Section 8.2
Fitting a line by least squares regression

Stats 7 Summer Session II 2022

A measure for the best line

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

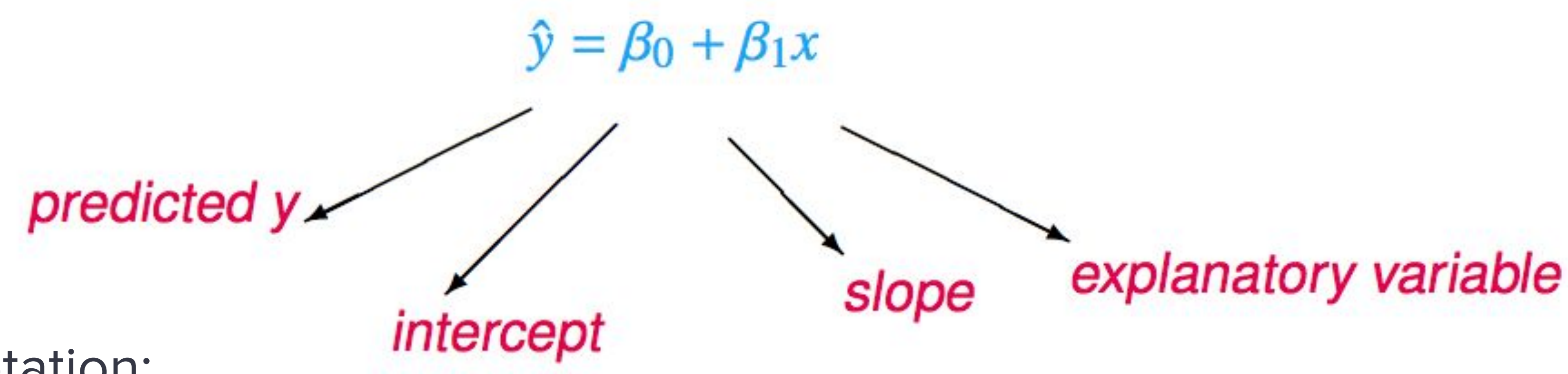
     $$|e_1| + |e_2| + \ldots + |e_n|$$

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

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

- 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



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

predicted y

intercept

slope

explanatory variable

Notation:

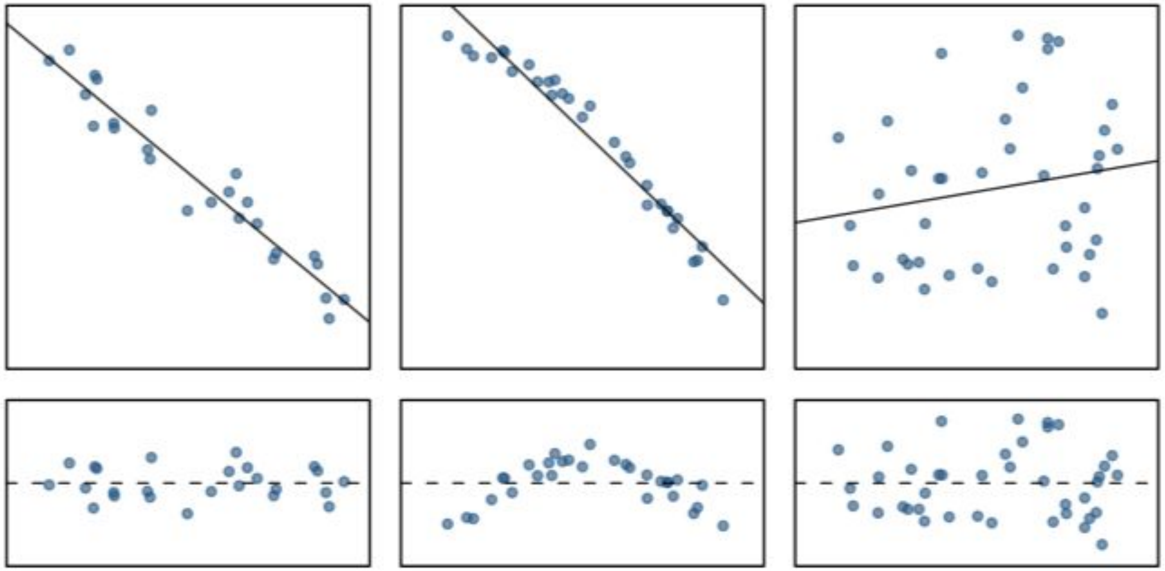
- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

Conditions for the least squares line

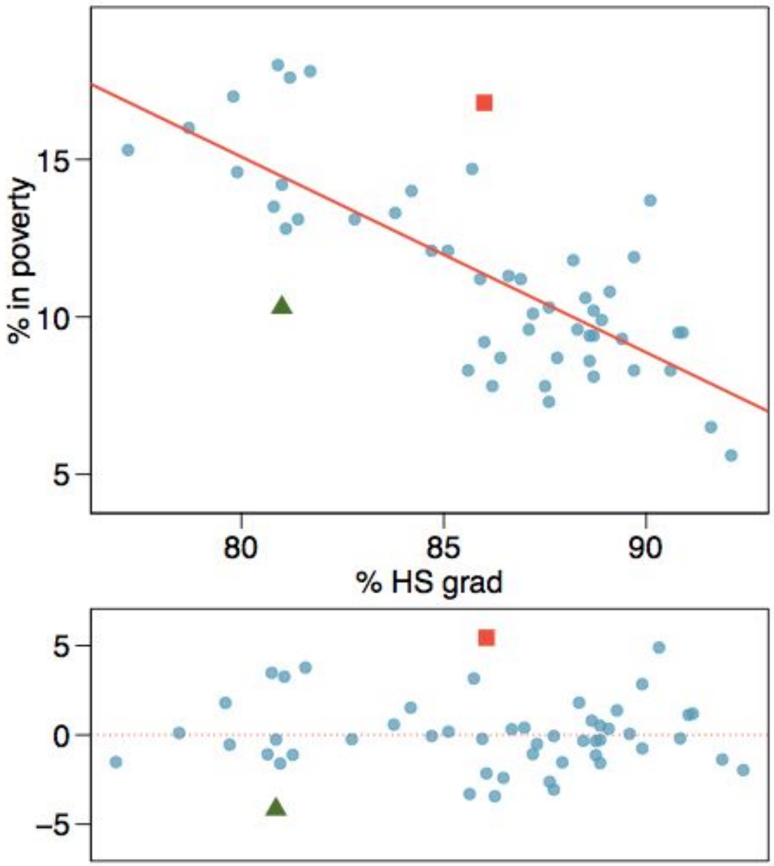
1. Linearity
2. Nearly normal residuals
3. Constant variability
4. Independent observations

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 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



