Elementary Statistics – Simple Linear Regression Pt 2

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Simple Linear Regression
  - Fitting a model
  - Assessing model fit
  - Issues to look out for
  - Binary predictors

Warm Up: Interpreting the Regression Line

In a linear regression line, Y_i represents an individual outcome or response, X_i represents an individual input, and ϵ_i represents error: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- β_0: the intercept term captures the average response given an input of 0
- β_1: the slope term captures the expected (average) change in response with a one-unit change in input

Suppose Y represents systolic blood pressure (in mm Hg) and X represents aspirin dosage (in mg). The relationship between these variables is modeled as:

$$Y_i = 120 + 10X_i$$

- What is the average blood pressure for someone with an aspirin dosage of 0 mg?
- How much would you expect blood pressure to change with a one mg increase in aspirin dosage?
- What is the average blood pressure for someone with an aspirin dosage of 100 mg?

Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

Using data from 2017-2018 summarizing the per-capita income (in dollars) and life expectancy in the US and Puerto Rico, we get this regression line:

$$lifeExp = 73.62 + 0.0000836 * income$$

Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

Using data from 2017-2018 summarizing the per-capita income (in dollars) and life expectancy in the US and Puerto Rico, we get this regression line:

$$lifeExp = 73.62 + 0.0000836 * income$$

What is the average life expectancy given an income of $0?
What is the average life expectancy given an income of $30,000

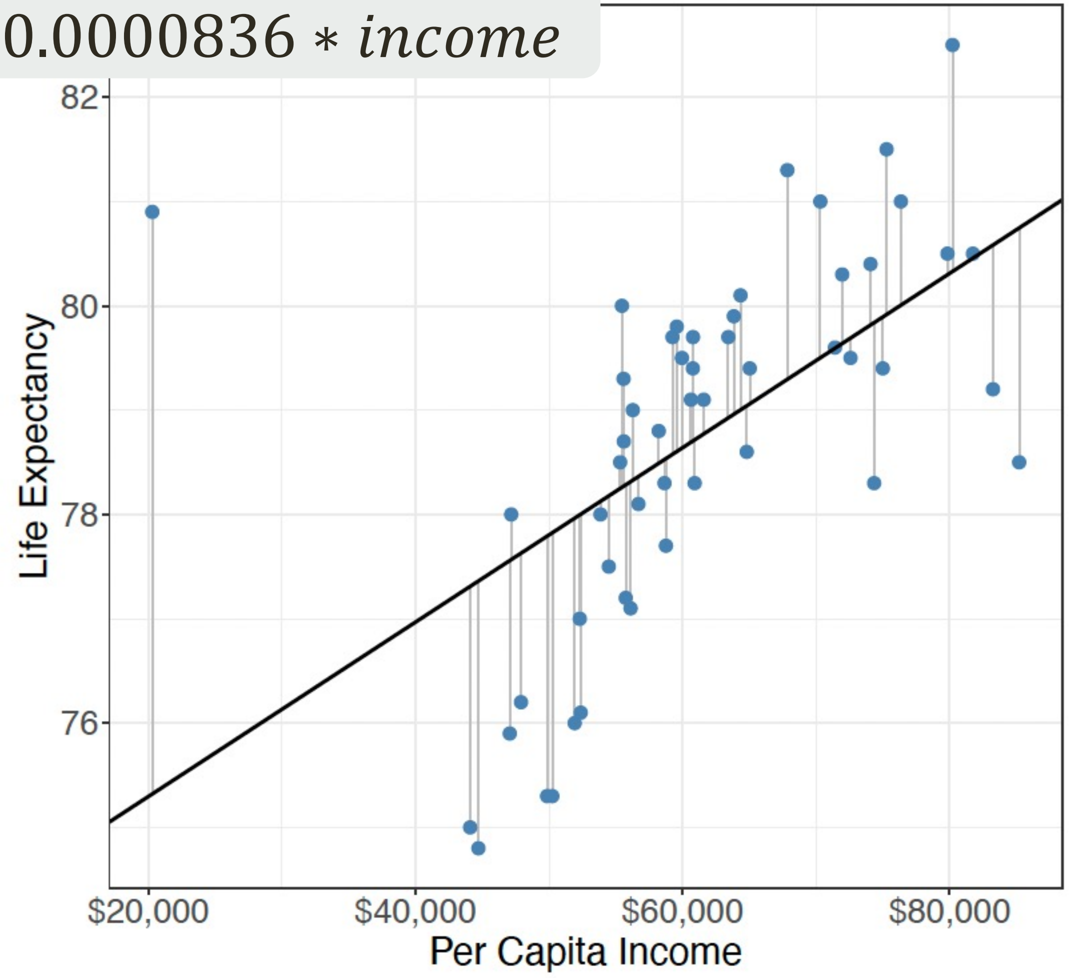
Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

- The line in this plot shows our regression
- The points show the actual data



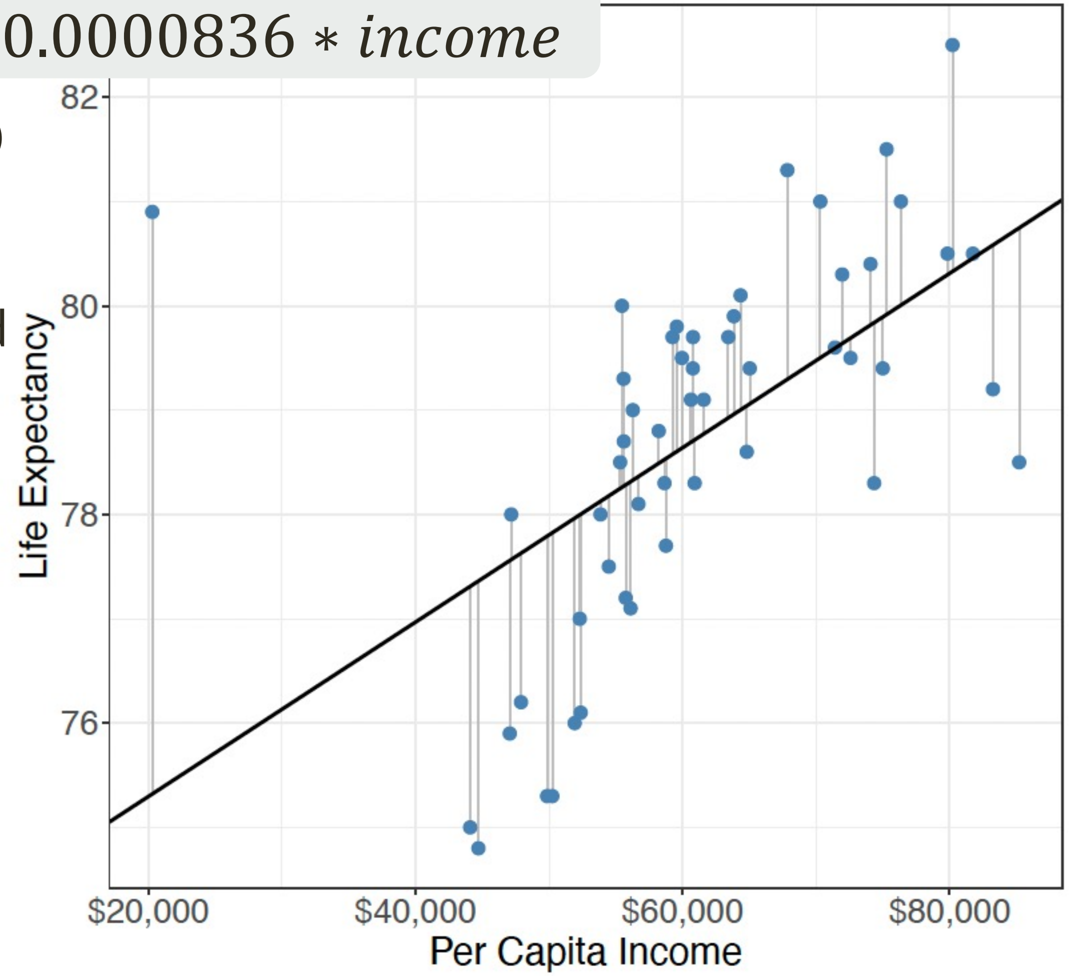
Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

For most observations (i) there is a difference between the predicted value ($\widehat{Y}_i$ aka the line) and the actual value (Y_i aka the point)



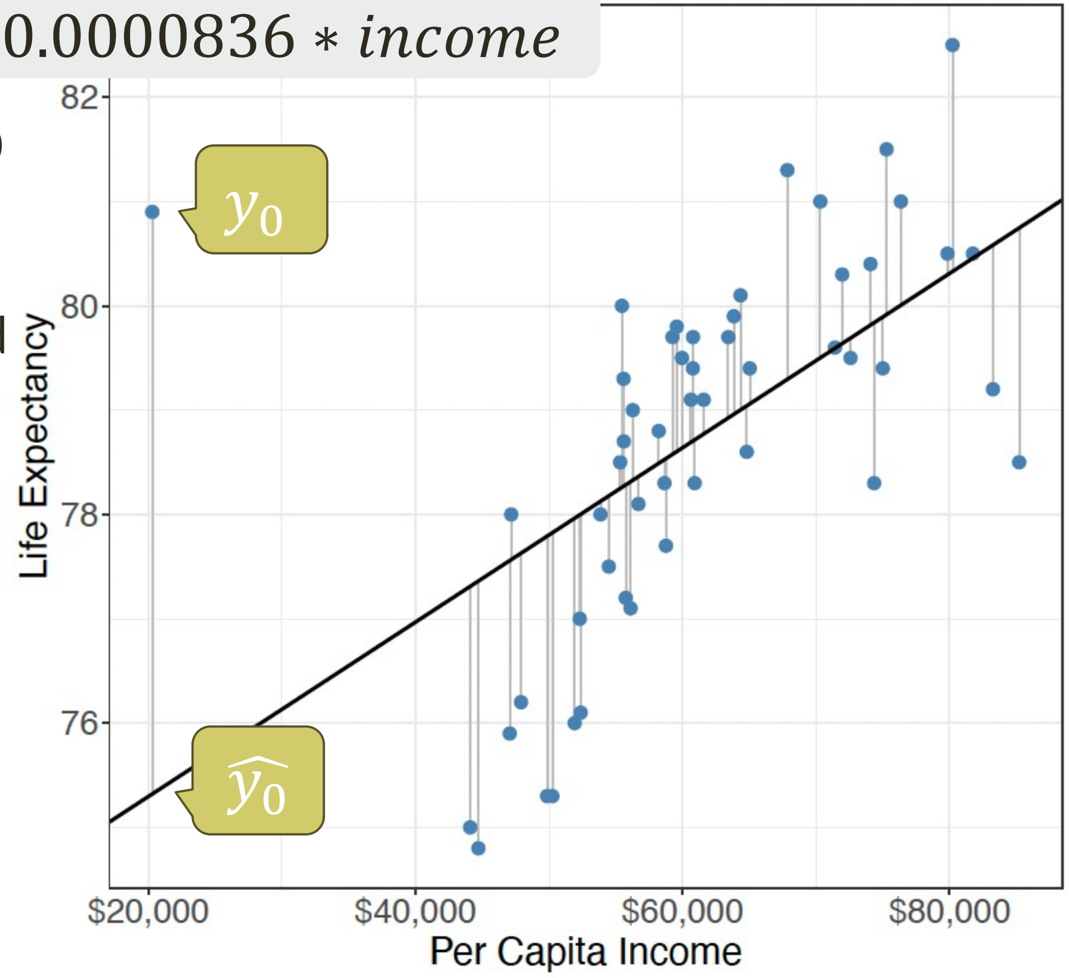
Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

