Simple Linear Regression

Nate Wells

Math 243: Stat Learning

September 13th, 2021

Outline

In today's class, we will...

- Discuss theoretical foundation for linear regression
- Perform inference for simple linear models
- Implement simple linear regression in R

Section 1

Foundations

Linear Regression

- Suppose we have one or more predictors $(X_1, X_2, \ldots, X_p)$ and a *quantitative* response variable Y, and that
$$Y = f(X_1, \ldots, X_p) + \epsilon$$

Linear Regression

- Suppose we have one or more predictors $(X_1, X_2, \dots, X_p)$ and a *quantitative* response variable Y, and that
$$Y = f(X_1, \dots, X_p) + \epsilon$$
- The function f could theoretically take many forms. But the simplest form assumes f is a linear function:

$$f(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Linear Regression

- Suppose we have one or more predictors $(X_1, X_2, \ldots, X_p)$ and a *quantitative* response variable Y, and that
$$Y = f(X_1, \ldots, X_p) + \epsilon$$
- The function f could theoretically take many forms. But the simplest form assumes f is a linear function:
$$f(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
    - Note: a change in f is constant per unit change in any of the inputs.

Linear Regression

- Suppose we have one or more predictors $(X_1, X_2, \dots, X_p)$ and a *quantitative* response variable Y, and that

$$Y = f(X_1, \dots, X_p) + \epsilon$$

- The function f could theoretically take many forms. But the simplest form assumes f is a linear function:

$$f(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - Note: a change in f is constant per unit change in any of the inputs.
- If Y depends on only 1 predictor X, then the linear model reduces to

$$y = \hat{f}(x) = \beta_0 + \beta_1 x$$

Linear Regression

- Suppose we have one or more predictors $(X_1, X_2, \ldots, X_p)$ and a *quantitative* response variable Y, and that
$$Y = f(X_1, \ldots, X_p) + \epsilon$$
- The function f could theoretically take many forms. But the simplest form assumes f is a linear function:
$$f(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
    - Note: a change in f is constant per unit change in any of the inputs.
- If Y depends on only 1 predictor X, then the linear model reduces to
$$y = \hat{f}(x) = \beta_0 + \beta_1 x$$
- We'll use **Simple Linear Regression** (SLR) to build intuition about all linear models

Approximations and Estimates

- In reality, the relationship f between Y and $X_1, \ldots, X_p$ may not be linear

Approximations and Estimates

- In reality, the relationship f between Y and $X_1, \ldots, X_p$ may not be linear
- But many functions can be well-approximated by linear ones (especially when inputs are restricted to a small range)

Approximations and Estimates

- In reality, the relationship f between Y and $X_1, \ldots, X_p$ may not be linear
- But many functions can be well-approximated by linear ones (especially when inputs are restricted to a small range)
- But even if f is truly linear, we still have problems: We do not know the parameters of the linear model.

Approximations and Estimates

- In reality, the relationship f between Y and $X_1, \dots, X_p$ may not be linear
- But many functions can be well-approximated by linear ones (especially when inputs are restricted to a small range)
- But even if f is truly linear, we still have problems: We do not know the parameters of the linear model.
- Based on data, we estimate the parameters to create an estimated linear model

$$\hat{f} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Approximations and Estimates

- In reality, the relationship f between Y and $X_1, \ldots, X_p$ may not be linear
- But many functions can be well-approximated by linear ones (especially when inputs are restricted to a small range)
- But even if f is truly linear, we still have problems: We do not know the parameters of the linear model.
- Based on data, we estimate the parameters to create an estimated linear model

$$\hat{f} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

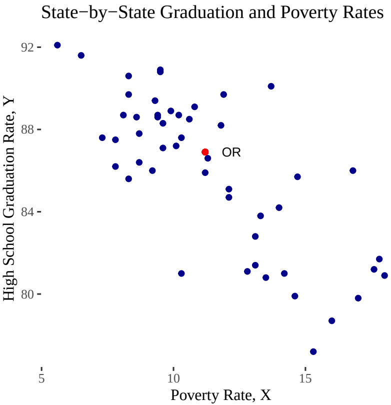
- So we are **estimating** an **approximation** to a relationship between response and predictors.

