Chapter 15: Linear Regression I

DSCC 462
Computational Introduction to Statistics

Anson Kahng
Fall 2022

Announcements

Announcements

- Midterm grades are out

Announcements

- Midterm grades are out
- Pickup tomorrow from 3-4 pm in my office (Wegmans 2401) or at office hours next Tuesday

Announcements

- Midterm grades are out
- Pickup tomorrow from 3-4 pm in my office (Wegmans 2401) or at office hours next Tuesday
- Project groups are due tomorrow! Datasets will be released tomorrow (no project proposal necessary)

Plan For Today

Plan For Today

- Explain how one variable affects another: simple linear regression!

Plan For Today

- Explain how one variable affects another: simple linear regression!
- Basics of regression

Plan For Today

- Explain how one variable affects another: simple linear regression!
- Basics of regression
- Inference for parameters

Plan For Today

- Explain how one variable affects another: simple linear regression!
- Basics of regression
- Inference for parameters
- Confidence intervals for true values

Simple Linear Regression

Simple Linear Regression

- Simple linear regression allows us to explore the relationship between two continuous random variables (think scatterplot)

Simple Linear Regression

- Simple linear regression allows us to explore the relationship between two continuous random variables (think scatterplot)
- Unlike correlation analyses, we can *directly* model how a change in one variable affects another variable

Simple Linear Regression

- Simple linear regression allows us to explore the relationship between two continuous random variables (think scatterplot)

- Unlike correlation analyses, we can *directly* model how a change in one variable affects another variable

  - *Explanatory variable* affects the *response variable*

X_i Y_i

Simple Linear Regression

- Simple linear regression allows us to explore the relationship between two continuous random variables (think scatterplot)
- Unlike correlation analyses, we can *directly* model how a change in one variable affects another variable
    - *Explanatory variable* affects the *response variable*
- Goal: Estimate the value of the response variable that is associated with a given value of the explanatory variable

Simple Linear Regression

- Simple linear regression allows us to explore the relationship between two continuous random variables (think scatterplot)
- Unlike correlation analyses, we can *directly* model how a change in one variable affects another variable
  - *Explanatory variable* affects the *response variable*
- Goal: Estimate the value of the response variable that is associated with a given value of the explanatory variable
  - Example: If a child is 9 years old, how tall do we expect them to be?

Simple Linear Regression

Simple Linear Regression

- In *linear* regression, we estimate the relationship between x (explanatory variable) and y (response variable) by a line

Simple Linear Regression

- In *linear* regression, we estimate the relationship between x (explanatory variable) and y (response variable) by a line

Var

- Regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1,\ldots,n$ and $\epsilon_i \sim N(0,\sigma^2)$ $\quad SD = \sqrt{\sigma^2} = \sigma$

Simple Linear Regression

- In *linear* regression, we estimate the relationship between x (explanatory variable) and y (response variable) by a line
- Regression model: $y_i = \boxed{\beta_0 + \beta_1 x_i} + \epsilon_i$ for $i = 1,\ldots,n$ and $\epsilon_i \sim N(0,\sigma^2)$
  - Expressed another way: $y_i | x_i \sim N(\underline{\beta_0 + \beta_1} x_i, \boxed{\sigma^2})$

Simple Linear Regression

- In *linear* regression, we estimate the relationship between x (explanatory variable) and y (response variable) by a line

- Regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, \ldots, n$ and $\epsilon_i \sim N(0, \sigma^2)$
    - Expressed another way: $y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$
    - β_0 is the y-intercept and β_1 is the slope for the population

$y = \overset{\beta_1}{m}x + \overset{\beta_0}{b}$

Simple Linear Regression

- In *linear* regression, we estimate the relationship between x (explanatory variable) and y (response variable) by a line
- Regression model: $y_i = \boxed{\beta_0} + \boxed{\beta_1} x_i + \epsilon_i$ for $i = 1,\ldots,n$ and $\epsilon_i \sim N(0,\sigma^2)$
  - Expressed another way: $y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$
  - β_0 is the y-intercept and β_1 is the slope for the population
- **Goal**: Estimate β_0 and β_1 based on a sample in order to model the relationship between y and x

Simple Linear Regression

Simple Linear Regression

- Assumptions:

Simple Linear Regression

- Assumptions:
  - Given x, the y's are independent

Simple Linear Regression

- Assumptions:
  - Given x, the y's are independent
  - There is a linear relationship between y and x (i.e., $E(\epsilon) = 0$)

Simple Linear Regression

- Assumptions:

  $\epsilon \sim N(0, \sigma^2)$

  - Given x, the y's are independent
  - There is a linear relationship between y and x (i.e., $E(\epsilon) = 0$)
  - The variance σ^2 is constant across all values of x (i.e., $Var(\epsilon) = \sigma^2$), known as homoscedasticity

Simple Linear Regression

- Assumptions:
  - Given x, the y's are independent
  - There is a linear relationship between y and x (i.e., $E(\epsilon) = 0$)
  - The variance σ^2 is constant across all values of x (i.e., $Var(\epsilon) = \sigma^2$), known as homoscedasticity
  - For a specified value of x, y is normally distributed

Simple Linear Regression

- Assumptions:
  - Given x, the y's are independent
  - There is a linear relationship between y and x (i.e., $E(\epsilon) = 0$)
  - The variance σ^2 is constant across all values of x (i.e., $Var(\epsilon) = \sigma^2$), known as homoscedasticity
  - For a specified value of x, y is normally distributed
  - x are fixed, known quantities

Simple Linear Regression

- Assumptions:
  - Given x, the y's are independent
  - There is a linear relationship between y and x (i.e., $E(\epsilon) = 0$)
  - The variance σ^2 is constant across all values of x (i.e., $Var(\epsilon) = \sigma^2$), known as homoscedasticity
  - For a specified value of x, y is normally distributed
  - x are fixed, known quantities
- When the regression assumptions are met, the use of linear regression is appropriate for describing the relationship between y and x

Simple Linear Regression

Simple Linear Regression

- Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate what y_i would be for a given x_i, under the model

Simple Linear Regression

- Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate what y_i would be for a given x_i, under the model
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

predicted value of y_i

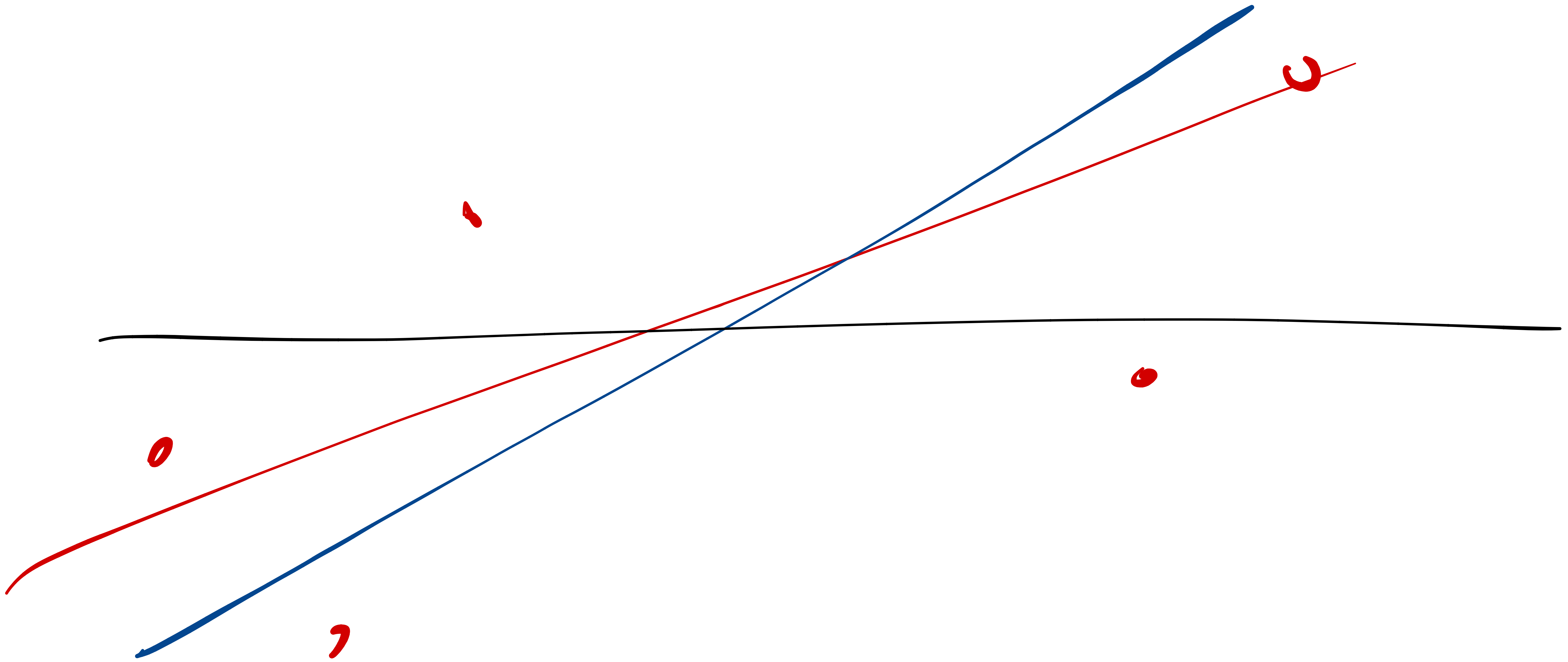
Simple Linear Regression

- Once we have estimates $\boxed{\hat{\beta}_0}$ and $\boxed{\hat{\beta}_1}$, we can estimate what y_i would be for a given x_i, under the model
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- But how do we fit a linear regression model?