For most observations (i) there is a difference between the predicted value ($\widehat{Y}_i$ aka the line) and the actual value (Y_i aka the point)



Least Squares Line

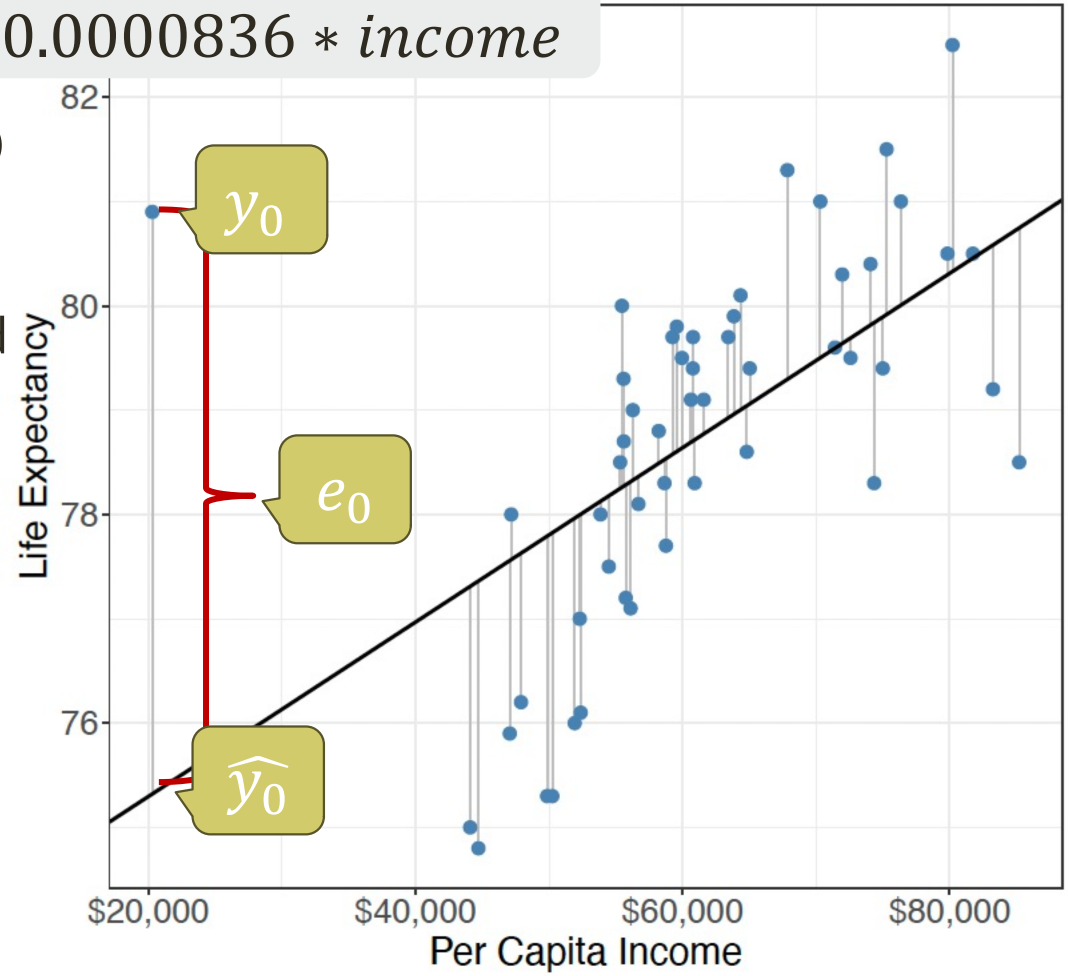
Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

For most observations (i) there is a difference between the predicted value ($\widehat{Y}_i$ aka the line) and the actual value (Y_i aka the point)

This difference is the ***residual*** (e_i) for observation i



Least Squares Line

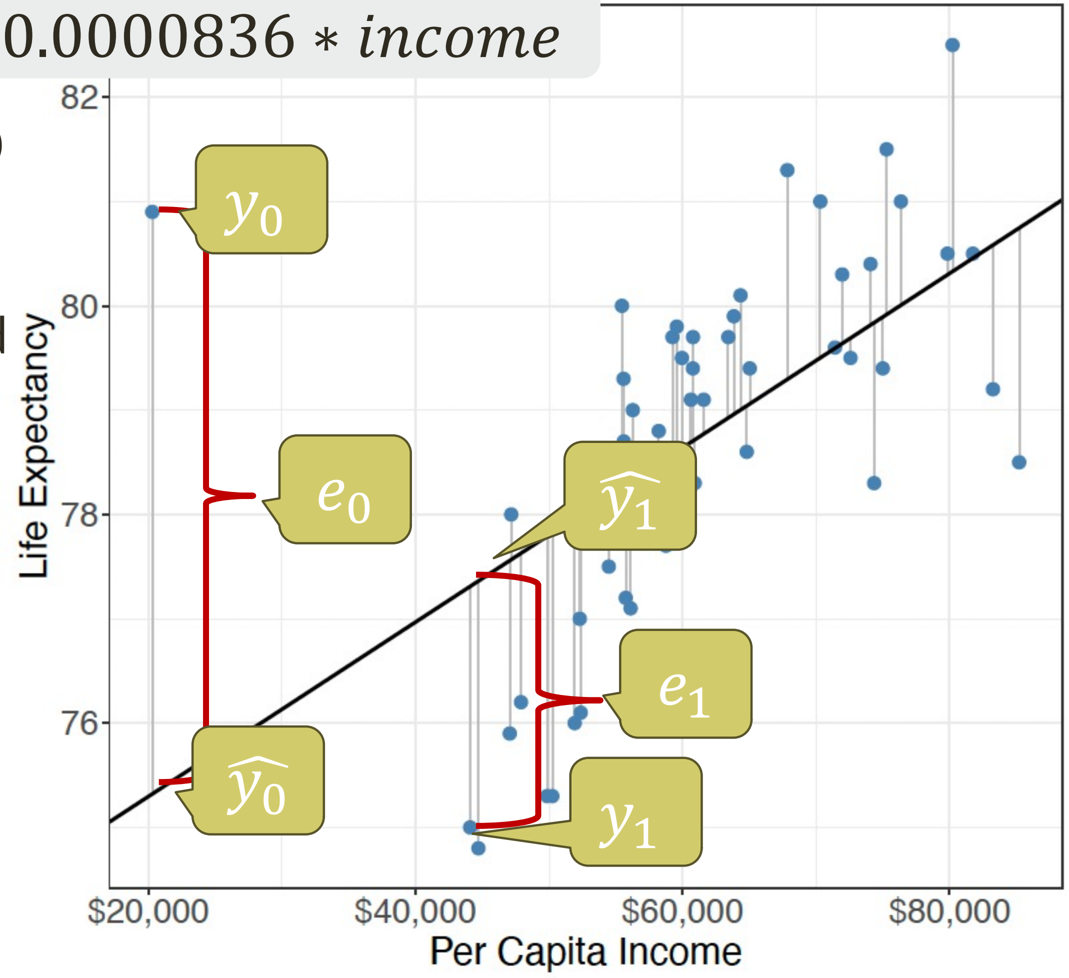
Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

For most observations (i) there is a difference between the predicted value ($\hat{Y}_i$ aka the line) and the actual value (Y_i aka the point)

This difference is the ***residual*** (e_i) for observation i



Least Squares Line

Motivating Question

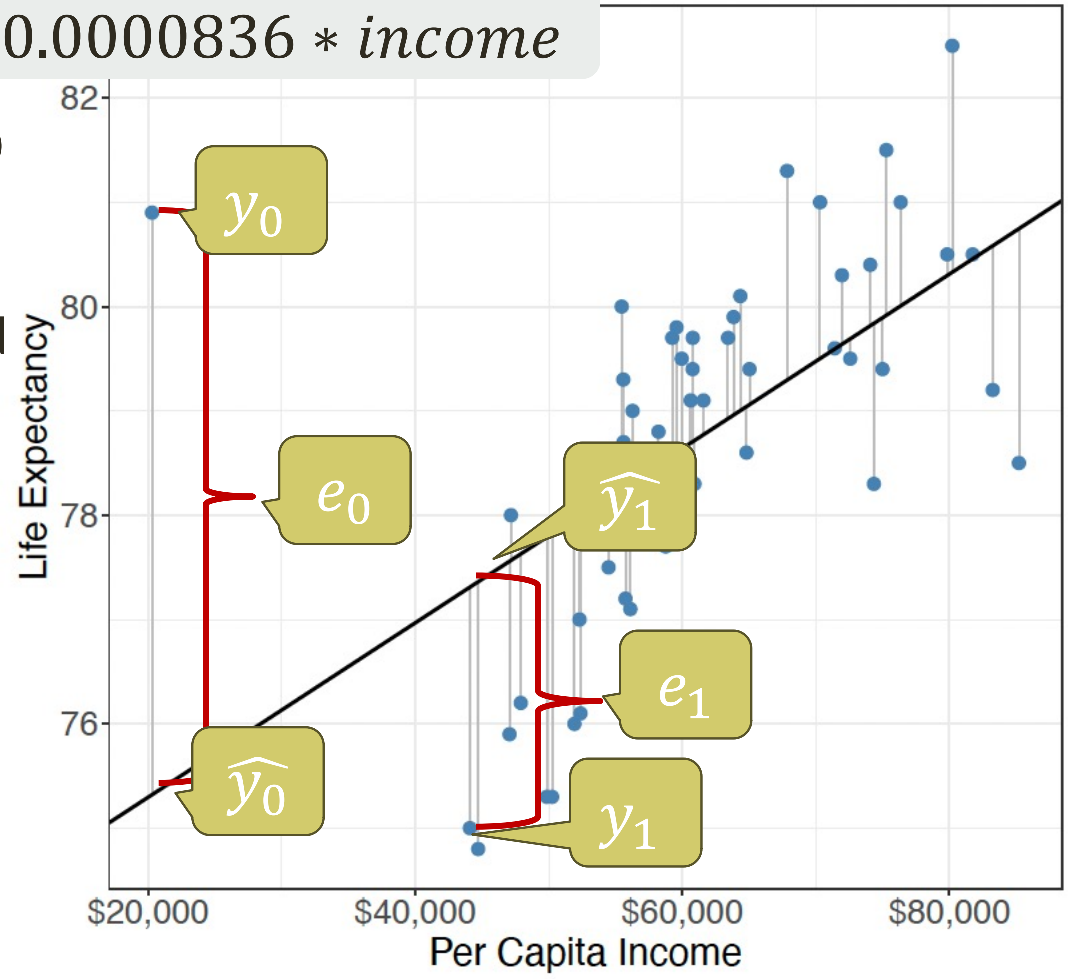
Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

