Stat 23400 Lecture 16: Analysis of Paired Data, Introduction to Simple Linear Regression

Outline for Today

- Topics:
  - Analysis of Paired Data (MMSA 10.3 and OIS 7.2)
  - Introduction to Simple Linear Regression (MMSA 12.1, 12.2 and OS4 8.1, 8.2)
- Lab09: Simple Linear Regression Analysis

Analysis of Paired Data

Problems with Paired Data

- In some experiments, multiple observations might be on the same sample unit. For instance,
  - measurements on both eyes: right eye, left eye
  - measurements at different times: morning, evening
  - measurements before and after treatment: pre medicine, post medicine
- In all these cases, there is only one set of sample units $\{X_i\}_{i=1}^n$, but for each of them we have two observations: $X_{i,A}$ and $X_{i,B}$. So all together we have $2n$ observations

$$\{X_{i,k}\}_{i=1,k\in\mathcal{K}}^n \qquad k \text{ is the type of measurement } A \text{ or } B\,.$$

- With this kind of data, we are still interested in testing if there is a difference between the two means, $\mu_A - \mu_B$.

Method for Analyzing Paired Data

- But now the two populations are **NOT independent**. We can't use the two sample t-procedure.
- However, if we take the differences between the two different observations, then we are back to having just ONE observation for each sample unit

$$D_i = X_{i,A} - X_{i,B}$$

- If the n sample units are independent, then $\{D_i\}_{i=1}^n$ are one-sample independent observations.

Example: Coffee & Blood Flow During Exercise

Doctors studying healthy men measured myocardial blood flow (MBF)[1] during bicycle exercise after giving the subjects a placebo or a dose of 200 mg of caffeine that was equivalent to drinking two cups of coffee[2].

There were 8 subjects, each was tested twice, 4 of them were randomly selected to receive caffeine in the first test and placebo in the second test; the other 4 received placebo first and caffeine second.

There was a 24-hour gap between the two tests (washout period).

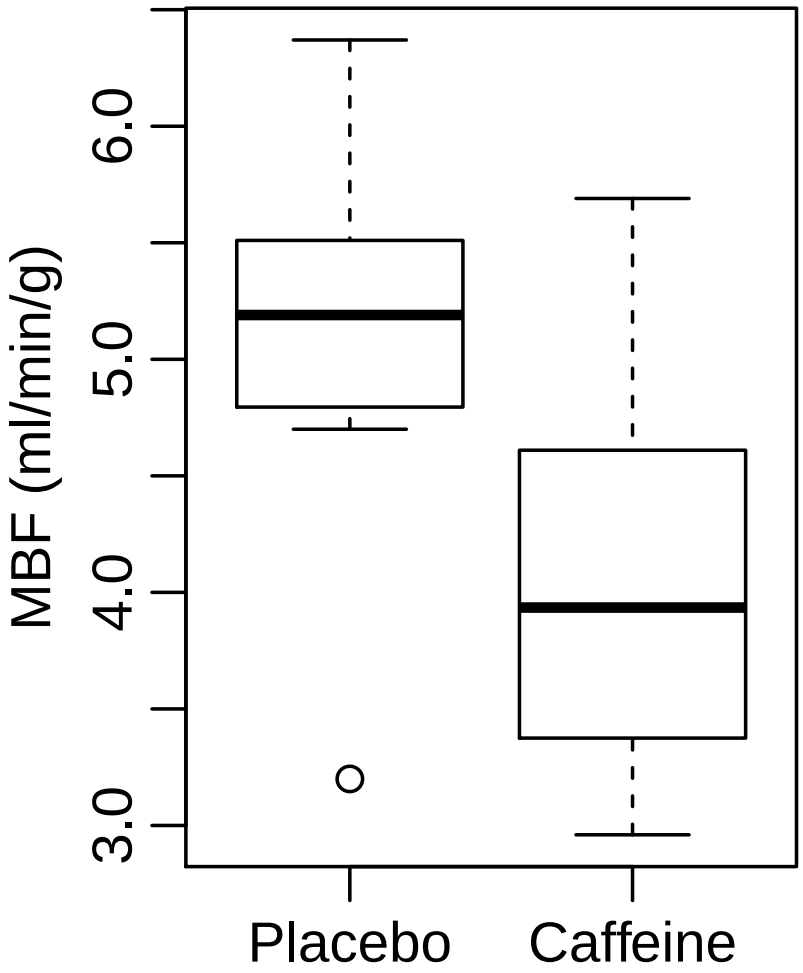
[1] MBF was measured by taking positron emission tomography (PET) images after oxygen-15 labeled water was infused in the patients.

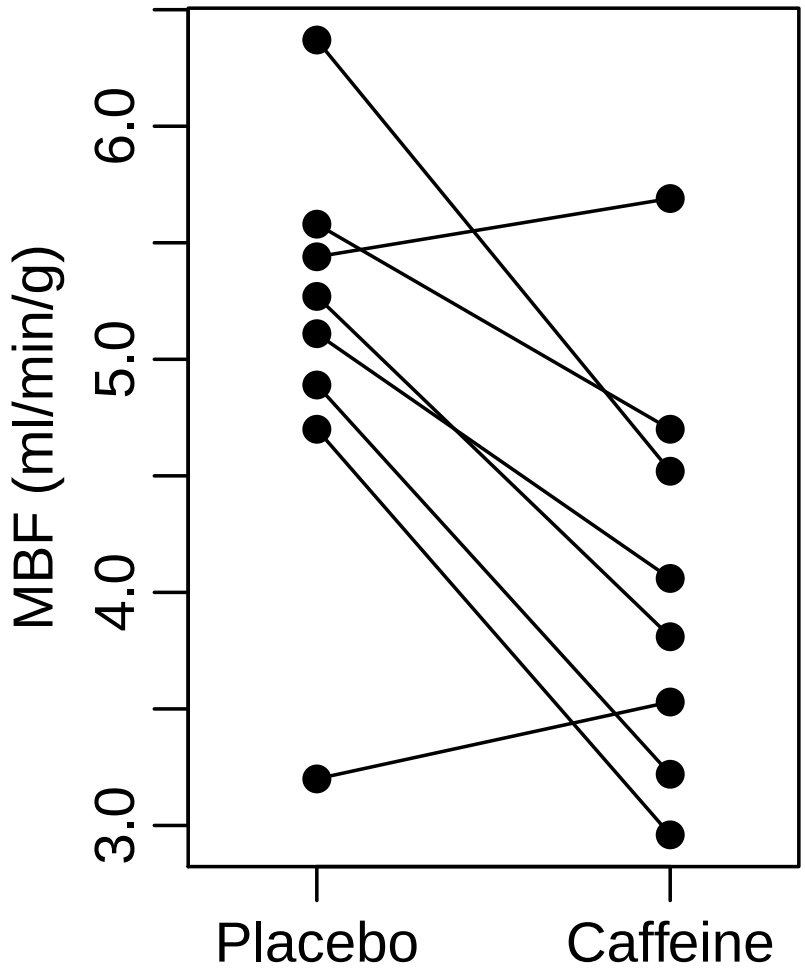
[2] Namdar et. al (2006). Caffeine decreases exercise-induced myocardial flow reserve. *Journal of the American College of Cardiology* **47**, 405-410.

