Chapter 8: Introduction to linear regression

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

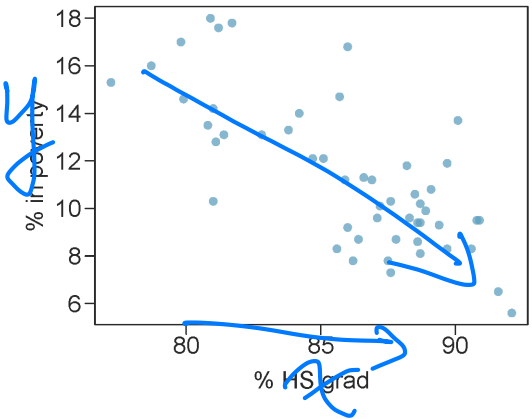
OpenIntro

Outline

Line fitting, residuals, and correlation

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

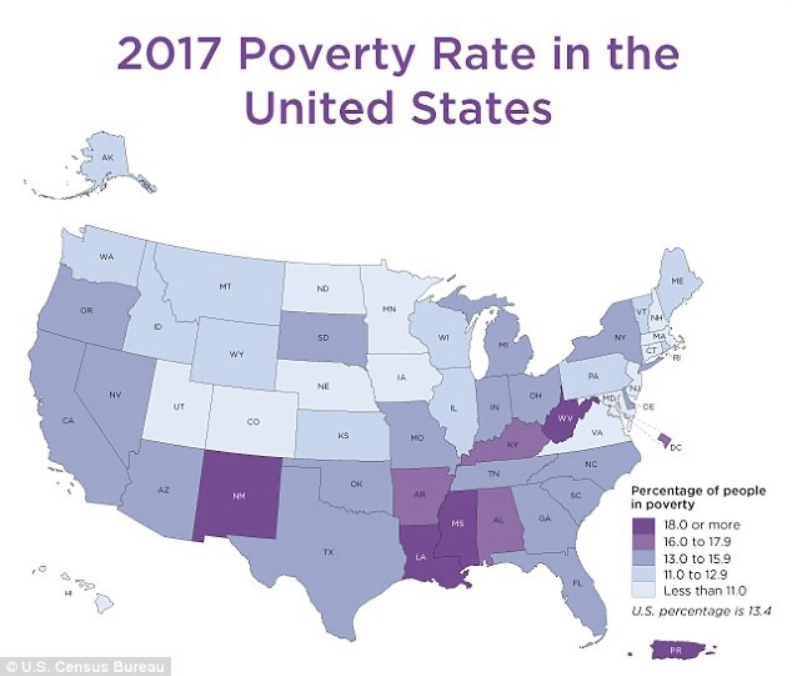
% in poverty

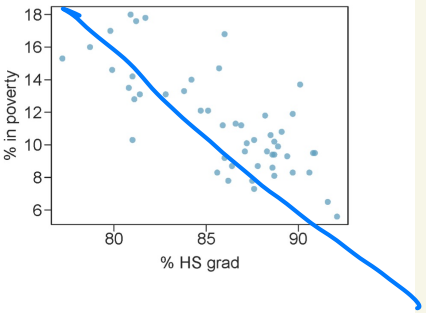
Explanatory variable?

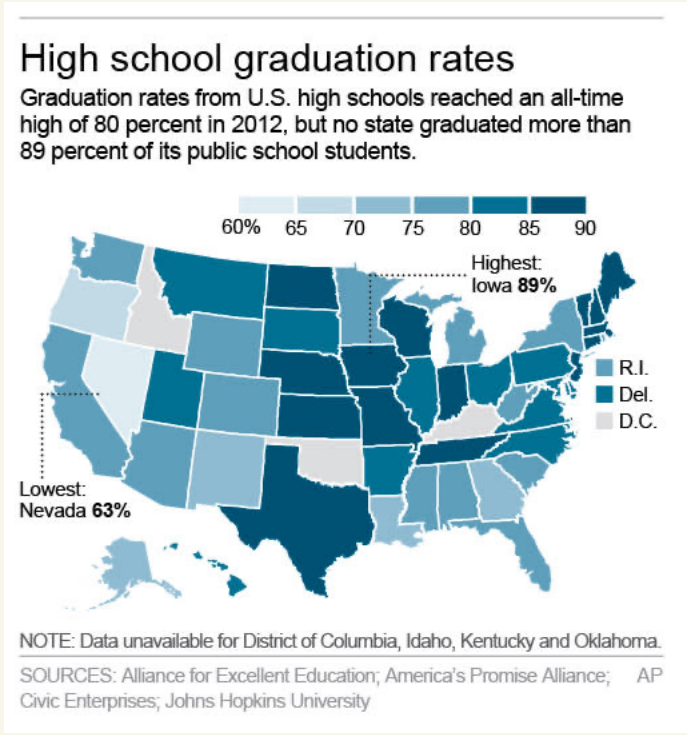
% HS grad

Relationship?

linear, negative, moderately strong


2017 Poverty Rate in the United States
Percentage of people in poverty
18.0 or more
16.0 to 17.9
13.0 to 15.9
11.0 to 12.9
Less than 11.0
U.S. percentage is 13.4
© U.S. Census Bureau


% in poverty
% HS grad
18
16
14
12
10
8
6
80
85
90


High school graduation rates
Graduation rates from U.S. high schools reached an all-time high of 80 percent in 2012, but no state graduated more than 89 percent of its public school students.
60%
65
70
75
80
85
90
Highest: Iowa 89%
Lowest: Nevada 63%
R.I.
Del.
D.C.
NOTE: Data unavailable for District of Columbia, Idaho, Kentucky and Oklahoma.
SOURCES: Alliance for Excellent Education; America's Promise Alliance; AP
Civic Enterprises; Johns Hopkins University

The linear model for predicting poverty from high school graduation rate in the US is

$$\hat{poverty} = 64.78 - 0.62 * HS_{grad}$$

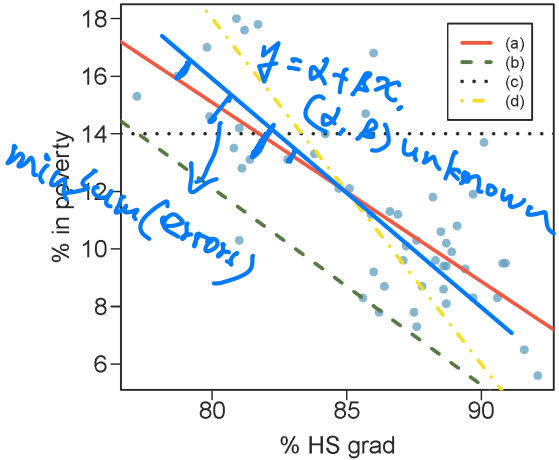
The "hat" is used to signify that this is an estimate.