For most observations (i) there is a difference between the predicted value ($\widehat{Y}_i$ aka the line) and the actual value (Y_i aka the point)

This difference is the ***residual*** (e_i) for observation i

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



Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

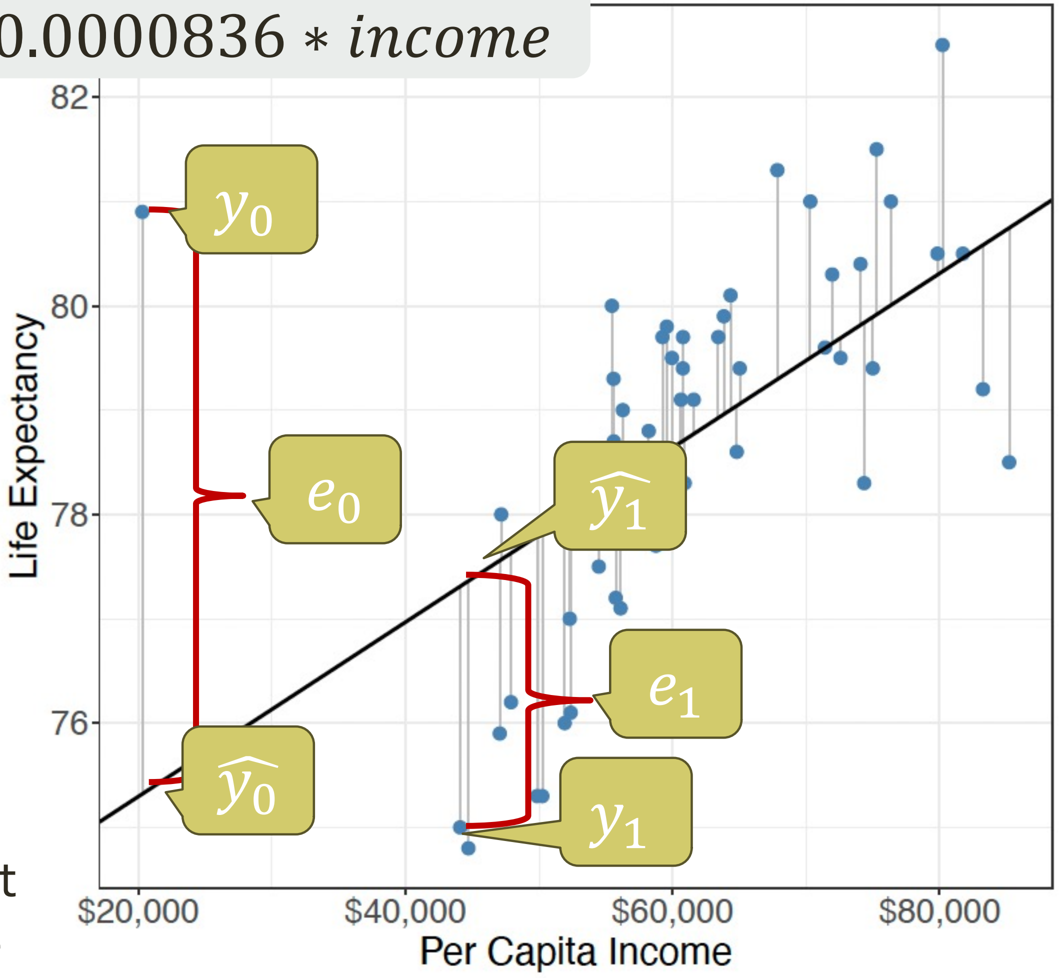
$$lifeExp = 73.62 + 0.0000836 * income$$

For most observations (i) there is a difference between the predicted value ($\widehat{Y}_i$ aka the line) and the actual value (Y_i aka the point)

This difference is the ***residual*** (e_i) for observation i

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

residual is the "error" that is unaccounted for by the regression line.



Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

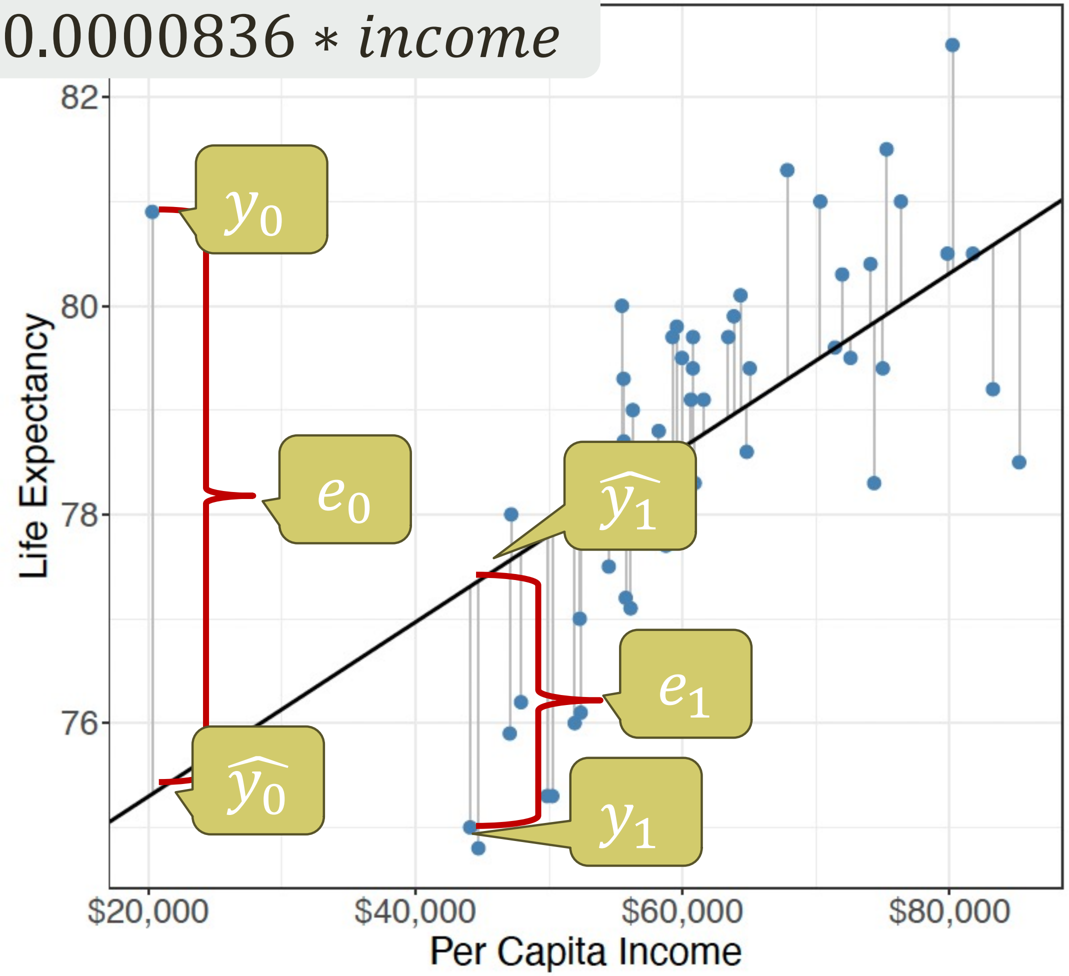
$$lifeExp = 73.62 + 0.0000836 * income$$

Residual (e_i) for observation i:

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

The best line minimizes residuals.

Why?



Least Squares Line

Motivating Question

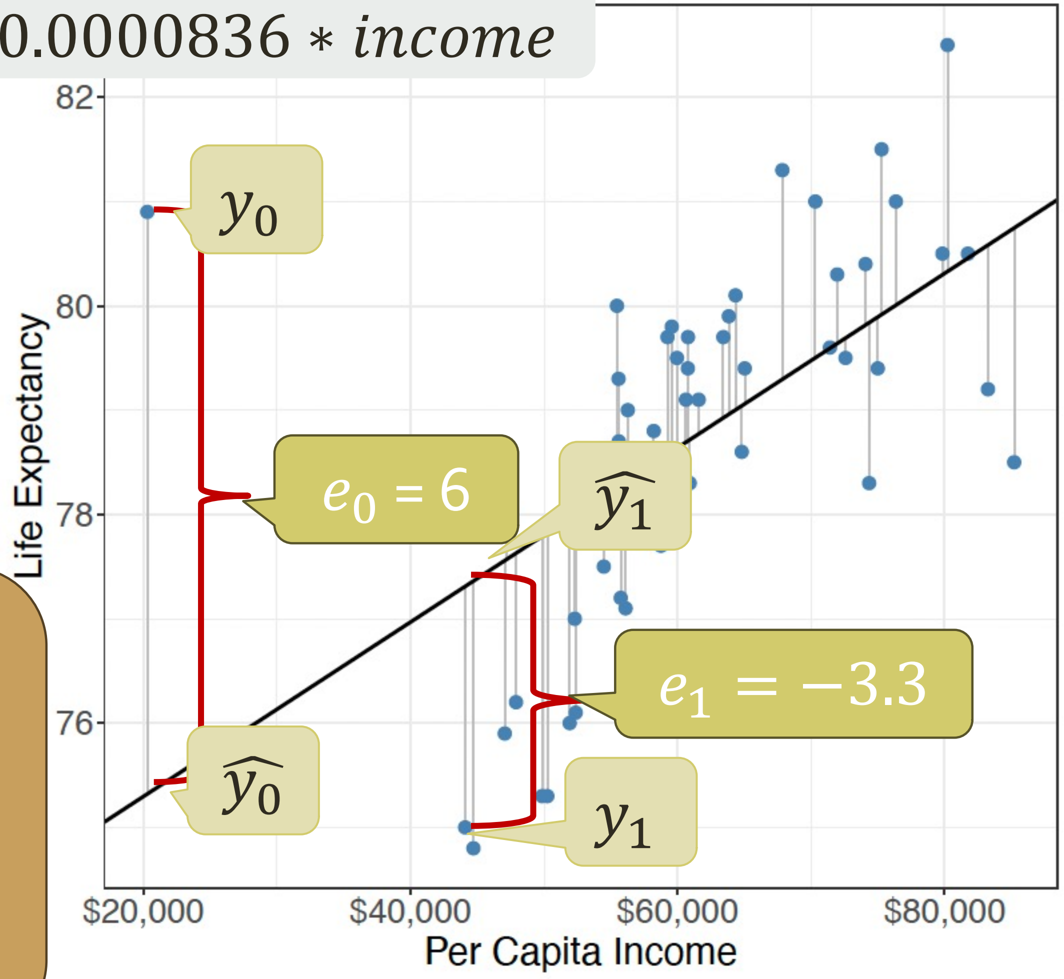
Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

Residual (e_i) for observation i:

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

The best line minimizes residuals (i.e. minimizes overall error).