Data for the Coffee & Blood Flow Experiment

	MBF (ml/min/g)	
Subject	Placebo	Caffeine
1	6.37	4.52
2	5.44	5.69
3	5.58	4.70
4	5.27	3.81
5	5.11	4.06
6	4.89	3.22
7	4.70	2.96
8	3.20	3.53
Mean	5.07	4.06
SD	0.91	0.89

Data for the Coffee & Blood Flow Experiment





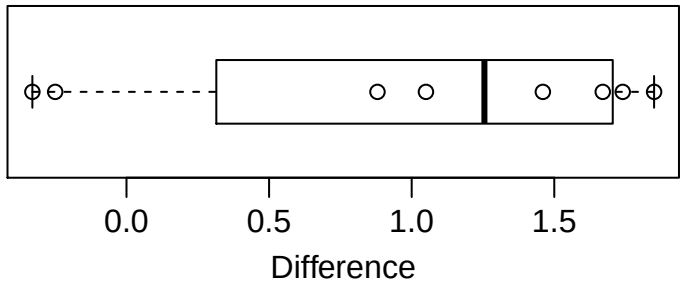
Discussion

- Can we analyze the data of the experiment like two independent samples?
- What is A and what is B?
- How many sample units we have?
- Why did 4 subjects caffeine first and placebo second and for the other 4 the order was reversed ?
- Why do we need a washout period (the 24 hour gap) between the two tests?

Hypothesis Tests for Paired Data

- Since measurements on different subjects can be reasonably assumed independent, we can take differences of the two measurements within each subject and analyze the differences like **one-sample data**.

	MBF (ml/min/g)		
Subject	Placebo	Caffeine	Diff
1	6.37	4.52	1.85
2	5.44	5.69	−0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.20	3.53	−0.33
Mean	5.07	4.06	1.01
SD	0.91	0.89	0.87



Hypothesis Tests for Paired Data

To test H_0: $\mu_A = \mu_B$ is equivalent to test $\mu_D = \mu_A - \mu_B = 0$.
What is the test statistic for testing H_0: $\mu_D = 0$?

Example: Coffee & Blood Flow During Exercise

	MBF (ml/min/g)		
Subject	Placebo	Caffeine	Difference
1	6.37	4.52	1.85
2	5.44	5.69	−0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.20	3.53	−0.33
Mean	5.07	4.06	1.01
SD	0.91	0.89	0.87

one tail two tails		0.1 0.2	0.05 0.10	0.025 0.050	0.01 0.02	0.005 0.010
df	7	1.41	1.89	2.36	3.00	3.50

Confidence Intervals for the Mean Difference in Paired Data

Find the 95% confidence interval for the mean difference.

Tests/CIs for Paired Data in R

```
> caffeine = c(4.52, 5.69, 4.70, 3.81, 4.06, 3.22, 2.96, 3.53)
> placebo = c(6.37, 5.44, 5.58, 5.27, 5.11, 4.89, 4.70, 3.20)
> t.test(placebo,caffeine, paired=T, conf.level=0.95)

Paired t-test

data:  placebo and caffeine
t = 3.2857, df = 7, p-value = 0.01338
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2827867 1.7347133
sample estimates:
mean of the differences
                1.00875
```

If paired data were analyzed like 2-sample data

	MBF (ml/min/g)		
subject	placebo	caffeine	diff.
1	6.37	4.52	1.85
2	5.44	5.69	−0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.20	3.53	−0.33
Mean	5.07	4.06	1.01
SD	0.91	0.89	0.87

If we ignore pairing, and analyze the caffeine data as two-sample data, the two-sample t-statistic

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} = \frac{5.07 - 4.06}{\sqrt{\frac{0.91^2}{8} + \frac{0.89^2}{8}}} \approx 2.244$$

would be less than the paired t-statistic

$$\frac{\bar{d}}{s_d/\sqrt{n}} = \frac{1.01}{0.87/\sqrt{8}} \approx 3.28,$$

- The p-value (6%) given by a two-sample t test is larger than the one given by a paired t-test (1.3%), less significant.
- 95% two-sample CI: $5.07 - 4.06 \pm 2.36\sqrt{\frac{0.91^2}{8} + \frac{0.89^2}{8}} \approx 1.01 \pm 1.06$
  95% paired CI: $1.01 \pm 2.36 \times 0.87/\sqrt{8} \approx 1.01 \pm 0.73$ (shorter)

A note on S_D^2

- Recall the two-sample t-statistic and paired t-statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}, \quad \text{and} \quad t_d = \frac{\bar{d}}{s_d/\sqrt{n}}.$$

- Note that $t < t_d$ if $s_1^2 + s_2^2 > s_d^2$ and $t > t_d$ if $s_1^2 + s_2^2 < s_d^2$.
- Recall that $E(S_D^2) = \sigma_d^2$, $E(S_1^2) = \sigma_1^2$ and $E(S_2^2) = \sigma_2^2$.

$$\begin{aligned} E(S_D^2 - S_1^2 - S_2^2) &= \sigma_d^2 - \sigma_1^2 - \sigma_2^2 = Var(X_A - X_B) - \sigma_1^2 - \sigma_2^2 \\ &= Var(X_A) + Var(X_B) - 2\,Cov(X_A, X_B) - \sigma_1^2 - \sigma_2^2 \\ &= \begin{cases} < 0 & \text{if} \quad Cov(X_A, X_B) > 0 \\ > 0 & \text{if} \quad Cov(X_A, X_B) < 0 \\ = 0 & \text{if} \quad Cov(X_A, X_B) = 0 \end{cases} \end{aligned}$$

- If $Cov(X_A, X_B) > 0$ as in this example, the sample variance S_D^2 is expected to be lower than $S_1^2 + S_2^2$.
  $\Longrightarrow$ **shorter confidence intervals** and **smaller p-values** than using two-sample t-test.

Benefits of Paired Designs

As we have seen, positive dependence within pairs will lead to more significant statistical results. Because of these benefits, experimenters have come up with all kinds of clever ways to get paired data with positive dependence within pairs, **even when the pair of observations can not be made on the same subjects (or same object)**.

- Pairs can be obtained by MATCHING individuals or objects on one or more characteristics considered to influence responses
  - patients with same weight, age, health habits
  - twin studies

In all these cases, we need $2n$ sample units, i.e. two samples $\{X_{i,A}\}_{i=1}^{n}$ and $\{Y_{i,B}\}_{i=1}^{n}$ and the analysis is done on

$$D_i = X_{i,A} - Y_{i,B} \qquad i = 1, \ldots, n$$

where i denotes the i^{th} pair of **matched subjects or objects**.

