Regression

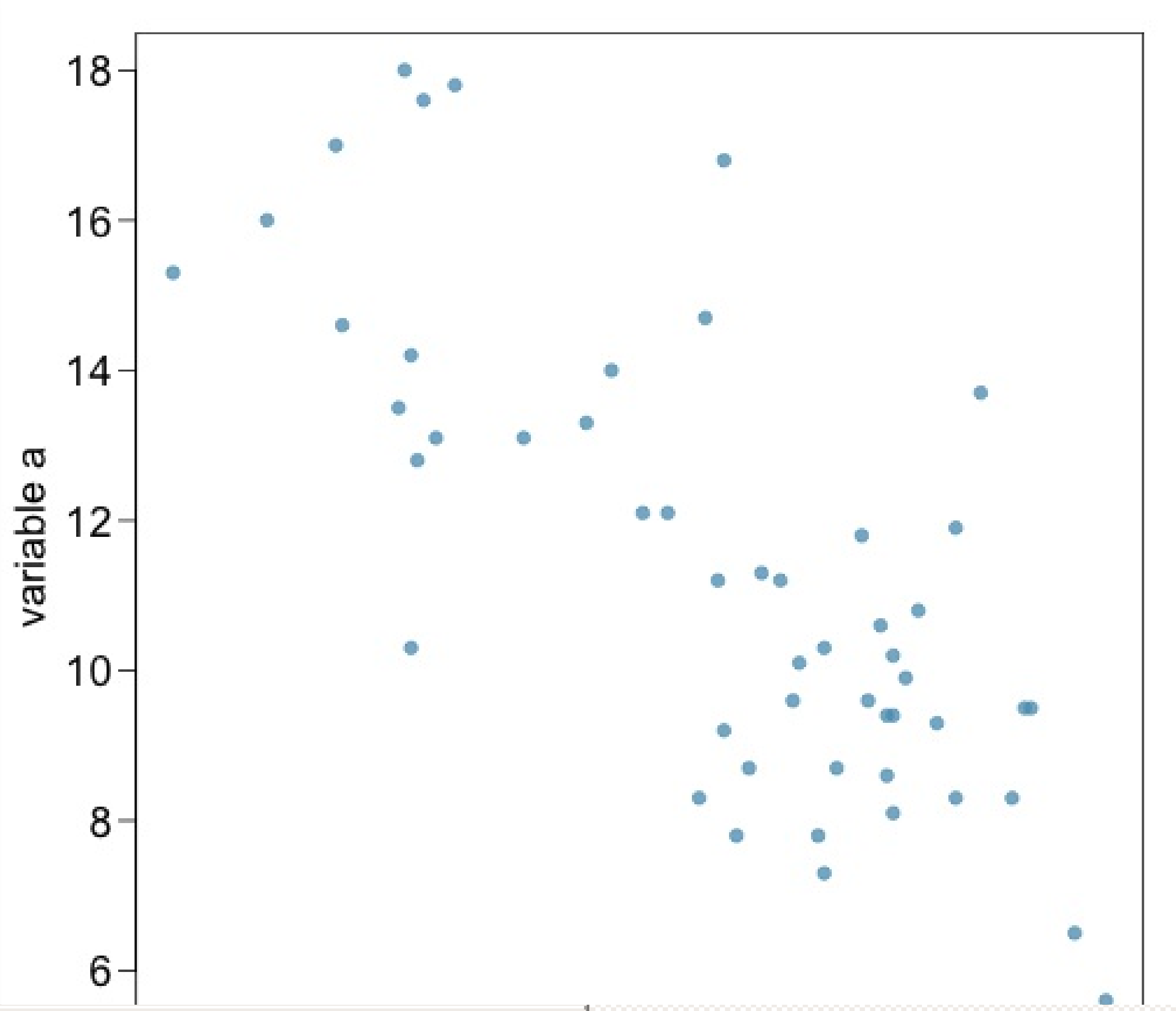
IS 665 Data Mining, Data Warehousing and Visualization

Agenda

- Linear Regression
  - Introduction
  - Prediction
  - R^2
  - Inference
- Multiple Regression