Simple Linear Regression

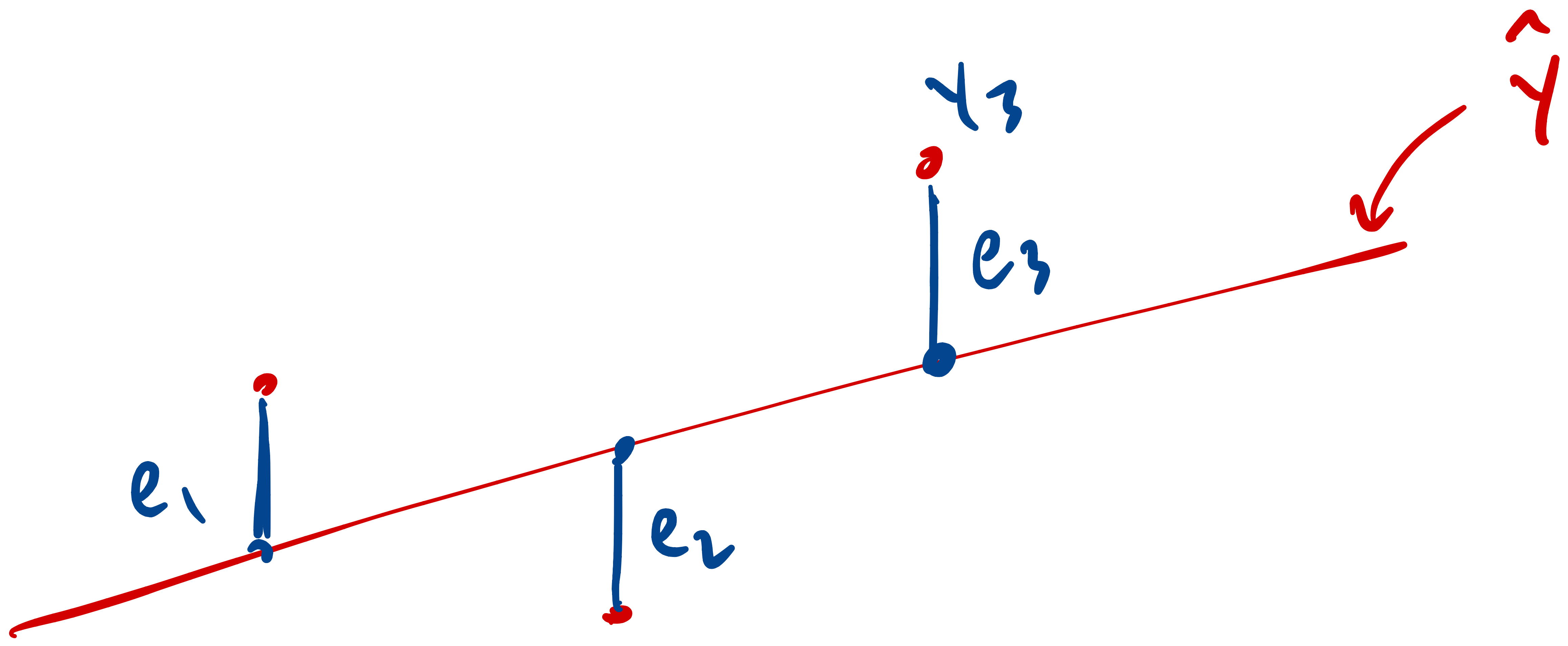
Simple Linear Regression

- Use the *method of least squares* to fit a straight line to a set of points (x_i, y_i)



Simple Linear Regression

- Use the *method of least squares* to fit a straight line to a set of points (x_i, y_i)
- If we look at how much each predicted $\hat{y}_i$ deviates from the true observed value y_i, we have the residual $e_i = y_i - \hat{y}_i$



Simple Linear Regression

- Use the *method of least squares* to fit a straight line to a set of points (x_i, y_i)
- If we look at how much each predicted $\hat{y}_i$ deviates from the true observed value y_i, we have the residual $e_i = y_i - \hat{y}_i$
  - R = A – P (residual = actual – predicted)

Simple Linear Regression

- Use the *method of least squares* to fit a straight line to a set of points (x_i, y_i)
- If we look at how much each predicted $\hat{y}_i$ deviates from the true observed value y_i, we have the residual $e_i = y_i - \hat{y}_i$
    - R = A – P (residual = actual – predicted)
- Residuals of $e_i = 0$ indicate that the observed point lies directly on the regression line

Simple Linear Regression

- Use the *method of least squares* to fit a straight line to a set of points (x_i, y_i)
- If we look at how much each predicted $\hat{y}_i$ deviates from the true observed value y_i, we have the residual $e_i = y_i - \hat{y}_i$
  - R = A – P (residual = actual – predicted)
- Residuals of $e_i = 0$ indicate that the observed point lies directly on the regression line
- Ideally, we would want every point to lie directly on the line

Simple Linear Regression

Simple Linear Regression

- Since points do not all lie on the regression line, we must determine the best criterion for fitting the line in such a way as to make these residuals as small as possible

Simple Linear Regression

- Since points do not all lie on the regression line, we must determine the best criterion for fitting the line in such a way as to make these residuals as small as possible
- *Residual sum of squares*:

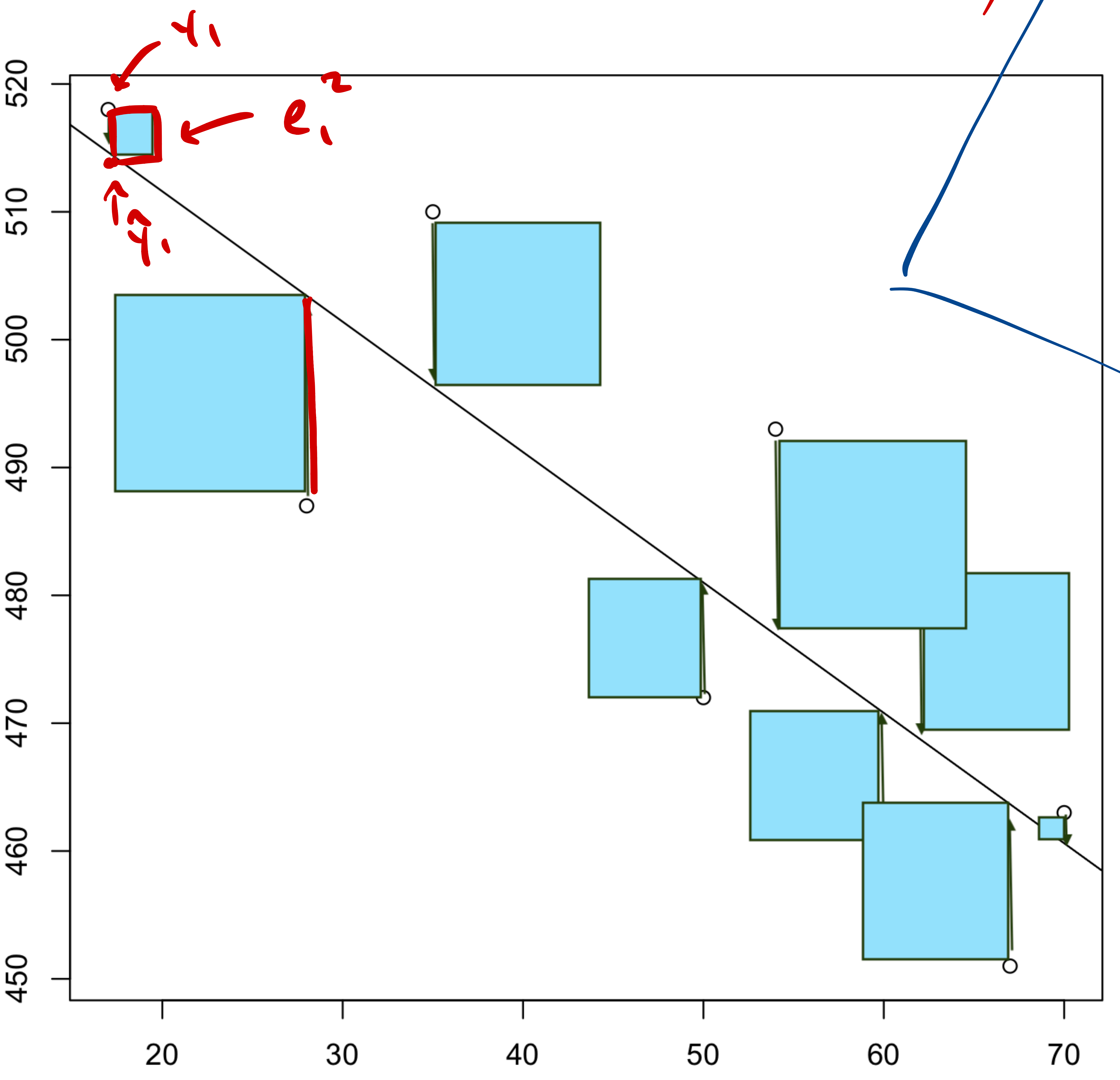
Simple Linear Regression

- Since points do not all lie on the regression line, we must determine the best criterion for fitting the line in such a way as to make these residuals as small as possible