Checking Conditions for Paired Data

As the inference problem for paired data is simply one-sample problem on the difference within each pair, we need to make sure that

1. the differences are independent: $D_i \perp\!\!\!\perp D_j$ for all $i \neq j$
2. the distribution (histogram) of the differences is not too skewed and has no outliers: $D_i \overset{approx}{\sim} N$

Whether the distributions of the two groups X_A and Y_B are skewed or have outliers DOES NOT matter.

Exercise 5.18. Paired or Not

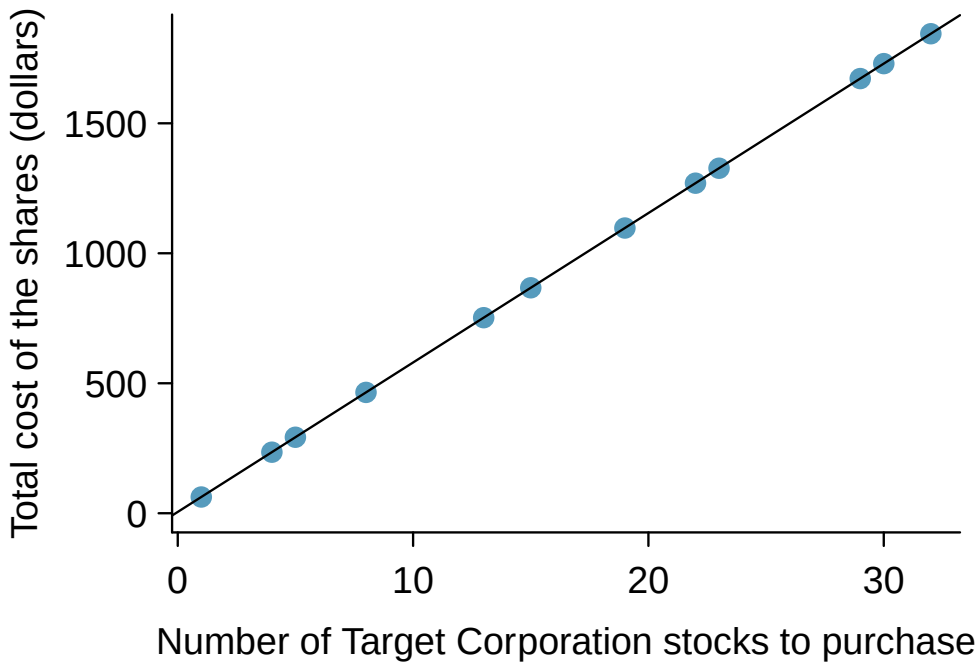
In each of the following scenarios, determine if the data are paired?

1. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days.
2. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.
3. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

Introduction to Simple Linear Regression

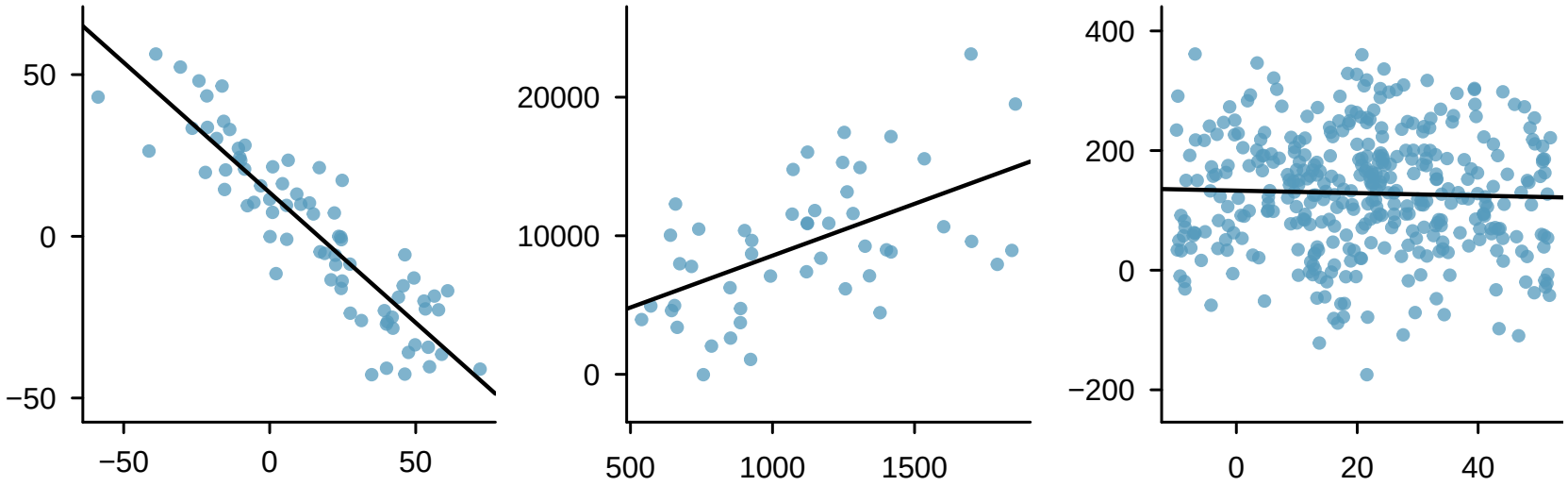
Linear Regression I

Linear regression is a very powerful statistical technique.
It can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.