Can we measure overall error by adding all residuals?
What is the sum of e_0 and e_1?
What if both were positive, is the sum different? Is the amount of error different?

Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

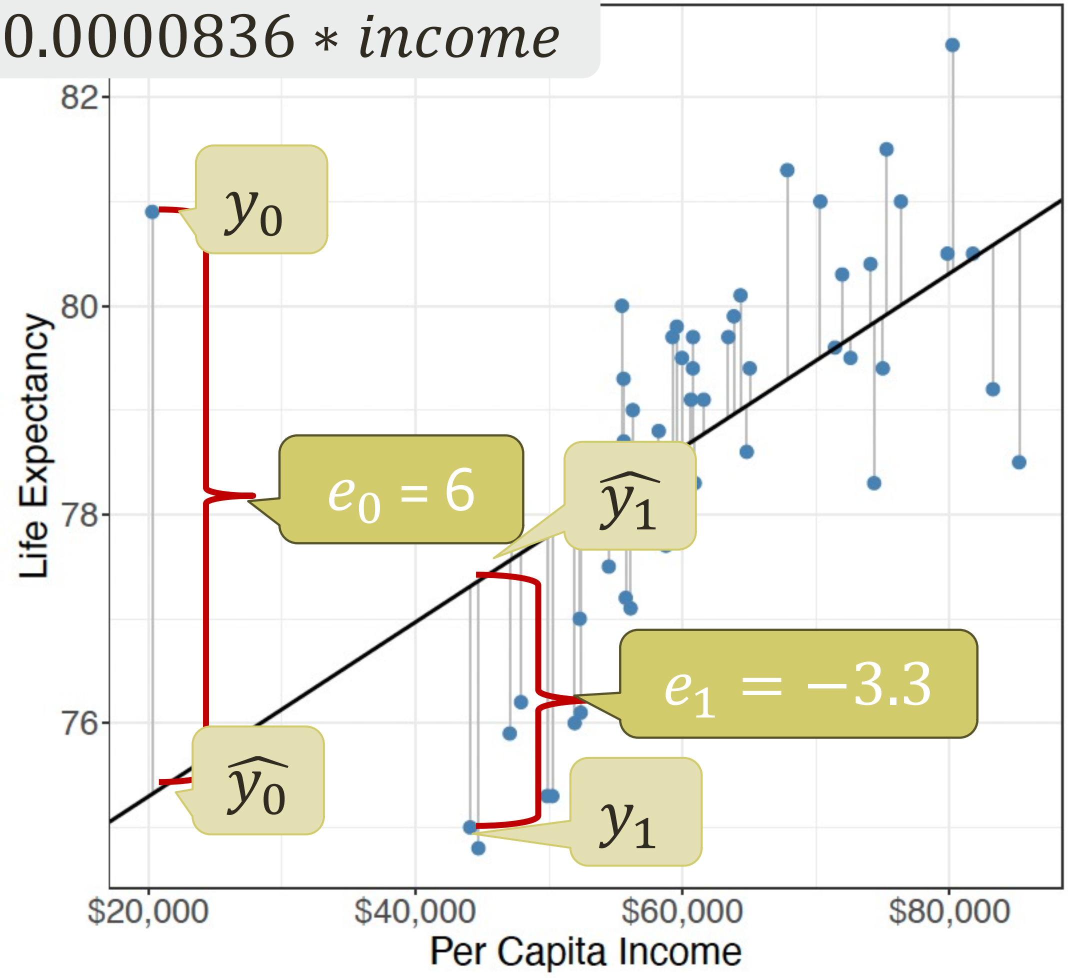
Residual (e_i) for observation i:

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

The best line minimizes residuals (i.e. minimizes overall error).

We measure overall error by looking at sum of squared residuals or error (SSE)

SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

Sum of squared residuals:

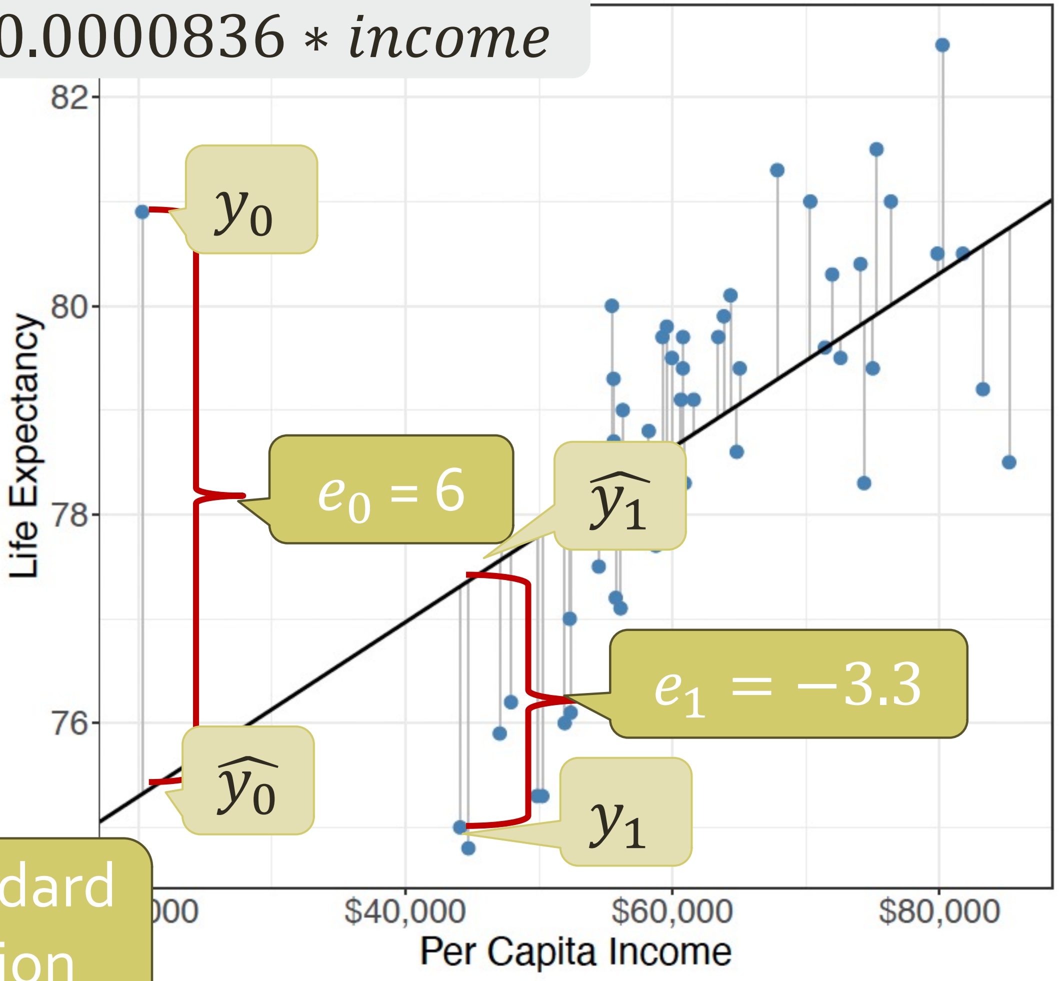
SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

We minimize SSE to get the best line. This gives us the following coefficients:

$$\widehat{\beta_1} = r\frac{s_y}{s_x}$$

r is correlation between x and y

s is standard deviation



Least Squares Line

Motivating Question

Is there an association between per-capita income and life expectancy in the United States?

$$lifeExp = 73.62 + 0.0000836 * income$$

Sum of squared residuals:

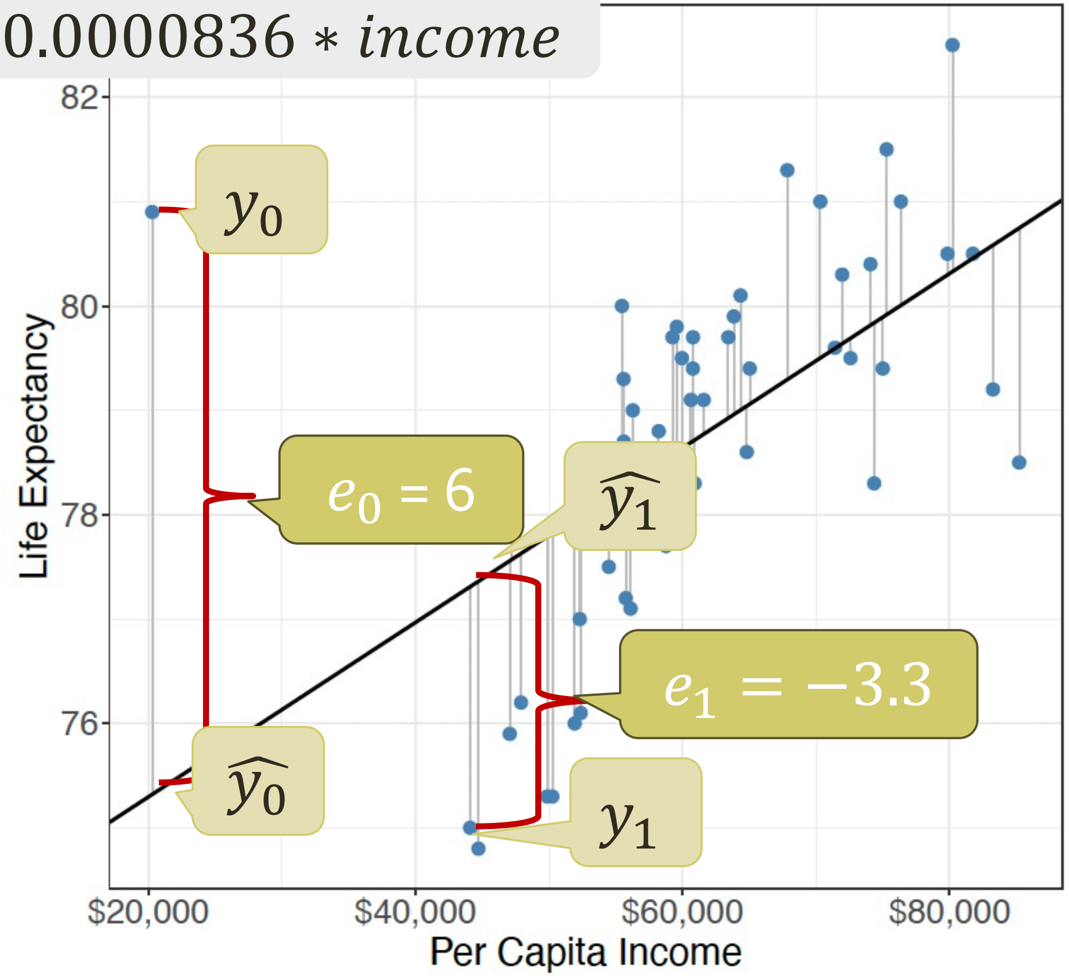
SSE = $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

We minimize SSE to get the best line. This gives us the following coefficients:

$$\widehat{\beta_1} = r\frac{s_y}{s_x}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

bar is mean



Least Squares Line

Practice: Let's fit a regression to represent the relationship between family income and gift aid for Elmhurst College in Il.