The high school graduate rate in Georgia is 85.1%. What poverty level does the model predict for this state?

$$64.78 - 0.62 * 85.1 = 12.018$$

Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one. (a)

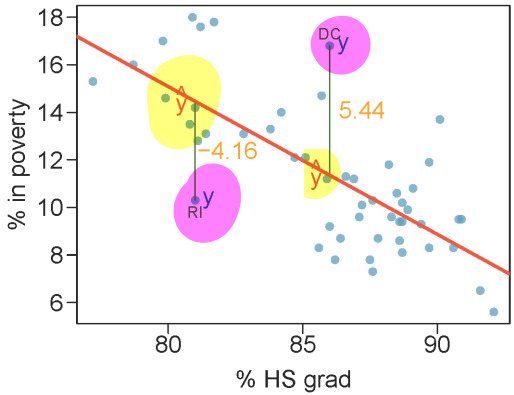


Residuals

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

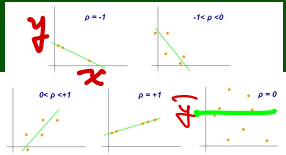
Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

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.
- $R = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$, where s_x, s_y are the sample standard deviations of x_i and y_i.
- Note that R represents the cosine of the angle between the vectors $\mathbf{a} = (x_1 - \bar{x}, \ldots, x_n - \bar{x})$ and $\mathbf{b} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$.

Geometric definition

In Euclidean space, a Euclidean vector is a geometric object that possesses both a magnitude and a direction. A vector can be pictured as an arrow. Its magnitude is its length, and its direction is the direction to which the arrow points. The magnitude of a vector $\mathbf{a}$ is denoted by $\|\mathbf{a}\|$. The dot product of two Euclidean vectors $\mathbf{a}$ and $\mathbf{b}$ is defined by[3][4][1]

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta,$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$.

In particular, if the vectors $\mathbf{a}$ and $\mathbf{b}$ are orthogonal (i.e., their angle is $\frac{\pi}{2}$ or 90°), then $\cos\frac{\pi}{2} = 0$, which implies that

$$\mathbf{a} \cdot \mathbf{b} = 0.$$

At the other extreme, if they are codirectional, then the angle between them is zero with $\cos 0 = 1$ and

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\|$$

This implies that the dot product of a vector $\mathbf{a}$ with itself is

$$\mathbf{a} \cdot \mathbf{a} = \|\mathbf{a}\|^2,$$

which gives

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}},$$

the formula for the Euclidean length of the vector.

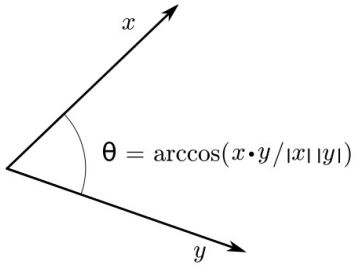
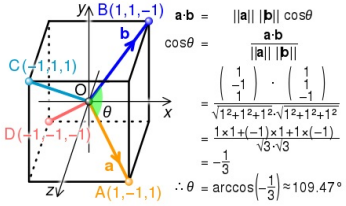


Illustration showing how to find the angle between vectors using the dot product



Calculating bond angles of a symmetrical tetrahedral molecular geometry using a dot product

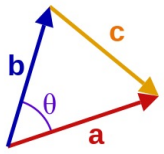
Application to the law of cosines

Main article: Law of cosines

Given two vectors $\mathbf{a}$ and $\mathbf{b}$ separated by angle θ (see image right), they form a triangle with a third side $\mathbf{c} = \mathbf{a} - \mathbf{b}$. Let a, b and c denote the lengths of $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, respectively. The dot product of this with itself is:

$$\begin{aligned}
\mathbf{c} \cdot \mathbf{c} &= (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) \\
&= \mathbf{a} \cdot \mathbf{a} - \mathbf{a} \cdot \mathbf{b} - \mathbf{b} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} \\
&= a^2 - \mathbf{a} \cdot \mathbf{b} - \mathbf{a} \cdot \mathbf{b} + b^2 \\
&= a^2 - 2\mathbf{a} \cdot \mathbf{b} + b^2 \\
c^2 &= a^2 + b^2 - 2ab\cos\theta
\end{aligned}$$

which is the law of cosines.



c = a − b

Triangle with vector edges **a** and **b**, separated by angle θ.

Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?

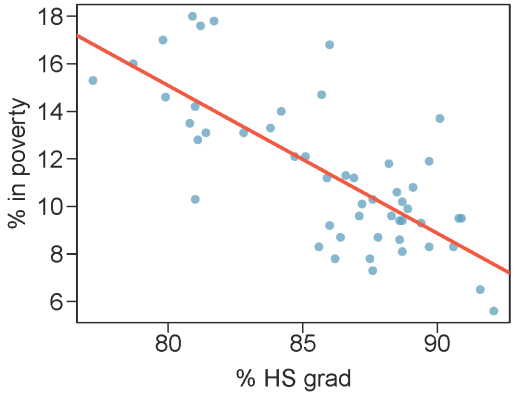
(a) 0.6

(b) -0.75

(c) -0.1

(d) 0.02

(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % female householder, no husband present?