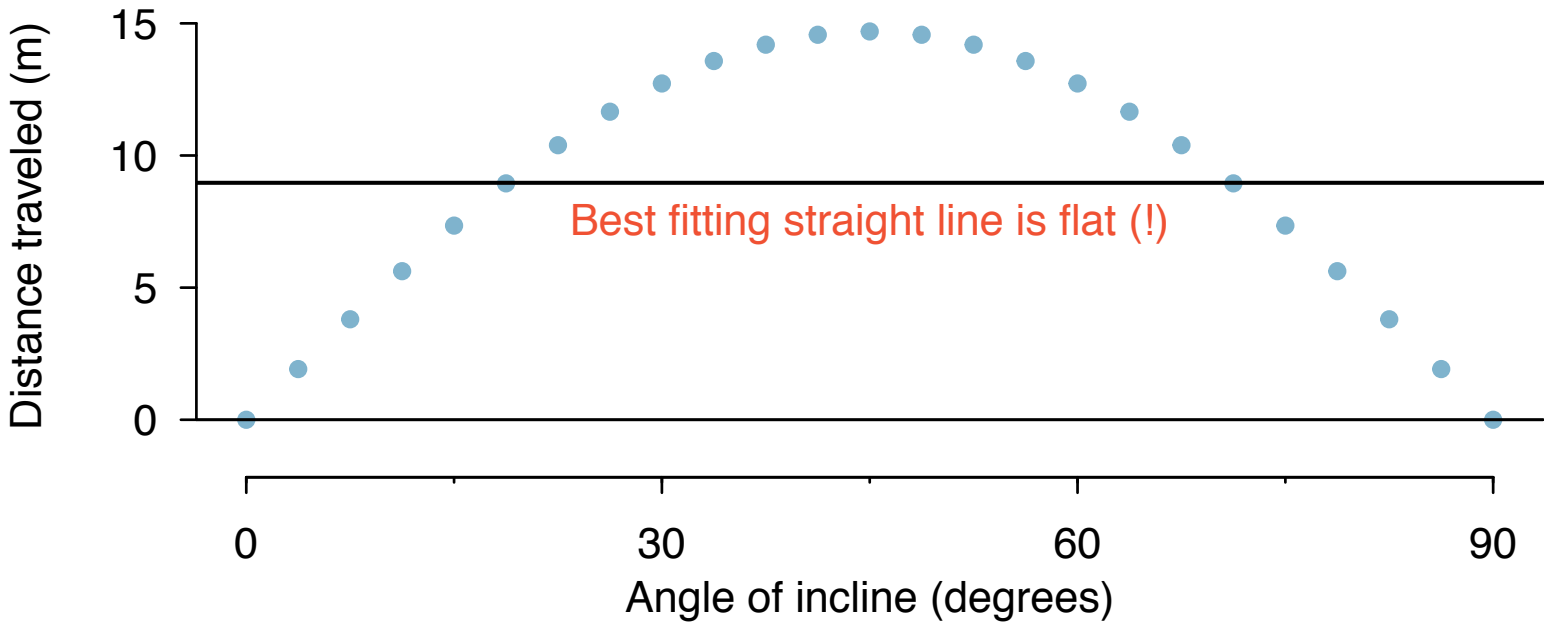
Linear Regression II

These are three data sets where a linear model may be useful even though the data do not fall exactly on the line.



Linear Regression III

Sometimes, fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful.



Describing Linear Relationship with Correlation I

- Recall we defined correlation for two random variables X and Y as

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

- Now, suppose we observe pairs (x_i, y_i) for $1 \leq i \leq n$ which are i.i.d generated from the joint distribution of (X, Y).
- How shall we estimate ρ using the data?

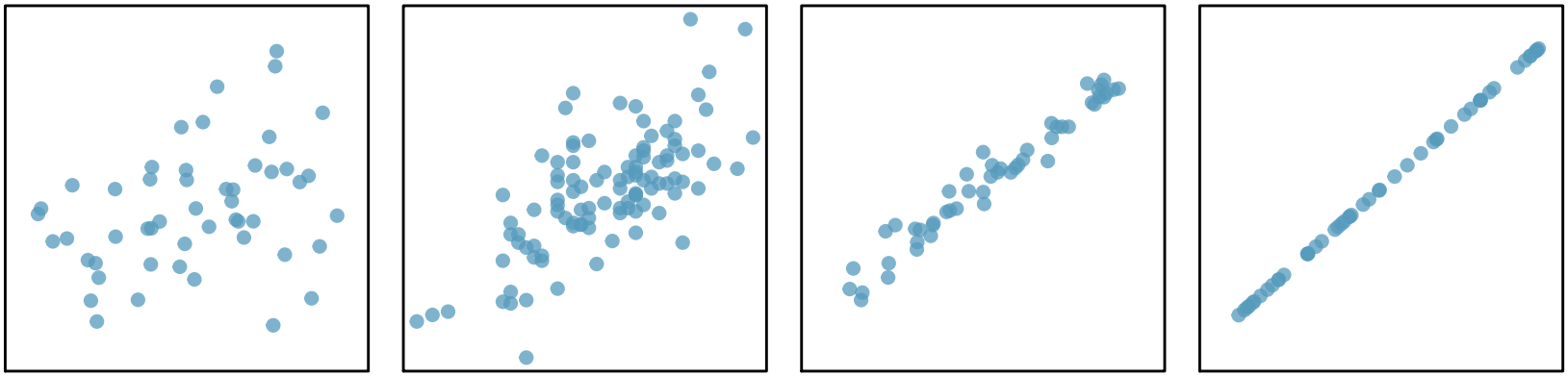
Sample Correlation

Let $\bar{x}$ and $\bar{y}$ be the sample means for x_i and y_i, and let SD_x and SD_y be sample standard deviation for x_i and y_i respectively. The **sample correlation** is defined as

$$\hat{\rho} = r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{SD_y} \right) \left(\frac{x_i - \bar{x}}{SD_x} \right) .$$

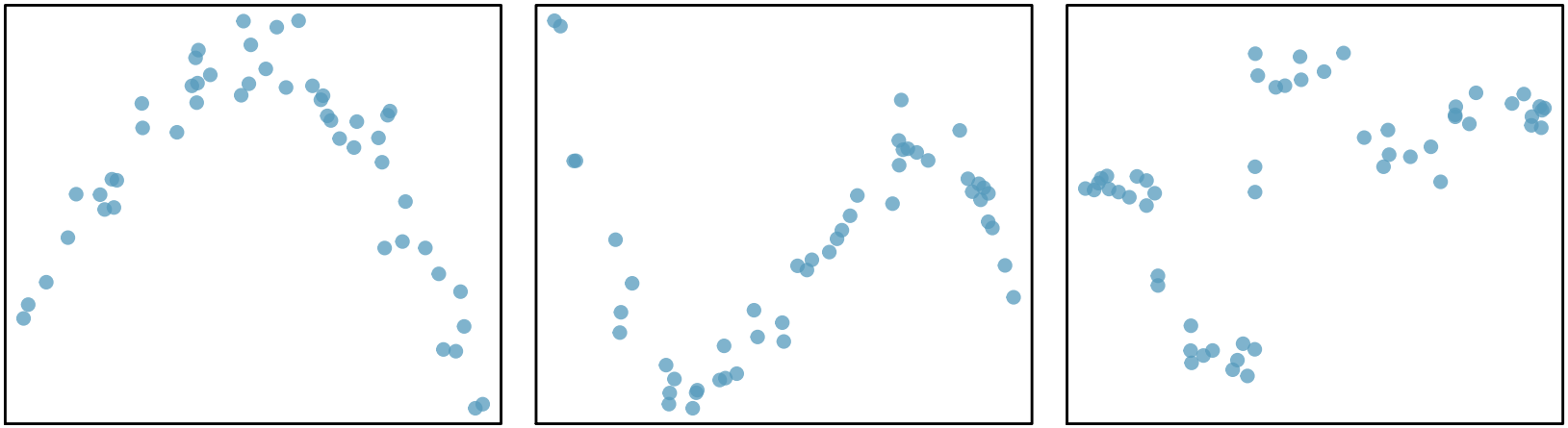
Note that r, the sample correlation always takes values between -1 and 1. We can use r to estimate ρ.

Sample Correlation



Sample Correlation

The correlation is intended to quantify the strength of a **linear** trend. Nonlinear trends, even when strong, sometimes produce small correlations.

