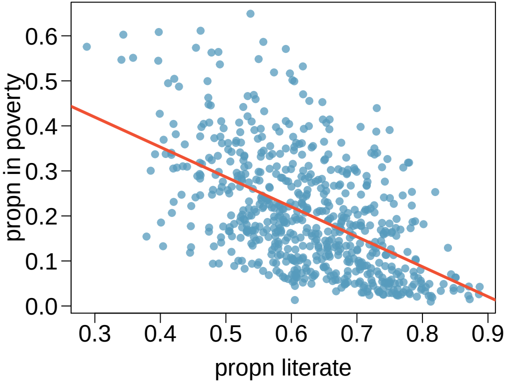
(a) 0.6

(b) -0.54

(c) -0.1

(d) 0.02

(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between propn in asset poverty and propn literate?

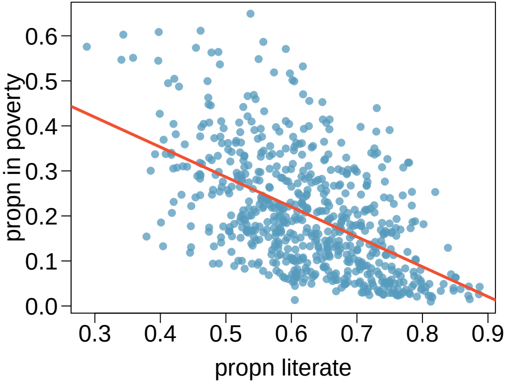
(a) 0.6

(b) *-0.54*

(c) -0.1

(d) 0.02

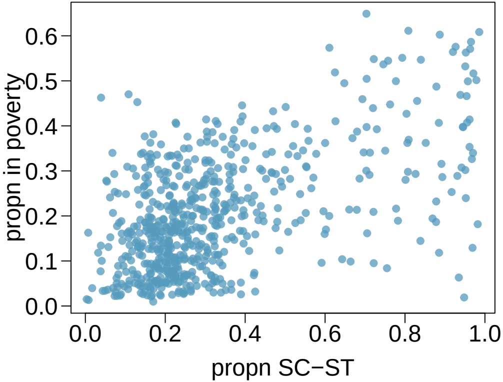
(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between propn in asset poverty and propn of SC-ST in a district?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.55



Guessing the correlation

Which of the following is the best guess for the correlation between propn in asset poverty and propn of SC-ST in a district?

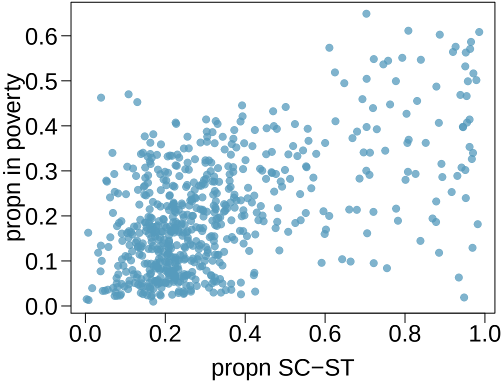
(a) 0.1

(b) -0.6

(c) -0.4

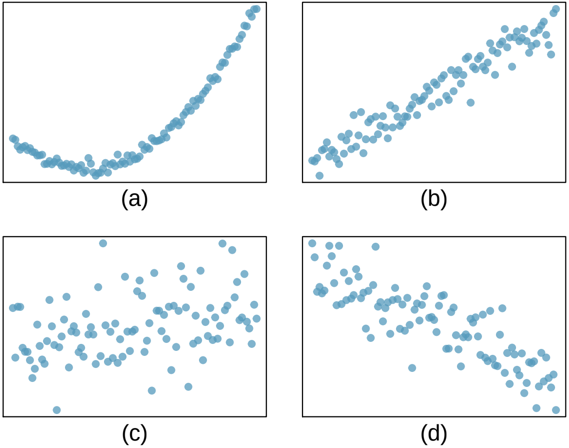
(d) 0.9

(e) *0.55*



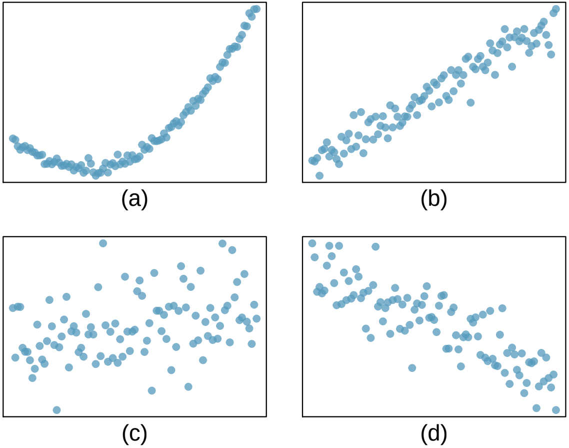
Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



(b) → correlation means linear association

Fitting a line by least squares regression

An objective measure for finding the best line

- We want a line that has small residuals:

An objective measure for finding the best line

- We want a line that has small residuals:
    1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  2. Option 2: Minimize the sum of squared residuals – *least squares*

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

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  2. Option 2: Minimize the sum of squared residuals – *least squares*

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

- Why least squares?