Use the table below to compute the slope and intercept of the line.

Family income, x		Gift aid, y		
mean	sd	mean	sd	r
102	63.2	19.9	5.46	-0.499

r is correlation between x and y

$$\beta_1 = r\frac{s_y}{s_x}$$

s is standard deviation

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

bar is mean

$$\widehat{aid} = \beta_0 + \beta_1 * family\_income$$

Least Squares Line

Practice: With your regression, what is the average aid for a family with an income of $30,000? How about $100,000?

Family income, x		Gift aid, y		
mean	sd	mean	sd	r
102	63.2	19.9	5.46	-0.499

r is correlation between x and y

$$\beta_1 = r\frac{s_y}{s_x}$$

s is standard deviation

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

bar is mean

$$\widehat{aid} = \beta_0 + \beta_1 * family\_income$$

Least Squares Line

Practice: Let's fit a regression to represent the relationship between possum head length and total length.

Use the table below to compute the slope and intercept of the line.

total_len, x (cm)		head_len, y (mm)		
mean	sd	mean	sd	r
87	15	92	7.5	0.44

r is correlation between x and y

$$\beta_1 = r\frac{s_y}{s_x}$$

s is standard deviation

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

bar is mean

$$\widehat{head\_len} = \beta_0 + \beta_1 * total\_len$$

Least Squares Line

Practice: With your regression, what is the average head_len for a raccoon with a total_len of 20cm? How about 100cm?

total_len, x (cm)		head_len, y (mm)		
mean	sd	mean	sd	r
87	15	92	7.5	0.44

r is correlation between x and y

$$\beta_1 = r\frac{s_y}{s_x}$$

s is standard deviation

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

bar is mean

$$\widehat{head\_len} = \beta_0 + \beta_1 * total\_len$$

Assessing Fit

Residual plots can help us to identify characteristics or patterns still apparent in data after fitting a model.

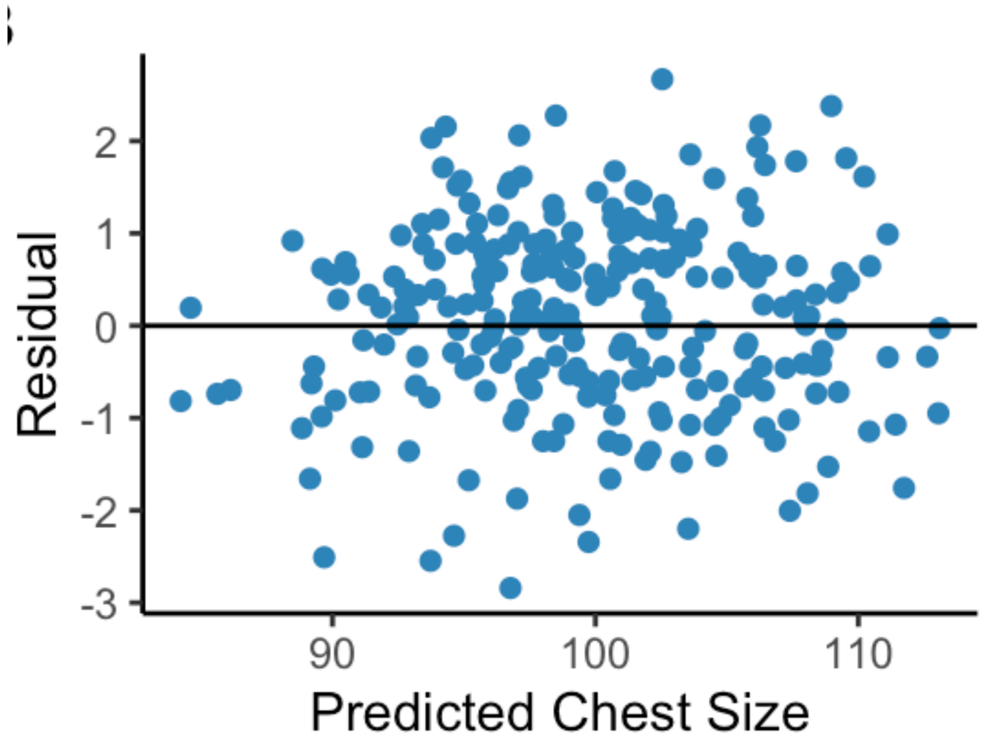
- To create a residual plot, plot predicted values, $\hat{y}_i$, on the x-axis and corresponding residuals, e_i on the y-axis

Assessing Fit

Residual plots can help us to identify characteristics or patterns still apparent in data after fitting a model.

- To create a residual plot, plot predicted values, $\hat{y}_i$, on the x-axis and corresponding residuals, e_i on the y-axis

Ex. For predicted chest-size from earlier:



If the fit is good, there will be no discernable pattern.

Why?

Assessing Fit

The coefficient of determination, written as R^2, measures the proportion of variation in the outcome variable, y, that our model is able to successfully explain.

Assessing Fit

The coefficient of determination, written as R^2, measures the proportion of variation in the outcome variable, y, that our model is able to successfully explain.

What is the range of possible values for R^2?

Is a bigger R^2 better?

Assessing Fit

The coefficient of determination, written as R^2, measures the proportion of variation in the outcome variable, y, that our model is able to successfully explain.

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

this is predicted y for observation i

- SST is the total sum of squares, which measures variability in y values
  - SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

this is mean observed y

Assessing Fit

The coefficient of determination, written as R^2, measures the proportion of variation in the outcome variable, y, that our model is able to successfully explain.

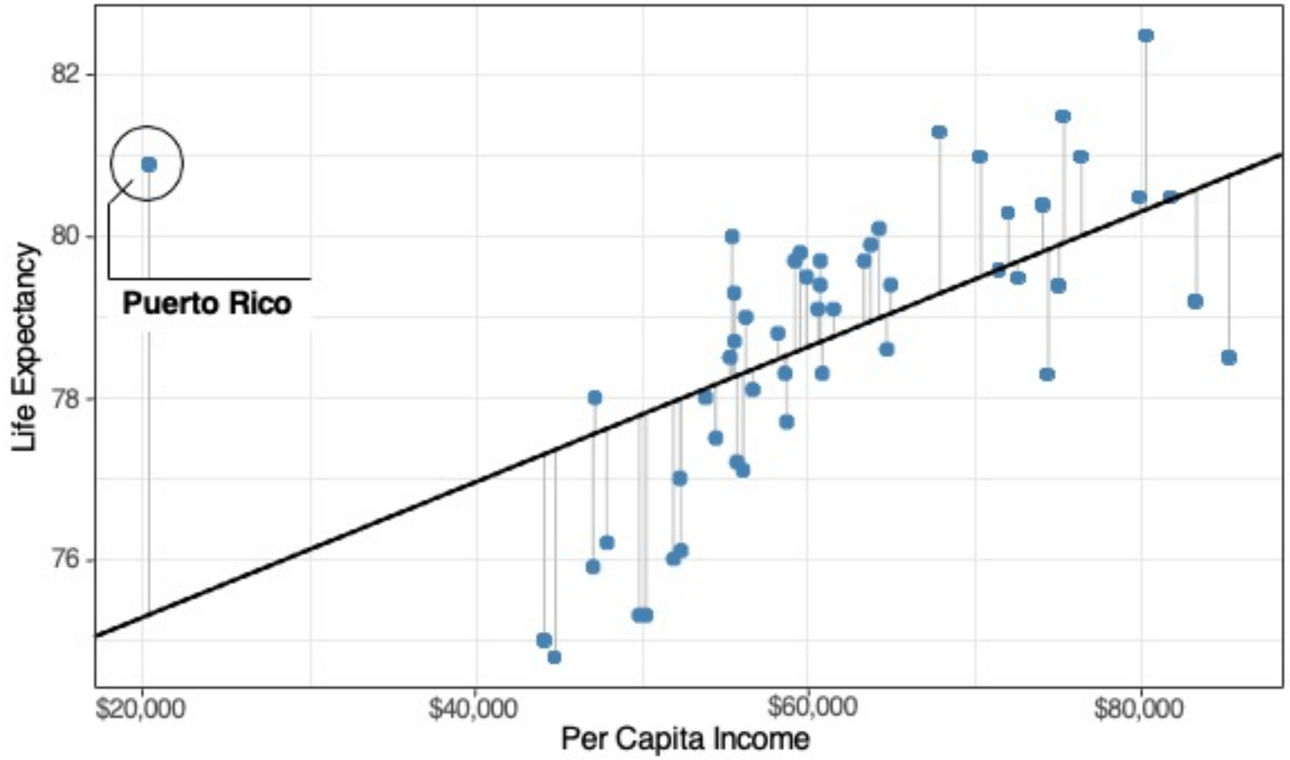
$$R^2 = \frac{SST\ - SSE}{SST} = 1 - \frac{SSE}{SST}$$

- SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- SST is the total sum of squares, which measures variability in y values
  - SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Practice: In the Elmhurst dataset SST = 1461, and SSE = 1098. What is R^2? What does it tell us?

Outliers and Influential Points

The observation on the far left side of the scatter plot lies substantially farther away from the “center” of the plot than any other point...