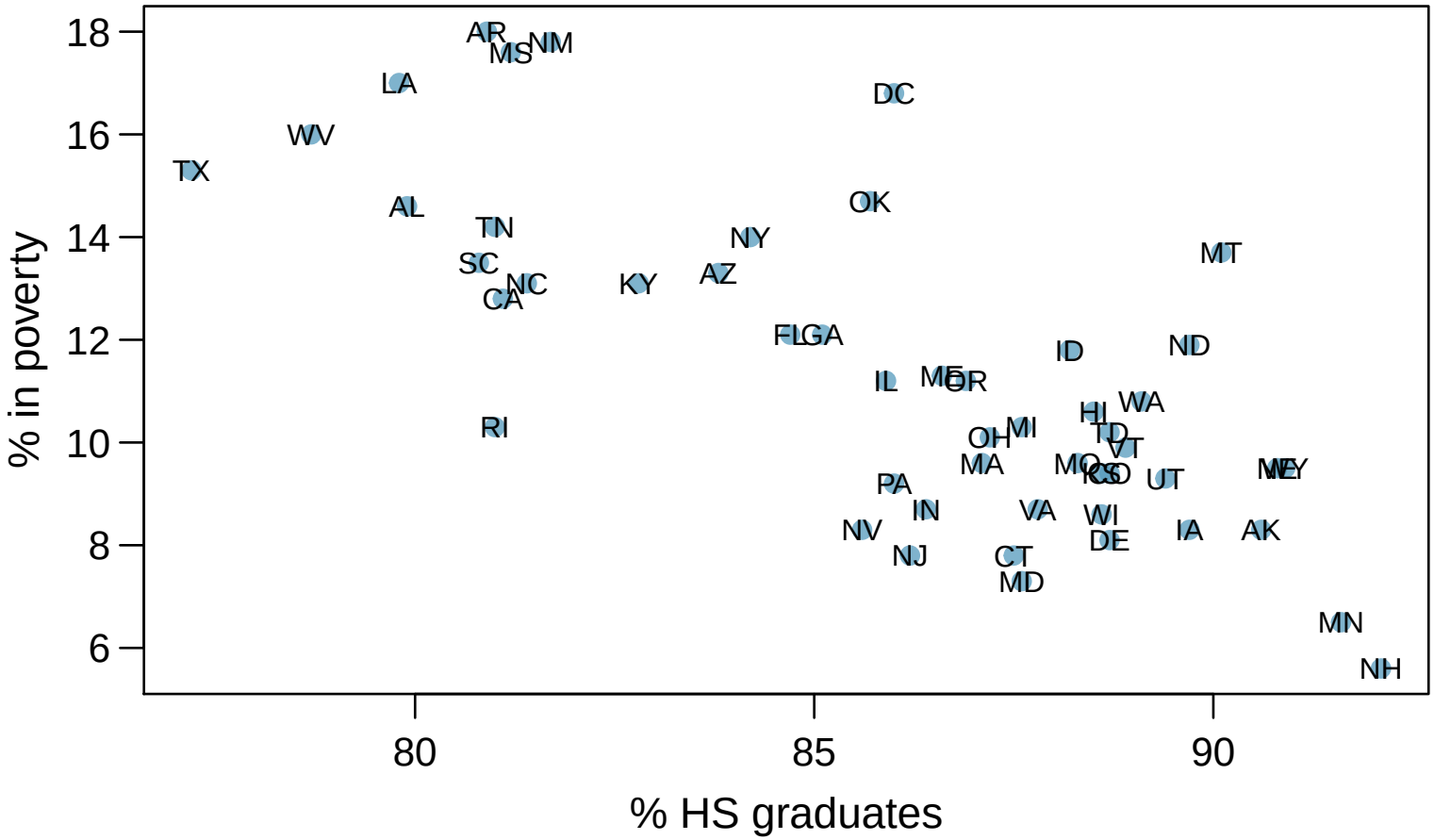


R = −0.23 R = 0.31 R = 0.50

“The Least Squares” Estimation

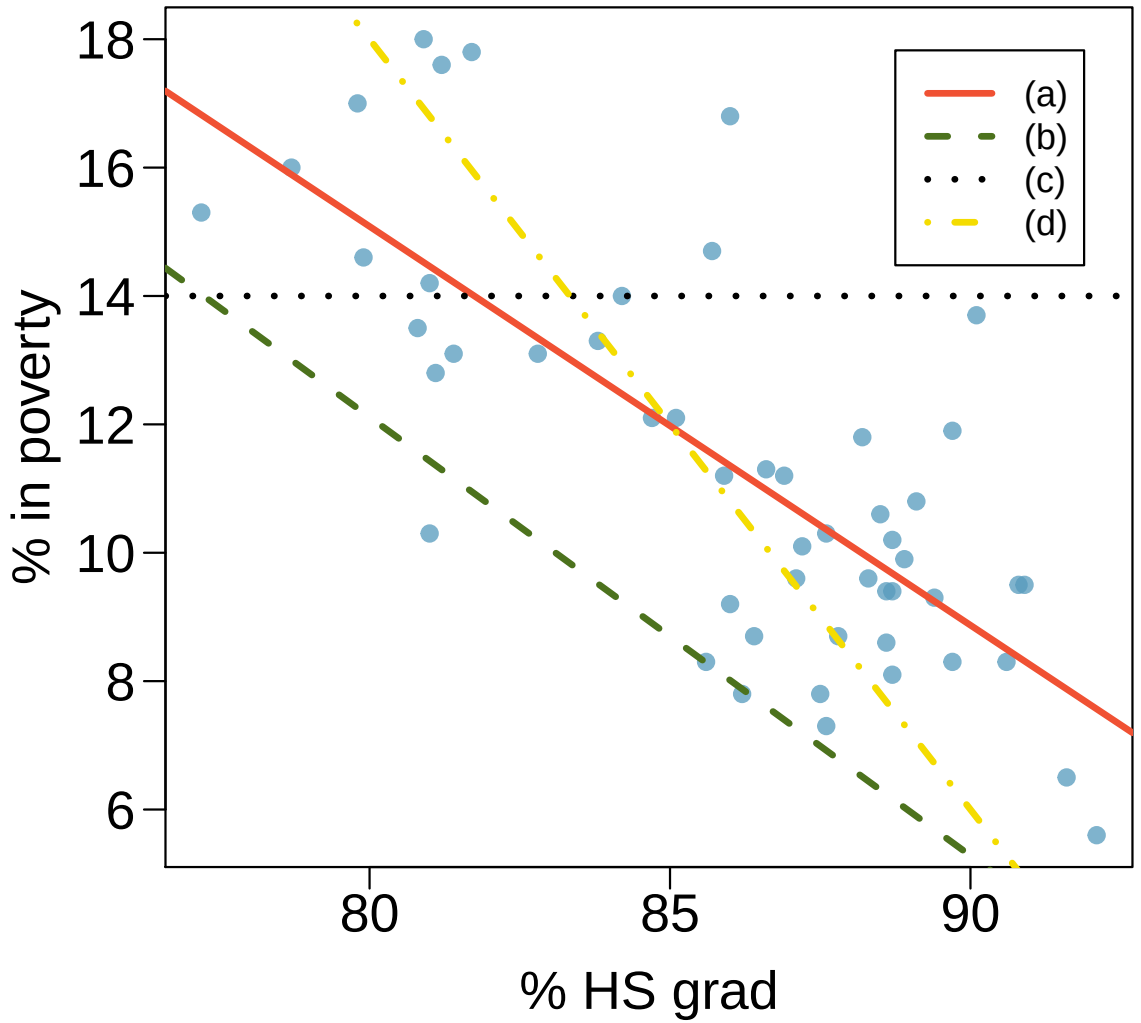
Case Study Example: Poverty vs. HS graduate rate

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).