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  2. Option 2: Minimize the sum of squared residuals – *least squares*

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

- Why least squares?
  1. Most commonly used

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  2. Option 2: Minimize the sum of squared residuals – *least squares*

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

- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

  2. Option 2: Minimize the sum of squared residuals – *least squares*

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

- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software
  3. In many applications, a residual twice as large as another is usually more than twice as bad

The least squares line

$$\hat{y} = \beta_0 + \beta_1 x$$

- $\hat{y}$: Predicted value of the outcome variable, y
- β_0: Intercept, parameter
  - b_0: Intercept, point estimate
- β_1: Slope, parameter
  - b_1: Slope, point estimate
- x: Predictor variable

Conditions for the least squares line

1. Linearity

Conditions for the least squares line

1. Linearity
2. Nearly normal residuals

Conditions for the least squares line

1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

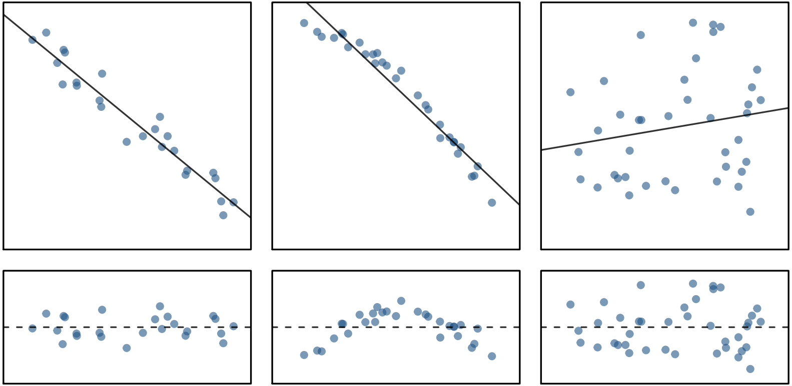
- The relationship between the predictor and the outcome variable should be linear.

Conditions: (1) Linearity

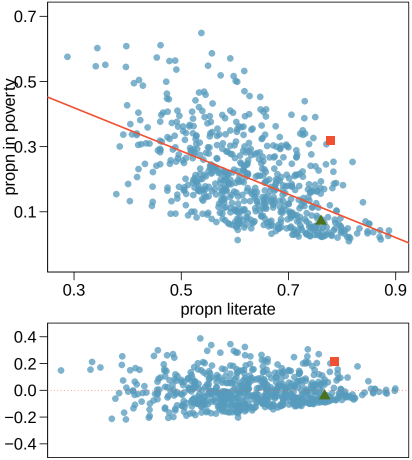
- The relationship between the predictor and the outcome variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.

Conditions: (1) Linearity

- The relationship between the predictor and the outcome variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.
- Check using a scatterplot of the data, or a *residuals plot*.



Anatomy of a residuals plot



Akola:

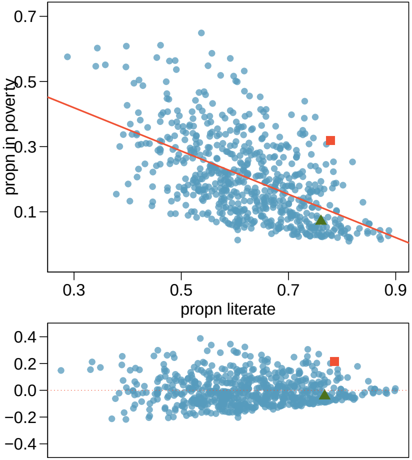
$$pLit = 0.778 \qquad pNoAssets = 0.318$$

$$\widehat{pNoAssets} = 0.619 - 0.664 * 0.778 = 0.102$$

$$e = pNoAssets - \widehat{pNoAssets}$$

$$= 0.318 - 0.108 = {\color{red}0.217}$$

Anatomy of a residuals plot



■ *Akola:*

$$pLit = 0.778 \qquad pNoAssets = 0.318$$

$$\widehat{pNoAssets} = 0.619 - 0.664 * 0.778 = 0.102$$

$$e = pNoAssets - \widehat{pNoAssets}$$

$$= 0.318 - 0.108 = 0.217$$

▲ *Nashik:*

$$pLit = 0.761 \qquad pNoAssets = 0.071$$

$$\widehat{pNoAssets} = 0.619 - 0.664 * 0.761 = 0.113$$

$$e = pNoAssets - \widehat{pNoAssets}$$

$$= 0.071 - 0.113 = -0.042$$

Conditions: (2) Nearly normal residuals

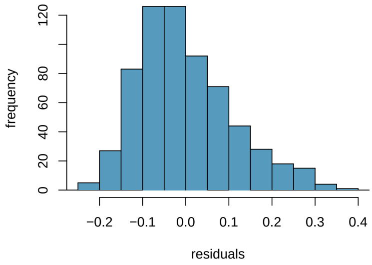
- The residuals should be nearly normal.