SLR Review

Consider the relationship between a state's high school grad rate Y and its poverty rate X.

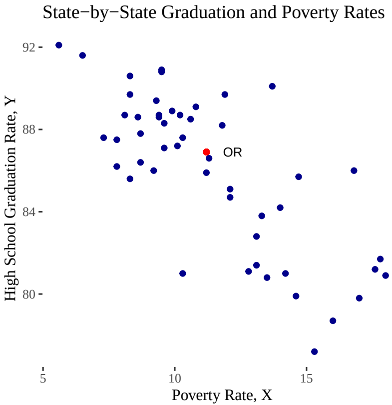
SLR Review

Consider the relationship between a state's high school grad rate Y and its poverty rate X.



SLR Review

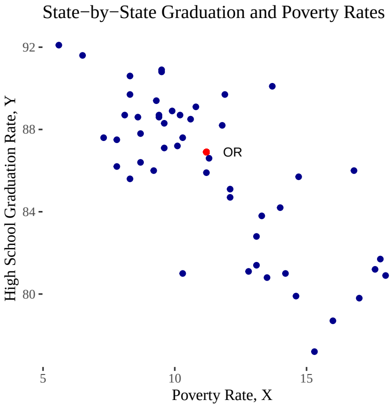
Consider the relationship between a state's high school grad rate Y and its poverty rate X.



- Suppose we want to model Y as a function of X

SLR Review

Consider the relationship between a state's high school grad rate Y and its poverty rate X.

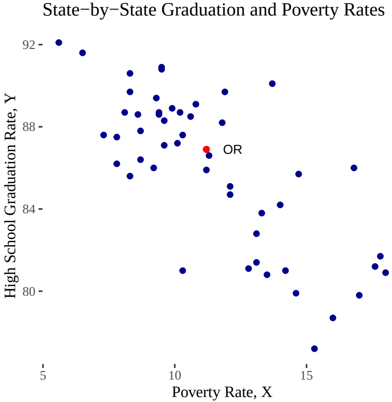


- Suppose we want to model Y as a function of X
- Let's assume a linear relationship

$$Y = \beta_0 + \beta_1 X + \epsilon$$

SLR Review

Consider the relationship between a state's high school grad rate Y and its poverty rate X.



- Suppose we want to model Y as a function of X
- Let's assume a linear relationship

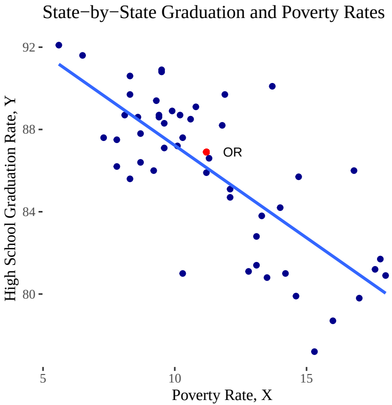
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Model (hand-fitted):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 96.2 - 0.9X$$

SLR Review

Consider the relationship between a state's high school grad rate Y and its poverty rate X.



- Suppose we want to model Y as a function of X
- Let's assume a linear relationship

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Model (hand-fitted):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 96.2 - 0.9X$$

Residuals

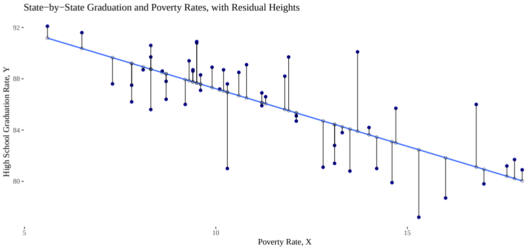
- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (x_i, y_i) has its own residual e_i, which is the difference between the observed (y_i) and predicted $(\hat{y}_i)$ value:

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

Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (x_i, y_i) has its own residual e_i, which is the difference between the observed (y_i) and predicted $(\hat{y}_i)$ value:

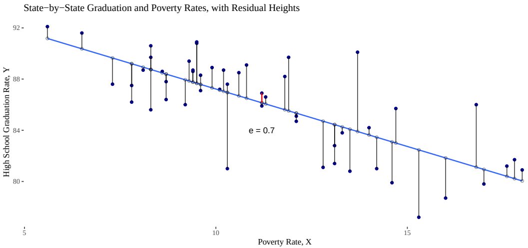
$$e_i = y_i - \hat{y}_i$$



Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (x_i, y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{y}_i)$ value:

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

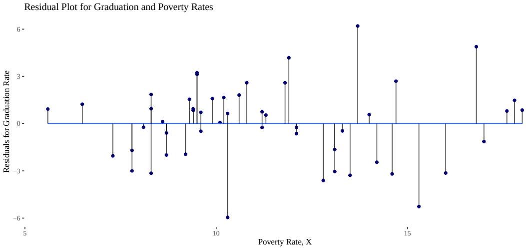


- Oregon's residual is

$$e = y - \hat{y} = 86.9 - 86.2 = 0.7$$

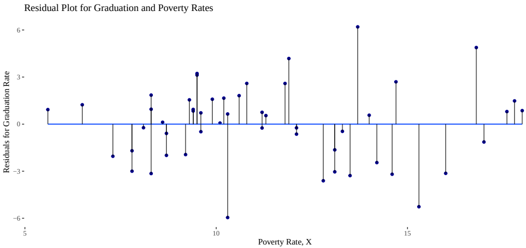
Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



Residual Plot

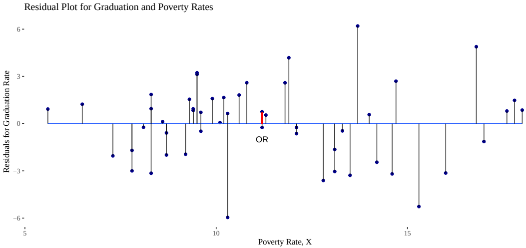
- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.

Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.

Residual Sum of Squares

- Define the **Residual Sum of Squares** (RSS) as

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e_1^2 + \cdots + e_n^2$$

Residual Sum of Squares

- Define the **Residual Sum of Squares** (RSS) as

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e_1^2 + \cdots + e_n^2$$

  - Note that $\mathrm{RSS} = n \cdot \mathrm{MSE}$.

Residual Sum of Squares

- Define the **Residual Sum of Squares** (RSS) as

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e_1^2 + \cdots + e_n^2$$

  - Note that $\mathrm{RSS} = n \cdot \mathrm{MSE}$.

- Using calculus or linear algebra, we can show that RSS is minimized when

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

Section 2

Inference for Linear Models

Statistical Inference

- **Goal**: Use *statistics* calculated from data to make estimates about unknown *parameters*

Statistical Inference

- **Goal**: Use *statistics* calculated from data to make estimates about unknown *parameters*
- **Parameters**: β_0, β_1

Statistical Inference

- **Goal**: Use *statistics* calculated from data to make estimates about unknown *parameters*
- **Parameters**: β_0, β_1
- **Statistics**: $\hat{\beta}_0$, $\hat{\beta}_1$

Statistical Inference

- **Goal**: Use *statistics* calculated from data to make estimates about unknown *parameters*
- **Parameters**: β_0, β_1
- **Statistics**: $\hat{\beta}_0$, $\hat{\beta}_1$
- **Tools**: confidence intervals, hypothesis tests