Eyeballing the line

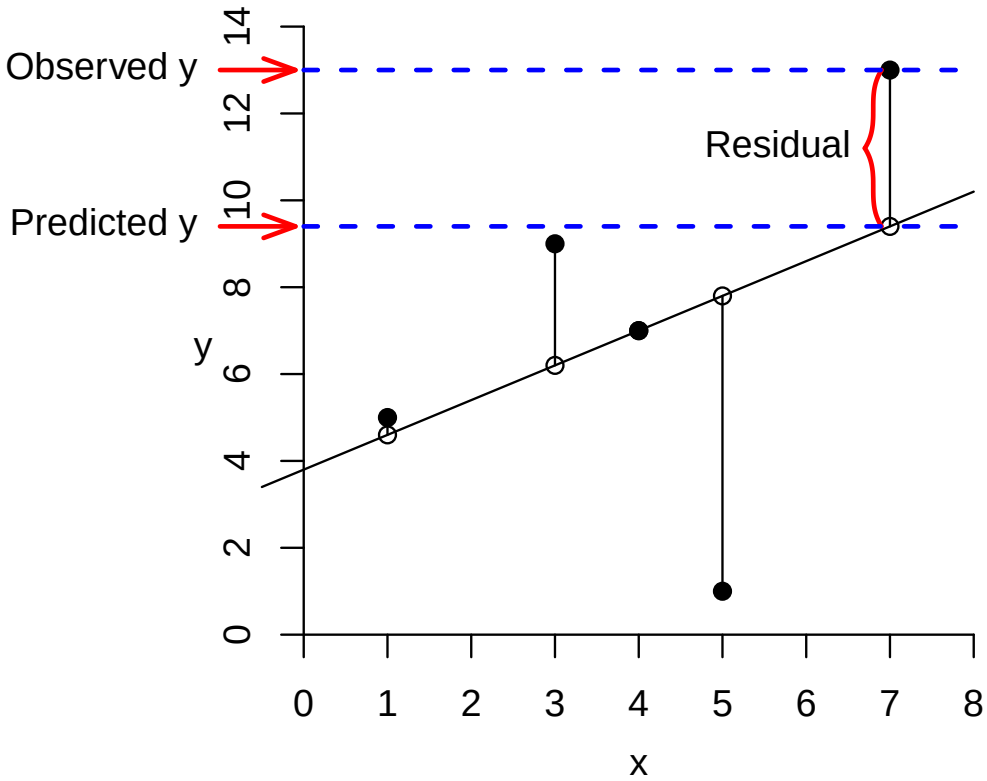
Which line appears to best fit the linear relationship between % in poverty and % HS grad?



Residuals (Prediction Errors)

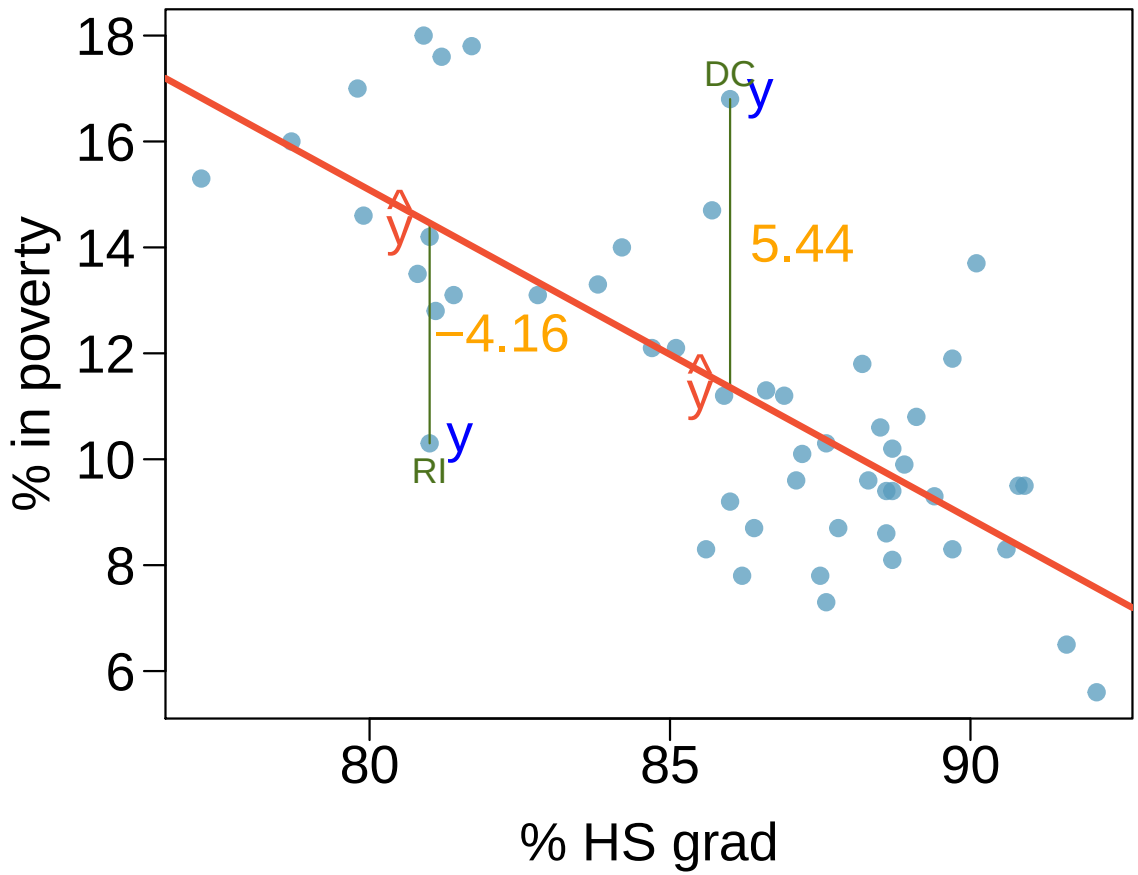
The residual (e_i) of the ith observation (x_i, y_i) is

$$\underset{\text{(Residual)}}{e_i} = \underset{\text{(Observed } y\text{)}}{y_i} - \underset{\text{(Predicted } y\text{)}}{\hat{y}_i}$$



- Residuals are the (signed) vertical distances from data points to model line, not the shortest distances.
- Points above/below the model line have positive/negative residuals.

Residuals (cont.)