Conditions: (2) Nearly normal residuals

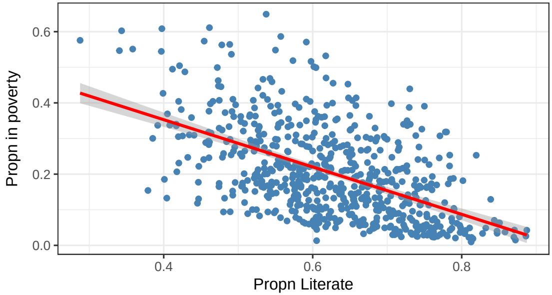
- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

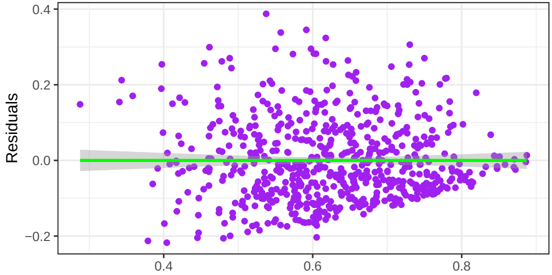
Conditions: (2) Nearly normal residuals

- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram.



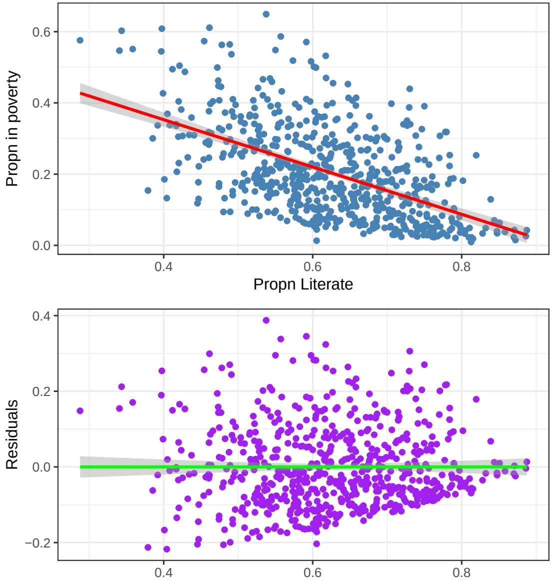
Conditions: (3) Constant variability





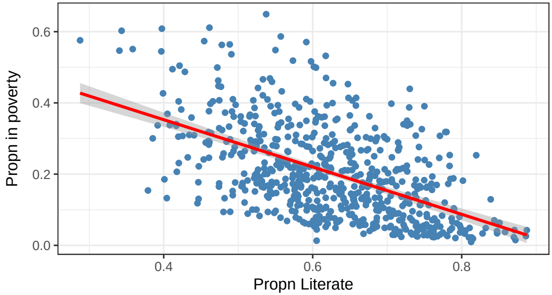
- The variability of points around the least squares line should be roughly constant.

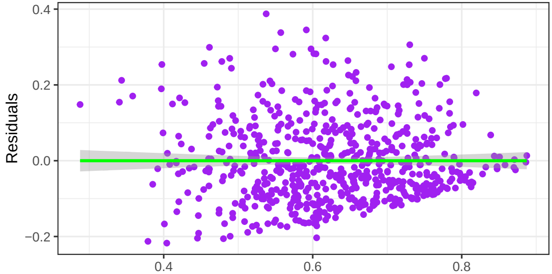
Conditions: (3) Constant variability



- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.

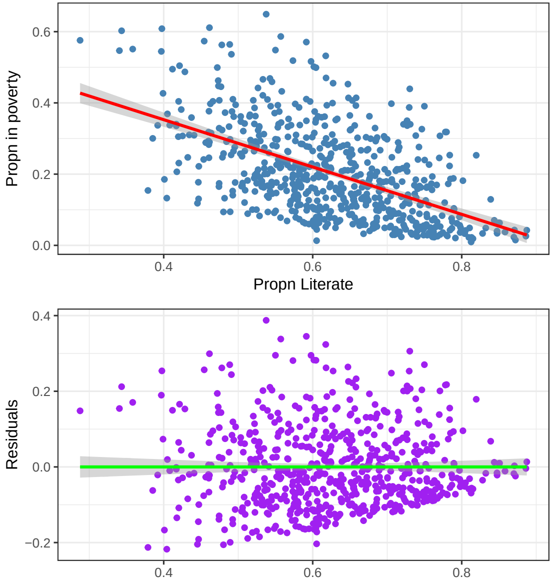
Conditions: (3) Constant variability





- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.

Conditions: (3) Constant variability

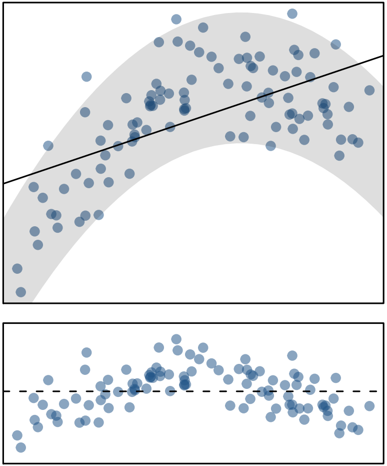


- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.
- Check using a residuals plot.

Checking conditions

What condition is this linear model obviously violating?