▲ *RI:*

$$\% \textit{ HS grad} = 81 \qquad \% \textit{ in poverty} = 10.3$$

$$\% \widehat{\textit{in poverty}} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \% \textit{ in poverty} - \% \widehat{\textit{in poverty}}$$

$$= 10.3 - 14.46 = -4.16$$

■ *DC:*

$$\% \textit{ HS grad} = 86 \qquad \% \textit{ in poverty} = 16.8$$

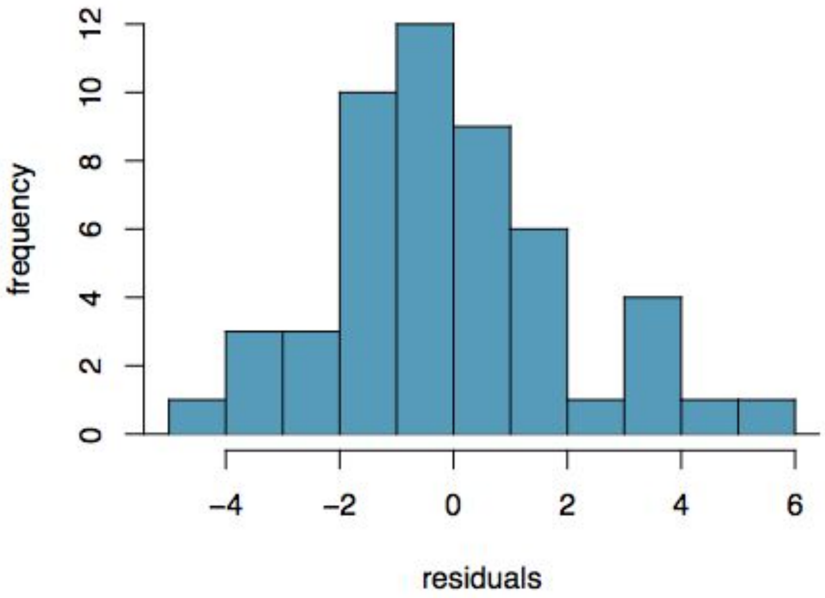
$$\% \widehat{\textit{in poverty}} = 64.68 - 0.62 * 86 = 11.36$$

$$e = \% \textit{ in poverty} - \% \widehat{\textit{in poverty}}$$

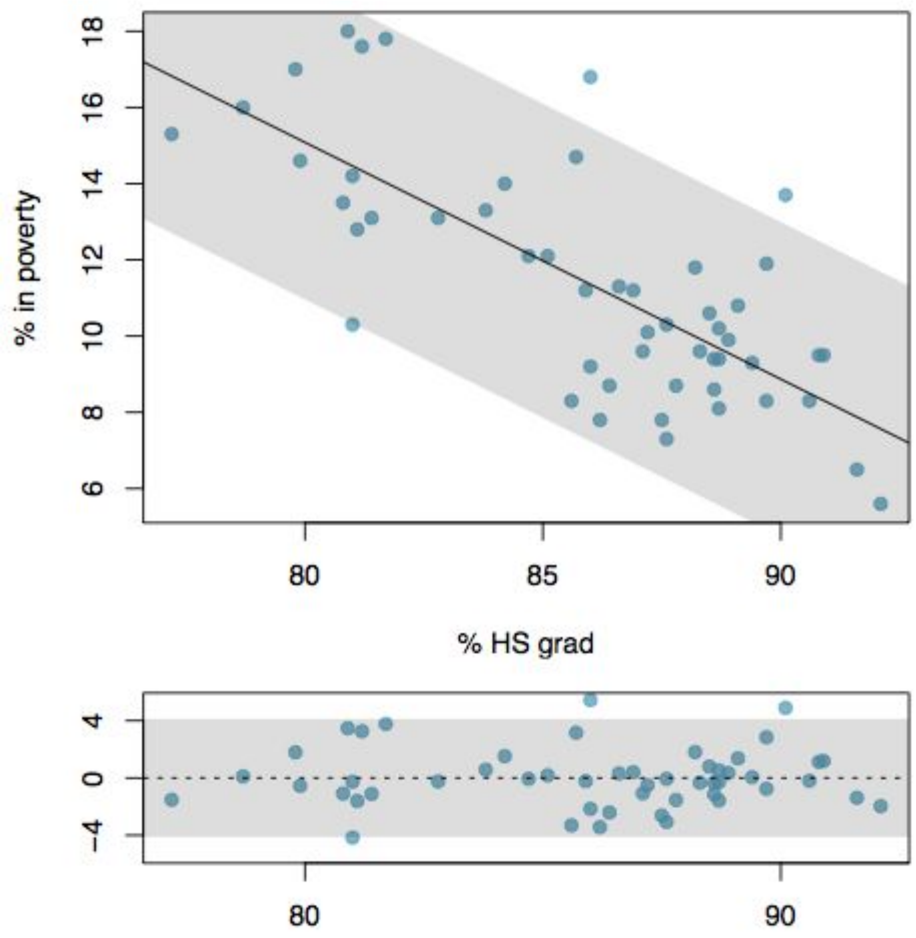
$$= 16.8 - 11.36 = 5.44$$

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 or normal probability plot of residuals.



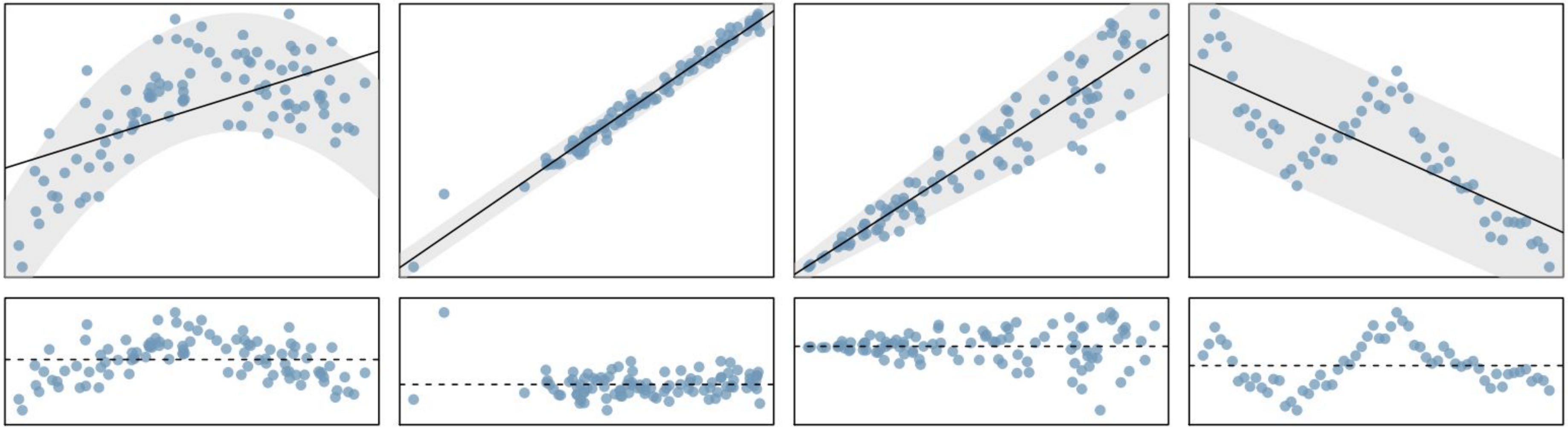
Conditions: (3) Constant variability



- 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 histogram or normal probability plot of residuals.