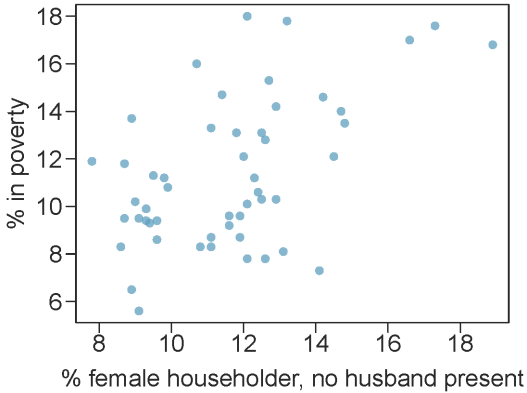
(a) 0.1

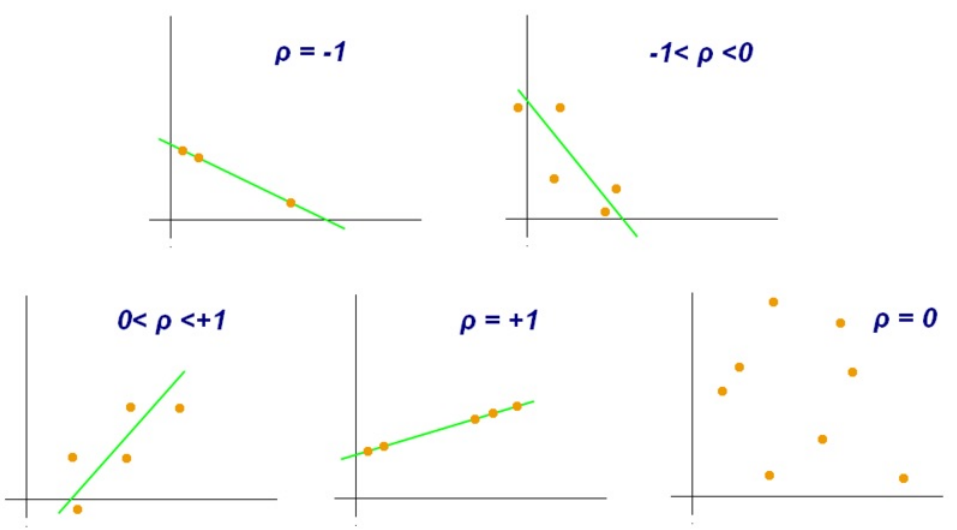
(b) -0.6

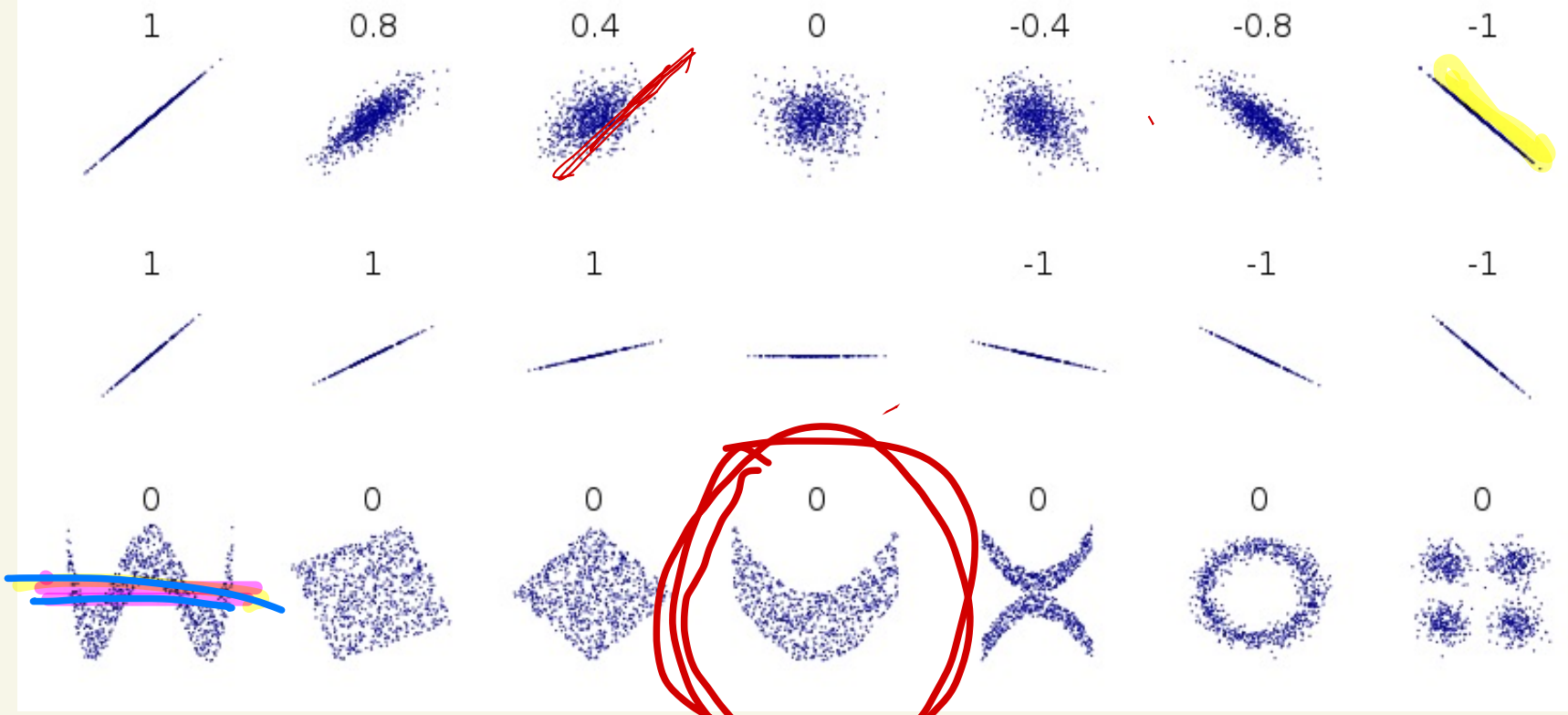
(c) -0.4

(d) 0.9

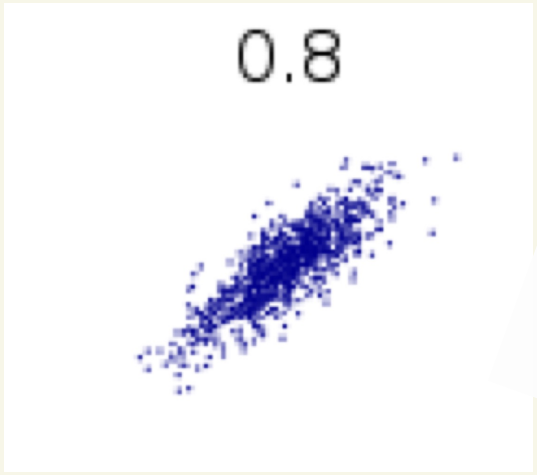
(e) 0.5







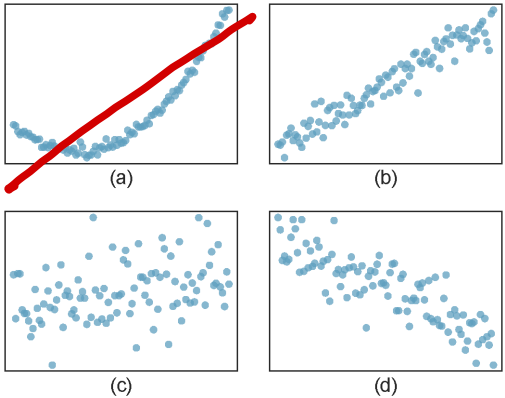
R: (linear) correlation coef.