- *Residual sum of squares*:

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

Simple Linear Regression



Simple Linear Regression

Simple Linear Regression

- Plugging in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we get:

Simple Linear Regression

- Plugging in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we get:

$$\sum e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \underline{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$$

Simple Linear Regression

- Plugging in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we get:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Using this, we get our parameter estimates as follows:

Simple Linear Regression

- Plugging in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we get:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Using this, we get our parameter estimates as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \boxed{r\frac{s_y}{s_x}}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

PCC

Simple Linear Regression

- Plugging in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we get:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Using this, we get our parameter estimates as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = r\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- Note, r is Pearson's correlation coefficient, s_y is the standard deviation of y, and s_x is the standard deviation of x

Simple Linear Regression

"Run LR on y given x"

```
> set.seed(223542)
> dat1 <- rmvnorm(10,c(11,10), sigma=matrix(c(1,.5, .5, 1),2,2))
> colnames(dat1) <- c("X","Y")
> x <- dat1[,1]
> y <- dat1[,2]
> model1 <- lm(y~x)
> plot(x,y, xlab="X", ylab="Y")
> abline(model1)
> model1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.3236       0.8578
```

$\hat{\beta}_0$ $\hat{\beta}_1$

```
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27555 -0.34855 -0.09534  0.52797  1.28676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3236     2.3820   0.136  0.89530
x             0.8578     0.2209   3.884  0.00465 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8009 on 8 degrees of freedom
Multiple R-squared:  0.6534,Adjusted R-squared:  0.6101
F-statistic: 15.08 on 1 and 8 DF,  p-value: 0.004651
```

SE($\hat{\beta}_0$)

SE($\hat{\beta}_1$)

Simple Linear Regression

Simple Linear Regression

- Consider the economic data presented in the longley dataset

Simple Linear Regression

- Consider the economic data presented in the longley dataset
- Include economic variables from mid-1900s America

Simple Linear Regression

- Consider the economic data presented in the longley dataset
- Include economic variables from mid-1900s America
- GNP: in billions of US dollars

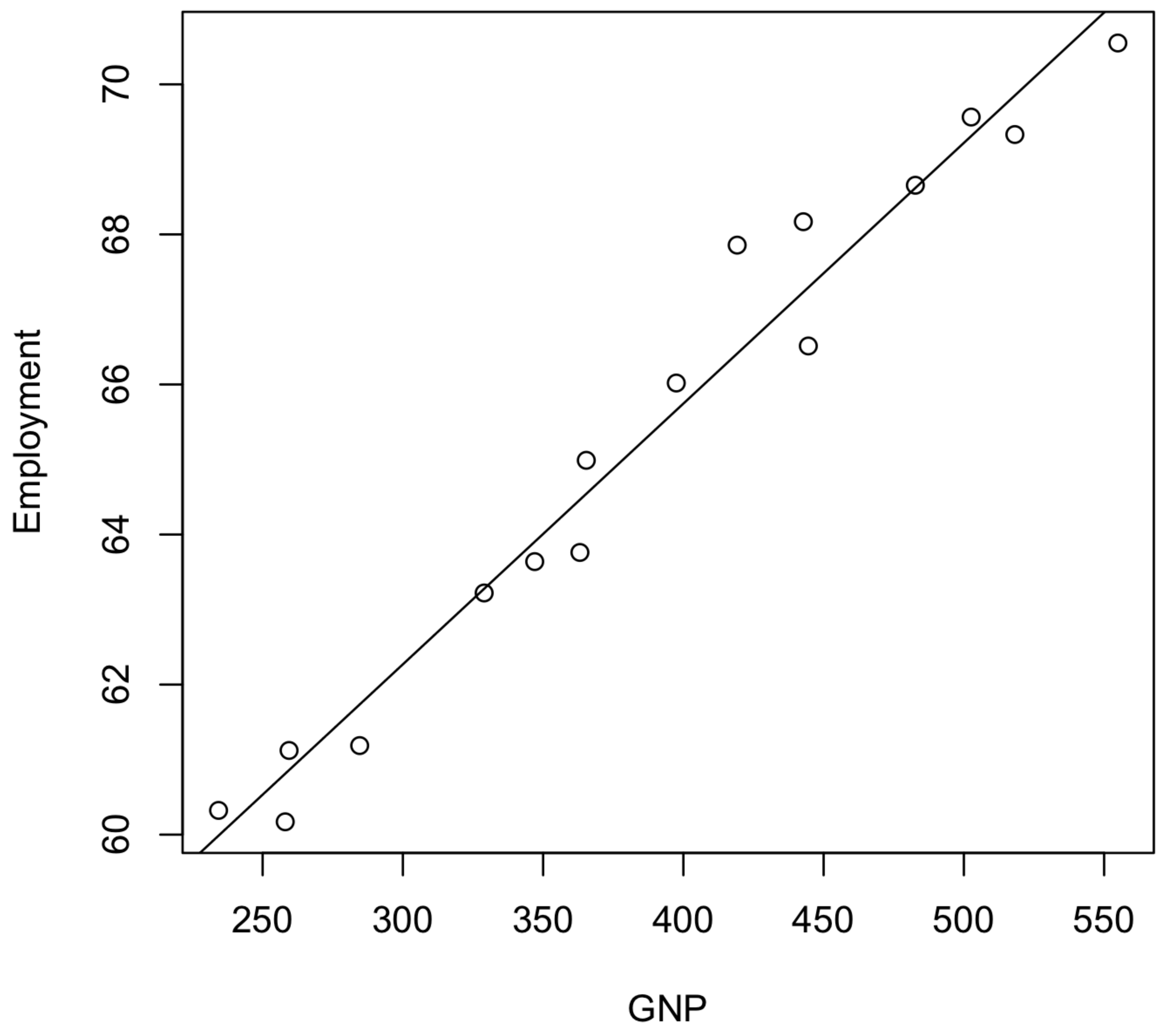
Simple Linear Regression

- Consider the economic data presented in the longley dataset
- Include economic variables from mid-1900s America
- GNP: in billions of US dollars
- Employment: in millions of people

$\hat{y}$ vs. y

$n = 16$

$\hat{y} = 51.844 + 0.03475x$



Simple Linear Regression

Simple Linear Regression

- cor(x,y) = $r = 0.98355$

Simple Linear Regression

- cor(x,y) = $r = 0.98355$
- sd(x) = $s_x = 99.395$

Simple Linear Regression

- cor(x,y) = $r = 0.98355$
- sd(x) = $s_x = 99.395$
- sd(y) = $s_y = 3.512$

Simple Linear Regression

- cor(x,y) = $r = 0.98355$
- sd(x) = $s_x = 99.395$
- sd(y) = $s_y = 3.512$
- mean(x) = $\bar{x} = 387.699$

Simple Linear Regression

- cor(x,y) = $r = 0.98355$
- sd(x) = $s_x = 99.395$
- sd(y) = $s_y = 3.512$
- mean(x) = $\bar{x} = 387.699$
- mean(y) = $\bar{y} = 65.317$

Simple Linear Regression

- cor(x,y) = $r = 0.98355$
- sd(x) = $s_x = 99.395$
- sd(y) = $s_y = 3.512$
- mean(x) = $\bar{x} = 387.699$
- mean(y) = $\bar{y} = 65.317$
- $\hat{\beta}_1 = 0.98355\left(\dfrac{3.512}{99.395}\right) = 0.03475$

Simple Linear Regression

- cor(x,y) = $r = 0.98355$
- sd(x) = $s_x = 99.395$
- sd(y) = $s_y = 3.512$
- mean(x) = $\bar{x} = 387.699$
- mean(y) = $\bar{y} = 65.317$

$r \cdot \frac{s_y}{s_x}$

- $\hat{\beta}_1 = 0.98355\left(\frac{3.512}{99.395}\right) = 0.03475$
- $\hat{\beta}_0 = 65.317 - 0.03475(387.699) = 51.844$

$\bar{y} - \hat{\beta}_1 \bar{x}$

Interpretation of Regression Estimates

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept
  - When $x = 0$ we expect y to be equal to $\hat{\beta}_0$

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept
  - When $x = 0$ we expect y to be equal to $\hat{\beta}_0$
  - Only makes sense if $x = 0$ is within the range of your data and has contextual meaning

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept
  - When $x = 0$ we expect y to be equal to $\hat{\beta}_0$
  - Only makes sense if $x = 0$ is within the range of your data and has contextual meaning
  - E.g., 51.844 million people are expected to be employed with the GNP is $0