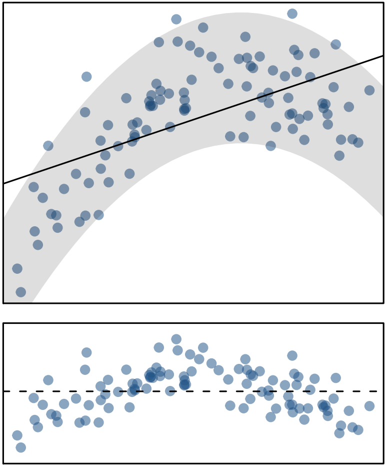
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

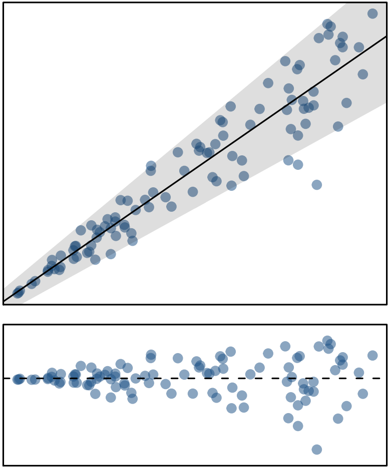
(a) Constant variability
(b) *Linear relationship*
(c) Normal residuals
(d) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

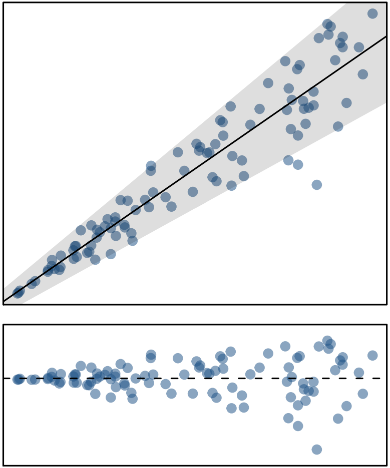
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



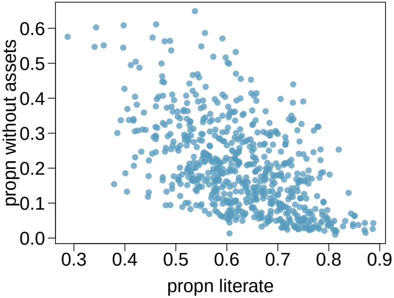
Checking conditions

What condition is this linear model obviously violating?

(a) *Constant variability*
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Given...



	propn literate (x)	propn in poverty (y)
mean	$\bar{x} = 0.6248$	$\bar{y} = 0.2036$
sd	$s_x = 0.105$	$s_y = 0.129$
	correlation	$R = -0.54$

Slope

Slope

The slope of the regression can be calculated as

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

Slope

Slope

The slope of the regression can be calculated as

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

In context...

$$b_1 = \frac{0.129}{0.105} \times -0.54 = -0.664$$

Slope

Slope

The slope of the regression can be calculated as

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

In context...

$$b_1 = \frac{0.129}{0.105} \times -0.54 = -0.664$$

Interpretation

For each additional percentage point in literacy rate, the asset poor rate would be lower on average by 0.66% points.

Intercept

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1\bar{x}$$

Intercept

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1\bar{x}$$

$$\begin{aligned} b_0 &= 0.2036 - (-0.664) \times 0.6248 \\ &= 0.619 \end{aligned}$$

Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in literacy rate, % living in poverty is expected to increase on average by 61.9%.

(b) For each % point decrease in literacy rate, % living in poverty is expected to increase on average by 61.9%.

(c) Having no literate person leads to 61.9% of households living without any assets.

(d) Districts with no literate population are expected on average to have 61.9% of households living in asset-poverty.

(e) In districts with no literate population % living in asset-poverty is expected to increase on average by 61.9%.

Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in literacy rate, % living in poverty is expected to increase on average by 61.9%.

(b) For each % point decrease in literacy rate, % living in poverty is expected to increase on average by 61.9%.

(c) Having no literate person leads to 61.9% of households living without any assets.

(d) *Districts with no literate population are expected on average to have 61.9% of households living in asset-poverty.*

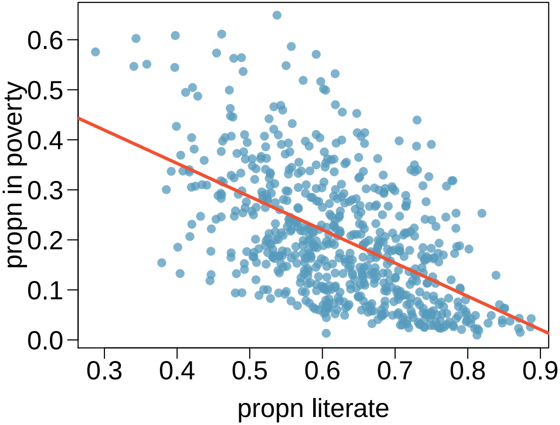
(e) In districts with no literate population % living in asset-poverty is expected to increase on average by 61.9%.

More on the intercept

Since there are no districts in the dataset with no literate population, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.

