Unit 4: Regression and Prediction

2. Residuals and Least-Squares

(Chapter 5.2)

11/16/2016

Recap from last time

1. Correlation is a measure of the linear relationship between two factors.
2. We can use linear regression to estimate this correlation
3. A regression line is the line that minimizes the residuals between each point and the line.

Key ideas

1. We can use the slope and intercept of a regression line to make predictions
2. We can also sometimes extrapolate, but this can be fraught
3. Just like t-tests and the other statistics we've explored so far, linear regression models are appropriate only when some conditions are met

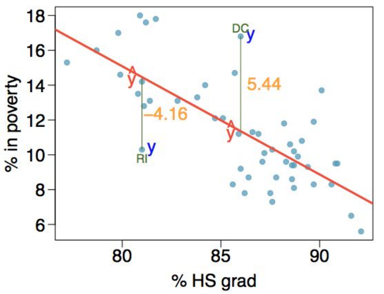
Residuals

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

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

For example, percent living in poverty in **DC** is 5.44% more than predicted.

Percent living in poverty in **RI** is 4.16% less than predicted.



Finding the best line

We want to find the line that has the smallest residuals

Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$|e_1| + |e_2| + \ldots + |e_n|$

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

$e_1^2 + e_2^2 + \ldots + e_1^2$

Why least squares?

- Easier to compute by hand and using software
- Often, a residual twice as large as another is more than twice as bad

The least-squares line

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

Predicted *y* **Intercept** **Slope** **Explanatory variable**

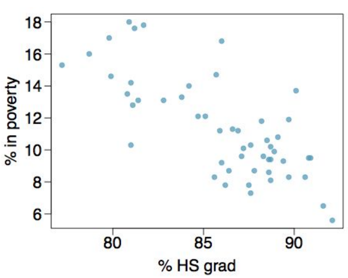
Intercept Notation

- Parameter: β_0
- Point estimate: b_0

Slope Notation

- Parameter: β_1
- Point estimate: b_1

In the context of the HS graduation data



	% 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		$R = -0.75$

Making sense of the model

The slope of the regression can be calculated as

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

In context...

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

	% 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		$R = -0.75$

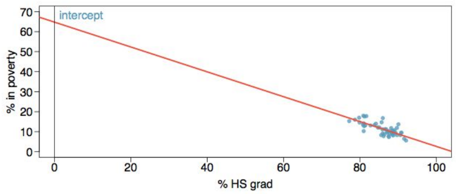
Interpretation

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

Making sense of the model

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

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



	% 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	$R = -0.75$

$b_0 = 11.35 - (-0.62) \times 86.01$
$= 64.68$

Practice Question 1

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

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

Practice Question 1

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

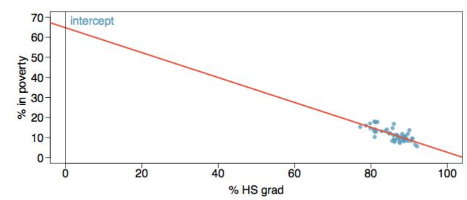
(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

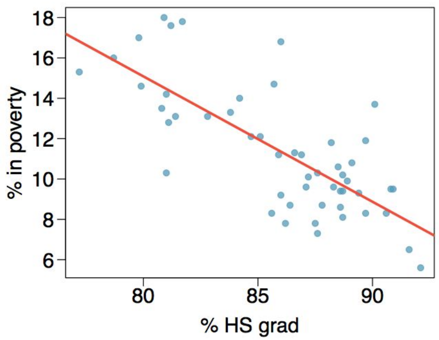
Do you believe this inference?



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.

Quick recap: The meaning of a regression line

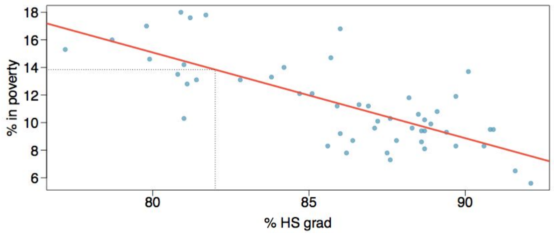
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Prediction from Regression Lines

Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called **prediction**.

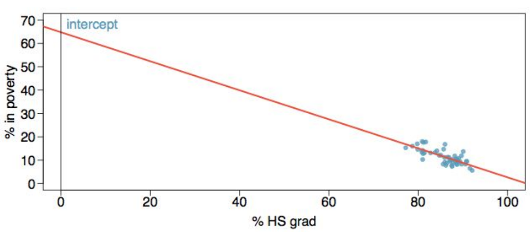
There will be some uncertainty associated with the predicted value.



Extrapolation from regression lines

Applying a model estimate to values outside of the realm of the original data is called **extrapolation**.

Sometimes the intercept might be an extrapolation.



An example 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.