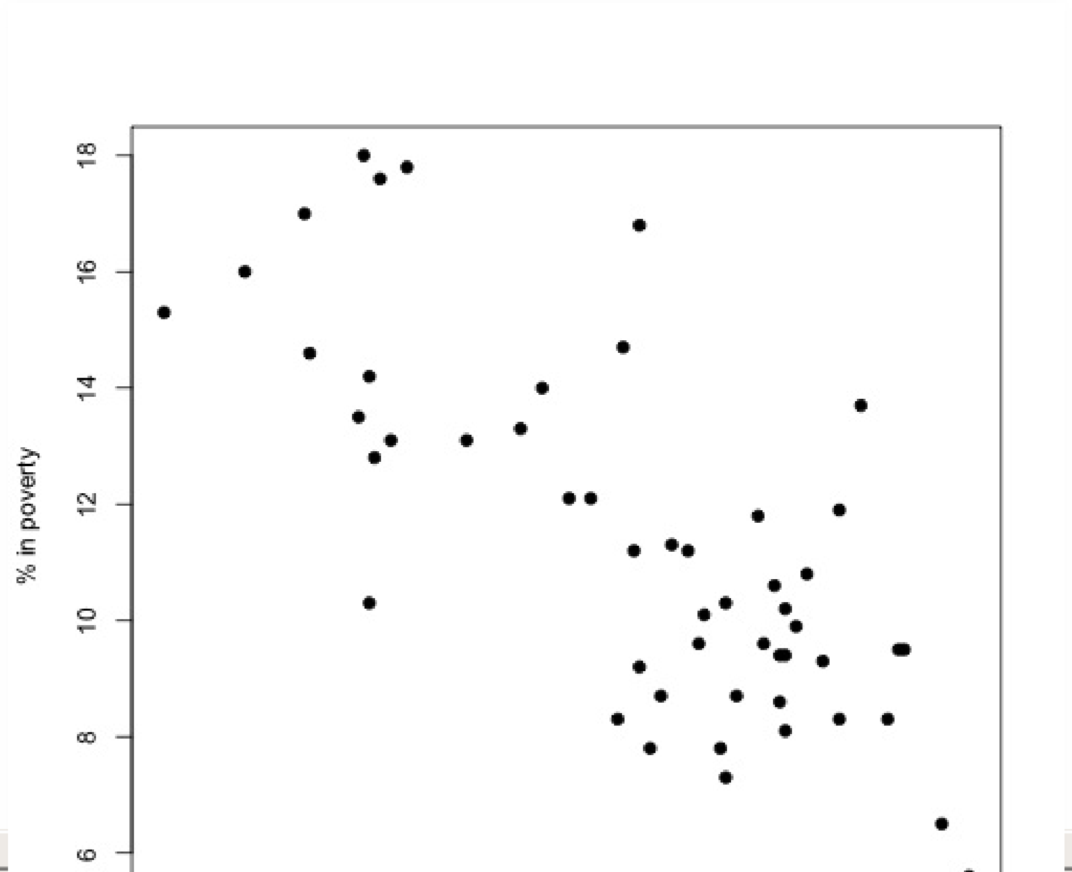
Modeling Numberical Values

Linear Regression is about quantifying the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable.



Poverty vs. High School Graduation

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

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.

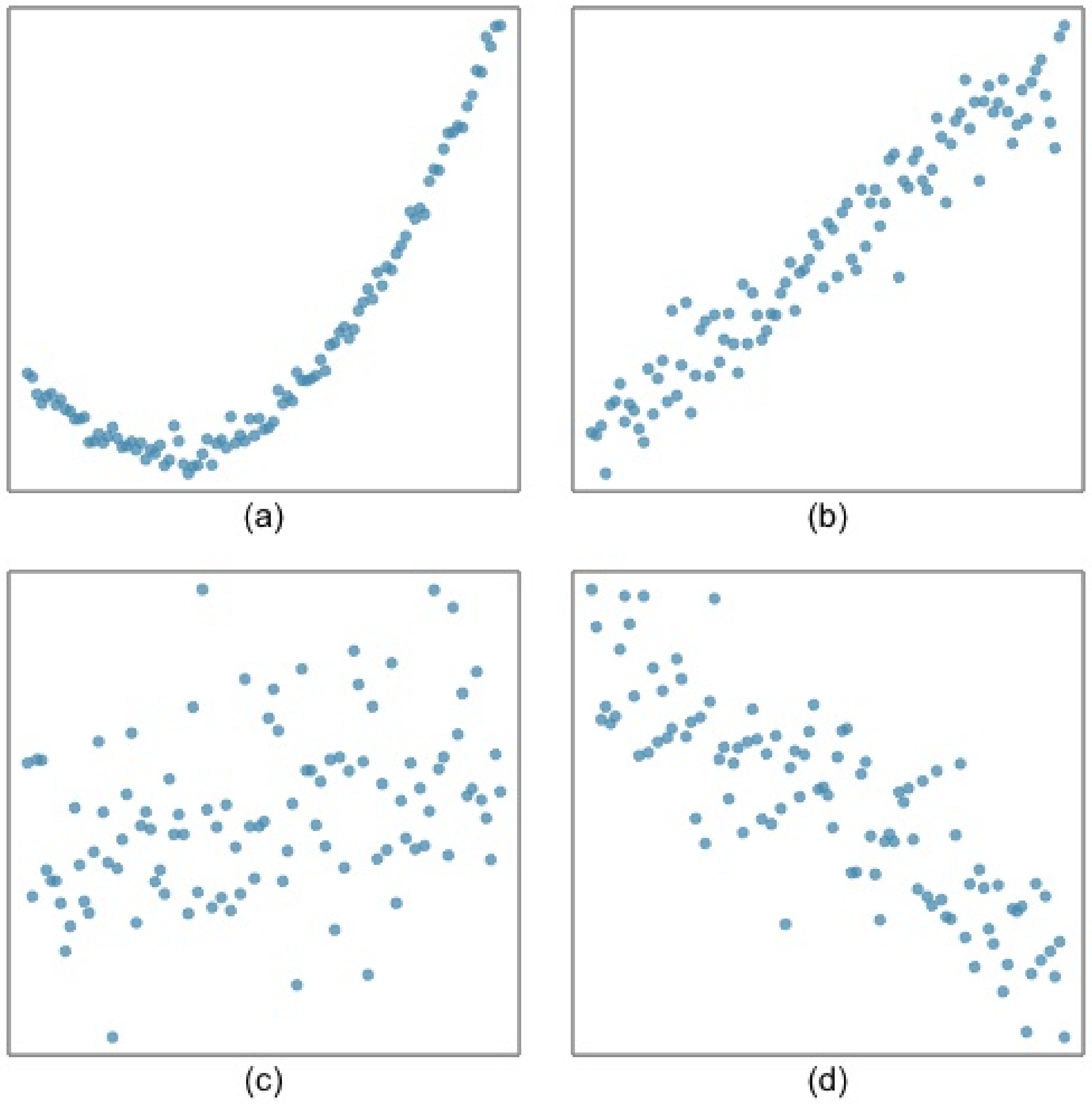
```
cor(poverty$Poverty , poverty$Graduates)
cov(poverty$Poverty , poverty$Graduates)

r=cov(poverty$Poverty , poverty$Graduates)
   /(sd(poverty$Poverty)*sd(poverty$Graduates))
```