Failure to meet conditions

Name the reason why each of these data scenarios do not meet the conditions necessary for linear regression



Linearity fails

Normality fails; extreme outlier

Constant variance fails

Independence fails; data are correlated

Checking conditions

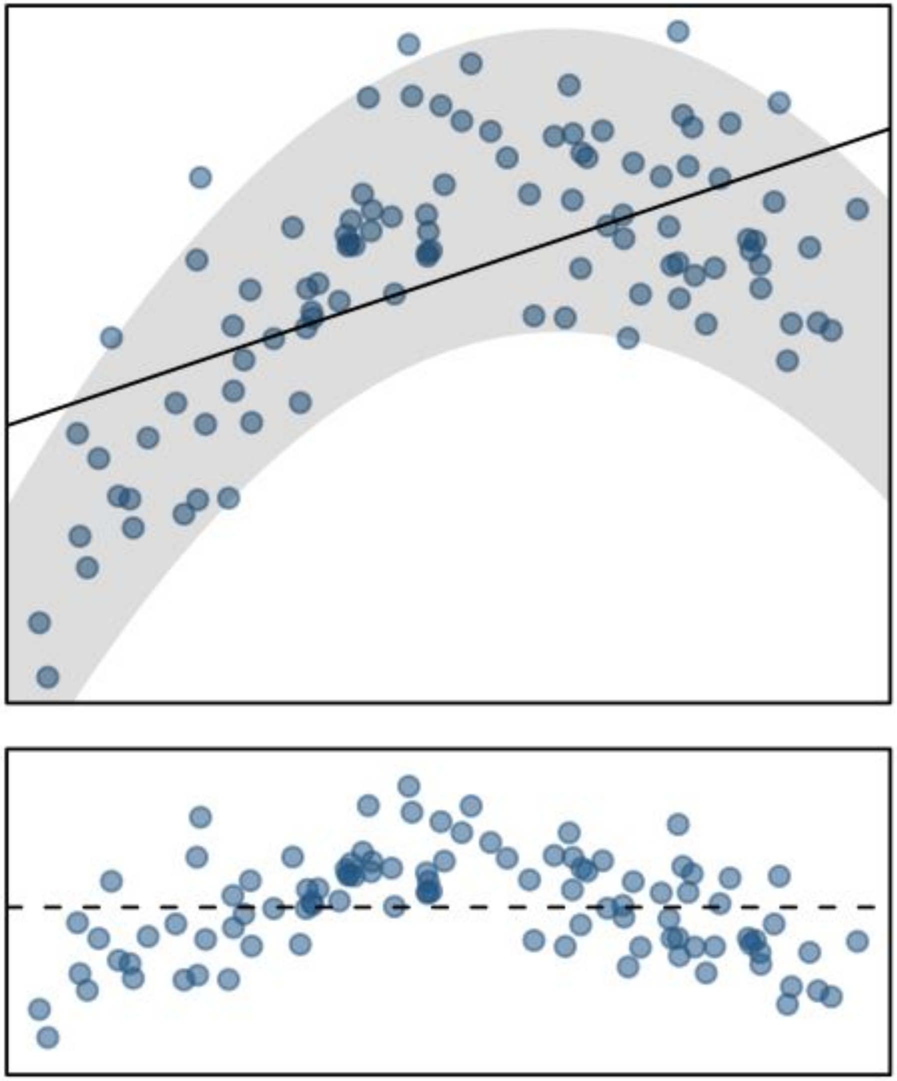
What condition is this linear model obviously violating?

(a) Constant variability

[✓ (b) Linear relationship]