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept
  - When $x = 0$ we expect y to be equal to $\hat{\beta}_0$
  - Only makes sense if $x = 0$ is within the range of your data and has contextual meaning
  - E.g., 51.844 million people are expected to be employed with the GNP is $0
- $\hat{\beta}_1$ is the slope

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept
  - When $x = 0$ we expect y to be equal to $\hat{\beta}_0$
  - Only makes sense if $x = 0$ is within the range of your data and has contextual meaning
  - E.g., 51.844 million people are expected to be employed with the GNP is $0
- $\hat{\beta}_1$ is the slope
  - For each 1 unit increase in x, we expect y to increase by $\hat{\beta}_1$ according to the model

Interpretation of Regression Estimates

- $\hat{\beta}_0$ is the y-intercept
  - When $x = 0$ we expect y to be equal to $\hat{\beta}_0$
  - Only makes sense if $x = 0$ is within the range of your data and has contextual meaning
  - E.g., 51.844 million people are expected to be employed with the GNP is $0
- $\hat{\beta}_1$ is the slope
  - For each 1 unit increase in x, we expect y to increase by $\hat{\beta}_1$ according to the model
  - E.g., for each $1 billion increase in GNP, we expect the number of people employed to increase by 0.03475 million

Prediction

Prediction

- We can use our regression line to make predictions

Prediction

$$\hat{Y} = 51.844 + 0.03475 \underset{\uparrow}{x} \quad 350$$

- We can use our regression line to make predictions
- Suppose we want to predict employment numbers when GNP is $350 billion

Prediction

- We can use our regression line to make predictions
- Suppose we want to predict employment numbers when GNP is \$350 billion
- $\hat{y} = 51.844 + 0.03475x = 51.844 + 0.03475 \cdot 350 = 64.0065$ ~~billion USD~~

mill people

Extrapolation

Extrapolation

- We can only use our regression line to make predictions over the set of values for which we have observations

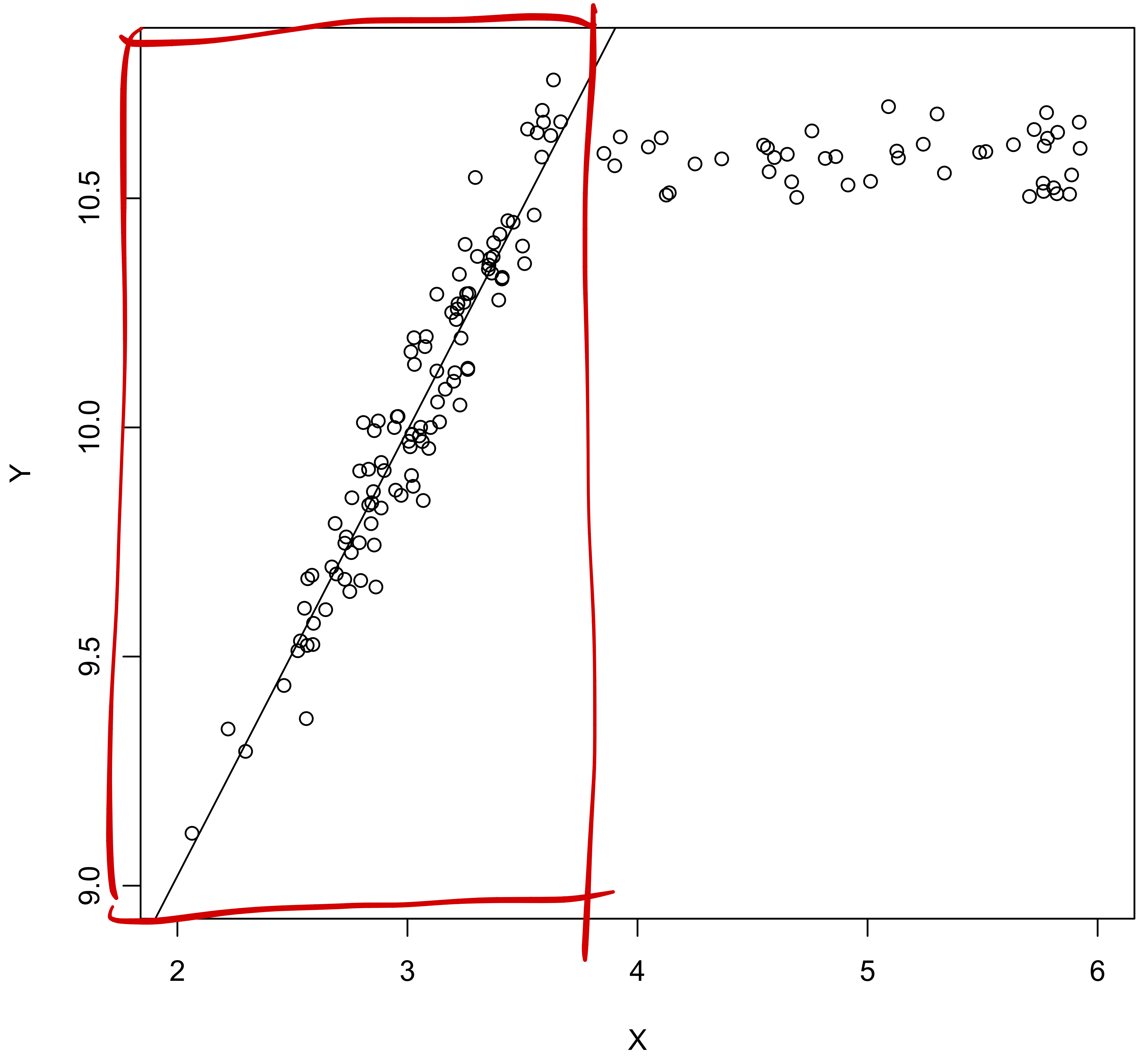
Extrapolation

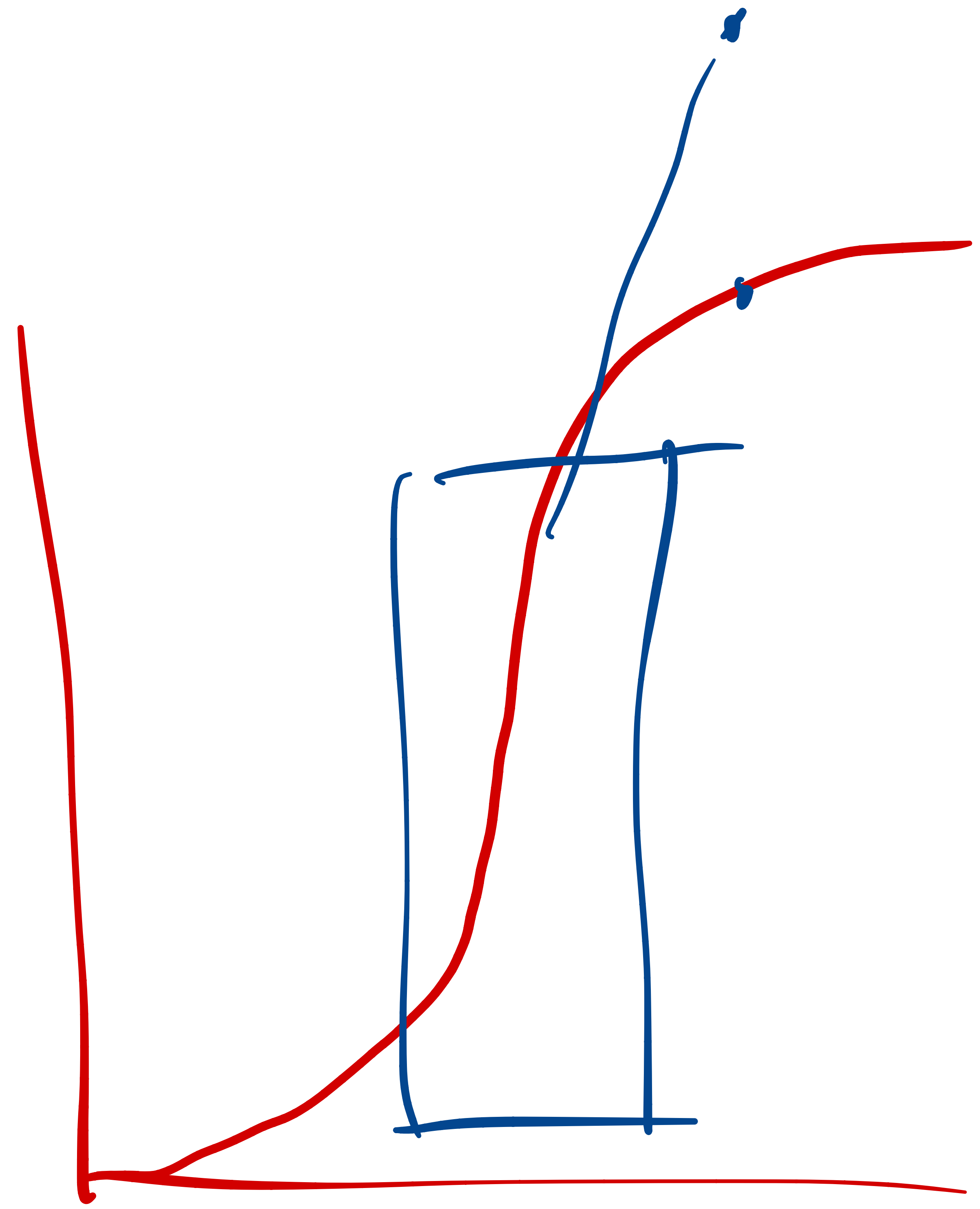
- We can only use our regression line to make predictions over the set of values for which we have observations
- The regression line should not be extended outside the range for which we have data