Regression line

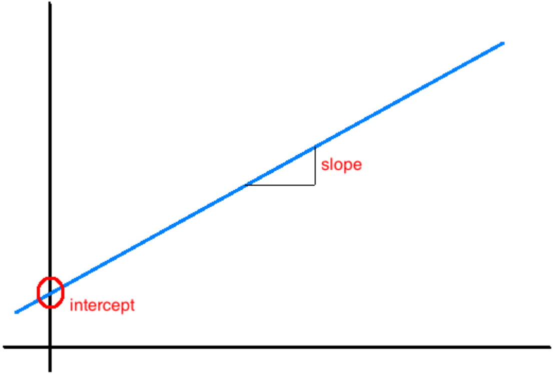
$$\widehat{pNoAssets} = 0.619 - 0.664 \times pLit$$



Interpretation of slope and intercept

- *Intercept:* When $x = 0$, y is expected to equal the intercept.

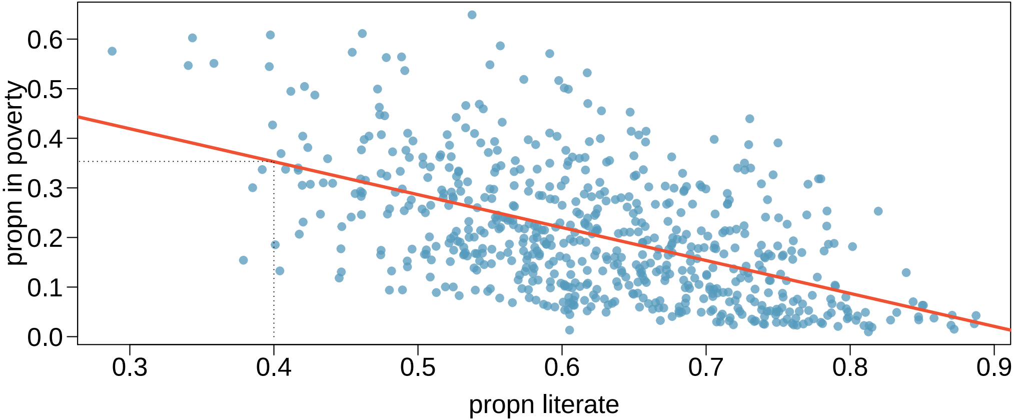
- *Slope:* For each unit in x, y is expected to increase / decrease on average by the slope.



Note: These statements are not causal, unless the study is a randomized controlled experiment.

Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the predicted value.



R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.

R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.

R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.

R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.

R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- For the model we've been working with, $R^2 = -0.54^2 = 0.29$.

Interpretation of R^2

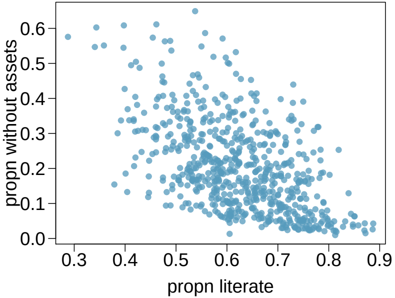
Which of the below is the correct interpretation of $R = -0.54$, $R^2 = 0.29$?

(a) 29% of the variability in the % of literates among the 640 districts is explained by the model.

(b) 29% of the variability in the % of households living in assets poverty among the 640 districts is explained by the model.

(c) 29% of the time % literates predict % living in asset poverty correctly.

(d) 71% of the variability in the % of households living in poverty among the 640 districts is explained by the model.



Interpretation of R^2

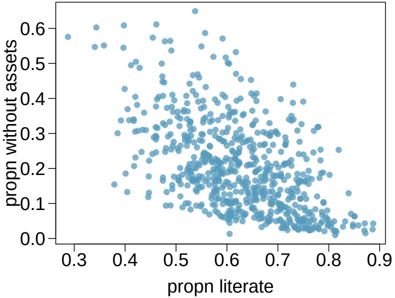
Which of the below is the correct interpretation of $R = -0.54$, $R^2 = 0.29$?

(a) 29% of the variability in the % of literates among the 640 districts is explained by the model.

(b) *29% of the variability in the % of households living in assets poverty among the 640 districts is explained by the model.*

(c) 29% of the time % literates predict % living in asset poverty correctly.

(d) 71% of the variability in the % of households living in poverty among the 640 districts is explained by the model.

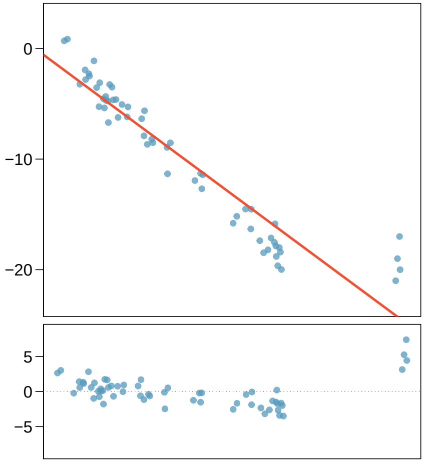


Types of outliers in linear regression

Types of outliers

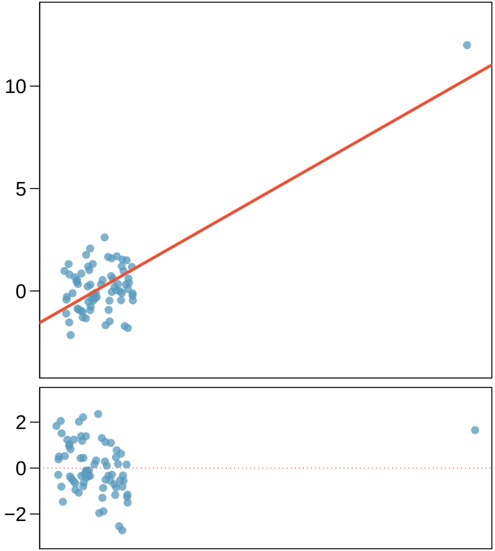
How do outliers influence the least squares line in this plot?