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$Cov(X,Y) = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

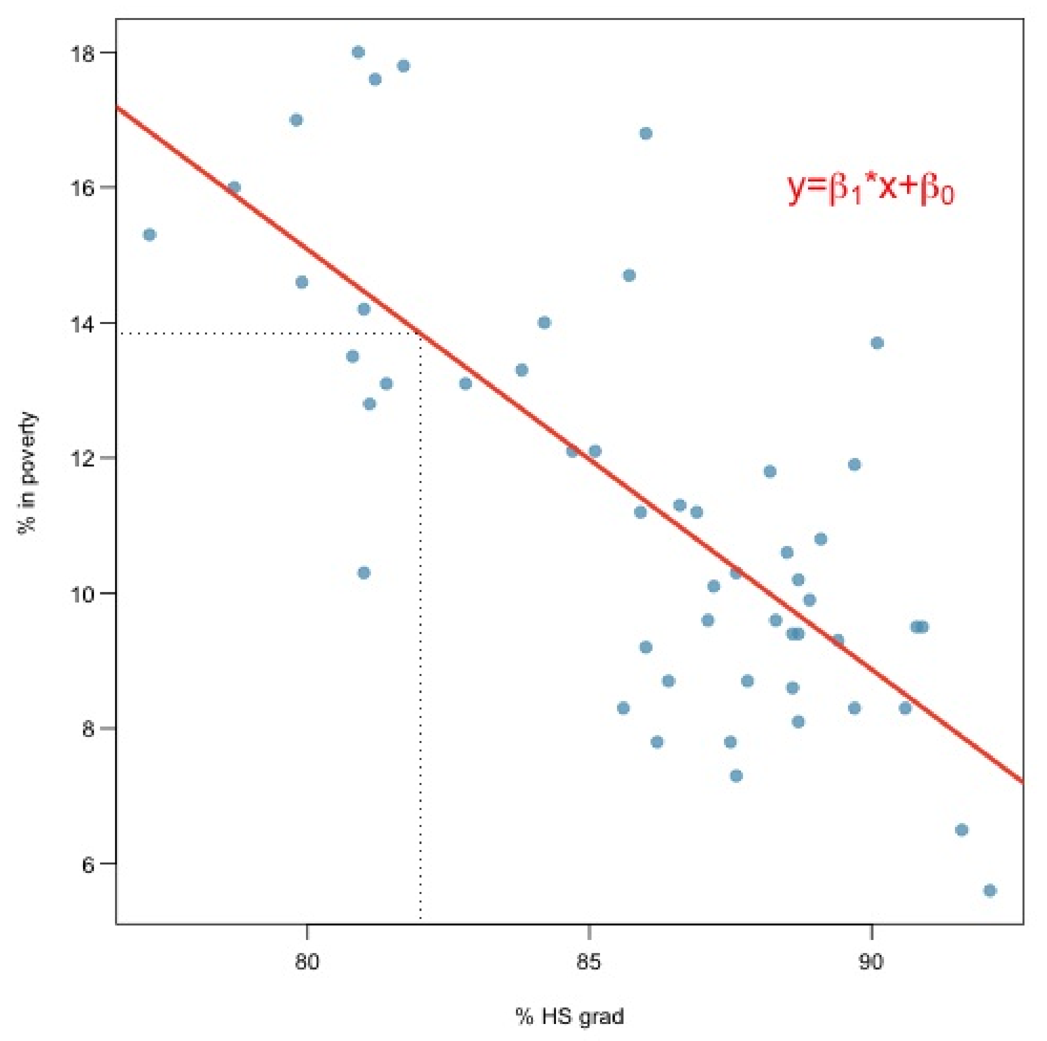
Assessing Correlation

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



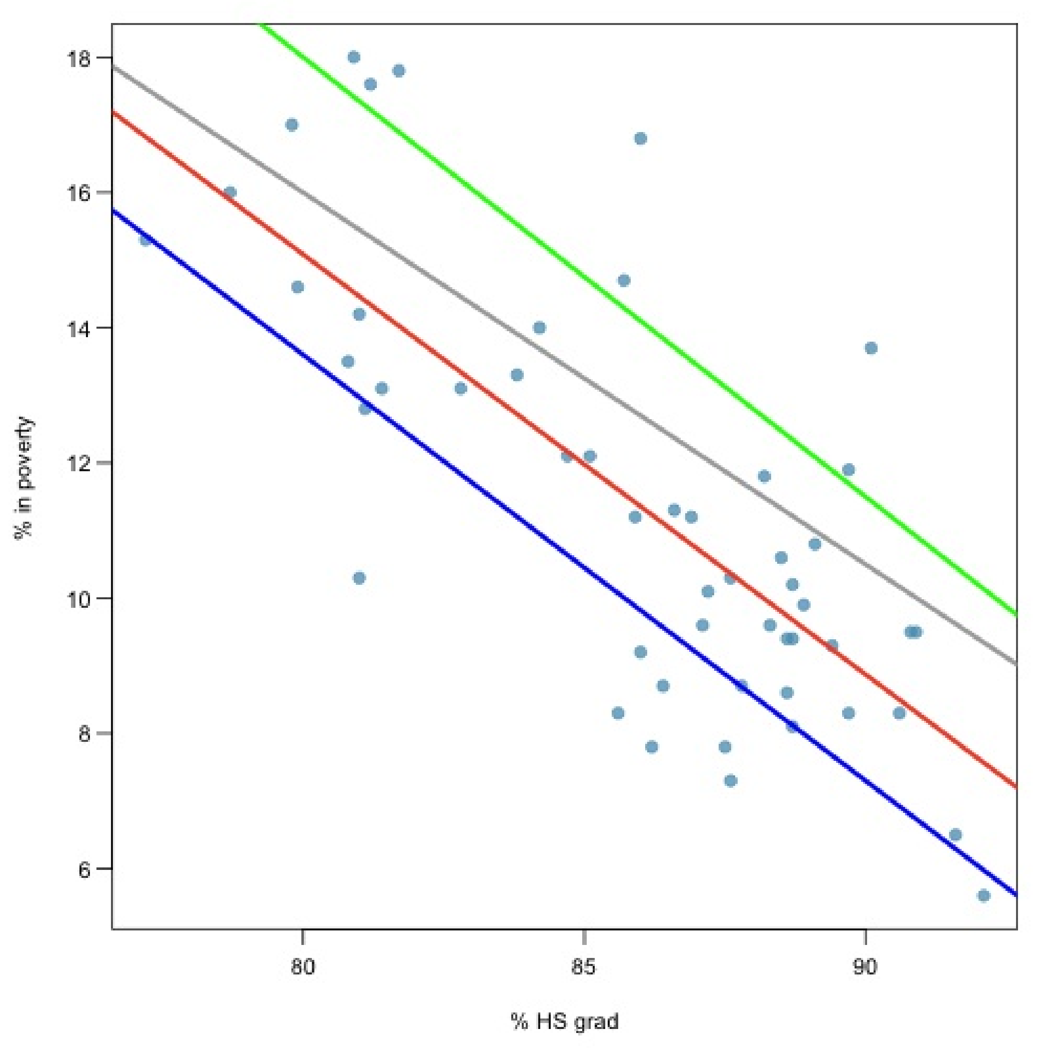
Linear Models

- In simple linear regression, the relationship between two variables is modeled using a straight line.
- Mathematically, it can be modeled as a linear function: $y = \beta_1 * X + \beta_0$ (this is also the function of the regression line)