(c) Normal residuals

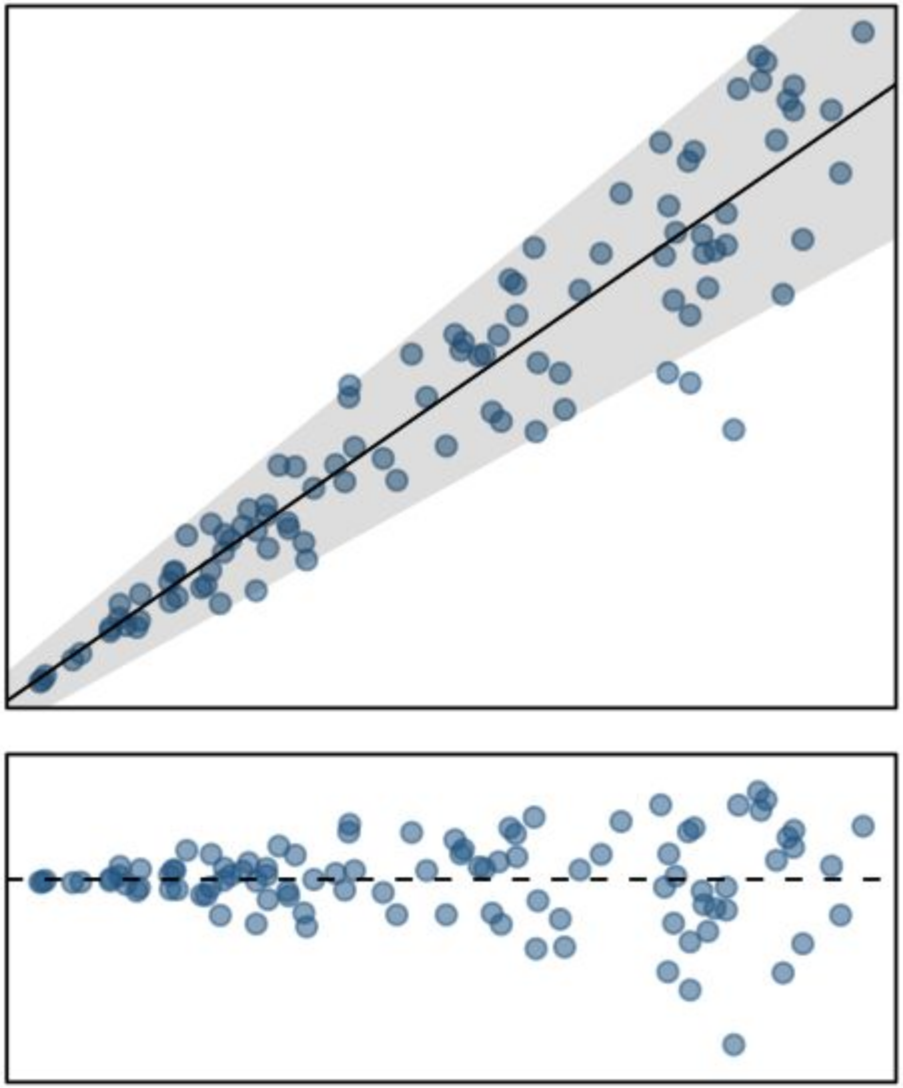
(d) No extreme outliers



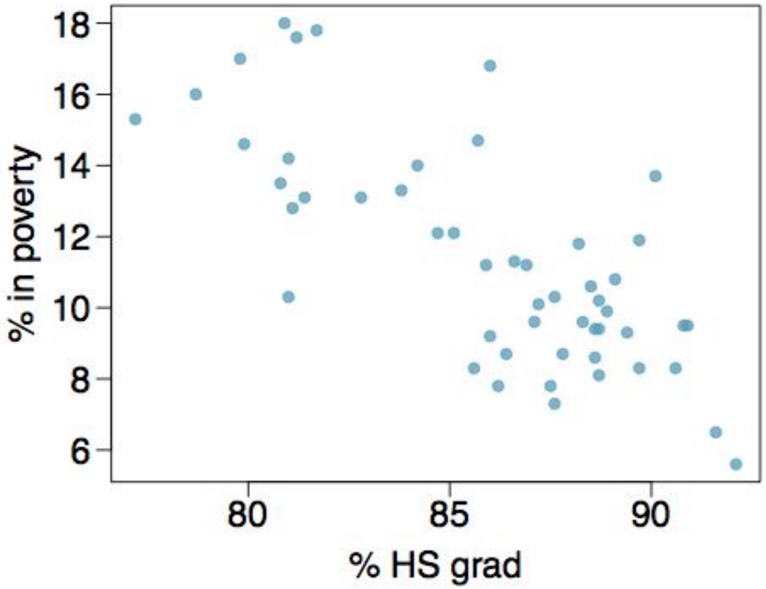
Checking conditions

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

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

In context…

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

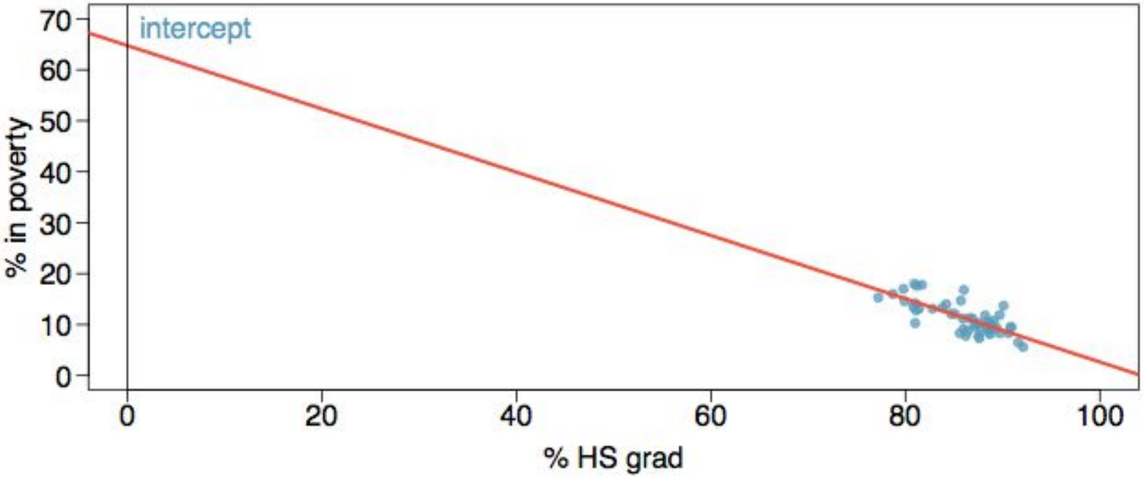
Interpretation

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

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



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

$$= 64.68$$

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%.

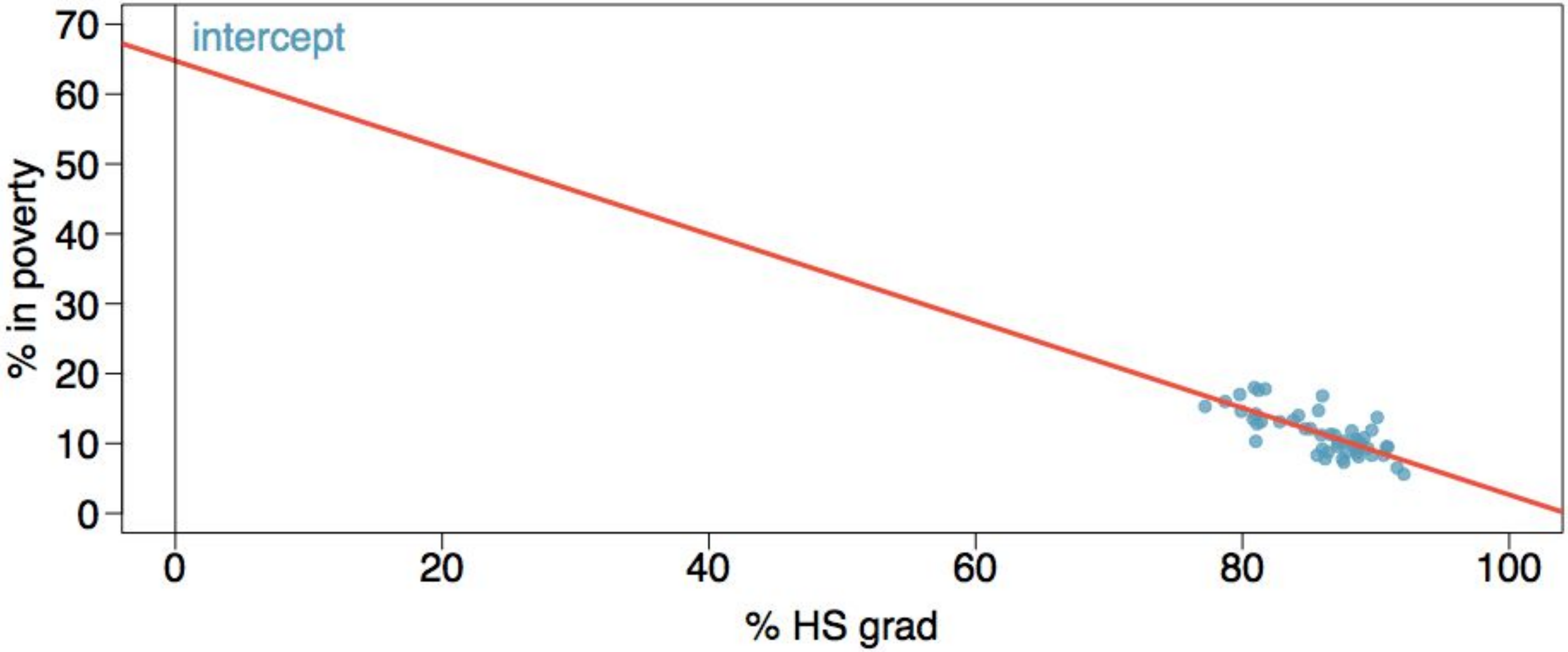
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