To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.



Types of outliers

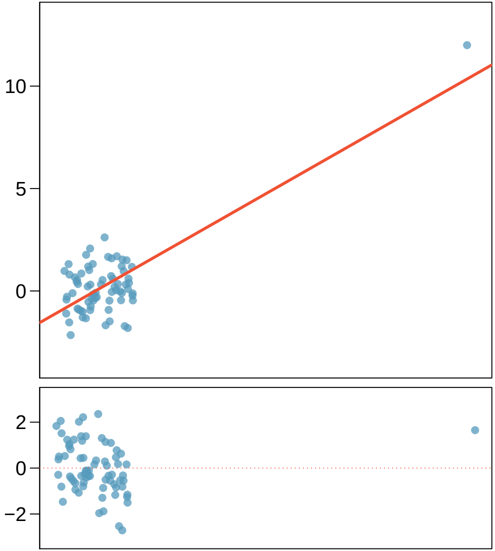
How do outliers influence the least squares line in this plot?



Types of outliers

How do outliers influence the least squares line in this plot?

Without the outlier there is no evident relationship between x and y.



Some terminology

- *Outliers* are points that lie away from the cloud of points.

Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.

Some terminology

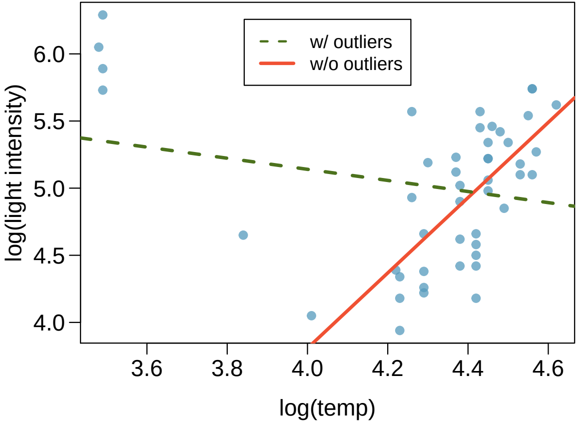
- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.

Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not an influential point.

Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



Types of outliers

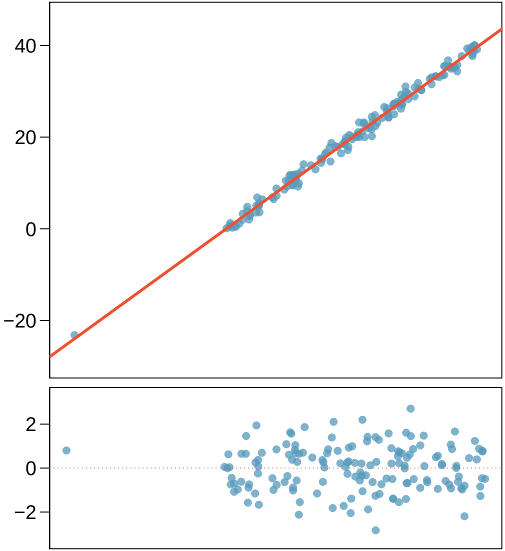
Which of the below best describes the outlier?

(a) influential

(b) high leverage

(c) none of the above

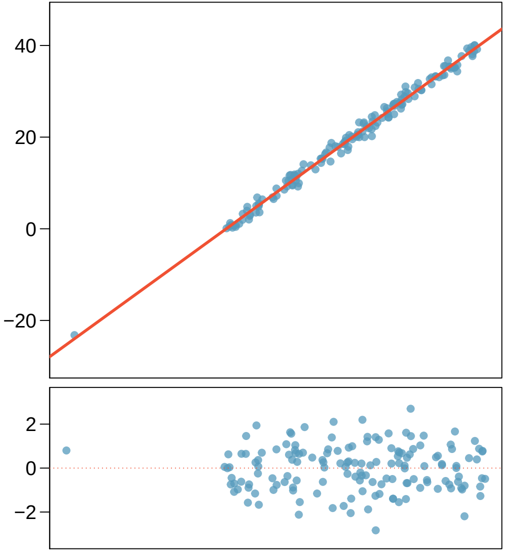
(d) there are no outliers



Types of outliers

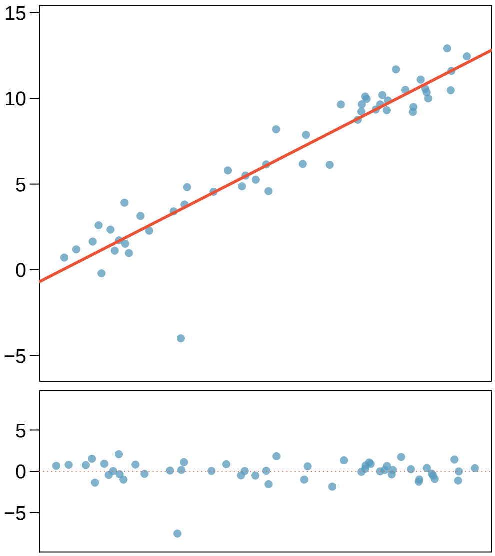
Which of the below best describes the outlier?