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

the Least-Square (LS) Regression Line

We want a line $y = b_0 + b_1 x$ having small residuals:

- **least square method:** to minimize the sum of squared residuals (RSS)

$$e_1^2 + e_2^2 + \cdots + e_n^2 = \sum\nolimits_i (y_i - \widehat{y}_i)^2 = \sum\nolimits_i (y_i - b_0 - b_1 x_i)^2$$

- The least-square (LS) regression line is the line $y = b_0 + b_1 x$ that minimizes the sum of squared errors.
- The slope and intercept of the LS regression line can be shown by math to be

$$b_1 = \text{slope} = r \cdot \frac{\text{SD}_y}{\text{SD}_x}$$

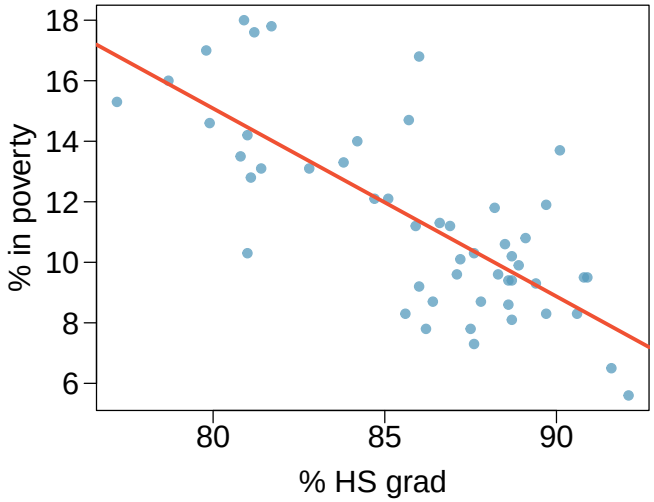
$$b_0 = \text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

Properties of the LS Regression Line

The LS regression line has the following properties:

- The LS regression line always passes through $(\bar{x}, \bar{y})$.
- As x goes up by 1 SD of x, the predicted value $\widehat{y}$ only goes up by $r \times$ (SD of y).
- If $r = 0$, the LS regression line is horizontal $y = \overline{y}$, and the predicted value $\widehat{y}$ is always the mean $\overline{y}$.

Poverty vs. HS graduate rate



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

- Find the slope and the intercept of the least square regression line.

- Write down the equation of the least square regression line.

Interpretation of Slope

The slope indicates how much the response changes associated with a unit change in x on average (may NOT be causal, unless the data are from an experiment).

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

- For each additional % point in HS graduation rate, we would expect the % in poverty to be lower on average by 0.62%.
- If a state manages to bring up its HS graduation rate by 1%, will its living-in-poverty rate lowers by 0.62%?

Interpretation of the Intercept

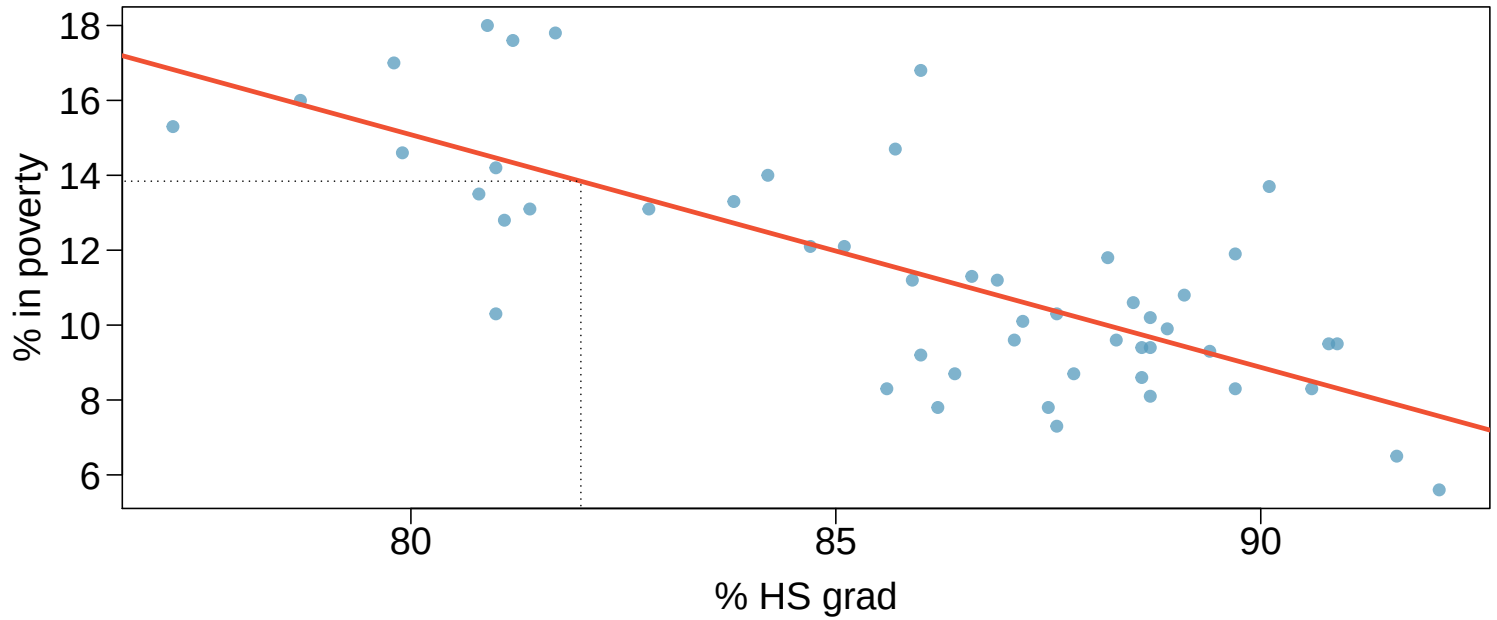
The intercept is the predicted value of response when $x = 0$, which might not have a practical meaning when $x = 0$ is not a possible value

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

- States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.
  - meaningless. There is no such state.