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

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

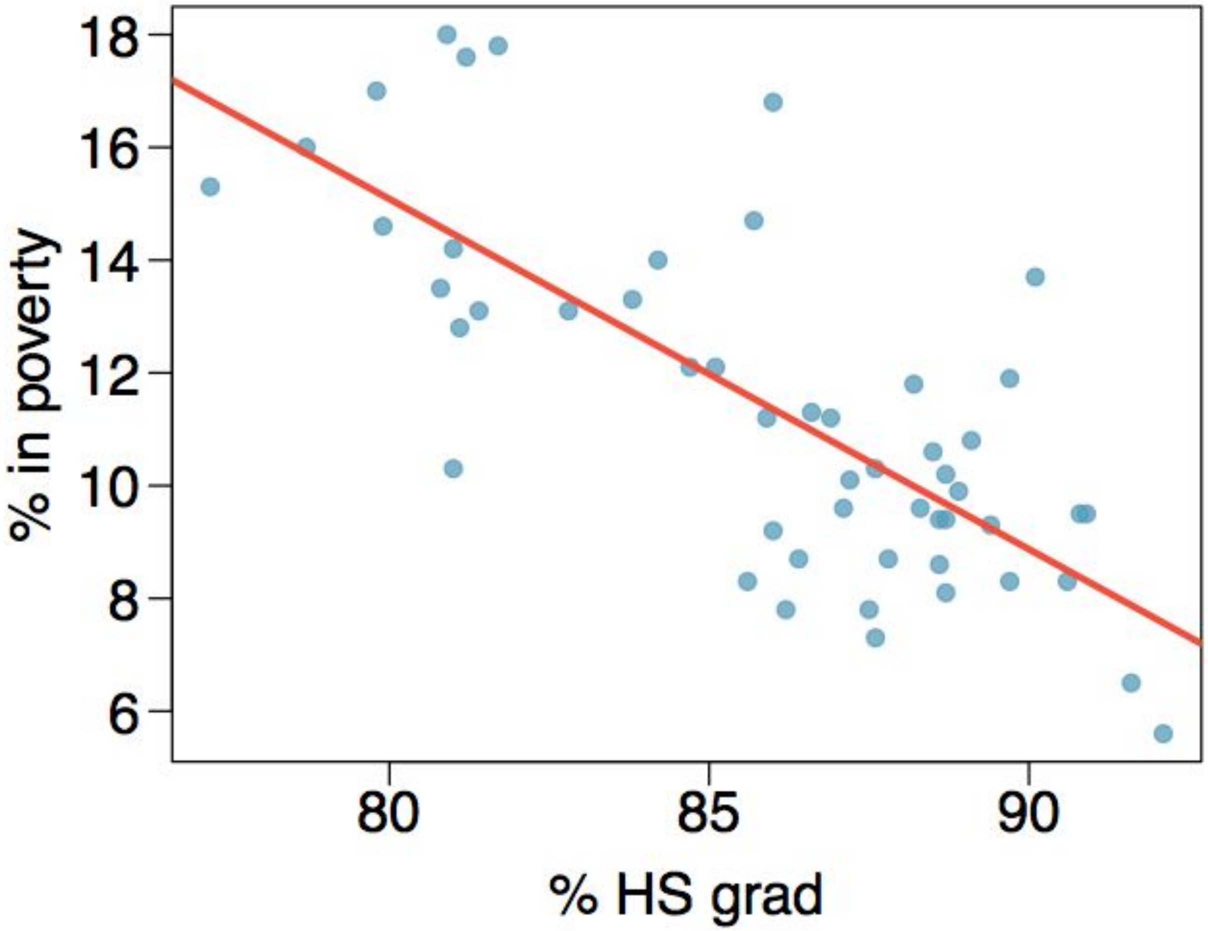
More on the intercept

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

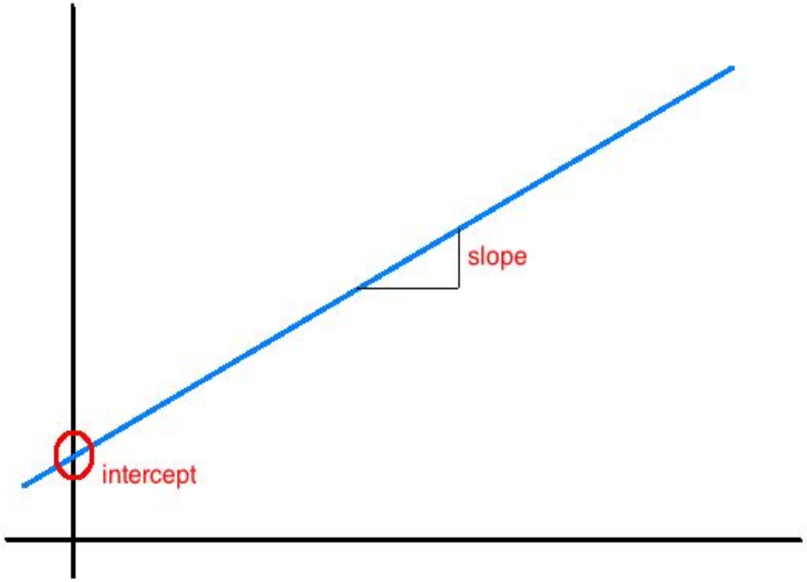
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Interpretation of slope and intercept

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

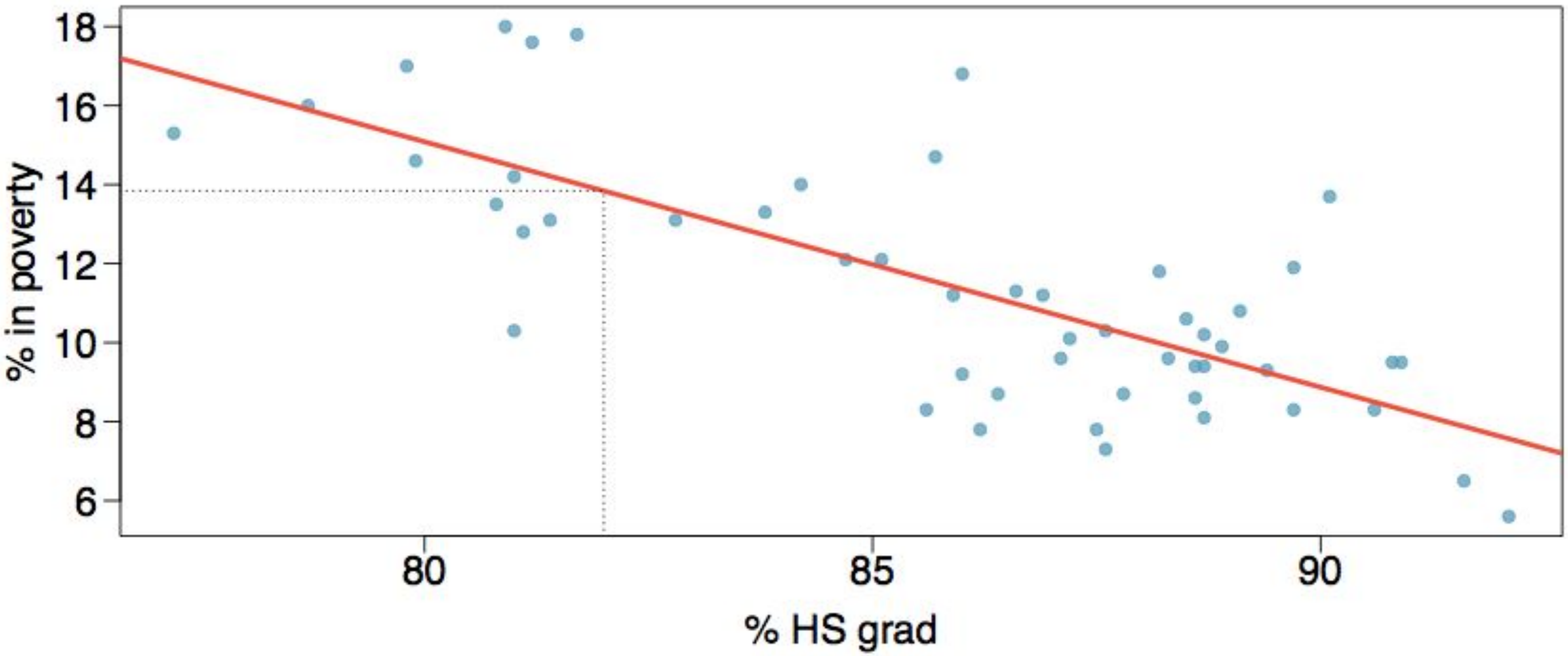
- *Slope*: For each unit in x, y is expected to increase / decrease on average by the slope.