0.8


R<0.8

Assessing the correlation

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



(b) $\rightarrow$ correlation means linear association

Fitting a line by least squares regression

$$\min \sum e_i^2$$

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-least absolute deviations

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

  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

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

$\varepsilon \sim N(0, \sigma^2)$

$\beta_0, \beta_1, \sigma^2$ unknown

$E(y|x) = \beta_0 + \beta_1 x$

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

- $\hat{y}$: Predicted value of the response variable, y
- β_0: Intercept, parameter
  - b_0 (or $\hat{\beta}_0$): Intercept, point estimate
- β_1 : Slope, parameter
  - b_1 (or $\hat{\beta}_1$): Slope, point estimate
- x: Explanatory variable

The method

Carl Friedrich Gauss

The first clear and concise exposition of the method of least squares was published by Legendre in 1805.[7] The technique is described as an algebraic procedure for fitting linear equations to data and Legendre demonstrates the new method by analyzing the same data as Laplace for the shape of the Earth. Within ten years after Legendre's publication, the method of least squares had been adopted as a standard tool in astronomy and geodesy in France, Italy, and Prussia, which constitutes an extraordinarily rapid acceptance of a scientific technique.[6]

In 1809 Carl Friedrich Gauss published his method of calculating the orbits of celestial bodies. In that work he claimed to have been in possession of the method of least squares since 1795.[8] This naturally led to a priority dispute with Legendre. However, to Gauss's credit, he went beyond Legendre and succeeded in connecting the method of least squares with the principles of probability and to the normal distribution. He had managed to complete Laplace's program of specifying a mathematical form of the probability density for the observations, depending on a finite number of unknown parameters, and define a method of estimation that minimizes the error of estimation. Gauss showed that the arithmetic mean is indeed the best estimate of the location parameter by changing both the probability density and the method of estimation. He then turned the problem around by asking what form the density should have and what method of estimation should be used to get the arithmetic mean as estimate of the location parameter. In this attempt, he invented the normal distribution.

Conditions for the least squares line

Data: $(x_1, y_1) \cdots (x_n, y_n)$

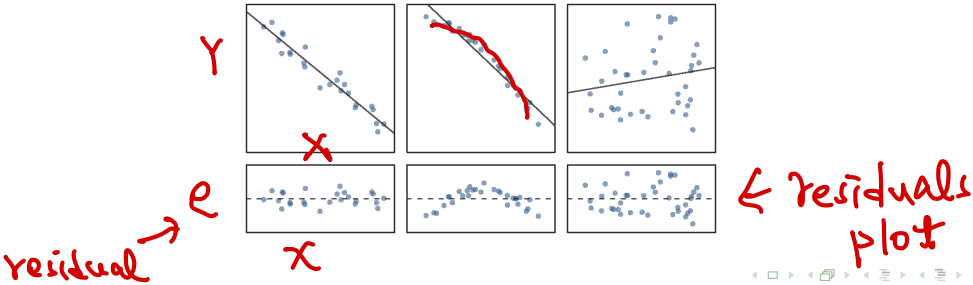
model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \cdots n$

ε_i 獨立 $N(0, \sigma^2)$

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

Conditions: (1) Linearity

- The relationship between the explanatory and the response 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 `http://www.openintro.org/download.php?file=os2_extra_nonlinear_relationships&referrer=/stat/textbook.php`Online Extra is available on openintro.org covering new techniques.
- Check using a scatterplot of the data, or a residuals plot.



Conditions: (2) Nearly normal residuals

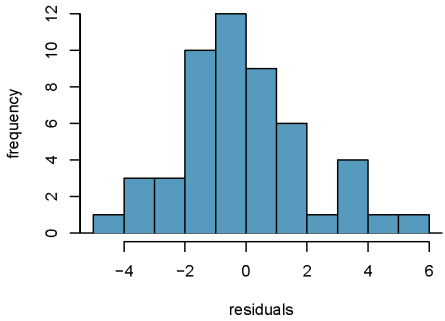
- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram, Q-Q plot

model: $y = \beta_0 + \beta_1 x_i + \varepsilon_i$

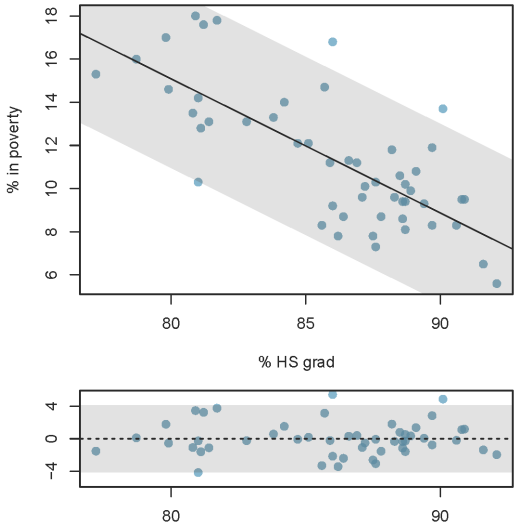
Fitted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

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

$e_i \approx \hat{\varepsilon}_i$



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.

← residuals plot

Checking conditions

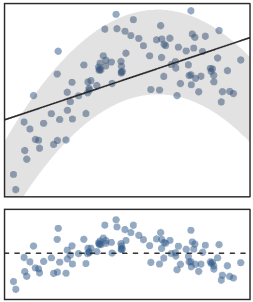
What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Checking conditions

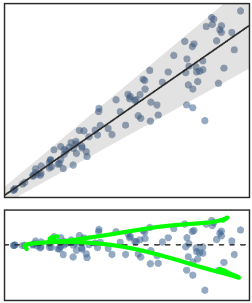
What condition is this linear model obviously violating?