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

Women are set to become the dominant sprinters

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

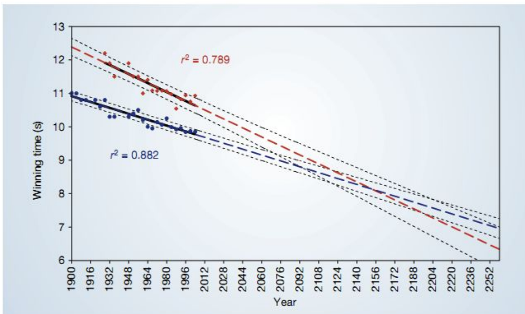
The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

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."

An example of extrapolation

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



Conditions for least-squares regression

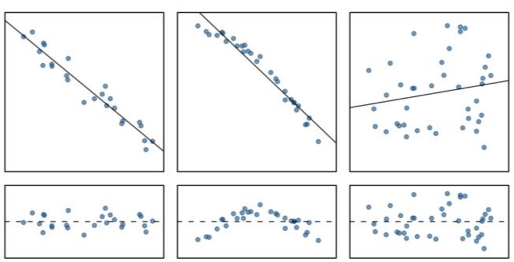
1. **Linearity**: The relationship between two variables must be linear
2. **Nearly normal residuals**: The errors between the line and the data are assumed to be drawn from a nearly-normal distribution
3. **Constant variability**: Assume that data are approximately equally variable at all ranges of x and y

Extrapolation: What could go wrong?

Condition 1: Linearity

(Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class)

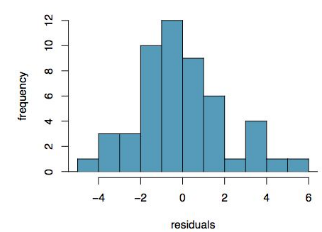
Check using a scatterplot of the data, or a **residuals plot**.

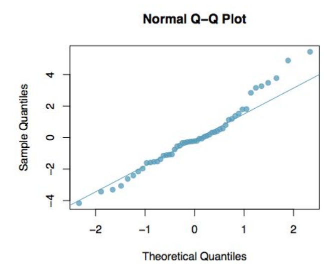


Condition 2: Nearly normal residuals

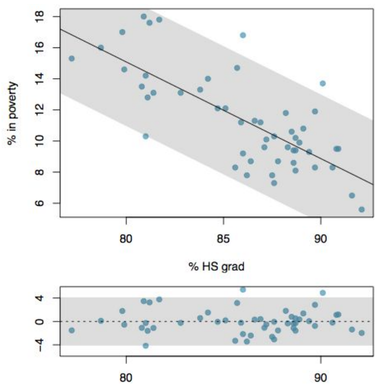
This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

Check using a histogram or normal probability plot of residuals.





Condition 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.

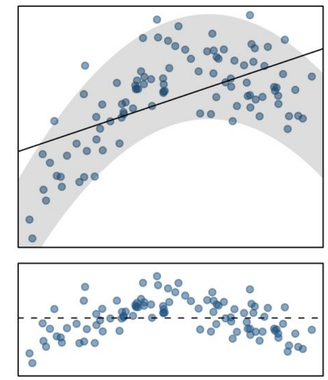
Also called **homoscedasticity.**

Check using a histogram or normal probability plot of residuals.

Practice Question 2: Checking conditions

Which condition is this model violating?

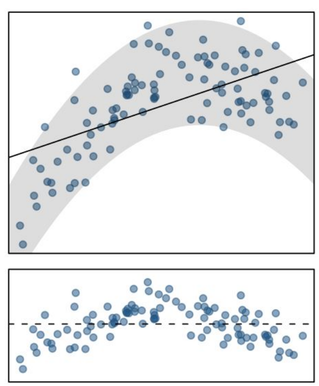
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Practice Question 2: Checking conditions

Which condition is this model violating?

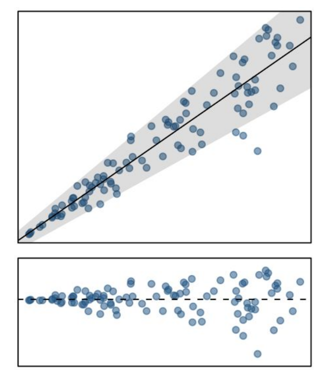
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Practice Question 3: Checking conditions

Which condition is this model violating?

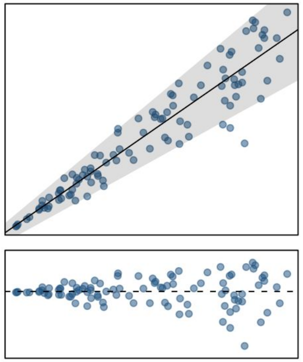
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Practice Question 3: Checking conditions

Which condition is this model violating?

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



How good is your model?

The strength of the fit of a linear model is most commonly evaluated using $\mathbf{R^2}$.

$\mathbf{R^2}$ 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.

Key ideas

1. We can use the slope and intercept of a regression line to make predictions
2. We can also sometimes extrapolate, but this can be fraught
3. Just like t-tests and the other statistics we've explored so far, linear regression models are appropriate only when some conditions are met