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

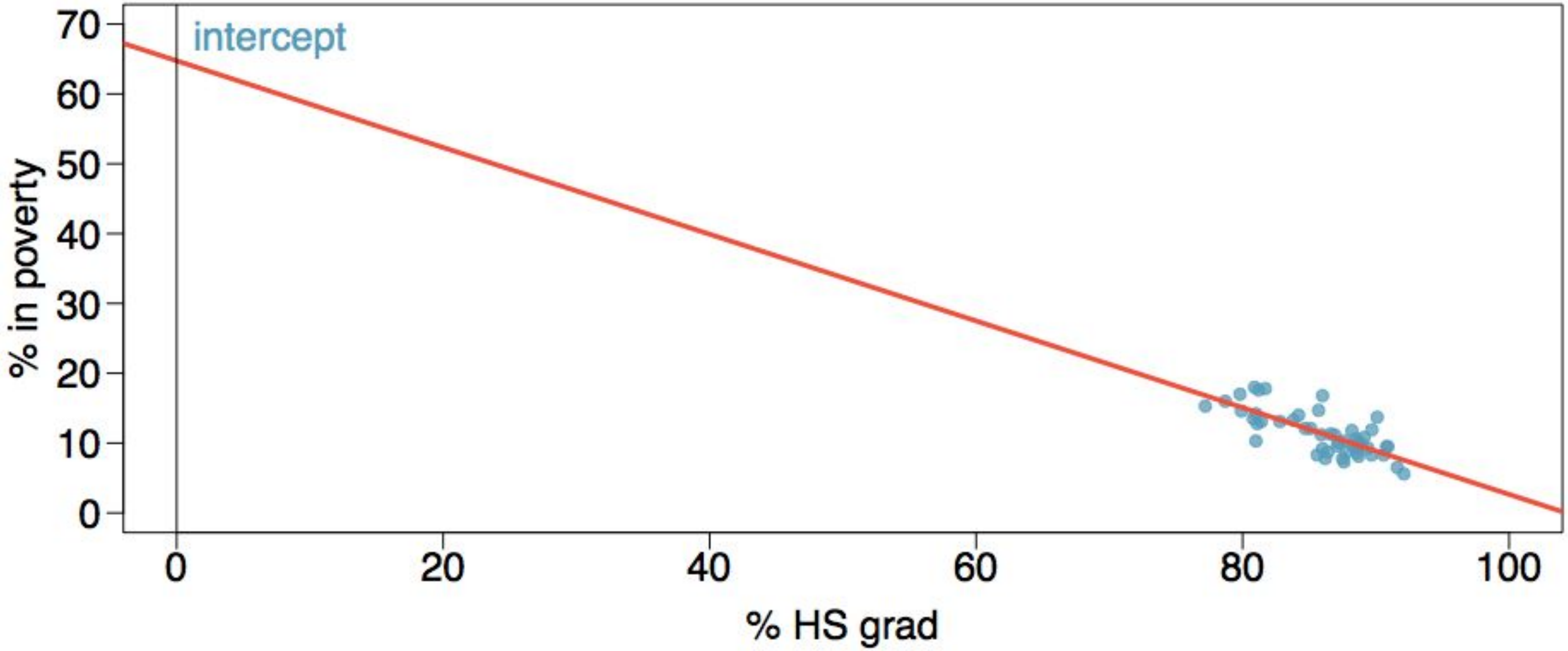
Prediction

- 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.