Our goal

- Finding the line that describes the relationship the best



- gray: y= 60 - 0.55 * X
- blue: y= 64 - 0.63 * X
- green: y= 70 - 0.65 * X

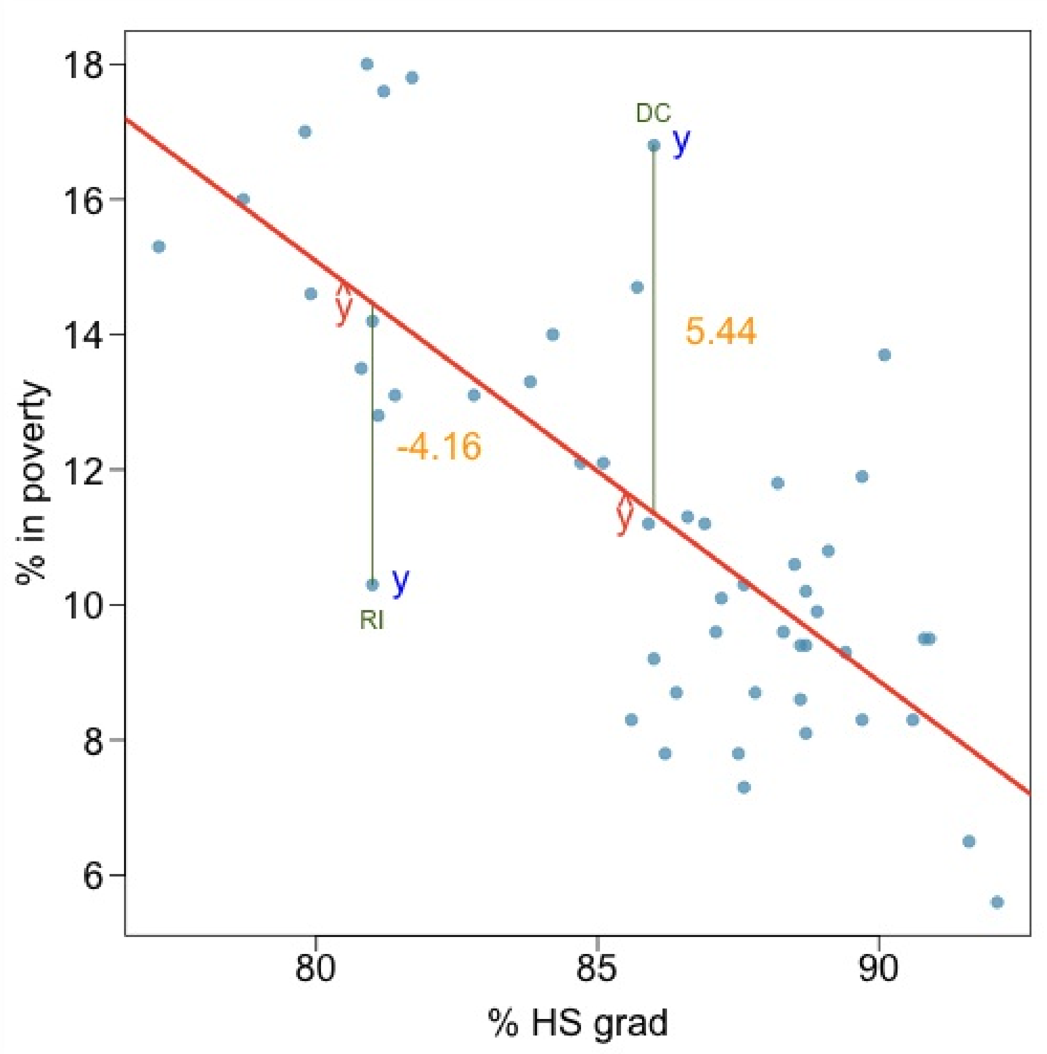
Residuals

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

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

Think of residuals as the leftovers from the model fit:

Data = Fit + Residual