Statistical Inference

- **Goal**: Use *statistics* calculated from data to make estimates about unknown *parameters*
- **Parameters**: β_0, β_1
- **Statistics**: $\hat{\beta}_0$, $\hat{\beta}_1$
- **Tools**: confidence intervals, hypothesis tests
- **The Problems**: Our model will change if built using a different random sample. So in addition to estimates, we need to know about variability

The Confidence Interval

- Confidence Intervals give estimates **and** express an amount of uncertainty we have about those estimates

The Confidence Interval

- Confidence Intervals give estimates **and** express an amount of uncertainty we have about those estimates
- A C-level confidence interval for a parameter θ using the statistic $\hat{\theta}$ takes the form

$$\hat{\theta} \pm t_C^* \cdot \mathrm{SE}(\hat{\theta})$$

The Confidence Interval

- Confidence Intervals give estimates **and** express an amount of uncertainty we have about those estimates
- A C-level confidence interval for a parameter θ using the statistic $\hat{\theta}$ takes the form

$$\hat{\theta} \pm t_C^* \cdot \mathrm{SE}(\hat{\theta})$$

- The value t_C^* is the $1 - (1 - C)/2$ quantile for the sampling distribution of $\hat{\theta}$
    - i.e. if $\hat{\theta}$ is approximately Normally distributed and $C = .95$, then $t_C^* \approx 2$.

The Confidence Interval

- Confidence Intervals give estimates **and** express an amount of uncertainty we have about those estimates
- A C-level confidence interval for a parameter θ using the statistic $\hat{\theta}$ takes the form

$$\hat{\theta} \pm t_C^* \cdot \mathrm{SE}(\hat{\theta})$$

- The value t_C^* is the $1-(1-C)/2$ quantile for the sampling distribution of $\hat{\theta}$
  - i.e. if $\hat{\theta}$ is approximately Normally distributed and $C = .95$, then $t_C^* \approx 2$.
- The value $\mathrm{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$, or the standard deviation of the sampling distribution

Common Regression Assumptions

In order to safely use simple linear regression, we require these assumptions:

1. Y is related to X by a simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Common Regression Assumptions

In order to safely use simple linear regression, we require these assumptions:

1. Y is related to X by a simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

2. The errors $e_1, e_2, \ldots, e_n$ are independent of one another.

Common Regression Assumptions

In order to safely use simple linear regression, we require these assumptions:

1. Y is related to X by a simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

2. The errors $e_1, e_2, \ldots, e_n$ are independent of one another.
3. The errors have a common variance $\mathrm{Var}(\epsilon) = \sigma^2$.

Common Regression Assumptions

In order to safely use simple linear regression, we require these assumptions:

1. Y is related to X by a simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

2. The errors $e_1, e_2, \ldots, e_n$ are independent of one another.
3. The errors have a common variance $\mathrm{Var}(\epsilon) = \sigma^2$.
4. The errors are normally distributed: $\epsilon \sim N(0, \sigma^2)$

Common Regression Assumptions

In order to safely use simple linear regression, we require these assumptions:

1. Y is related to X by a simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

2. The errors $e_1, e_2, \ldots, e_n$ are independent of one another.
3. The errors have a common variance $\mathrm{Var}(\epsilon) = \sigma^2$.
4. The errors are normally distributed: $\epsilon \sim N(0, \sigma^2)$

If one or more of these conditions do not hold, our predictions may not be accurate and we should be skeptical of inferential claims.

The Sampling Distribution of $\hat{\beta}_1$

Assume the following true model:

$$f(x) = 12 + .7x; \epsilon \sim N(0, 4)$$

The Sampling Distribution of $\hat{\beta}_1$

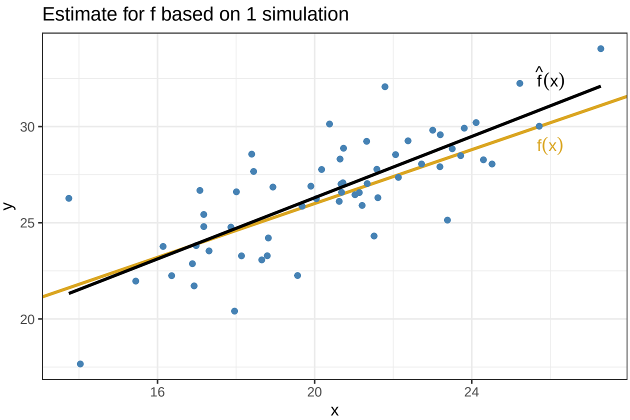
Assume the following true model:

$$f(x) = 12 + .7x; \epsilon \sim N(0, 4)$$

The Sampling Distribution of $\hat{\beta}_1$

Assume the following true model:

$$f(x) = 12 + .7x; \epsilon \sim N(0, 4)$$

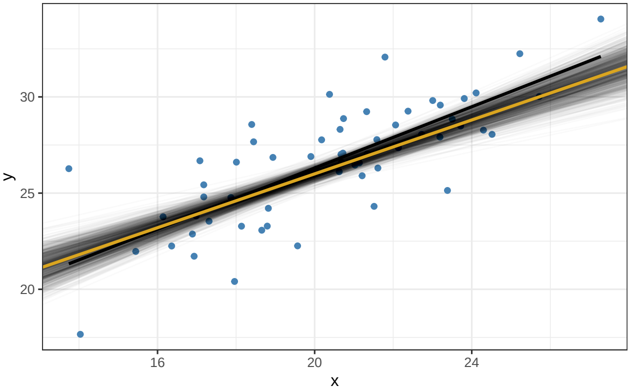


The Sampling Distribution of $\hat{\beta}_1$

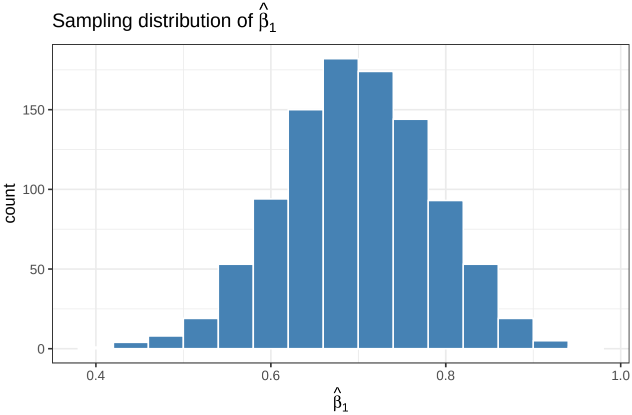
Assume the following true model:

$$f(x) = 12 + .7x; \epsilon \sim N(0, 4)$$

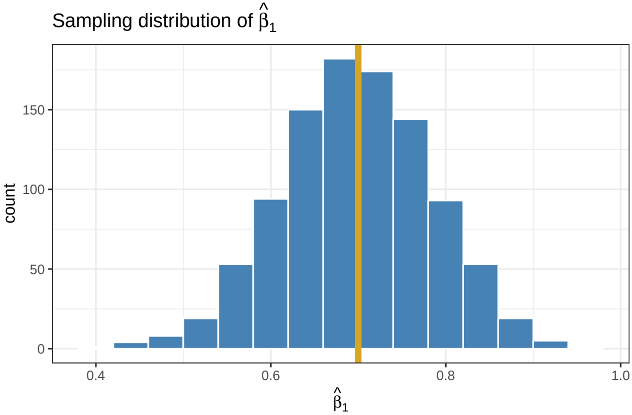
Estimates for f based on 1000 simulations