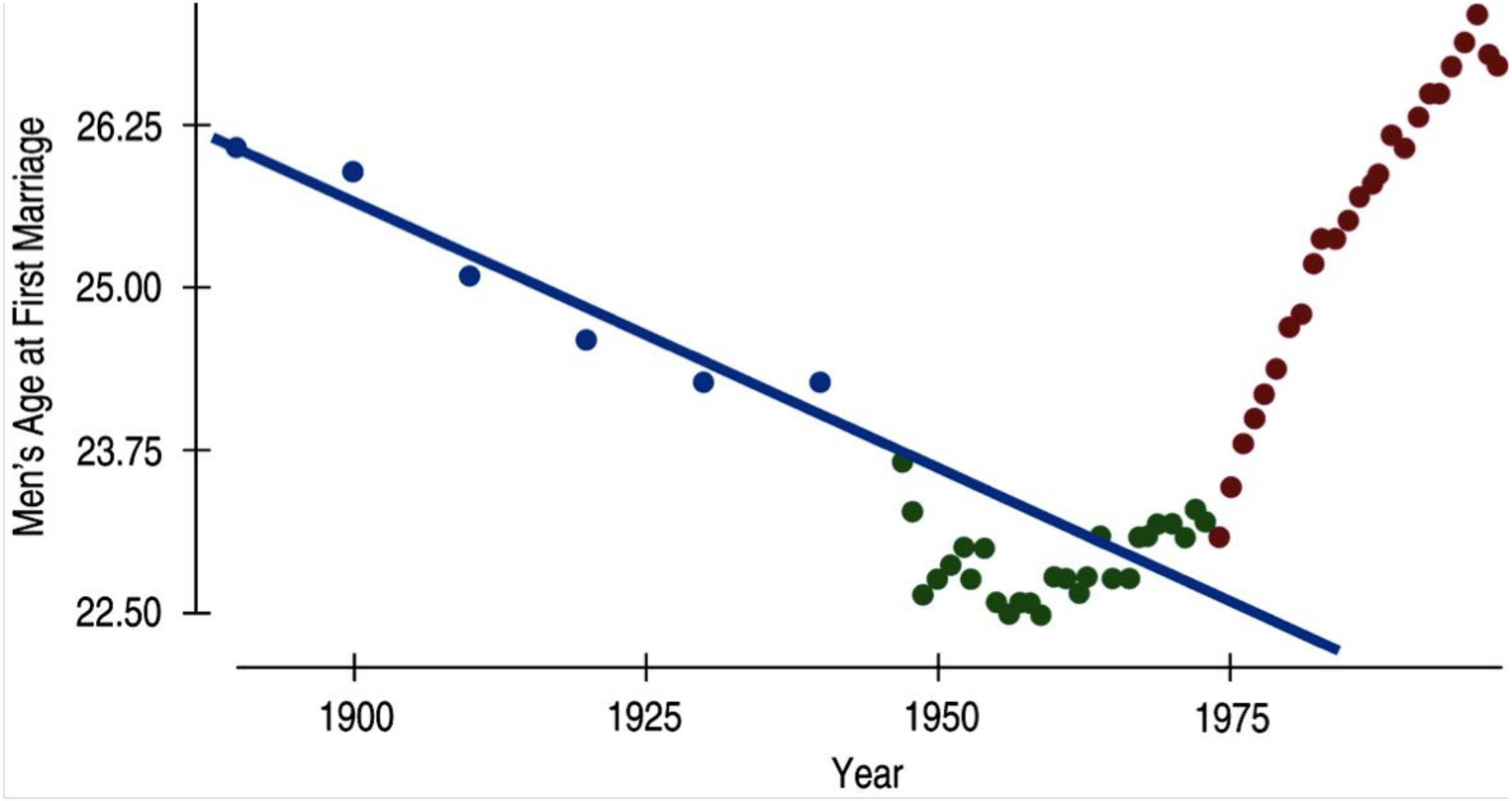


Extrapolation

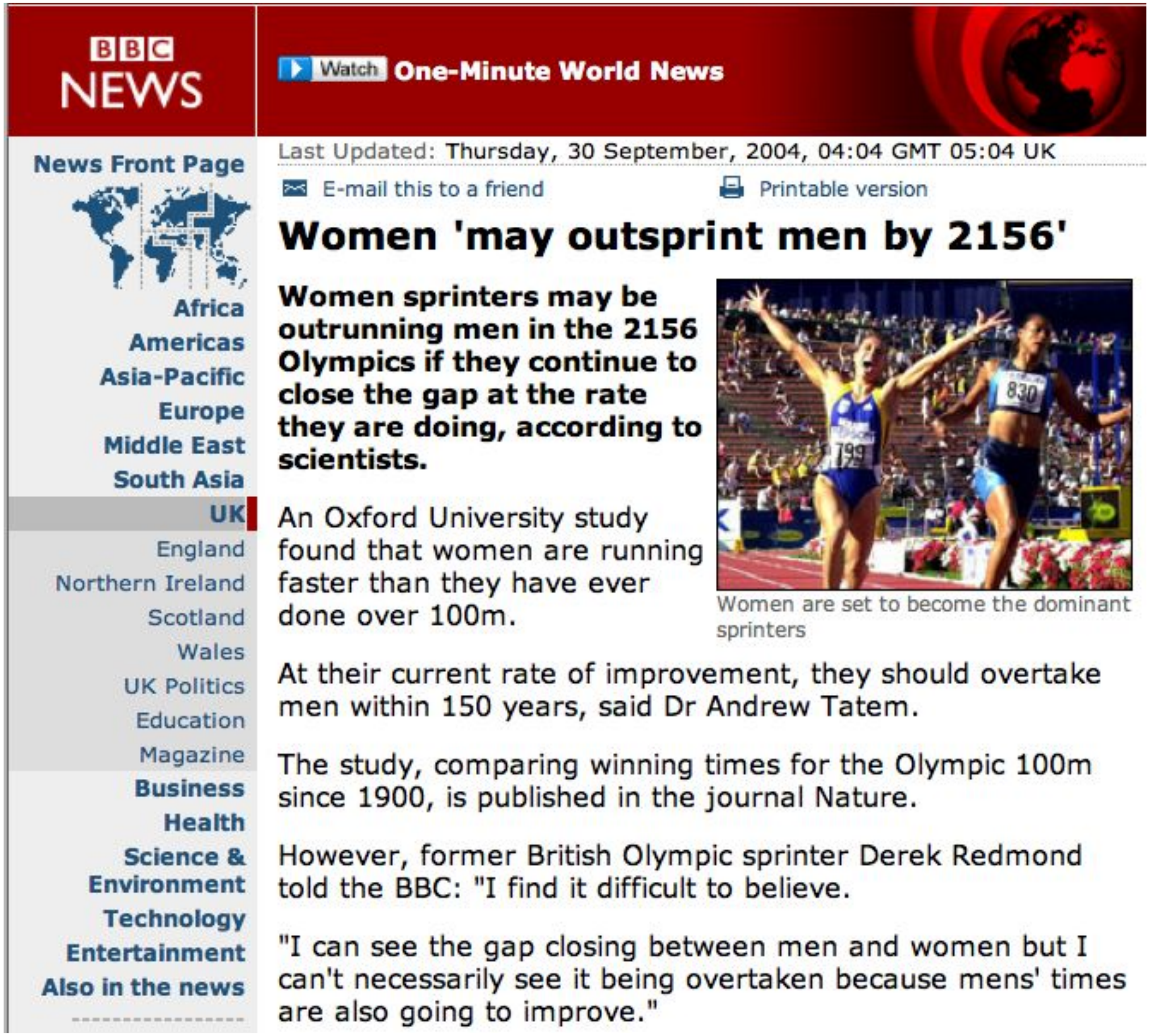
- Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.
- Sometimes the intercept might be an extrapolation.



Examples of extrapolation



Examples of extrapolation



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

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

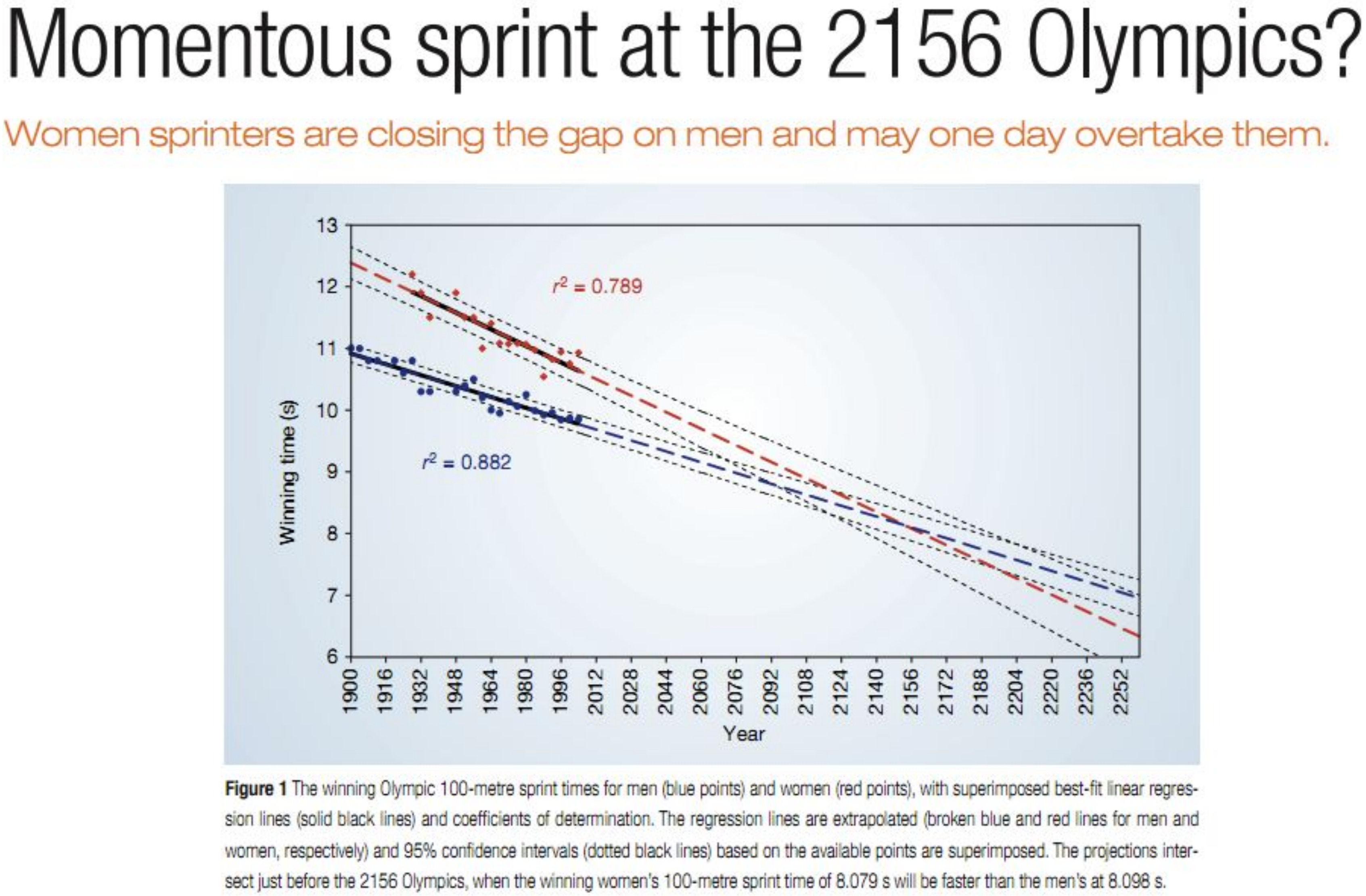
At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

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."

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 8.098 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.
- 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, $R^2 = -0.62^2 = 0.38$.

Interpretation of R^2

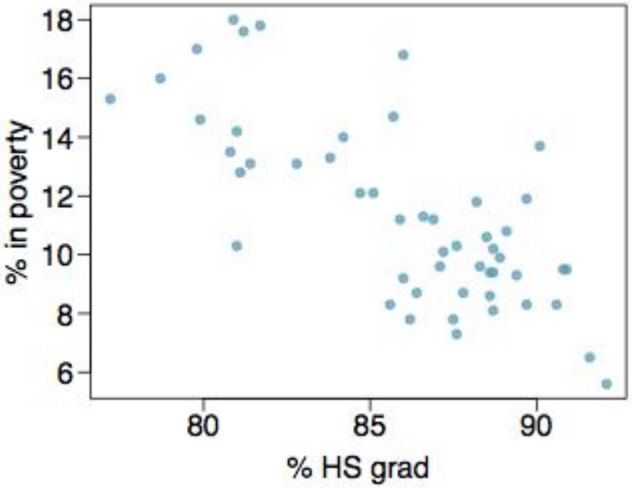
Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

(a) 38% of the variability in the % of HG graduates among the 51 states is explained by the model.