(a) Constant variability

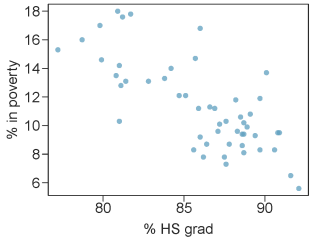
(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Given...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Slope

The slope of the regression can be calculated as

$$b_1 = \frac{s_y}{s_x} R \quad (\hat{\beta}_1)$$

In context...

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

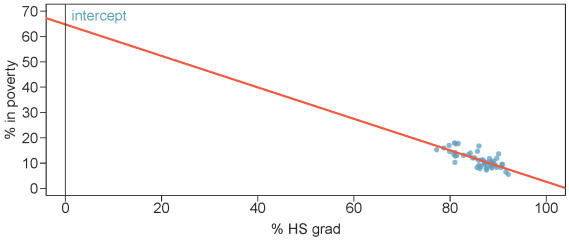
Interpretation

For each additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

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 $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x} \quad (\hat{\beta}_0)$$



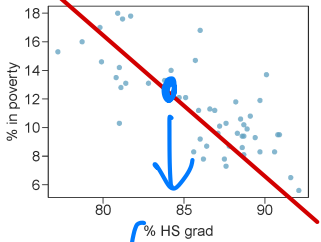
$$\begin{aligned} b_0 &= 11.35 - (-0.62) \times 86.01 \\ &= 64.68 \end{aligned}$$

① $\hat{y} = b_0 + b_1 x$, ② $b_1 = \frac{s_y}{s_x} R$

③ $\hat{y} = \bar{y} + b_1 (x - \bar{x})$

Given...

$\hat{y} = \bar{y} + b_1(x - \bar{x})$



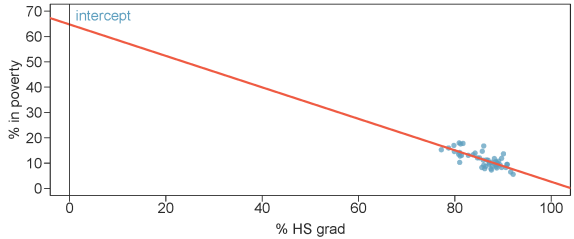
$(\bar{x}, \bar{y})$

	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

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

More on the intercept

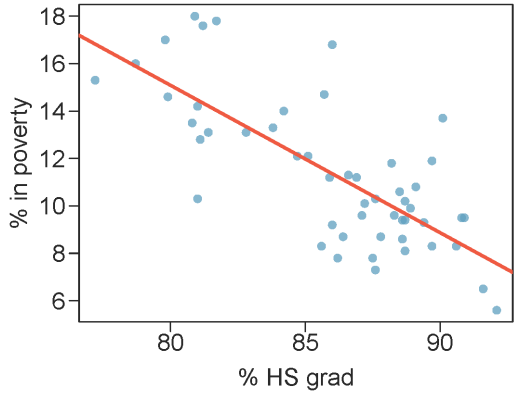
Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.



Regression line

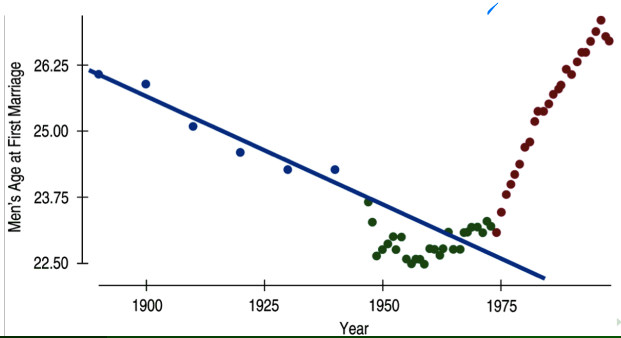
Predictions work best where we have data

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Prediction and extrapolation

- 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.
- Applying a model estimate to values outside of the realm of the original data is called extrapolation.
- Extrapolation can be very unreliable



Examples of extrapolation

BBC NEWS

Watch One-Minute World News

News Front Page

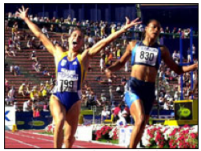
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.



Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

Examples of extrapolation

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

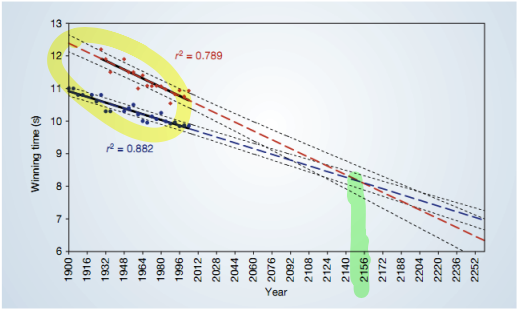
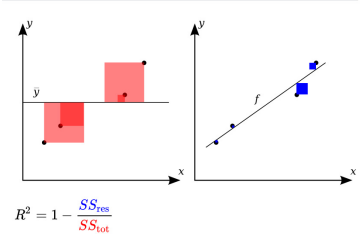


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.096 s.

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.
- For the model we've been working with, $R^2 = (-0.75)^2 = 0.56$.
- $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\text{Residuals sum of squares}}{\text{Total sum of squares}}$



Interpretation of R^2

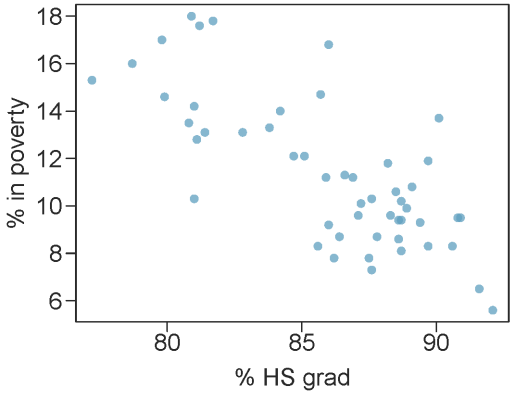
Which of the below is the correct interpretation of $R = -0.75$, $R^2 = 0.56$?