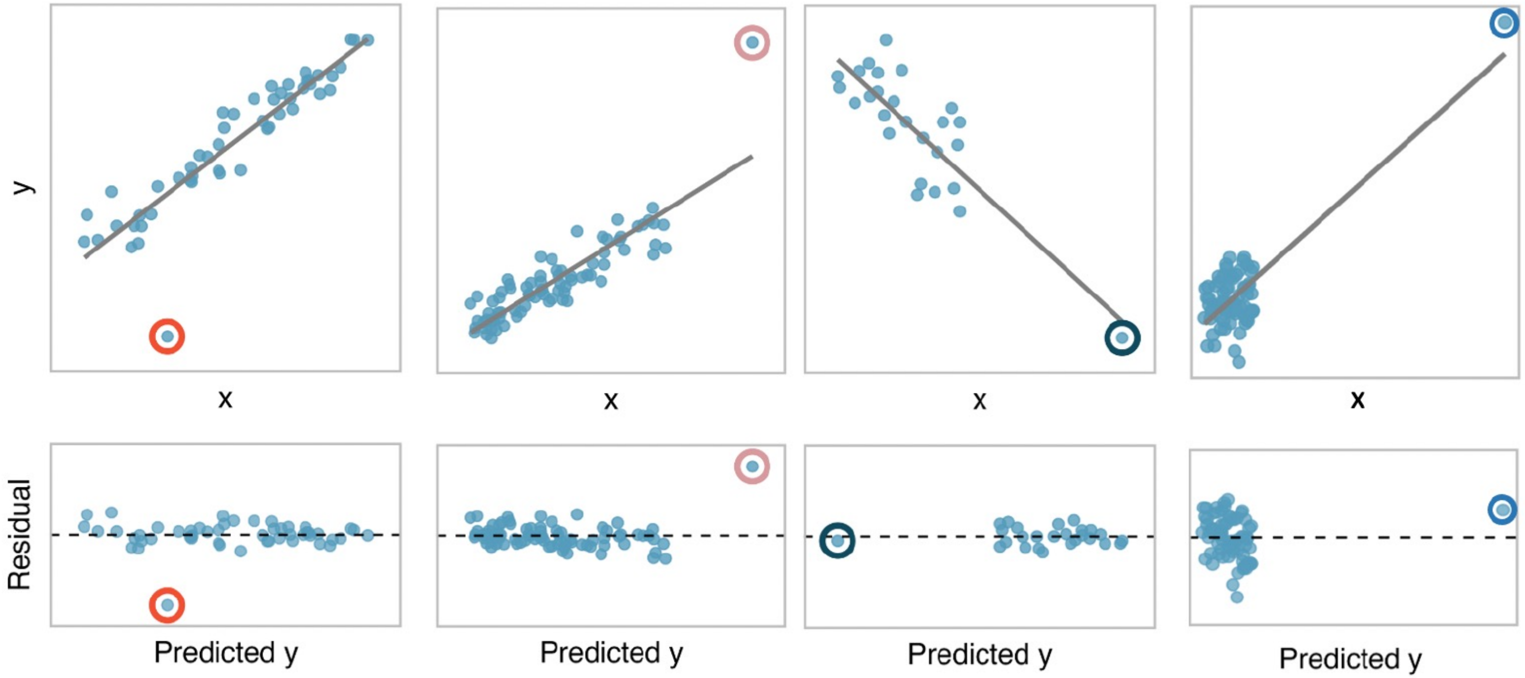
Should we be worried?

Outliers and Influential Points

Outliers are observations that fall far from the majority of data points. *They can have a strong influence on the least squares line!*

Leverage: Outliers that fall horizontally away from the center of the cloud of data points are called leverage points.

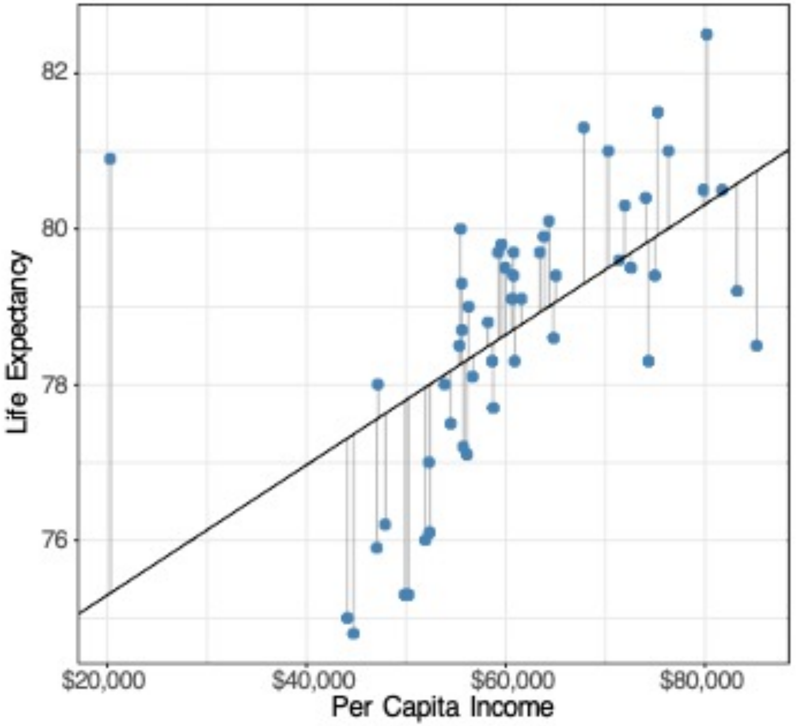
Influential points: Leverage points that influence the slope of the line.



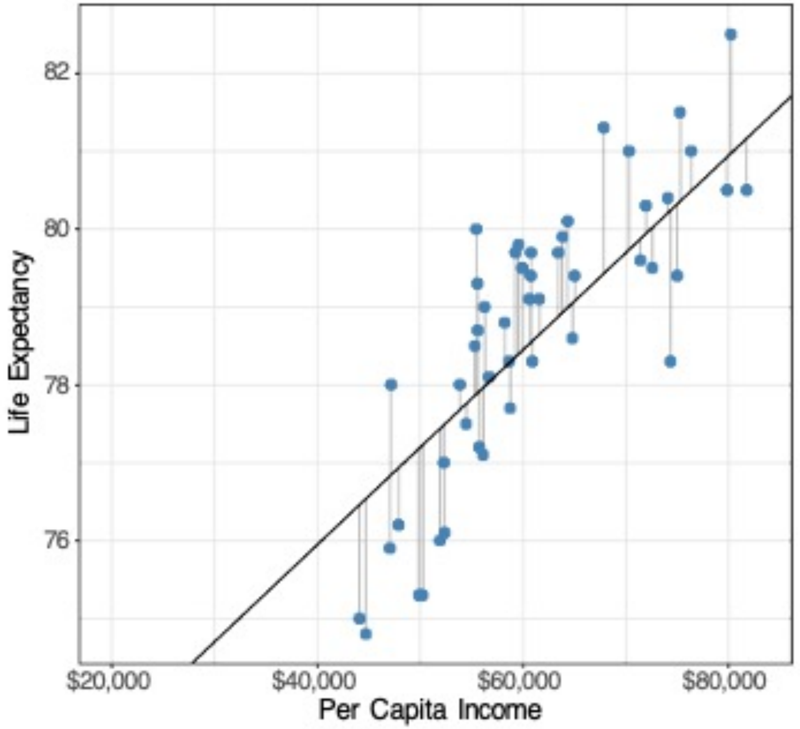
Outliers and Influential Points

Should we exclude Puerto Rico from our analysis? Why or why not? What would you want to check first?

Model Fit With Puerto Rico



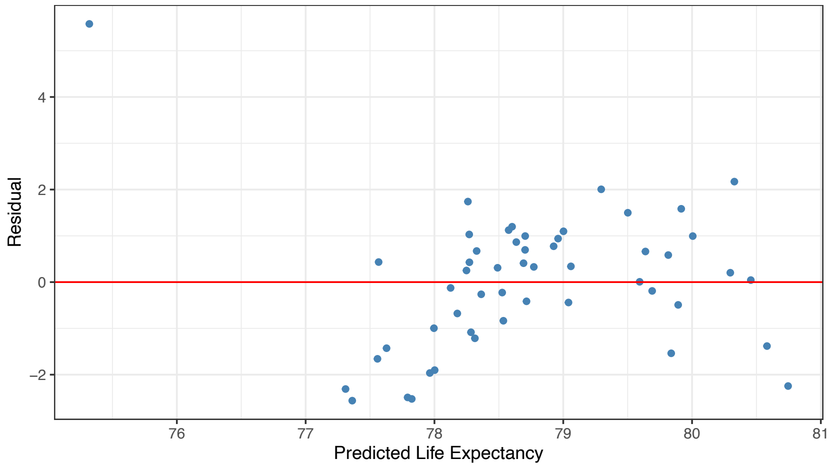
Model Fit Without Puerto Rico



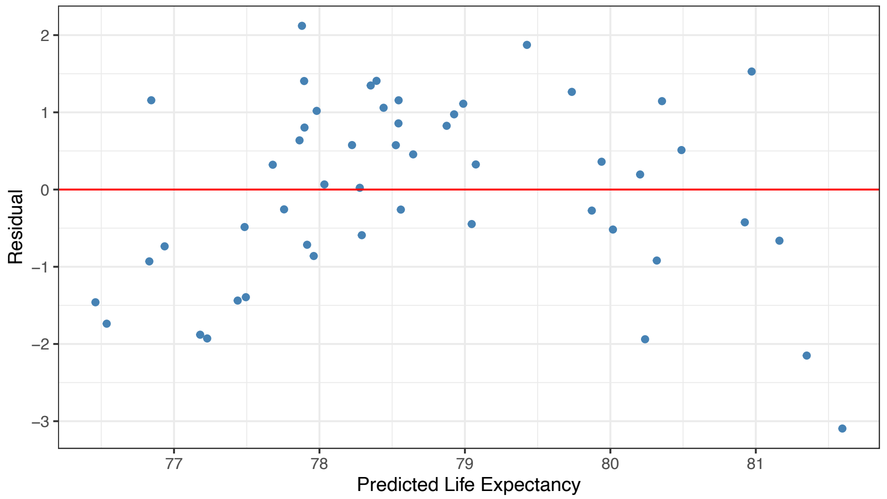
Outliers and Influential Points

Should we exclude Puerto Rico from our analysis? Why or why not? What would you want to check first?

Residual Plot With Puerto Rico



Residual Plot Without Puerto Rico



With Puerto Rico

$$R^2 = 0.32$$

Without Puerto Rico

$$R^2 = 0.56$$

What are the implications of removing Puerto Rico on our research question?

Binary Predictors

What if we have a binary predictor, instead of a continuous one?

Ex. Instead of looking at how income effects life expectancy, we want to understand...