[✓ (b) 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.]

(c) 38% of the time % HS graduates predict % living in poverty correctly.

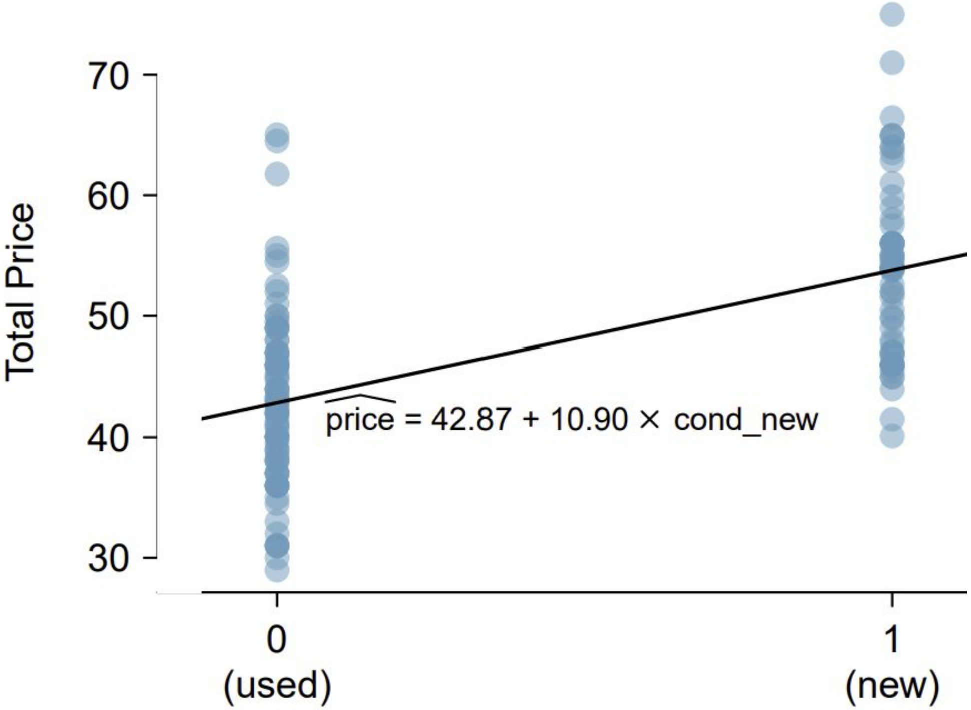
(d) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Categorical predictors with two levels

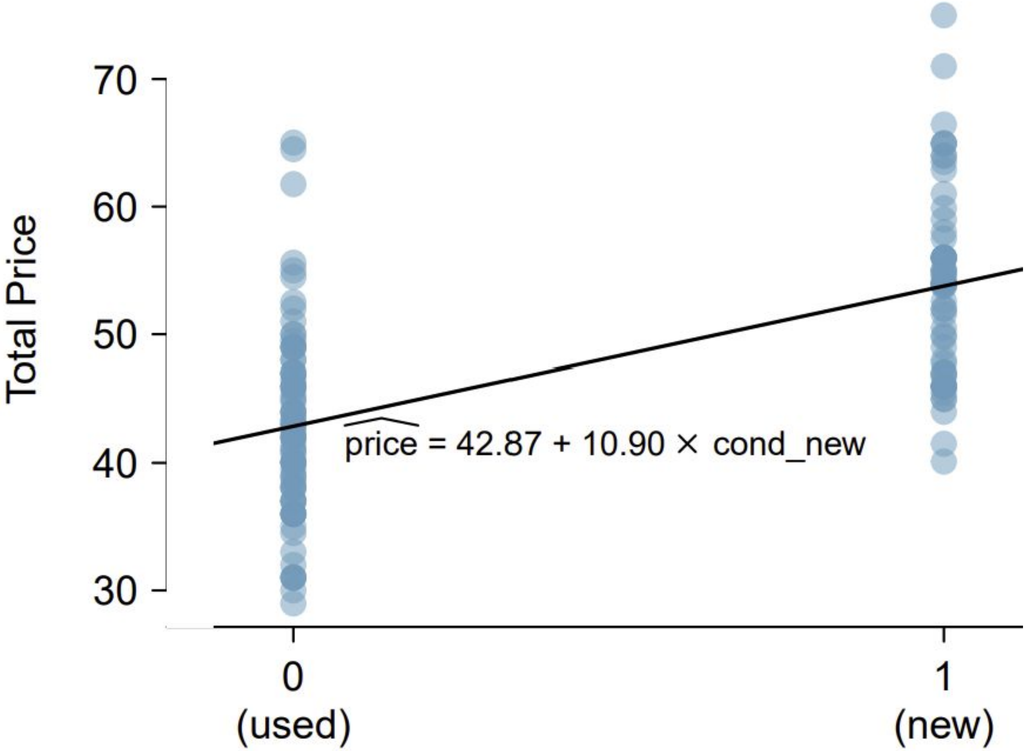
Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a level is the same as a category).

We'll consider Ebay auctions for a video game, Mario Kart for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded. Here we want to predict total price based on game condition, which takes values used and new.



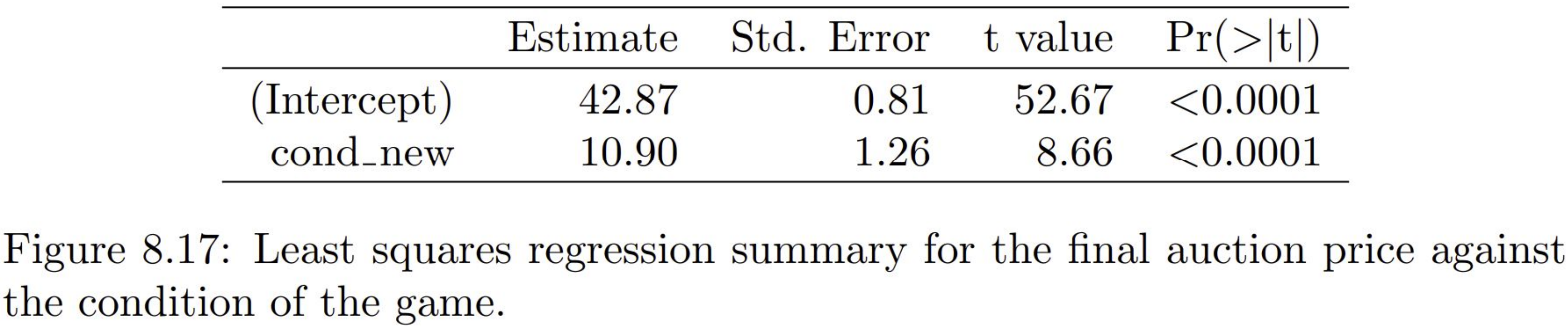
Categorical predictors with two levels

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an indicator variable called cond new, which takes value 1 when the game is new and 0 when the game is used.



Categorical predictors with two levels

Below is a table summarizing the linear regression fit results.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	<0.0001
cond_new	10.90	1.26	8.66	<0.0001

Figure 8.17: Least squares regression summary for the final auction price against the condition of the game.

From this we can identify the estimates for our intercept (42.87) and predictor (10.90).

We will discuss the other values in the table later.

Categorical predictors with two levels

The Mario Kart price model is:

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

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

The intercept is the estimated price when cond new takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.