e.x % living in poverty in DC is 5.44 % more than predicted.

A measure for the best line

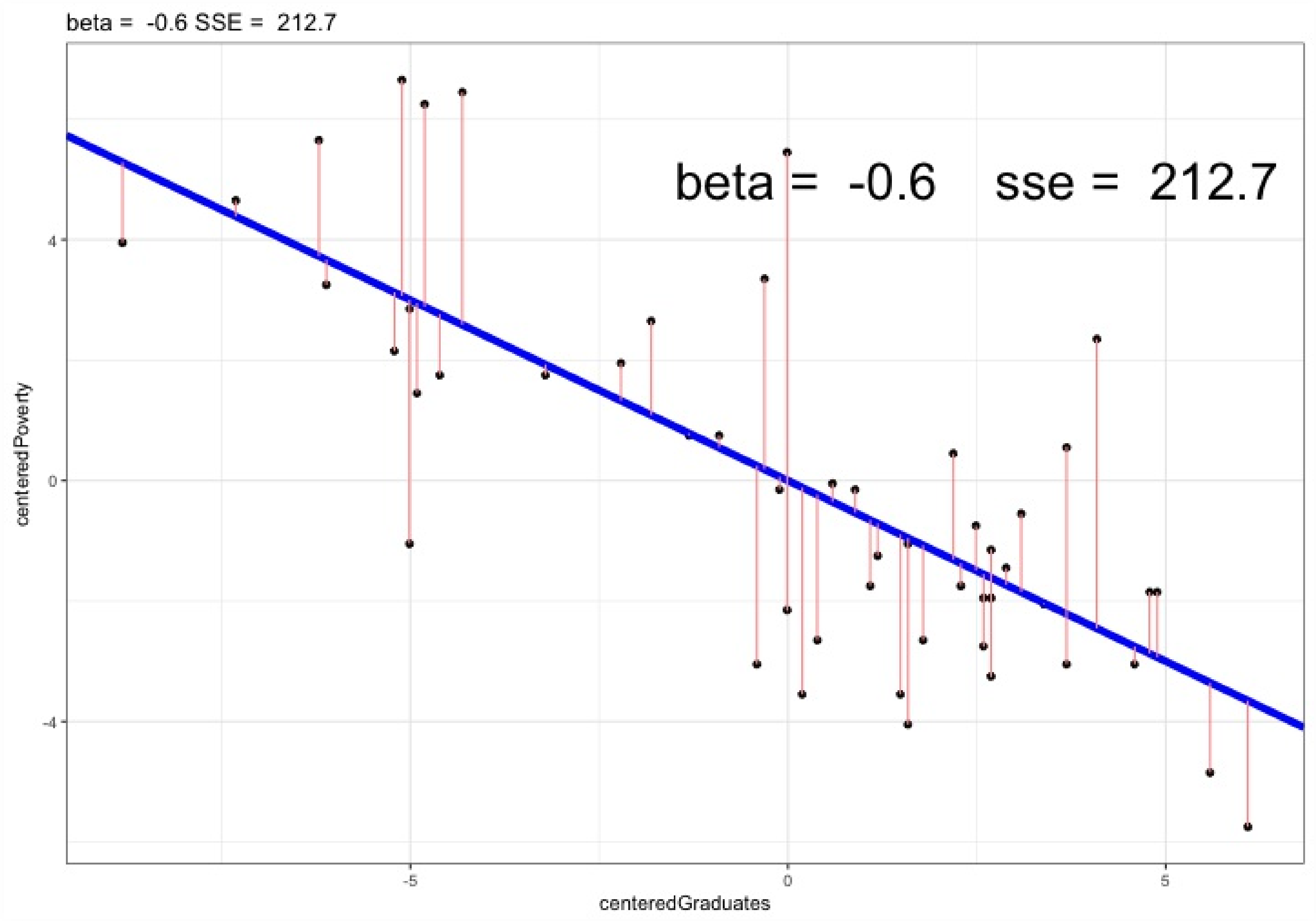
We want a line that has small residuals:

- Option 1: Minimize the sum of magnitudes (absolute values) of residuals
$$|e_1| + |e_2| + \cdots + |e_n|$$
- Option 2: Minimize the sum of squared residuals: Sum of Squared Errors (SSE)
$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Why least squares (SSE)?