Extrapolation

- We can only use our regression line to make predictions over the set of values for which we have observations
- The regression line should not be extended outside the range for which we have data
- Intuition: Our model was created only for our range of data, and we do not know what happens outside of this range

Extrapolation





Inference for Regression Coefficients

Inference for Regression Coefficients

- Regression line is based on a sample, but we want to make a conclusion about a population

Inference for Regression Coefficients

- Regression line is based on a sample, but we want to make a conclusion about a population

β_0 β_1

- Create confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$

Inference for Regression Coefficients

- Regression line is based on a sample, but we want to make a conclusion about a population
- Create confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$
- Perform hypothesis tests

Inference for Regression Coefficients

- Regression line is based on a sample, but we want to make a conclusion about a population
- Create confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$
- Perform hypothesis tests
    - If $\beta_1 = 0$, this implies that a change in x has no impact on y

Inference for Regression Coefficients

- Regression line is based on a sample, but we want to make a conclusion about a population
- Create confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$
- Perform hypothesis tests
  - If $\beta_1 = 0$, this implies that a change in x has no impact on y
  - Hypothesis tests for β_0 are often unimportant and have little meaning, so we will focus only on β_1

Inference for Regression Coefficients

Inference for Regression Coefficients

- Since different samples will lead to different values of $\hat{\beta}_0$ and $\hat{\beta}_1$, this implies that there is variability to these estimates

Inference for Regression Coefficients

- Since different samples will lead to different values of $\hat{\beta}_0$ and $\hat{\beta}_1$, this implies that there is variability to these estimates

- $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Inference for Regression Coefficients

- Since different samples will lead to different values of $\hat{\beta}_0$ and $\hat{\beta}_1$, this implies that there is variability to these estimates

- $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $\leftarrow$ $\varepsilon \sim N(0, \sigma^2)$

- $Var(\hat{\beta}_0) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$ $\leftarrow$ x

Inference for Regression Coefficients

- Since different samples will lead to different values of $\hat{\beta}_0$ and $\hat{\beta}_1$, this implies that there is variability to these estimates

- $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

- $Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$

- Note that σ^2 is the variance of the residuals around the predicted regression line

Inference for Regression Coefficients

- Since different samples will lead to different values of $\hat{\beta}_0$ and $\hat{\beta}_1$, this implies that there is variability to these estimates

- $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

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

- $Var(\hat{\beta}_0) = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$

$df = n-2$

- Note that σ^2 is the variance of the residuals around the predicted regression line

- Estimate σ^2 with sample $s^2 = \underline{Var(e_i)} = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n} e_i^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Inference for Regression Coefficients

Inference for Regression Coefficients

- Using this estimate s^2 for unknown σ^2, we get the following

Inference for Regression Coefficients

- Using this estimate s^2 for unknown σ^2, we get the following
- $Var(\hat{\beta}_1) = \dfrac{s^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$

Inference for Regression Coefficients

- Using this estimate s^2 for unknown σ^2, we get the following

σ^2

- $$Var(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$SE(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)}$

- $$Var(\hat{\beta}_0) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$SE(\hat{\beta}_0)$

σ^2

Inference for Regression Coefficients

Inference for Regression Coefficients

- We typically are interested only in hypothesis tests regarding $\hat{\beta}_1$

Inference for Regression Coefficients

- We typically are interested only in hypothesis tests regarding $\hat{\beta}_1$
- The slope tells us of the relationship between x and y

Inference for Regression Coefficients

- We typically are interested only in hypothesis tests regarding $\hat{\beta}_1$
- The slope tells us of the relationship between x and y
- Test the null hypothesis $H_0 : \beta_1 = \beta_1^*$ vs. $H_1 : \beta_1 \neq \beta_1^*$, where β_1^* is some population slope value, at the α significance level

Inference for Regression Coefficients

- We typically are interested only in hypothesis tests regarding $\hat{\beta}_1$
- The slope tells us of the relationship between x and y
- Test the null hypothesis $H_0 : \beta_1 = \beta_1^*$ vs. $H_1 : \beta_1 \neq \beta_1^*$, where β_1^* is some population slope value, at the α significance level
    - Generally interested in the case of $\beta_1^* = 0$

Inference for Regression Coefficients

- We typically are interested only in hypothesis tests regarding $\hat{\beta}_1$
- The slope tells us of the relationship between x and y
- Test the null hypothesis $H_0 : \beta_1 = \beta_1^*$ vs. $H_1 : \beta_1 \neq \beta_1^*$, where β_1^* is some population slope value, at the α significance level
  - Generally interested in the case of $\beta_1^* = 0$
  - No relationship vs. some relationship

Inference for Regression Coefficients

Inference for Regression Coefficients

- Consider the test statistic $t = \dfrac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$

Inference for Regression Coefficients

- Consider the test statistic $t = \dfrac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$
    - $SE(\hat{\beta}_1) = \sqrt{\dfrac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

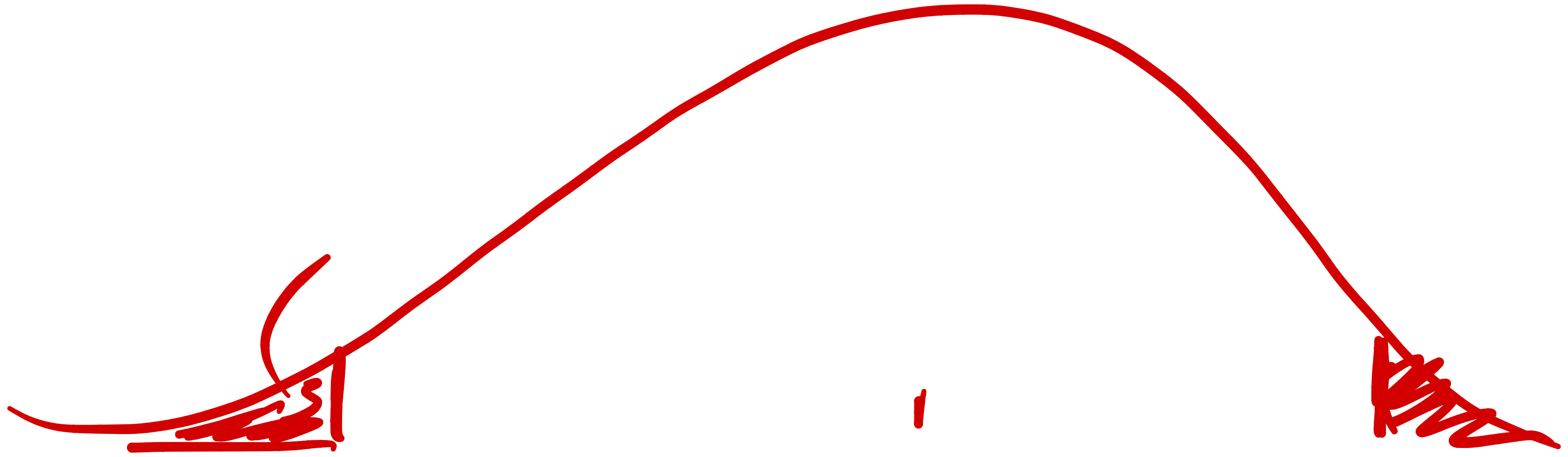
Inference for Regression Coefficients

- Consider the test statistic $t = \dfrac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$
  - $SE(\hat{\beta}_1) = \sqrt{\dfrac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$
- Under null hypothesis H_0, t follows a t-distribution with $n - 2$ degrees of freedom

Inference for Regression Coefficients

- Consider the test statistic $t = \dfrac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$

- $SE(\hat{\beta}_1) = \sqrt{\dfrac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$



- Under null hypothesis H_0, t follows a t-distribution with $n-2$ degrees of freedom

2x (1− pt(abs(t), df = n-2))

''

- Our p-value is thus `2*pt(-abs(t), df=n-2)`

Inference for Regression Coefficients

- Consider the test statistic $t = \dfrac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)}$
- $SE(\hat{\beta}_1) = \sqrt{\dfrac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
- Under null hypothesis H_0, t follows a t-distribution with $n - 2$ degrees of freedom
- Our p-value is thus `2*pt(-abs(t), df=n-2)`
- If the p-value $p \leq \alpha$, we reject the null hypothesis

Inference for Regression Coefficients

Inference for Regression Coefficients

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

- In this simple regression case, if $\beta_1 = 0$, then $\rho = 0$

Inference for Regression Coefficients

- In this simple regression case, if $\beta_1 = 0$, then $\rho = 0$
- The test of the null hypothesis $H_0 : \beta_1 = 0$ is the same as the test of $H_0 : \rho = 0$