(a) 56% of the variability in the % of HG graduates among the 51 states is explained by the model.

(b) 56% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

(c) 56% of the time % HS graduates predict % living in poverty correctly.

(d) 75% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



LSE · $\hat{y} = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x})$

The goal is to find estimated values $\widehat{\alpha}$ and $\widehat{\beta}$ for the parameters α and β which would provide the "best" fit in some sense for the data points. As mentioned in the introduction, in this article the "best" fit will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals (see also Errors and residuals) $\hat{\varepsilon}_i$ (differences between actual and predicted values of the dependent variable *y*), each of which is given by, for any candidate parameter values α and β,

$$\hat{\varepsilon}_i = y_i - \alpha - \beta x_i .$$

In other words, $\widehat{\alpha}$ and $\widehat{\beta}$ solve the following minimization problem:

$$(\hat{\alpha},\, \hat{\beta}) = \operatorname{argmin}(Q(\alpha, \beta)),$$

where the objective function Q is:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

$\frac{\partial Q}{\partial \alpha} = 0$, $\frac{\partial Q}{\partial \beta} = 0$

By expanding to get a quadratic expression in α and β, we can derive minimizing values of the function arguments, denoted $\widehat{\alpha}$ and $\widehat{\beta}$:[6]

$$\widehat{\alpha} = \bar{y} - (\widehat{\beta}\,\bar{x}),$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} \Delta x_i \Delta y_i}{\sum_{i=1}^{n} \Delta x_i^2} = r \cdot \frac{s_y}{s_x}$$

Here we have introduced

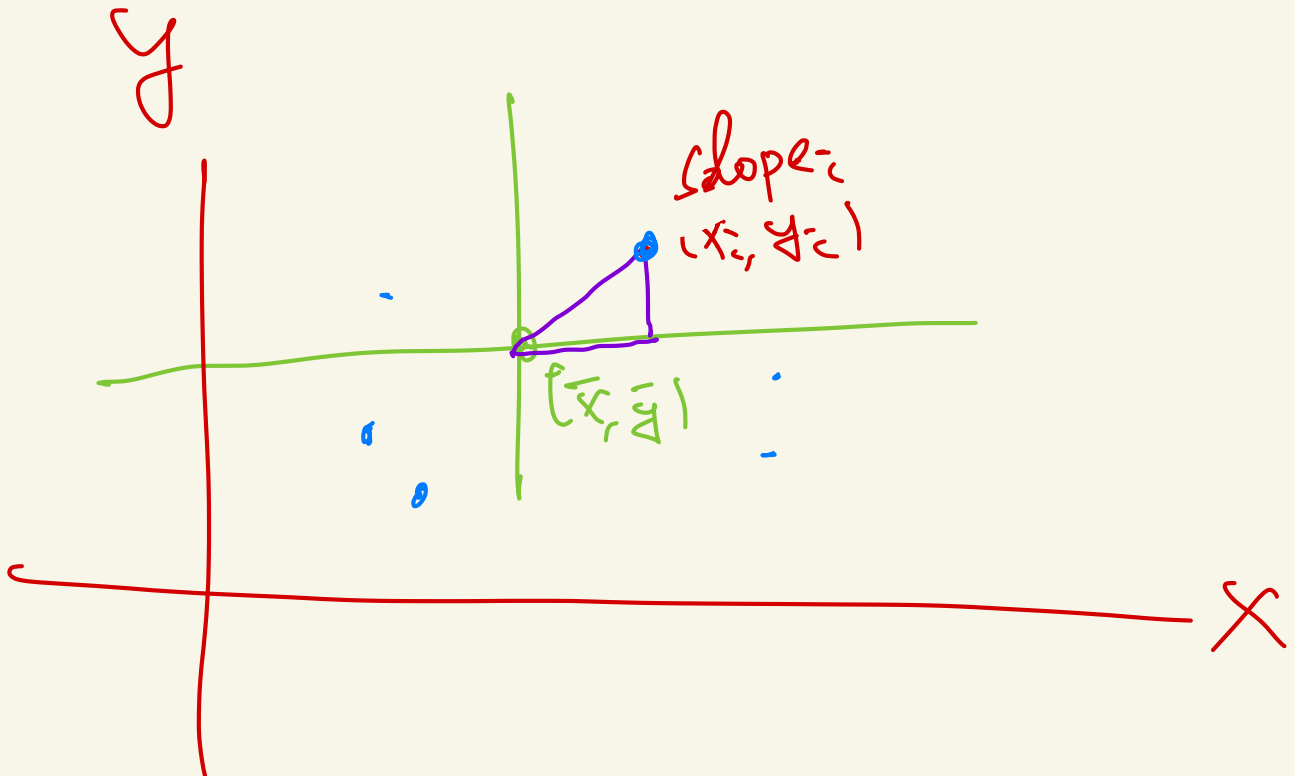
- $\bar{x}$ and $\bar{y}$ as the average of the x_i and y_i, respectively
- Δx_i and Δy_i as the deviations in x_i and y_i with respect to their respective means.

Interpretation about the slope

By multiplying all members of the summation in the numerator by : $\frac{(x_i - \bar{x})}{(x_i - \bar{x})} = 1$ (thereby not changing it):

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{(y_i - \bar{y})}{(x_i - \bar{x})}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

We can see that the slope (tangent of angle) of the regression line is the weighted average of $\frac{(y_i - \bar{y})}{(x_i - \bar{x})}$ that is the slope (tangent of angle) of the line that connects the i-th point to the average of all points, weighted by $(x_i - \bar{x})^2$ because the further the point is the more "important" it is, since small errors in its position will affect the slope connecting it to the center point more.



Definitions

A data set has n values marked $y_1, \ldots, y_n$ (collectively known as y_i or as a vector $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$), each associated with a fitted (or modeled, or predicted) value $f_1, \ldots, f_n$ (known as f_i, or sometimes $\hat{y}_i$, as a vector $\boldsymbol{f}$).

Define the residuals as $e_i = y_i - f_i$ (forming a vector $\boldsymbol{e}$).