(a) influential
(b) *high leverage*
(c) none of the above
(d) there are no outliers



Types of outliers

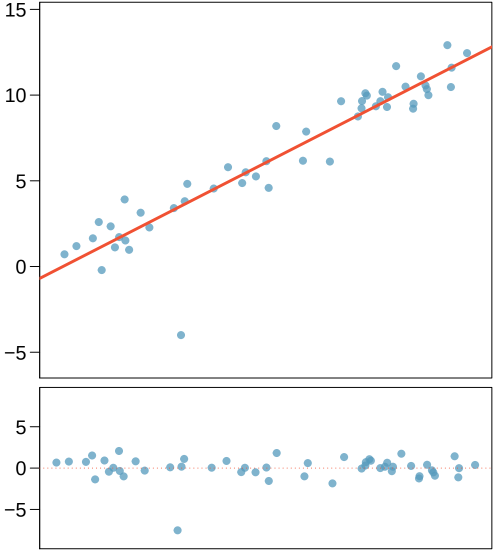
Does this outlier influence the slope of the regression line?



Types of outliers

Does this outlier influence the slope of the regression line?

Not much...



Recap

Which of following is true?

(a) Influential points always change the intercept of the regression line.

(b) Influential points always reduce R^2.

(c) It is much more likely for a low leverage point to be influential, than a high leverage point.

(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

(e) None of the above.

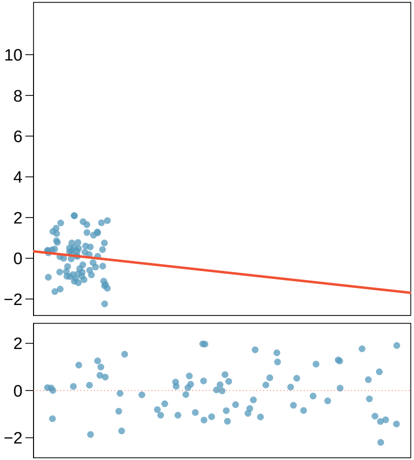
Recap

Which of following is true?

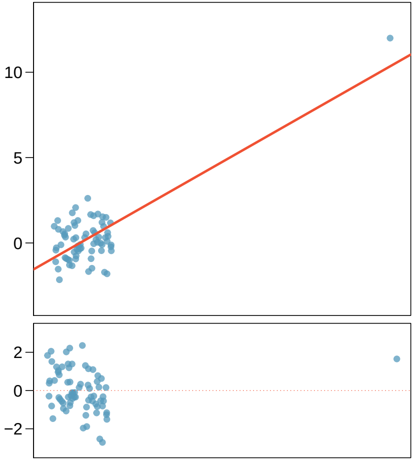
(a) Influential points always change the intercept of the regression line.
(b) Influential points always reduce R^2.
(c) It is much more likely for a low leverage point to be influential, than a high leverage point.
(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
(e) *None of the above.*

Recap (cont.)

$R = 0.08$, $R^2 = 0.0064$



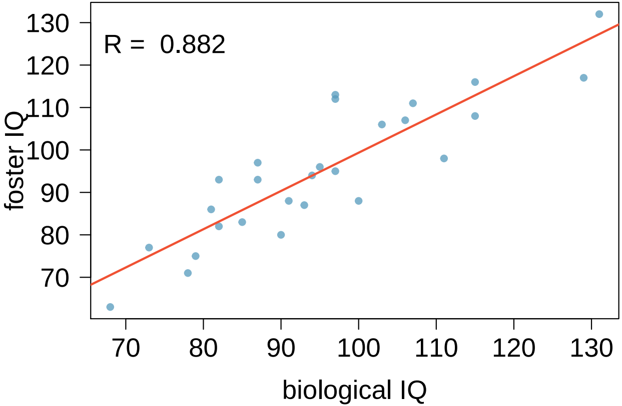
$R = 0.79$, $R^2 = 0.6241$



Inference for linear regression

Nature or nurture?

In 1966 Cyril Burt published a paper called "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?" The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Which of the following is false?

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.20760    9.29990   0.990    0.332
bioIQ             0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.

(b) Roughly 78% of the foster twins' IQs can be accurately predicted by the model.

(c) The linear model is $\widehat{fosterIQ} = 9.2 + 0.9 \times bioIQ$.

(d) Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Which of the following is false?

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.20760    9.29990   0.990    0.332
bioIQ             0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.

(b) *Roughly 78% of the foster twins' IQs can be accurately predicted by the model.*

(c) The linear model is $\widehat{fosterIQ} = 9.2 + 0.9 \times bioIQ$.