Inference for Regression Coefficients

- In this simple regression case, if $\beta_1 = 0$, then $\rho = 0$
- The test of the null hypothesis $H_0 : \beta_1 = 0$ is the same as the test of $H_0 : \rho = 0$
  - Both hypotheses claim that y does not change as x increases

Inference for Regression Coefficients

Inference for Regression Coefficients

- Consider our employment example

Inference for Regression Coefficients

- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level

Inference for Regression Coefficients

- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level
- We calculated $\hat{\beta}_1 = 0.03475$

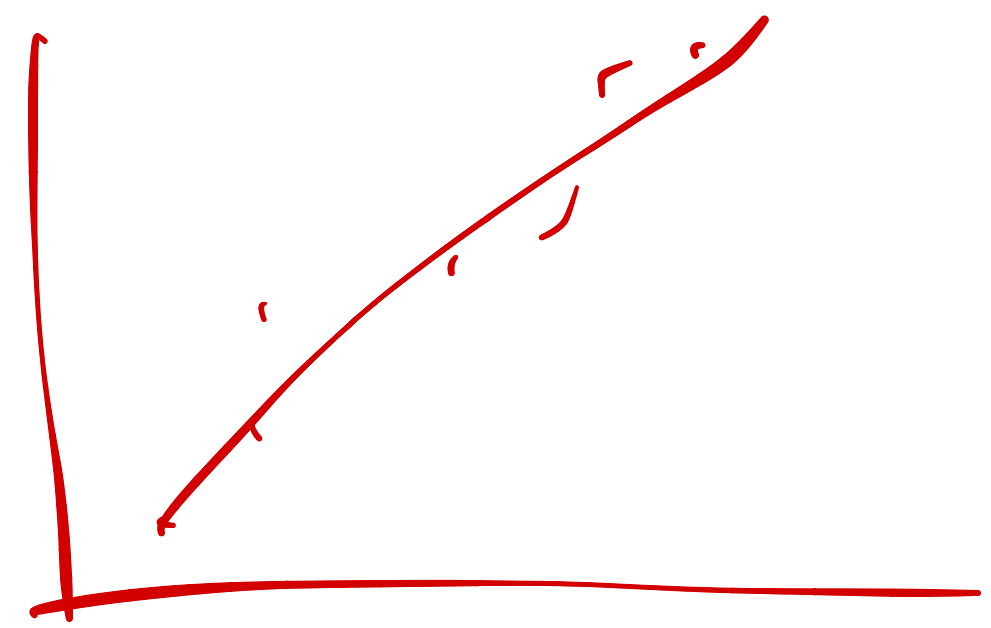
Inference for Regression Coefficients

- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level
- We calculated $\hat{\beta}_1 = 0.03475$

σ

- $s = 0.6566$

Inference for Regression Coefficients



- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level
- We calculated $\hat{\beta}_1 = 0.03475$
- $s = 0.6566$
- $SE(\hat{\beta}_1) = \dfrac{0.6566}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}} = 0.001706$

$n = 16$

$df = n - 2 = 14$

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{0.03475}{0.001706} = 20.$$

$2 \times pt(-20, df = 14) = 0.$

Inference for Regression Coefficients

- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level
- We calculated $\hat{\beta}_1 = 0.03475$
- $s = 0.6566$
- $SE(\hat{\beta}_1) = \dfrac{0.6566}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}} = 0.001706$
- $t = \dfrac{0.03475}{0.001706} = 20.37$

Inference for Regression Coefficients

- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level
- We calculated $\hat{\beta}_1 = 0.03475$
- $s = 0.6566$
- $SE(\hat{\beta}_1) = \dfrac{0.6566}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}} = 0.001706$
- $t = \dfrac{0.03475}{0.001706} = 20.37$
- $p =$ 2*pt(-20.37, df=14) $= 8.4 \times 10^{-12}$

Inference for Regression Coefficients

- Consider our employment example
- Test the hypotheses $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level
- We calculated $\hat{\beta}_1 = 0.03475$
- $s = 0.6566$
- $SE(\hat{\beta}_1) = \dfrac{0.6566}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = 0.001706$
- $t = \dfrac{0.03475}{0.001706} = 20.37$
- $p =$ 2*pt(-20.37, df=14) $= 8.4 \times 10^{-12}$
- Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant linear relationship between GNP and employment

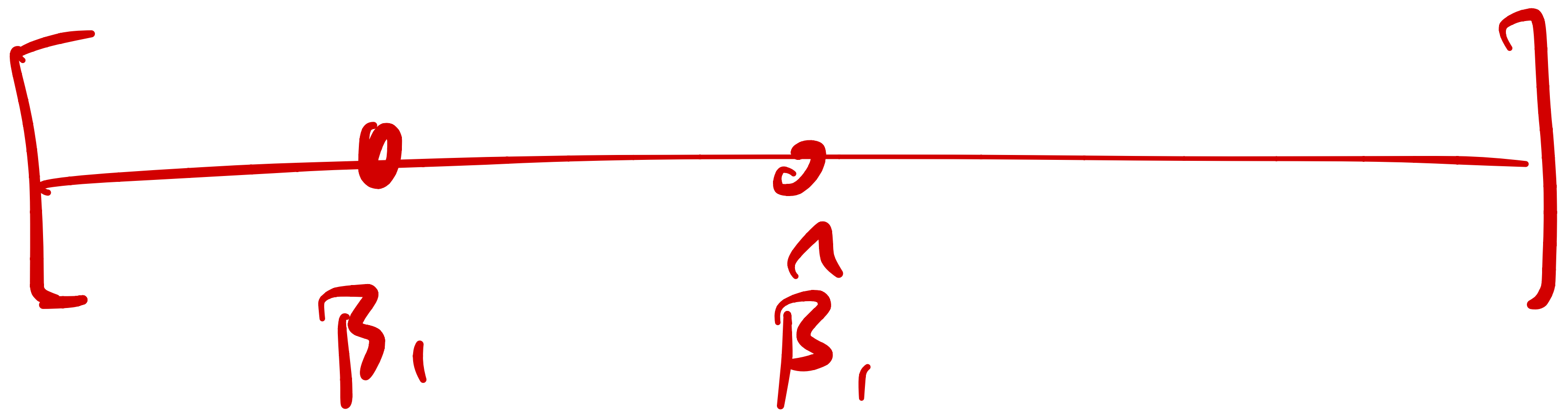
Confidence Intervals for Regression Coefficients

Confidence Intervals for Regression Coefficients

- We can also create confidence intervals for regression coefficients

Confidence Intervals for Regression Coefficients

- We can also create confidence intervals for regression coefficients
- A $(1-\alpha)\times 100\,\%$ confidence interval for $\hat{\beta}_1$ is given as
  $\left(\hat{\beta}_1 - t_{\alpha/2}SE(\hat{\beta}_1), \hat{\beta}_1 + t_{\alpha/2}SE(\hat{\beta}_1)\right)$



Confidence Intervals for Regression Coefficients

Confidence Intervals for Regression Coefficients

- Applied to our example:

Confidence Intervals for Regression Coefficients

- Applied to our example:

$\hat{\beta}_1$ t $SE(\hat{\beta}_1)$

$$
\begin{aligned}
CI &= 0.03475 \pm \mathrm{qt}(0.975, 14)(0.001706) \\
&= 0.03475 \pm 2.145(0.001706) \\
&= \underbrace{(0.03109,\ 0.03841)}
\end{aligned}
$$

Confidence Intervals for Regression Coefficients

- Applied to our example:

$$
\begin{aligned}
CI &= 0.03475 \pm \text{qt}(0.975, 14)(0.001706) \\
&= 0.03475 \pm 2.145(0.001706) \\
&= (0.03109,\ 0.03841)
\end{aligned}
$$

- We are 95% confident that the interval $(0.03109,\ 0.03841)$ contains the true population slope

Confidence Intervals for Regression Coefficients

lm(y ~ x)

```
> confint(lm1)
                  2.5 %   97.5 %
(Intercept) -5.1694143 5.816569
dat1[, 1]    0.3484624 1.367178
```