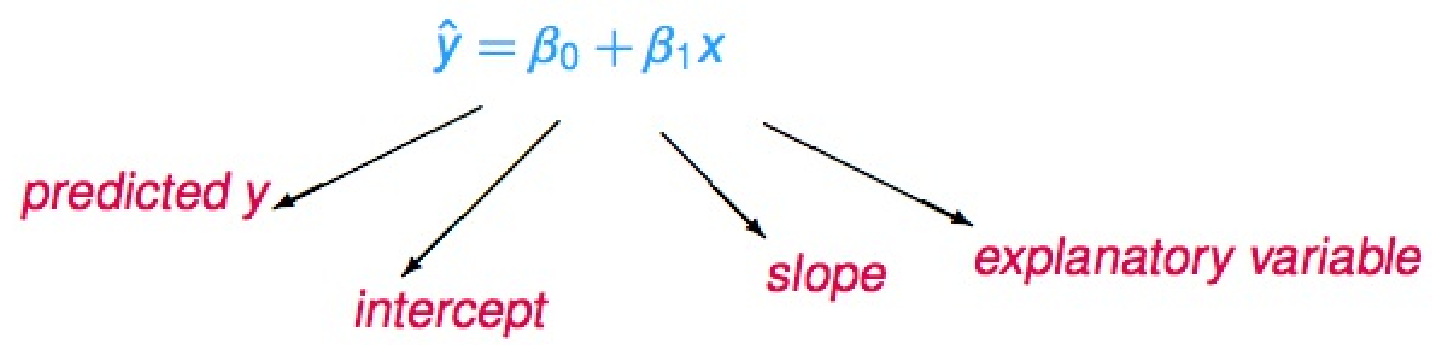
- Most commonly used
- Easier to compute by hand and using software
- In many applications, a residual twice as large as another is usually more than twice as bad

A measure of best line



The least square line

The mathematical model of the relationship (also the function for the best line)



Notation

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

Finding the Least Squares Line

The line is:

$$y = b_0 + b_1 x$$

We need to find b_0 and b_1 by looking at the data. (aka by fitting a linear model to the data)

```
my_lm = lm(Poverty ~ Graduates, data = poverty)
names(my_lm)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