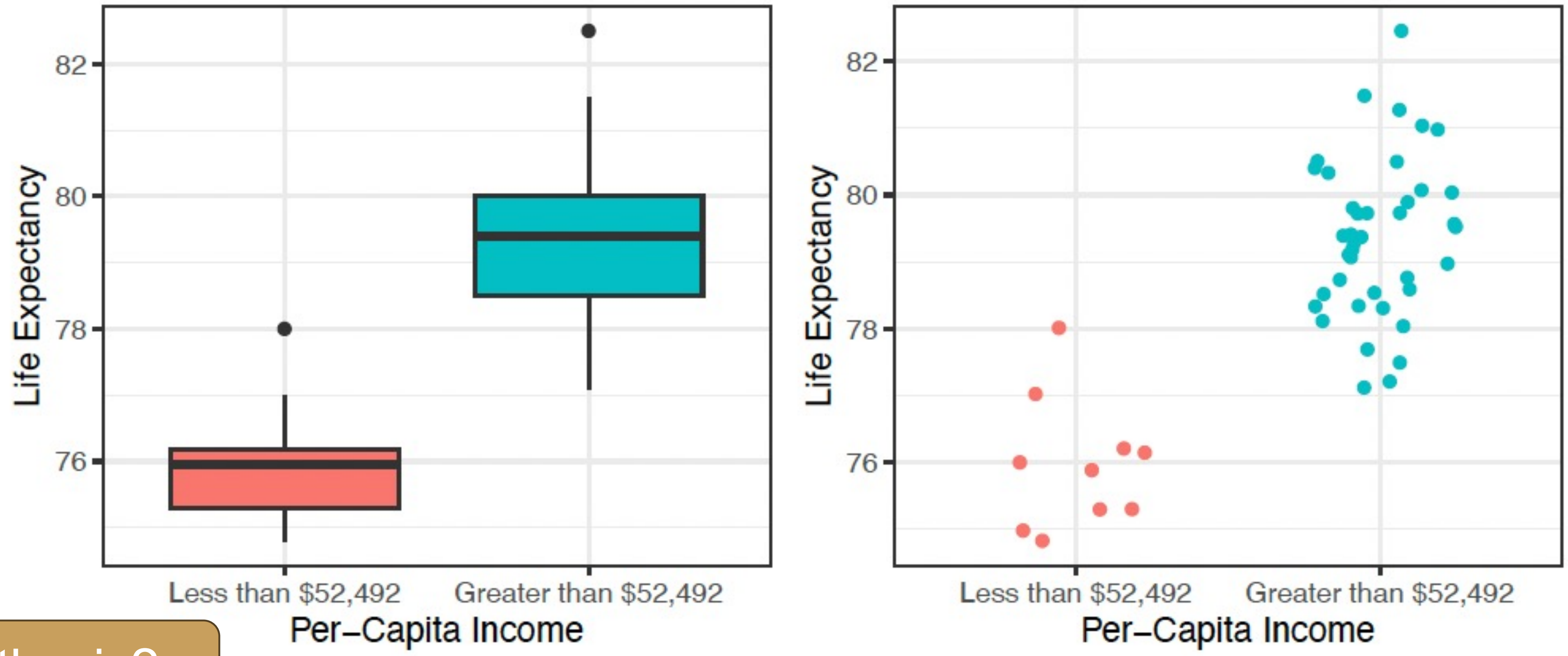
Motivating Question

How does the life expectancy in a US state associate with whether or not its average per-capita income is below the poverty threshold?

Binary Predictors

Motivating Question

How does the life expectancy in a US state associate with whether or not its average per-capita income is below the poverty threshold?



What is your hypothesis?

Binary Predictors

What if we have a binary predictor, instead of a continuous one?

Ex. Instead of looking at how income effects life expectancy, we want to look at if above or below the poverty line effects life expectancy.

We do this by transforming our categorical variable into a numerical one with an indicator variable.

$$povertyLine_i = \begin{cases} 1 \ if \ \ state \ i \ has \ percapita \ income \ above \ povery \ line \\ 0 \ if \ \ state \ i \ has \ percapita \ income \ below \ povery \ line \end{cases}$$

Binary Predictors

What if we have a binary predictor, instead of a continuous one?

Ex. Instead of looking at how income effects life expectancy, we want to look at if above or below the poverty line effects life expectancy.

We do this by transforming our categorical variable into a numerical one with an indicator variable.

$$povertyLine_i = \begin{cases} 1\ if\ \ state\ i\ has\ percapita\ income\ above\ povery\ line \\ 0\ if\ \ state\ i\ has\ percapita\ income\ below\ povery\ line \end{cases}$$

The value for which povertyLine is 0 is called the baseline

Binary Predictors

$$povertyLine_i = \begin{cases} 1\ if\ \ state\ i\ has\ percapita\ income\ above\ povery\ line \\ 0\ if\ \ state\ i\ has\ percapita\ income\ below\ povery\ line \end{cases}$$

Now, our model is

$$lifeExp = \ \beta_0 + \ \beta_1 povertyLine$$

Binary Predictors

$$povertyLine_i = \begin{cases} 1\ if\ \ state\ i\ has\ percapita\ income\ above\ povery\ line \\ 0\ if\ \ state\ i\ has\ percapita\ income\ below\ povery\ line \end{cases}$$

Now, our model is

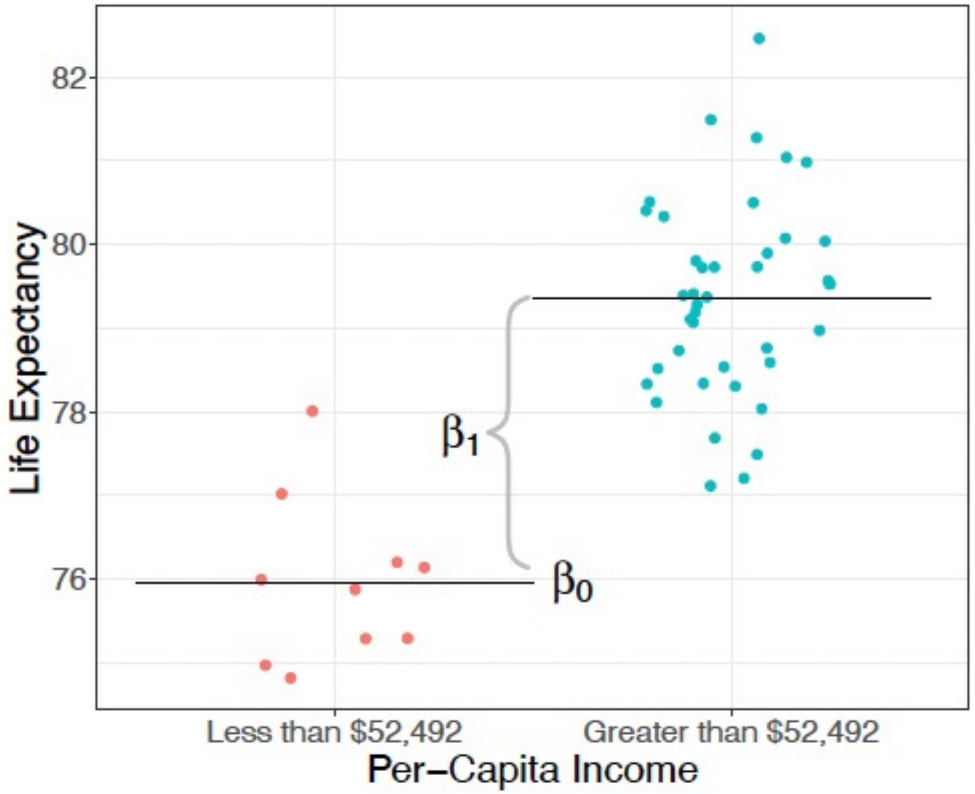
$$lifeExp = \beta_0 + \beta_1 povertyLine$$

What do β_0 and β_1 represent in the context of this model?

Binary Predictors

$$povertyLine_i = \begin{cases} 1\ if\ \ state\ i\ has\ percapita\ income\ above\ povery\ line \\ 0\ if\ \ state\ i\ has\ percapita\ income\ below\ povery\ line \end{cases}$$

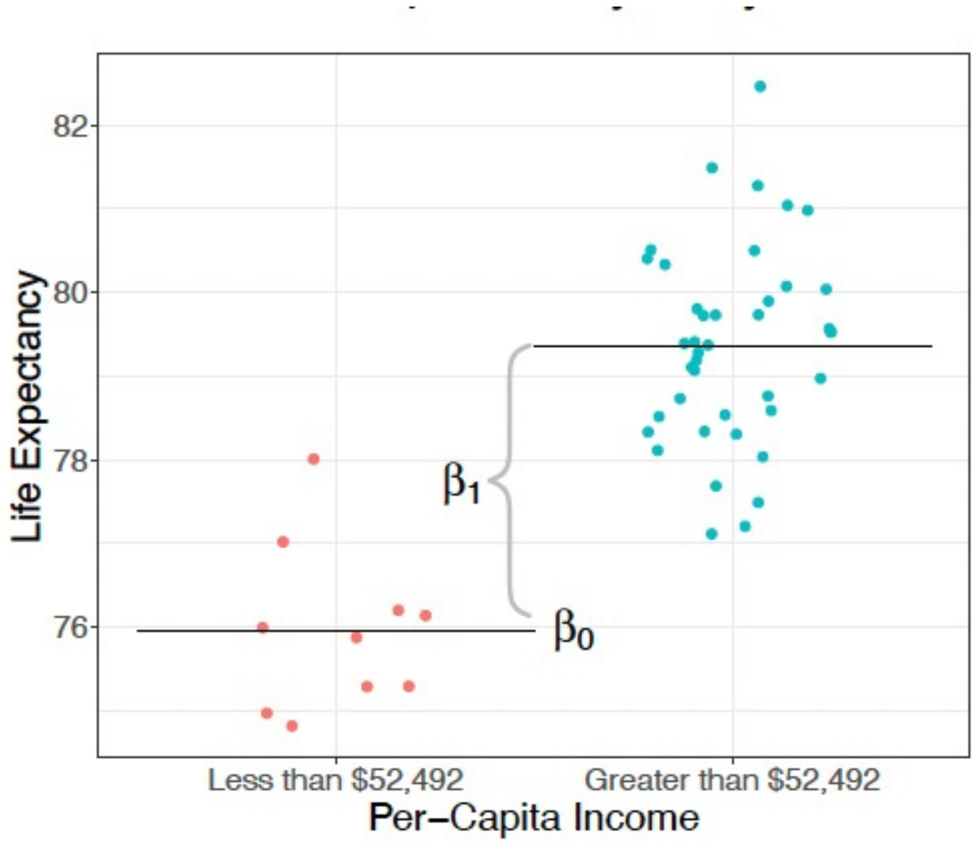
$$lifeExp = \ \beta_0 + \ \beta_1 povertyLine$$



Binary Predictors

$$povertyLine_i = \begin{cases} 1\ if\ state\ i\ has\ percapita\ income\ above\ povery\ line \\ 0\ if\ state\ i\ has\ percapita\ income\ below\ povery\ line \end{cases}$$