(d) Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

(a) $H_0 : b_0 = 0$; $H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0$; $H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

(a) $H_0 : b_0 = 0; H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0; H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0; H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{point\ estimate - null\ value}{SE}$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate = b_1 is the observed slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: Test statistic, $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

  Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

$$df = 27 - 2 = 25$$

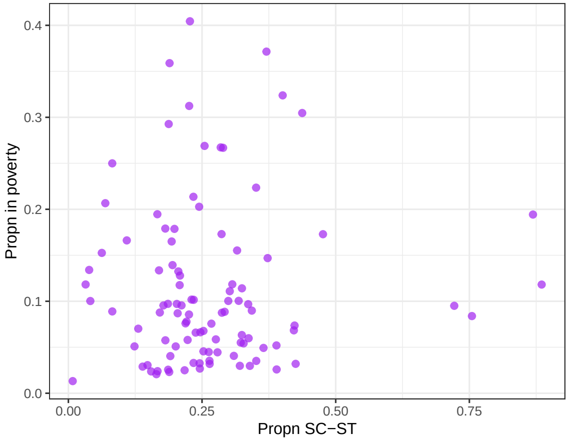
Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$
\begin{aligned}
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25 \\
p - value &= P(|T| > 9.36) < 0.01
\end{aligned}
$$

% in Poverty vs. % SC-ST

What can you say about the relationship between % in poverty and % SC-ST in a sample of 100 districts in India?



% in Poverty vs % SC-ST - linear model

Which of the below is the best interpretation of the slope?

	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.10	0.02	5.75	0.00
2	pSCST	0.03	0.06	0.50	0.62

(a) A 1% increase in SC-ST population in a district is associated with a 3% increase in % of asset poor.

(b) A 1% increase in SC-ST population in a district is associated with a 0% increase in % of asset poverty.

(c) An additional 1% of SC-ST population increases the % of asset poor in a district by 10%.

(d) In districts with zero SC-ST population, % of asset poor is expected to be 75%.

% in Poverty vs % SC-ST - linear model

Which of the below is the best interpretation of the slope?

	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.10	0.02	5.75	0.00
2	pSCST	0.03	0.06	0.50	0.62

(a) A 1% increase in SC-ST population in a district is associated with a 3% increase in % of asset poor.

(b) *A 1% increase in SC-ST population in a district is associated with a 0% increase in % of asset poverty.*

(c) An additional 1% of SC-ST population increases the % of asset poor in a district by 10%.

(d) In districts with zero SC-ST population, % of asset poor is expected to be 75%.

% in Poverty vs. % SC-ST - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % SC-ST and % asset poor in randomly chosen Indian districts?

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	0.1049	0.0182	5.75	0.0000
pSCST	0.0301	0.0602	0.50	0.6179

How reliable is this p-value if these zip code areas are not randomly selected?

% in Poverty vs. % SC-ST - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % SC-ST and % asset poor in randomly chosen Indian districts?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1049	0.0182	5.75	0.0000
pSCST	0.0301	0.0602	0.50	0.6179

No, the p-value for % SCST is low, indicating that the data does not provide convincing evidence that the slope parameter is different than 0.

How reliable is this p-value if these zip code areas are not randomly selected?

% in Poverty vs. % SC-ST - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % SC-ST and % asset poor in randomly chosen Indian districts?

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	0.1049	0.0182	5.75	0.0000
pSCST	0.0301	0.0602	0.50	0.6179

No, the p-value for % SCST is low, indicating that the data does not provide convincing evidence that the slope parameter is different than 0.

How reliable is this p-value if these zip code areas are not randomly selected?

Not very...

Confidence interval for the slope

Remember that a confidence interval is calculated as $point\ estimate \pm ME$ and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

Confidence interval for the slope

Remember that a confidence interval is calculated as $point\ estimate \pm ME$ and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$n = 27 \qquad df = 27 - 2 = 25$

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

$$n = 27 \qquad df = 27 - 2 = 25$$

$$95\% : \; t^{\star}_{25} = 2.06$$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

$$n = 27 \qquad df = 27 - 2 = 25$$

$$95\% : t^{\star}_{25} = 2.06$$

$$0.9014 \pm 2.06 \times 0.0963$$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

$$
\begin{aligned}
n &= 27 \qquad df = 27 - 2 = 25 \\
95\% :\ t^{\star}_{25} &= 2.06 \\
0.9014 &\pm 2.06 \times 0.0963 \\
(0.7 &,\ 1.1)
\end{aligned}
$$

Recap

- Inference for the slope for a single-predictor linear regression model:

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- The null value is often 0 since we are usually checking for *any* relationship between the explanatory and the response variable.

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- The null value is often 0 since we are usually checking for *any* relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and *two-tailed* p-value for the t-test for the slope where the null value is 0.

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- The null value is often 0 since we are usually checking for *any* relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and *two-tailed* p-value for the t-test for the slope where the null value is 0.
- We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.
- If you have a sample that is non-random (biased), inference on the results will be unreliable.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.
- If you have a sample that is non-random (biased), inference on the results will be unreliable.
- The ultimate goal is to have independent observations.