```
my_lm$coefficients
```

```
(Intercept)   Graduates
 64.7809658  -0.6212167
```

Model coefficients

```
my_lm$coefficients
```

```
(Intercept)   Graduates
 64.7809658  -0.6212167
```

Slope 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.

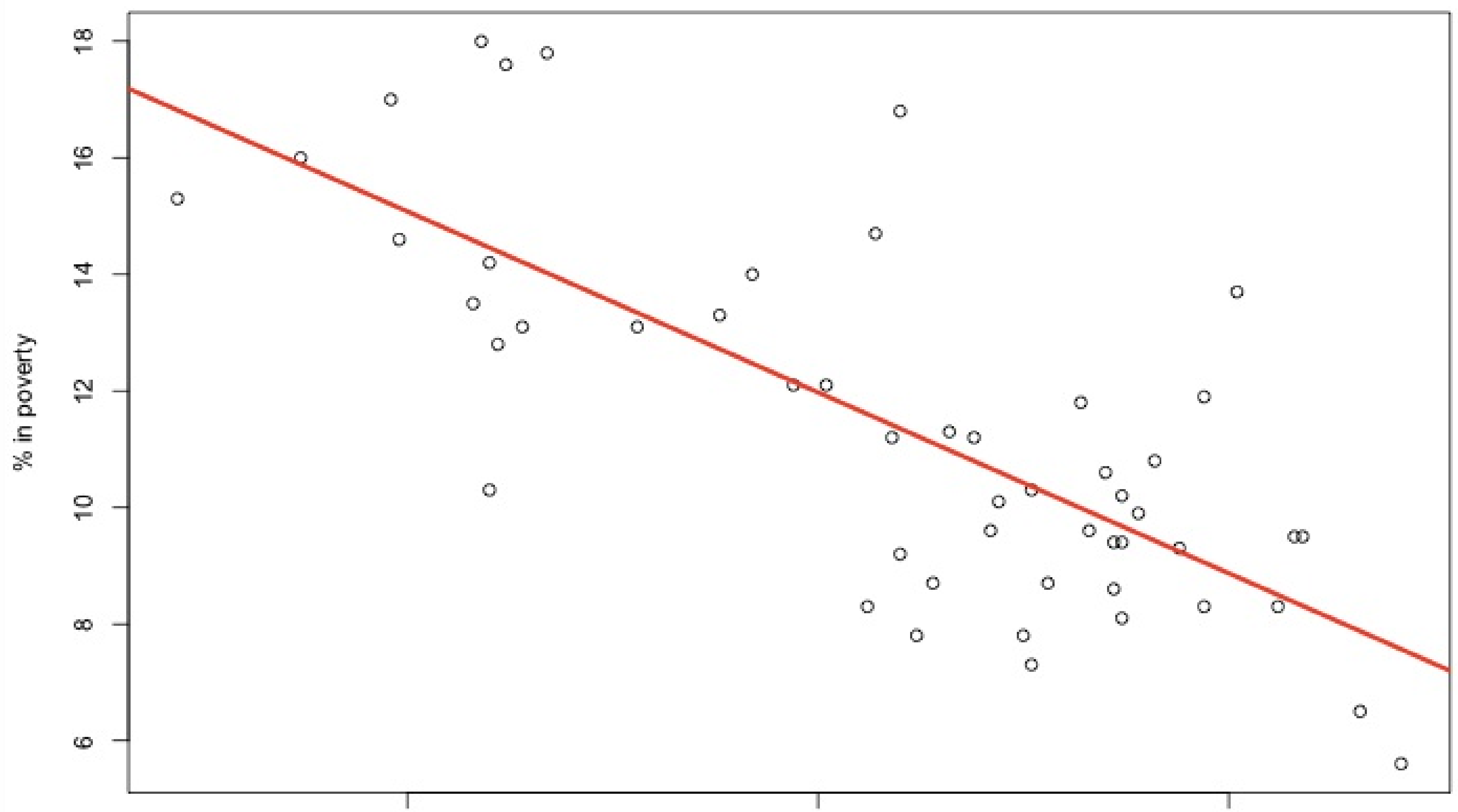
Intercept

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

Regression Line

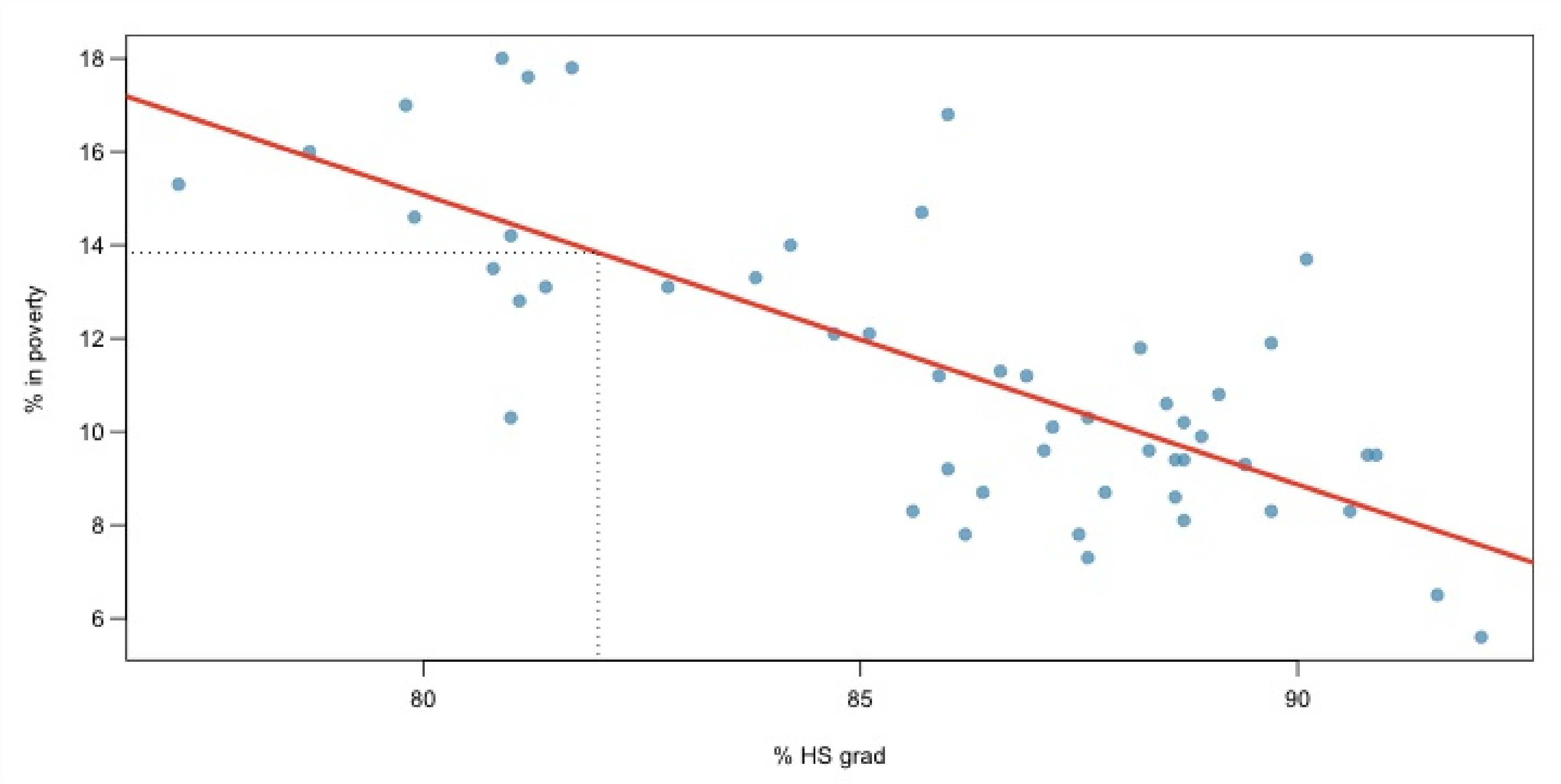
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$