β_0

β_1

ANOVA Approach to Regression

ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA

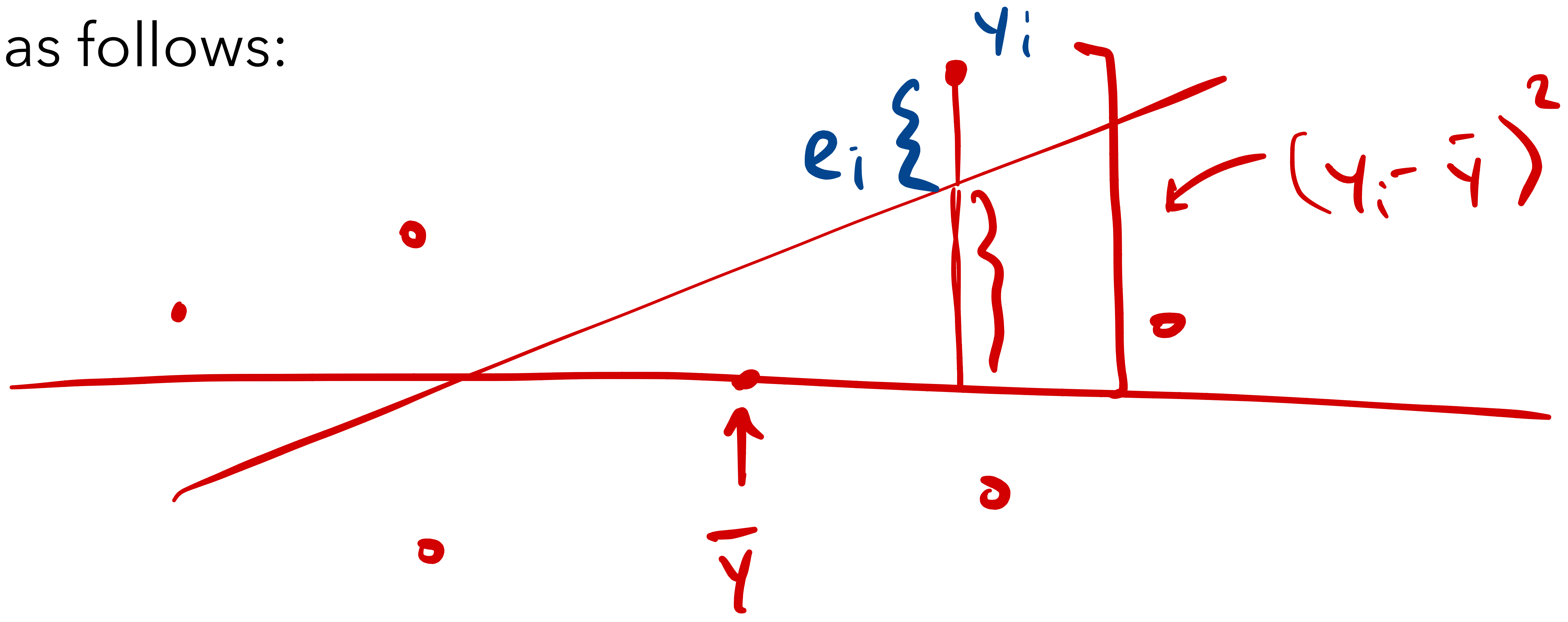
ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA
- The sum of squared errors (SSE) is defined as follows:

ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA
- The sum of squared errors (SSE) is defined as follows:

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



$(y_i - \hat{y})$

$(\hat{y} - \bar{y})$

ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA
- The sum of squared errors (SSE) is defined as follows:

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

- This is also known as the residual sum of squares and describes the random variability about the regression line

ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA
- The sum of squared errors (SSE) is defined as follows:

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

- This is also known as the residual sum of squares and describes the random variability about the regression line
- The mean squared error is then

ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA
- The sum of squared errors (SSE) is defined as follows:

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

- This is also known as the residual sum of squares and describes the random variability about the regression line
- The mean squared error is then

$$MSE = \frac{SSE}{n-2}$$

ANOVA Approach to Regression

- We can decompose the variability in a regression model in a way similar to what we did with ANOVA
- The sum of squared errors (SSE) is defined as follows:

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

- This is also known as the residual sum of squares and describes the random variability about the regression line
- The mean squared error is then

$$MSE = \frac{SSE}{n-2}$$

- Since we have two unknown parameters (β_0 and β_1), this is analogous to the $n - k$ degrees of freedom with the MSE in ANOVA

ANOVA Approach to Regression

ANOVA Approach to Regression

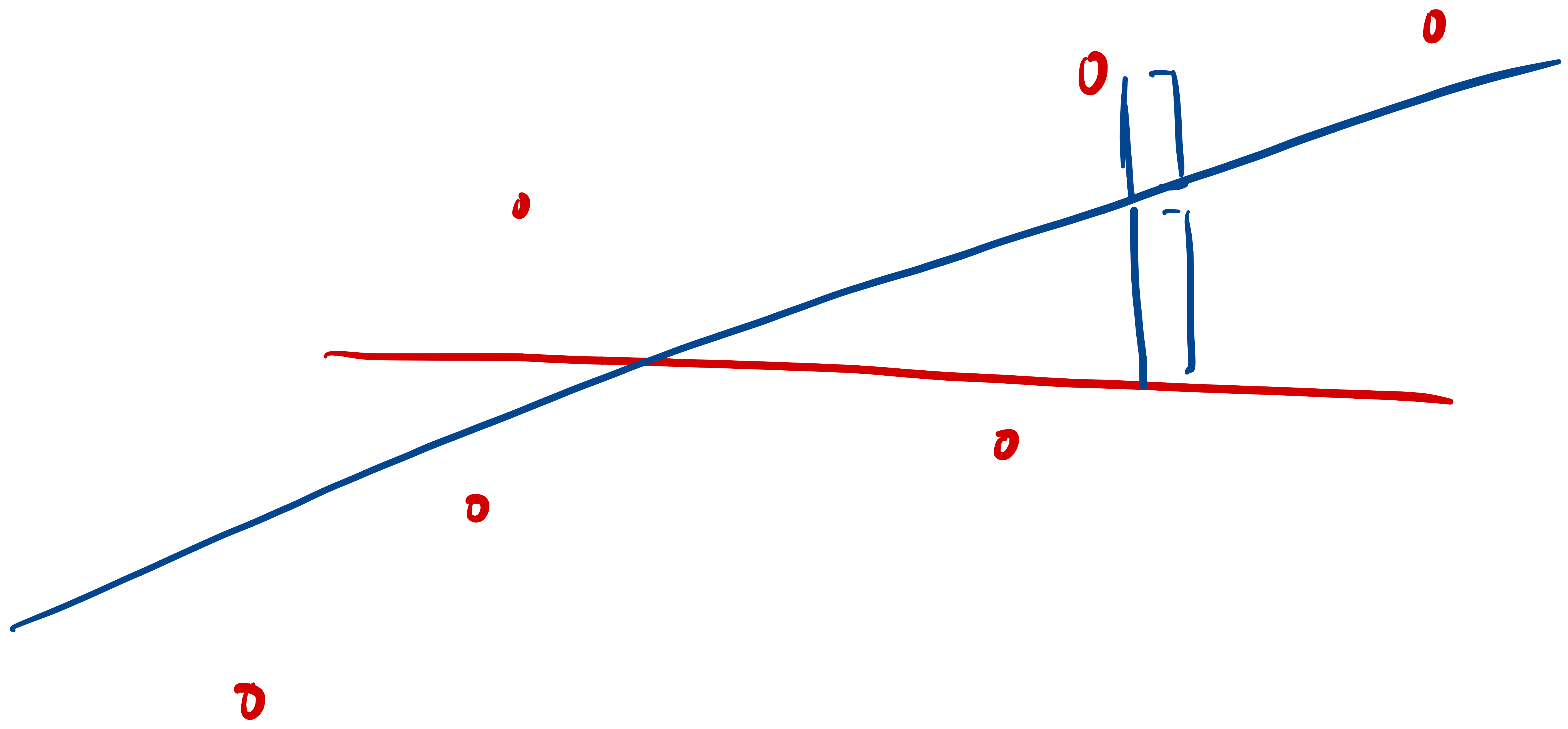
- The treatment sum of squares (SST) gets redefined as the regression sum of squares (the explained variability):

ANOVA Approach to Regression

- The treatment sum of squares (SST) gets redefined as the regression sum of squares (the explained variability):

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

$(y_i - \bar{y})^2$



ANOVA Approach to Regression

- The treatment sum of squares (SST) gets redefined as the regression sum of squares (the explained variability):

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

- This describes variability that is explained due to the regression line

ANOVA Approach to Regression

- The treatment sum of squares (SST) gets redefined as the regression sum of squares (the explained variability):

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

- This describes variability that is explained due to the regression line

- The total sum of squares is then $SSTo = SSR + SSE$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

ANOVA Approach to Regression

ANOVA Approach to Regression

- We can compile this information into an ANOVA table

ANOVA Approach to Regression

- We can compile this information into an ANOVA table

Source	SS	df	MS	F
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
Total	SSTo	n-1		

ANOVA Approach to Regression

- We can compile this information into an ANOVA table

Source	SS	df	MS	F
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
Total	SSTo	n-1		

- F can be used to test the hypothesis $H_0 : \beta_1 = 0$

ANOVA Approach to Regression

- We can compile this information into an ANOVA table

Source	SS	df	MS	F
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
Total	SSTo	n-1		

- F can be used to test the hypothesis $H_0 : \beta_1 = 0$
- F has an F distribution with 1 and $n-2$ degrees of freedom

Inference for Mean Response

Inference for Mean Response

- Regression parameter estimates have sampling distributions

Inference for Mean Response

- Regression parameter estimates have sampling distributions
  - If we selected a different sample, the regression line would be slightly different than the line we got from our sample

Inference for Mean Response

- Regression parameter estimates have sampling distributions
    - If we selected a different sample, the regression line would be slightly different than the line we got from our sample
- We can use the regression line to estimate the mean value of y corresponding to a particular value of $x = x*$