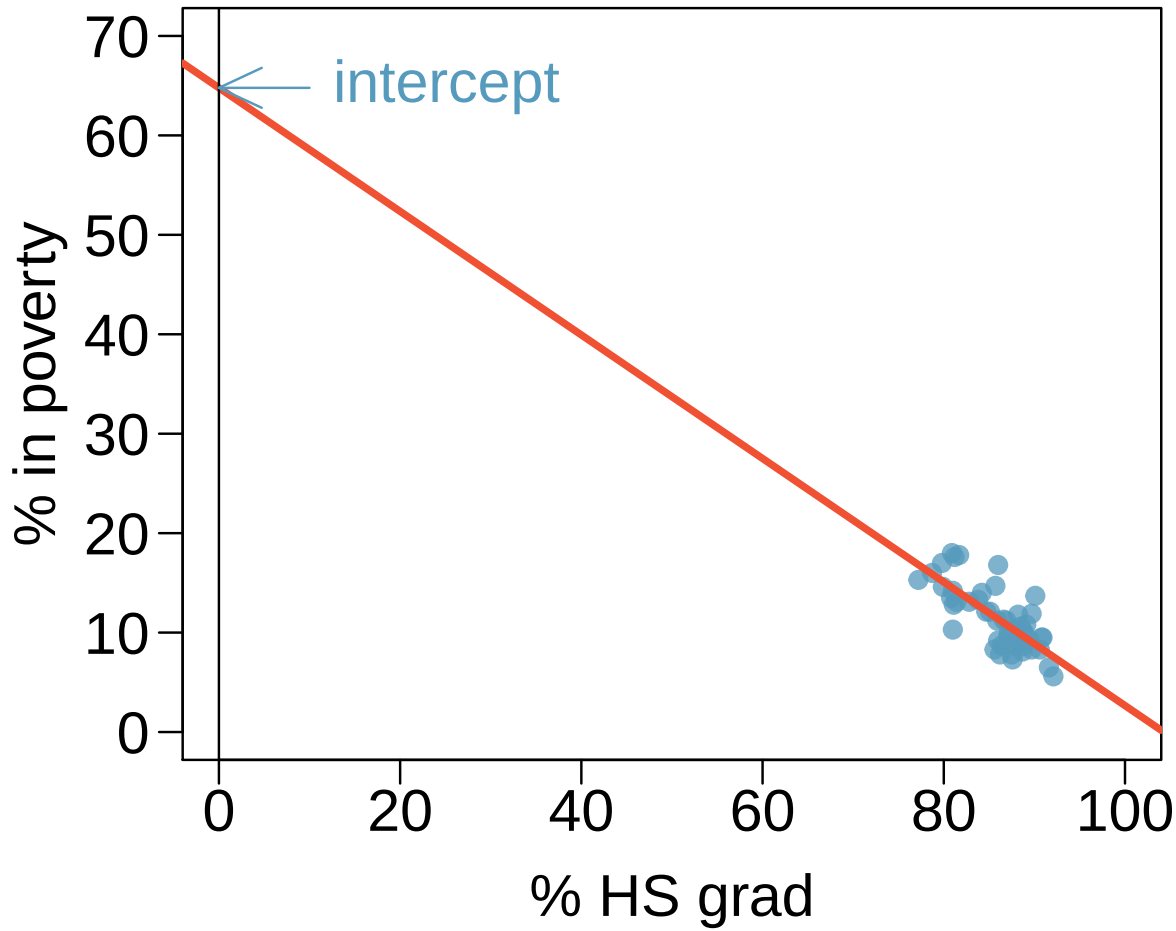
Prediction

Using the least square line 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.



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.



Properties of Residuals

Residuals for a least square regression line have the following properties.

1. Residuals always **sum to zero**, $\sum_{i=1}^{n} e_i = 0$.
   - If the sum > 0, can you improve the prediction?
2. Residuals and the explanatory variable x_i's have **zero correlation**.
   - Residuals are the part in the response that CANNOT be explained or predicted linearly by the explanatory variables.
   - If non-zero, the residuals can be predicted by x_i's, not the best prediction.

R^2 = R-squared = r^2

Moreover, one can show that

$$r^2 = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = \frac{\text{Sample Variance of predicted } y\text{'s}}{\text{Sample Variance of observed } y\text{'s}}$$

That is,

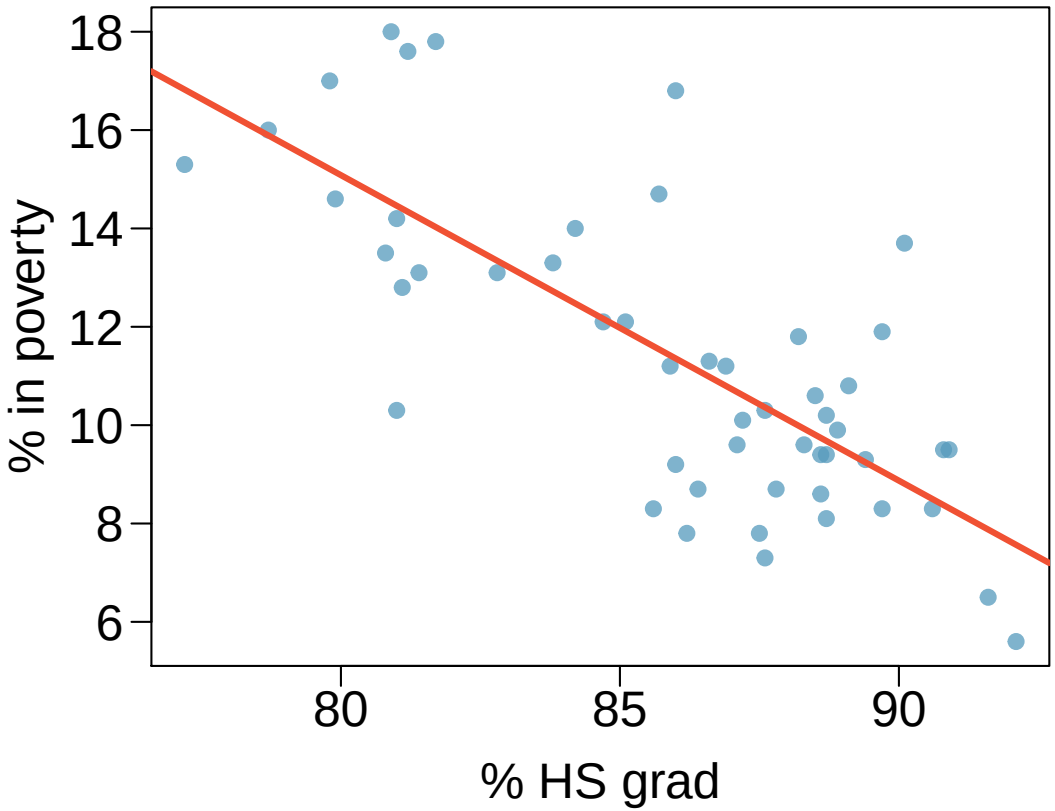
$$\begin{aligned} R^2 = r^2 &= \text{the square of the correlation coefficient} \\ &= \text{the proportion of variation in the response} \\ &\quad \text{explained by the explanatory variable} \end{aligned}$$

The remainder is explained by variables not included in the model or by inherent randomness in the data.

$$1 - r^2 = \frac{\sum_i \hat{e}_i^2}{\sum_i (y_i - \overline{y})^2} = \frac{\text{Sample Variance of Residuals}}{\text{Sample Variance of } y}$$

Interpretation of R-squared

For the model we've been working with,
$R^2 = r^2 = (-0.75)^2 \approx 0.56$, which means — 56% of the variability in the % of residents living in poverty among the 51 states is explained by the variable "% of HS grad".



Summary: Least Square Regression Line

- The least-square (LS) regression line is the line $y = b_0 + b_1 x$ that minimizes the sum of squared errors:

$$\sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - b_0 - b_1 x_i)^2$$

- The LS estimates for the slope and intercept are

$$b_1 = \text{slope} = r \cdot \frac{\text{SD}_y}{\text{SD}_x}, \quad b_0 = \text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

- For LS regression, residuals add up to zero, and have 0 correlation with the explanatory variable x.
- R^2 = R-squared = r^2 = proportion of variation in the response y that can be explained by the explanatory variable
- Regression treats x and y differently.
  The LS regression line that predicts y from x can only predict y from x, not x from y.