If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

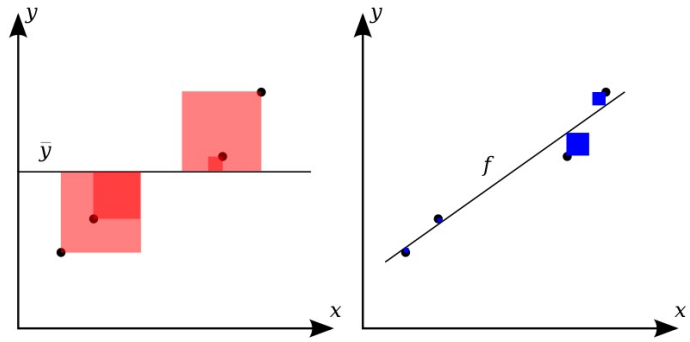
then the variability of the data set can be measured with two sums of squares formulas:

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

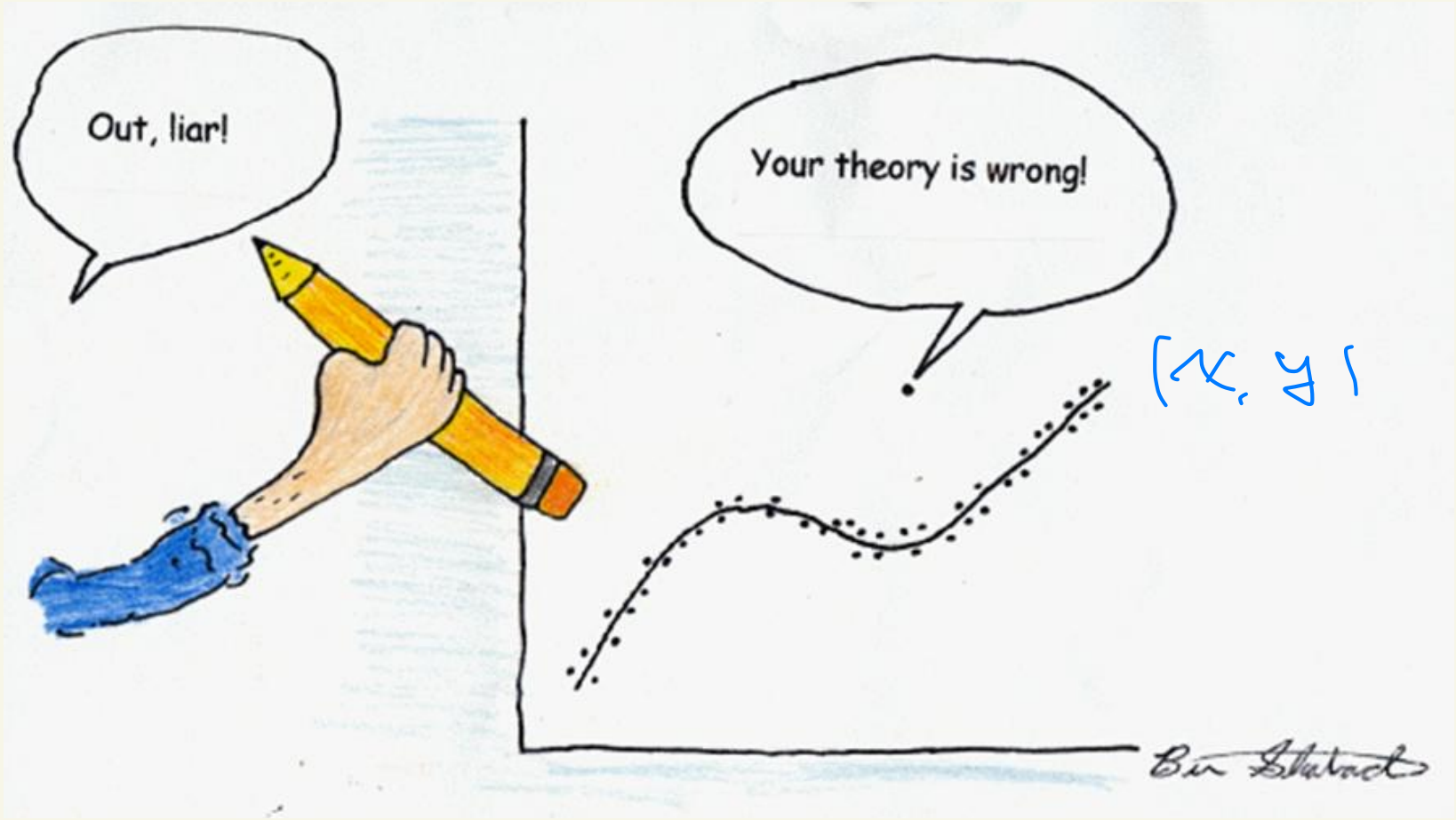
The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R^2 is to 1. The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average value.

The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

In the best case, the modeled values exactly match the observed values, which results in $SS_{\text{res}} = 0$ and $R^2 = 1$. A baseline model, which always predicts $\bar{y}$, will have $R^2 = 0$. Models that have worse predictions than this baseline will have a negative R^2.

Outliers in linear regression

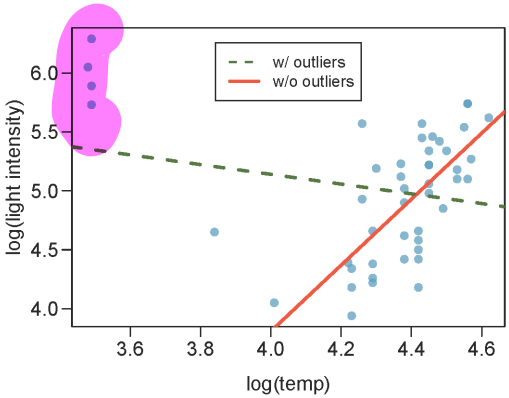

Out, liar!
Your theory is wrong!
(x, y)

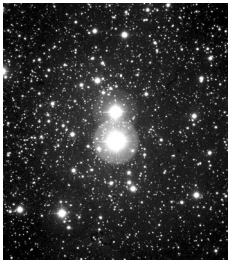
Some terminology

- 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

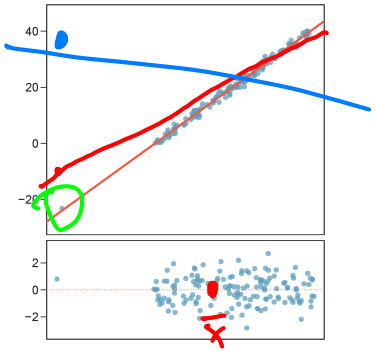
Which of the below best describes the outlier?