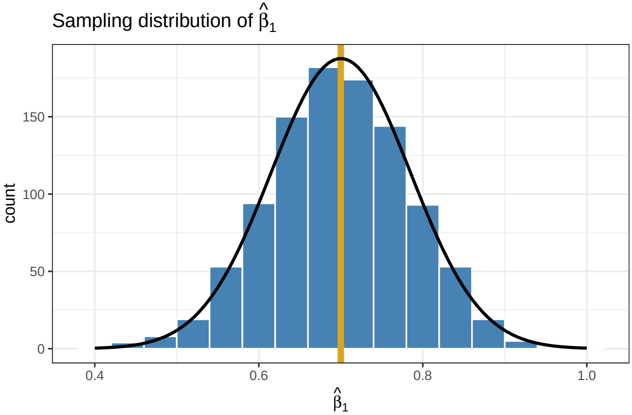
The Sampling Distribution of $\hat{\beta}_1$



The Sampling Distribution of $\hat{\beta}_1$



The Sampling Distribution of $\hat{\beta}_1$



The Sampling Distribution of $\hat{\beta}_1$

The Sampling Distribution has the following characteristics:

1. Centered at β_1, i.e. $E(\hat{\beta}_1) = \beta$.

The Sampling Distribution of $\hat{\beta}_1$

The Sampling Distribution has the following characteristics:

1. Centered at β_1, i.e. $E(\hat{\beta}_1) = \beta$.
2. $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$.
    - where $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

The Sampling Distribution of $\hat{\beta}_1$

The Sampling Distribution has the following characteristics:

1. Centered at β_1, i.e. $E(\hat{\beta}_1) = \beta$.
2. $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$.
   - where $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$
3. $\hat{\beta}_1 | X \sim N(\beta_1, \frac{\sigma^2}{S_{XX}})$.

Approximating the Sampling Dist. of $\hat{\beta}_1$

- Our best estimate of β_1 is $\hat{\beta}_1$ (since the expected value $\hat{\beta}_1$ is β_1)

Approximating the Sampling Dist. of $\hat{\beta}_1$

- Our best estimate of β_1 is $\hat{\beta}_1$ (since the expected value $\hat{\beta}_1$ is β_1)
- However, since we have to estimate σ with the Residual Standard Error $\hat{\sigma} = \mathrm{RSE} = \sqrt{\mathrm{RSS}/n-2}$, the distribution of $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}$ isn't Normal...

Approximating the Sampling Dist. of $\hat{\beta}_1$

- Our best estimate of β_1 is $\hat{\beta}_1$ (since the expected value $\hat{\beta}_1$ is β_1)
- However, since we have to estimate σ with the Residual Standard Error $\hat{\sigma} = \mathrm{RSE} = \sqrt{\mathrm{RSS}/n-2}$, the distribution of $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}$ isn't Normal...
- Instead, it is the t-distribution with $n-2$ degrees of freedom.

Approximating the Sampling Dist. of $\hat{\beta}_1$

- Our best estimate of β_1 is $\hat{\beta}_1$ (since the expected value $\hat{\beta}_1$ is β_1)
- However, since we have to estimate σ with the Residual Standard Error $\hat{\sigma} = \mathrm{RSE} = \sqrt{\mathrm{RSS}/n-2}$, the distribution of $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}$ isn't Normal. . .
- Instead, it is the t-distribution with $n-2$ degrees of freedom.
- Our confidence interval for $\hat{\beta}_1$ is thus

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1) \qquad \text{where } SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{XX}}}$$

Approximating the Sampling Dist. of $\hat{\beta}_1$

- Our best estimate of β_1 is $\hat{\beta}_1$ (since the expected value $\hat{\beta}_1$ is β_1)
- However, since we have to estimate σ with the Residual Standard Error $\hat{\sigma} = \mathrm{RSE} = \sqrt{\mathrm{RSS}/n-2}$, the distribution of $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}$ isn't Normal. . .
- Instead, it is the t-distribution with $n-2$ degrees of freedom.
- Our confidence interval for $\hat{\beta}_1$ is thus

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1) \qquad \text{where } SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{XX}}}$$

Interpretation We are *95% confident* that the true slope relating x and y lies between lower and upper bound of this interval.