```
plot(Poverty ~ Graduates, data = poverty, ylab = "% in poverty",
   xlab = "% HS grad")
lm_pov_grad = lm(poverty$Poverty ~ poverty$Graduates)
abline(lm_pov_grad, col = COL[4], lwd = 3)
```



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.



Prediction

```
my_prediction = function(x, my_model) {
    pr = my_model$coefficients[1] + x * my_model$coefficients[2]
    return(as.numeric(pr))
}

my_prediction(84, lm_pov_grad)
```

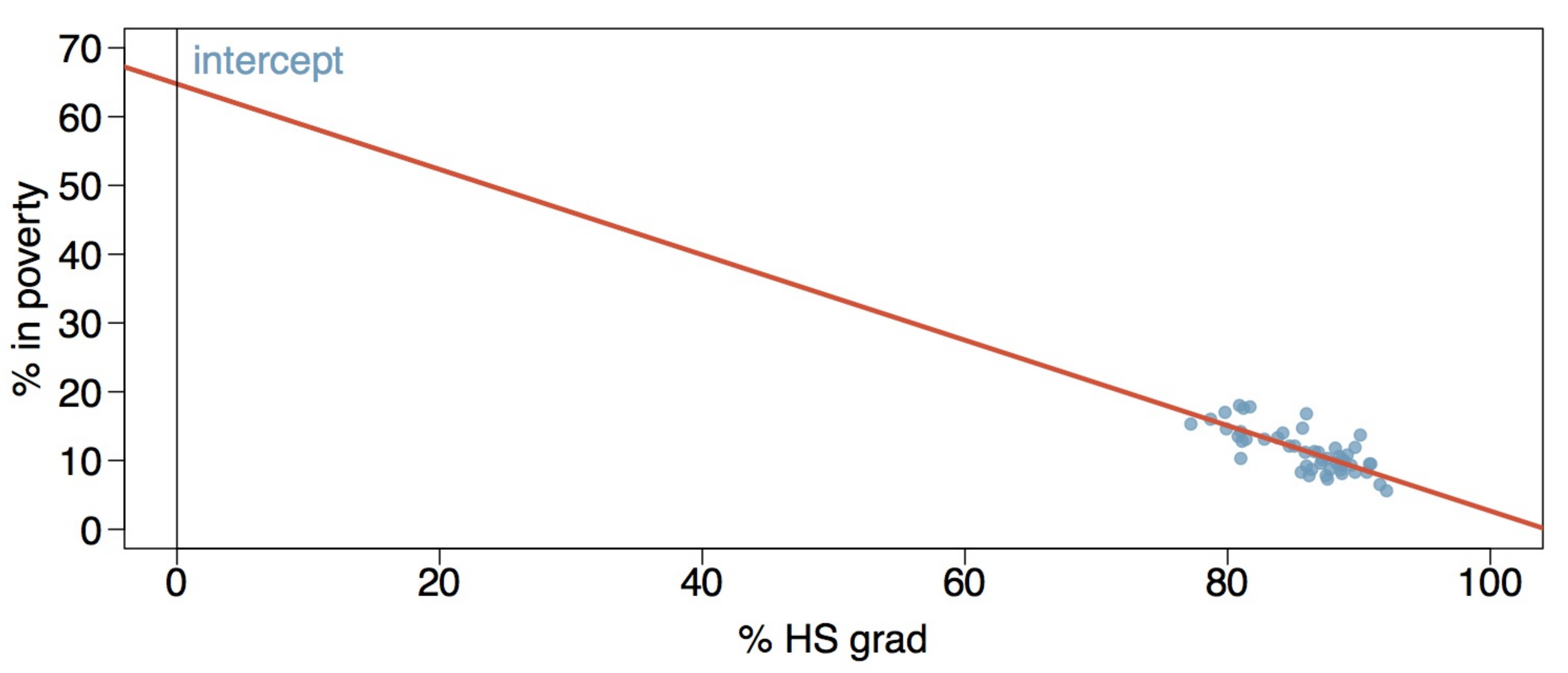
```
[1] 12.59876
```

```
newdata = data.frame(Graduates = c(84, 80, 75))
predict(lm_pov_grad, newdata)
```

```
       1        2        3
12.59876 15.08363 18.18971
```

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.



R^2 :R-Squared

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

Inference

$$\widehat{\%\ in\ poverty} = \beta_0 - \beta_1 * \ \%\ HS\ grad$$

Does the data provide convincing evidence that graduation rate is a significant predictor of poverty? What are the appropriate hypotheses?

- What would β_1 be if graduation is NOT a predictor of poverty

Inference

$$\widehat{\%\ in\ poverty} = \beta_0 - \beta_1 * \%\ HS\ grad$$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

```
my_lm=lm(Poverty~Graduates, data=poverty)
summary(my_lm)
```

low p value means significant results

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***

Residual standard error: 2.082 on 49 degrees of freedom
Multiple R-squared:  0.5578,  Adjusted R-squared:  0.5488
F-statistic: 61.81 on 1 and 49 DF,  p-value: 3.109e-10
```