(a) influential

(b) high leverage

(c) none of the above

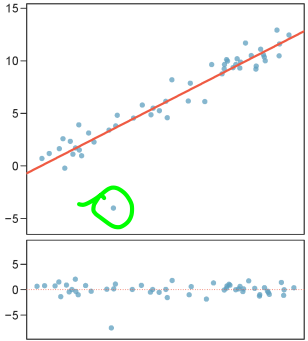
(d) there are no outliers



• High leverage ≒ far away from $\bar{x}$

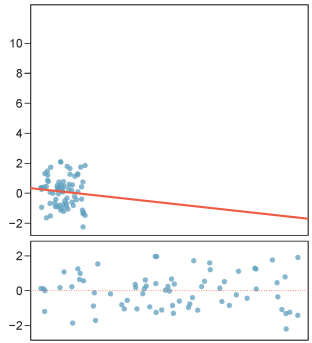
Types of outliers

Does this outlier influence the slope of the regression line? Not much...

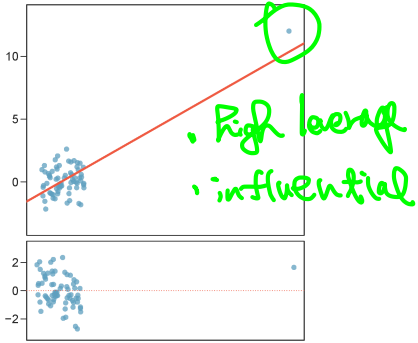


Recap

$R = 0.08, R^2 = 0.0064$



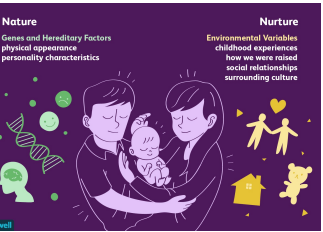
$R = 0.79, R^2 = 0.6241$

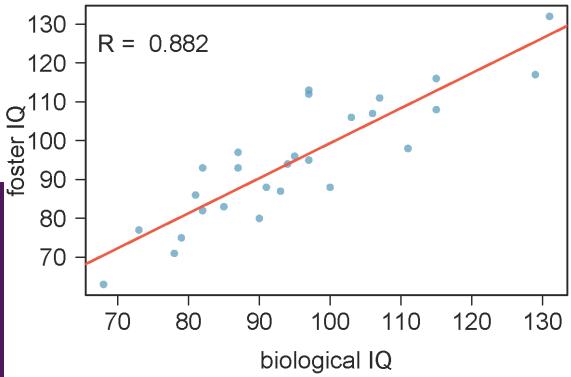


Inference for linear regression

Nature or nurture?

In 1966 Cyril Burt published a paper called "The genetic determination of differences in intelligence: A study of monozygotic twins reared together and apart". The data consist of IQ scores for 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.

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: $\beta_1 \neq 0$

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

$\beta_0, \beta_1, \varepsilon \sim N(0, \sigma^2)$

unknown

$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$

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

p-value < α(=0.01)

⇒ reject

$H_0: \beta_1 = 0$

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

$$df = 27 - 2 = 25$$

$$p - value = P(|\ T\ | > 9.36) < 0.01$$

In practice, $T = 9.36$

By Normal Dist.

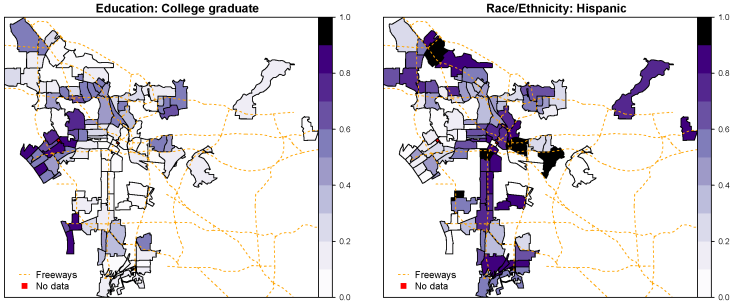
95% CI

$0.9014 \pm 2 \cdot 0.0963$

$t_{\alpha/2}$

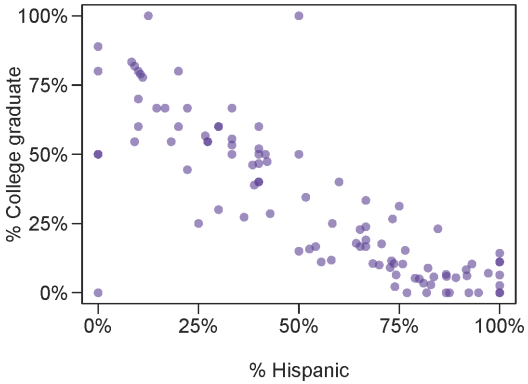
% College graduate vs. % Hispanic in LA

What can you say about the relationship between % college graduate and % Hispanic in a sample of 100 zip code areas in LA?



% College educated vs. % Hispanic in LA - another look

What can you say about the relationship between of % college graduate and % Hispanic in a sample of 100 zip code areas in LA?



% College educated vs. % Hispanic in LA - linear model

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
%Hispanic	-0.7527	0.0501	-15.01	0.0000

(a) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.

(b) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.

(c) An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.

(d) In zip code areas with no Hispanic residents, % of college graduates is expected to be 75%.

% College educated vs. % Hispanic in LA - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % Hispanic and % college graduates in zip code areas in LA?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
hispanic	-0.7527	0.0501	-15.01	0.0000

Yes, the p-value for % Hispanic is low, indicating that the data provide convincing evidence that the slope parameter is different than 0.

Reject $H_0 : \beta_1 = 0$

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) $0.9014 \pm 1.65 \times 0.0963$
(d) $9.2076 \pm 2.06 \times 0.0963$

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

$97.5\%: \; t^{\star}_{25} = 2.06$

$0.9014 \pm 2.06 \times 0.0963 = (0.7, 1.1)$

In practice, use $z_{0.25} = 1.96$ or ≈ 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:
$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$
- 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.
- The intercept is rarely the focus of inference as it often lacks meaningful interpretation outside the data range.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- If you have a sample that is non-random (biased), inference on the results will be unreliable.