Hypothesis test for $\hat{\beta}_1$

Suppose we are interested in testing the claim that the slope is zero.

$$H_0 : \beta_1^0 = 0 \qquad \text{vs} \quad H_A : \beta_1^0 \neq 0$$

Hypothesis test for $\hat{\beta}_1$

Suppose we are interested in testing the claim that the slope is zero.

$$H_0 : \beta_1^0 = 0 \qquad \text{vs} \quad H_A : \beta_1^0 \neq 0$$

- Consider the statistic t given by

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

- Then t will be t-distributed with $n-2$ degrees of freedom and $SE(\hat{\beta}_1)$ calculated the same as in the CI.

Hypothesis test for $\hat{\beta}_1$

Suppose we are interested in testing the claim that the slope is zero.

$$H_0 : \beta_1^0 = 0 \qquad \text{vs} \quad H_A : \beta_1^0 \neq 0$$

- Consider the statistic t given by

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

- Then t will be t-distributed with $n - 2$ degrees of freedom and $SE(\hat{\beta}_1)$ calculated the same as in the CI.
- The p-value for an observed test statistic t is the probability that a randomly chosen value from the t-dist is larger in absolute value than $|t|$.

Hypothesis test for $\hat{\beta}_1$

Suppose we are interested in testing the claim that the slope is zero.

$$H_0 : \beta_1^0 = 0 \qquad \text{vs} \quad H_A : \beta_1^0 \neq 0$$

- Consider the statistic t given by

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

- Then t will be t-distributed with $n-2$ degrees of freedom and $SE(\hat{\beta}_1)$ calculated the same as in the CI.
- The p-value for an observed test statistic t is the probability that a randomly chosen value from the t-dist is larger in absolute value than $|t|$.
- An observed t with p-value less than a desired significance level (often $\alpha = 0.05$) gives good evidence against the null-hypothesis.

Inference for other parameters in the linear model

- We can also perform inference for β_0, although it is often less interesting in practice (why?)

Inference for other parameters in the linear model

- We can also perform inference for β_0, although it is often less interesting in practice (why?)
  - We proceed as before, using a t distribution to estimate the sampling distribution of $\hat{\beta}_0$.
  - However, the SE of $\hat{\beta}_0$ is

$$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{S_{XX}} \right]$$

Inference for other parameters in the linear model

- We can also perform inference for β_0, although it is often less interesting in practice (why?)
  - We proceed as before, using a t distribution to estimate the sampling distribution of $\hat{\beta}_0$.
  - However, the SE of $\hat{\beta}_0$ is

$$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{S_{XX}} \right]$$

- Inference is even possible for combinations of β_0 and β_1 (i.e $\beta_0 + \beta_1 x$ for any fixed value of x)

Inference for other parameters in the linear model

- We can also perform inference for β_0, although it is often less interesting in practice (why?)
  - We proceed as before, using a t distribution to estimate the sampling distribution of $\hat{\beta}_0$.
  - However, the SE of $\hat{\beta}_0$ is
$$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{S_{XX}}\right]$$
- Inference is even possible for combinations of β_0 and β_1 (i.e $\beta_0 + \beta_1 x$ for any fixed value of x)
  - Why might we want to obtain a confidence interval for $\beta_0 + \beta_1 x$?

Inference for other parameters in the linear model

- We can also perform inference for β_0, although it is often less interesting in practice (why?)
    - We proceed as before, using a t distribution to estimate the sampling distribution of $\hat{\beta}_0$.
    - However, the SE of $\hat{\beta}_0$ is
$$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{S_{XX}} \right]$$
- Inference is even possible for combinations of β_0 and β_1 (i.e $\beta_0 + \beta_1 x$ for any fixed value of x)
    - Why might we want to obtain a confidence interval for $\beta_0 + \beta_1 x$?
    - The associated statistic is again t-distributed, although with more complicated SE.
    - For details, see DeGroot and Schervish "Probability and Statistics" (or take Math 392)