Inference for Mean Response

- Regression parameter estimates have sampling distributions
  - If we selected a different sample, the regression line would be slightly different than the line we got from our sample
- We can use the regression line to estimate the mean value of y corresponding to a particular value of $x = x^*$
  - If we took many samples for a particular x^* and found their average response y, it would be equal to the estimated response from our regression

$x^* \rightarrow \hat{y}$

Inference for Mean Response

Inference for Mean Response

- Recall that we predicted, on average, employment to be 64.0065 million given GNP was 350 billion USD

Inference for Mean Response

- Recall that we predicted, on average, employment to be 64.0065 million given GNP was 350 billion USD
- In actuality, when the GNP is 350 billion USD, the employment will not necessarily equal exactly 64.0065 million, but instead that will be the average response

Inference for Mean Response

Inference for Mean Response

- We can create a confidence interval for this mean value

Inference for Mean Response

- We can create a confidence interval for this mean value
- The $(1-\alpha)\overset{\times 100}{\%}$ confidence interval for the true mean $\bar{y}$ for the regression line at a particular point x^* is given as $\left(\hat{y} - t_{\alpha/2}SE(\hat{y}), \hat{y} + t_{\alpha/2}SE(\hat{y})\right)$

Inference for Mean Response

- We can create a confidence interval for this mean value

- The $(1-\alpha)\%$ confidence interval for the true mean $\bar{y}$ for the regression line at a particular point x^* is given as $\left(\hat{y} - t_{\alpha/2}SE(\hat{y}), \hat{y} + t_{\alpha/2}SE(\hat{y})\right)$

- Here, $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{\boxed{(x^* - \bar{x})^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$

Inference for Mean Response

- We can create a confidence interval for this mean value
- The $(1-\alpha)\,\%$ confidence interval for the true mean $\bar{y}$ for the regression line at a particular point x^* is given as $\left(\hat{y} - t_{\alpha/2}SE(\hat{y}), \hat{y} + t_{\alpha/2}SE(\hat{y})\right)$
- Here, $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
- The standard error depends on the particular estimate of x^*

Inference for Mean Response

- We can create a confidence interval for this mean value
- The $(1-\alpha)\,\%$ confidence interval for the true mean $\bar{y}$ for the regression line at a particular point x^* is given as $\left(\hat{y} - t_{\alpha/2}SE(\hat{y}), \hat{y} + t_{\alpha/2}SE(\hat{y})\right)$
- Here, $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
- The standard error depends on the particular estimate of x^*
- The closer x^* is to $\bar{x}$, the smaller the variability around the line

Inference for Mean Response

- We can create a confidence interval for this mean value
- The $(1-\alpha)\%$ confidence interval for the true mean $\bar{y}$ for the regression line at a particular point $x*$ is given as $\left(\hat{y} - t_{\alpha/2}SE(\hat{y}), \hat{y} + t_{\alpha/2}SE(\hat{y})\right)$
- Here, $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
- The standard error depends on the particular estimate of $x*$
- The closer $x*$ is to $\bar{x}$, the smaller the variability around the line
    - At the mean, there is the most information and thus this is the best estimated point

Inference for Mean Response

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[\frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.17629$

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$
- Recall that $s = 0.6566,\ \bar{x} = 387.698,\ s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[\frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.17629$
- `qt(0.975, df=14)` = 2.145

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$
- Recall that $s = 0.6566,\ \bar{x} = 387.698,\ s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[\frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.17629$
- `qt(0.975, df=14)` = 2.145
- CI: $64.0065 \pm 2.145 \cdot 0.17629$

Inference for Mean Response

- Create a confidence interval for $\bar{y}$ at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(\hat{y}) = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[\frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.17629$
- `qt(0.975, df=14)` = 2.145
- CI: $64.0065 \pm 2.145 \cdot 0.17629$
- Conclusion: I am 95% confident that the interval (63.628, 64.385) million contains the true mean employment number when the GNP is 350 billion USD

Inference for Predicted Response

Inference for Predicted Response

- Instead of considering the mean response of y for a given value of x, perhaps we are interested in the response of a single observation of x^*

Inference for Predicted Response

- Instead of considering the mean response of y for a given value of x, perhaps we are interested in the response of a single observation of $x*$
  - "You find the GNP of a given area to be 350. What do we expect the area's employment numbers to be?"

Inference for Predicted Response

- Instead of considering the mean response of y for a given value of x, perhaps we are interested in the response of a single observation of $x*$
    - "You find the GNP of a given area to be 350. What do we expect the area's employment numbers to be?"
- The best estimate we have is still the predicted value from the regression line

Inference for Predicted Response

- Instead of considering the mean response of y for a given value of x, perhaps we are interested in the response of a single observation of $x*$
    - "You find the GNP of a given area to be 350. What do we expect the area's employment numbers to be?"
- The best estimate we have is still the predicted value from the regression line
    - $y* = \hat{\beta}_0 + \hat{\beta}_1 x* = \hat{y}$

Inference for Predicted Response

- Instead of considering the mean response of y for a given value of x, perhaps we are interested in the response of a single observation of $x*$
    - "You find the GNP of a given area to be 350. What do we expect the area's employment numbers to be?"
- The best estimate we have is still the predicted value from the regression line
    - $y* = \hat{\beta}_0 + \hat{\beta}_1 x* = \hat{y}$
- We are less certain in this estimate; we know that on average it is good, but for one point, it is probably going to be a bit off

Inference for Predicted Response

Inference for Predicted Response

- In creating the regression line, we have variability based on our sample

Inference for Predicted Response

- In creating the regression line, we have variability based on our sample
    - The variability of the regression line around the mean response at $x = x*$ is what we calculated as $Var(\hat{y})$

Inference for Predicted Response

- In creating the regression line, we have variability based on our sample
    - The variability of the regression line around the mean response at $x = x*$ is what we calculated as $Var(\hat{y})$
- Now, we have added variability of $y*$ around the regression line

Inference for Predicted Response

- In creating the regression line, we have variability based on our sample
    - The variability of the regression line around the mean response at $x = x^*$ is what we calculated as $Var(\hat{y})$
- Now, we have added variability of y^* around the regression line
    - The variability of y^* around the regression line is given as σ^2, since the outcomes of y are assumed to be normally distributed at a given value of $x = x^*$ with variance σ^2

Inference for Predicted Response

- In creating the regression line, we have variability based on our sample
    - The variability of the regression line around the mean response at $x = x^*$ is what we calculated as $Var(\hat{y})$
- Now, we have added variability of y^* around the regression line
    - The variability of y^* around the regression line is given as σ^2, since the outcomes of y are assumed to be normally distributed at a given value of $x = x^*$ with variance σ^2
- This means that the total standard error of y^* can be written as $SE(y^*) = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]}$

Inference for Predicted Response

- In creating the regression line, we have variability based on our sample
  - The variability of the regression line around the mean response at $x = x^*$ is what we calculated as $Var(\hat{y})$
- Now, we have added variability of y^* around the regression line
  - The variability of y^* around the regression line is given as σ^2, since the outcomes of y are assumed to be normally distributed at a given value of $x = x^*$ with variance σ^2
- This means that the total standard error of y^* can be written as $SE(y^*) = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
- The $(1 - \alpha)\,\%$ confidence interval for the true observed y^* (not the mean $\bar{y}$!) for the regression line at a particular point x^* is given as $\left(\hat{y} - t_{\alpha/2}SE(y^*), \hat{y} + t_{\alpha/2}SE(y^*)\right)$

Inference for Predicted Response

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$
- Recall that $s = 0.6566,\ \bar{x} = 387.698,\ s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(y^*) = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[1 + \frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.6799$

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(y^*) = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[1 + \frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.6799$
- `qt(0.975, df=14)` = 2.145

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$
- Recall that $s = 0.6566,\ \bar{x} = 387.698,\ s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $$SE(y^*) = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[1 + \frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.6799$$
- `qt(0.975, df=14)` = 2.145
- CI: $64.0065 \pm 2.145 \cdot 0.6799$

Inference for Predicted Response

- Create a confidence interval for y^* at $x^* = 350$
- Recall that $s = 0.6566$, $\bar{x} = 387.698$, $s_x = 99.395$, and $n = 16$
- $\hat{y} = 64.0065$
- $SE(y^*) = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} = 0.6566\sqrt{\left[1 + \frac{1}{16} + \frac{(350 - 387.698)^2}{15 * 99.395^2}\right]} = 0.6799$
- `qt(0.975, df=14)` = 2.145
- CI: $64.0065 \pm 2.145 \cdot 0.6799$
- Conclusion: I am 95% confident that the interval (62.548, 65.465) million contains the true employment number when the GNP is 350 billion USD

Inference for Mean and Predicted Response

x^* $1-\alpha$ mean

```
> set.seed(223542)
> x <- rnorm(10, 5, 2)
> y <- rnorm(10, 13, 3)
> predict(lm(y~x), data.frame(x=6), conf.level=0.95, interval="confidence")
       fit      lwr      upr
1 10.41062 7.617309 13.20392
> predict(lm(y~x), data.frame(x=6), conf.level=0.95, interval="prediction")
       fit     lwr      upr
1 10.41062 2.83408 17.98715
```

predicted