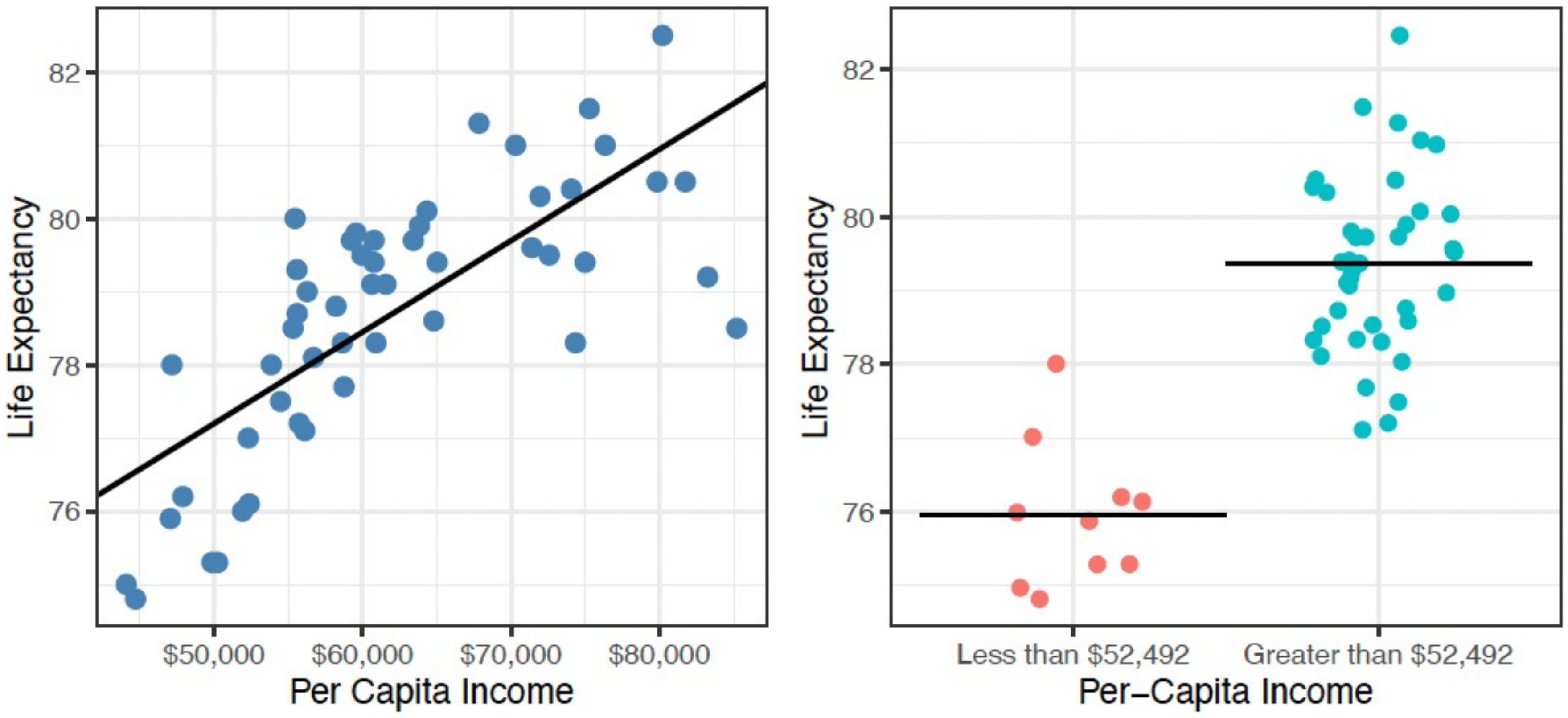
$$\widehat{lifeExp_i} = 7.60 + 3.41povertyLine_i$$



What is the predicted life expectancy for a state with percapita income above the poverty line? How about below? What is the difference between these estimates?

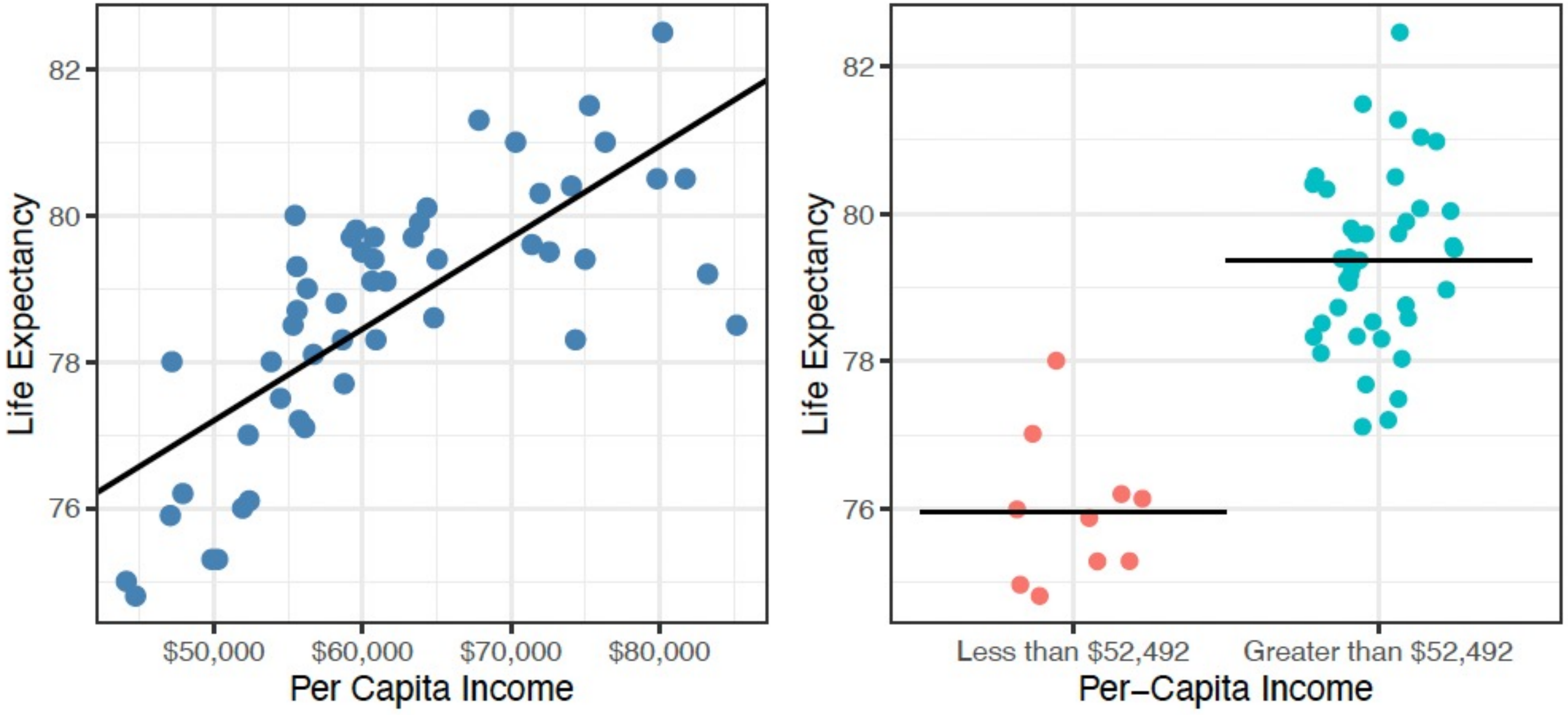
Comparison of Models

We made two models for the same phenomenon. Which is better?



Comparison of Models

We made two models for the same phenomenon. Which is better?



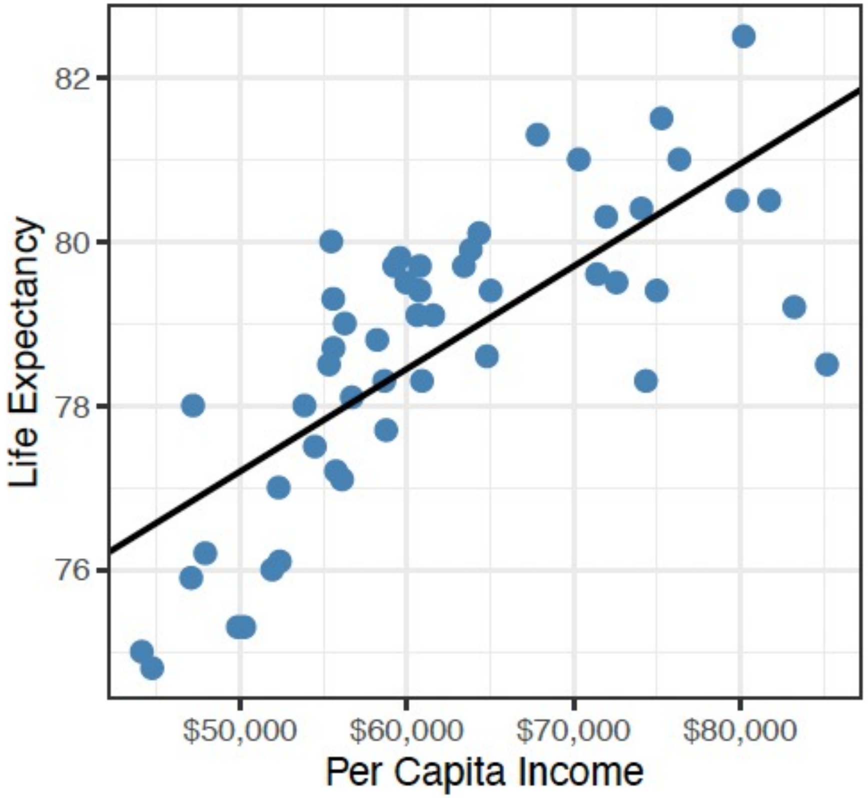
$R^2 = 0.557$

$R^2 = 0.592$

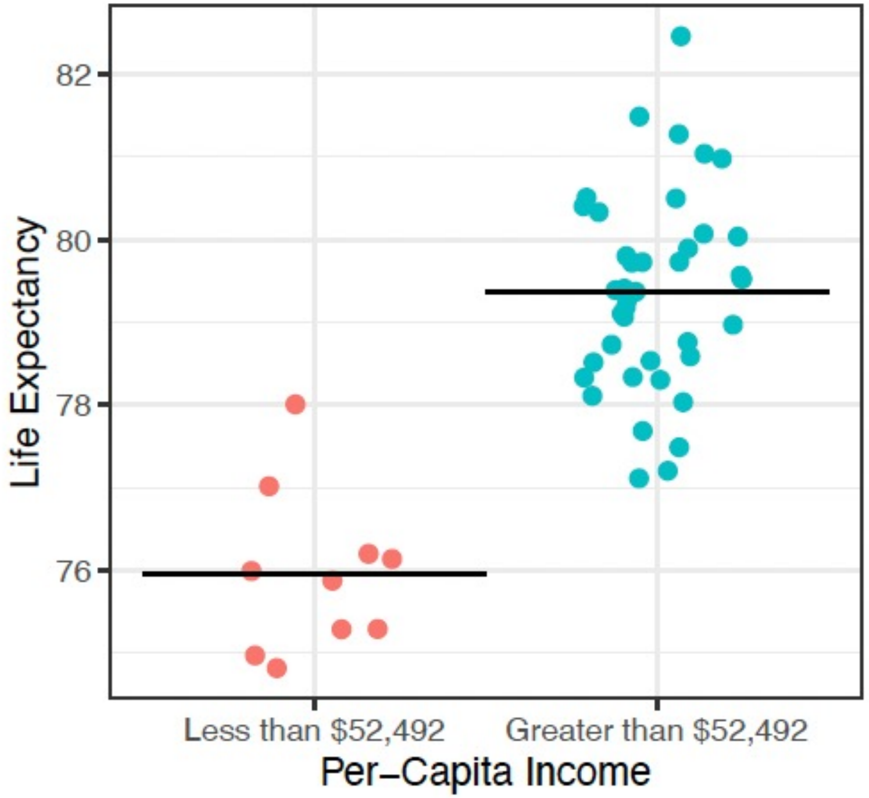
Which model explains more of the variance in life expectancy? Is the difference enough to matter?

Comparison of Models

We made two models for the same phenomenon. Which is better?



$R^2 = 0.557$



$R^2 = 0.592$

Is a smaller R^2 always better?