Lesson 18:
Chapter 10 Section 1

Caleb Moxley

BSC Mathematics

9 November 15

§10.1 Simple Linear Regression

In §2, we looked at the least-squares regression line.

§10.1 Simple Linear Regression

In §2, we looked at the least-squares regression line. In this chapter, we'll investigate how well the least-squares regression line models the data it is fitted to.

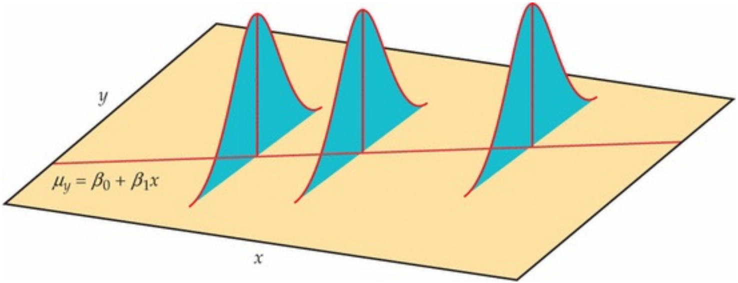
§10.1 Simple Linear Regression

In §2, we looked at the least-squares regression line. In this chapter, we'll investigate how well the least-squares regression line models the data it is fitted to. We work under the prevailing assumption that **for each value of the explanatory variable x, the response variable y is normally distributed with a mean that depends on x.**

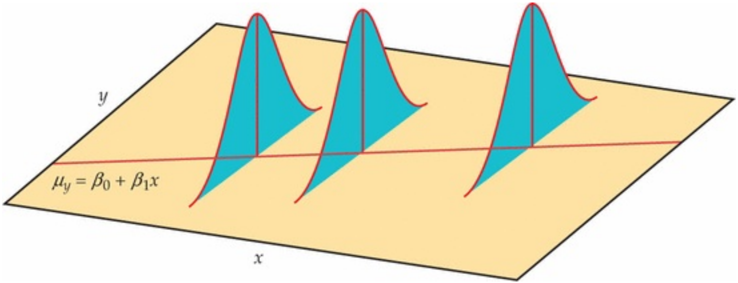
§10.1 Simple Linear Regression

In §2, we looked at the least-squares regression line. In this chapter, we'll investigate how well the least-squares regression line models the data it is fitted to. We work under the prevailing assumption that **for each value of the explanatory variable x, the response variable y is normally distributed with a mean that depends on x.** We further assume that the variances of these distributions are all the same and that the means of these distributions depend on x in a linear fashion.

§10.1 Simple Linear Regression

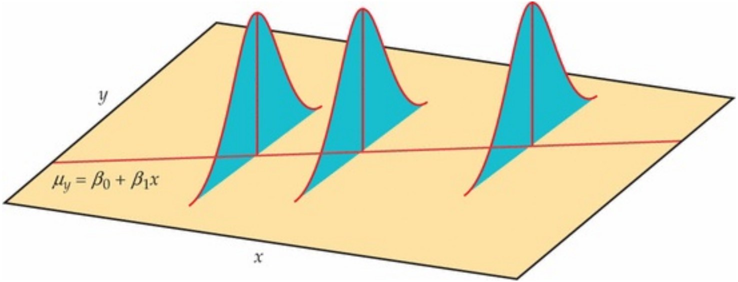


§10.1 Simple Linear Regression



Population regression line: $\mu_y = \beta_0 + \beta_1 x$.

§10.1 Simple Linear Regression



Population regression line: $\mu_y = \beta_0 + \beta_1 x$.
Sample regression line: $\hat{y} = b_0 + b_1 x$.

§10.1 Simple Linear Regression

Just as when we were first talking about regression, it's important to start with a graphical display of the data to guarantee that there is basic visual evidence to support a linear relationship between the data.

§10.1 Simple Linear Regression

Just as when we were first talking about regression, it's important to start with a graphical display of the data to guarantee that there is basic visual evidence to support a linear relationship between the data. If the data exhibits a non-linear relationship, don't perform inference for linear regression!

§10.1 Simple Linear Regression

Definition (simple linear regression model)

Given n observations of the explanatory variable x and response variable y,

$$(x_1, y_1), \ldots (x_n, y_n),$$

the **statistical model for simple linear regression** states that the responses y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Here, $\beta_0 + \beta_1 x_i = \mu_{y_i}$. The deviations ϵ_i are assumed to be independent $N(0, \sigma)$. The parameters of the model are β_0, β_1, and σ.

§10.1 Simple Linear Regression

A few warnings and tips:

- The regression line and the inferential result is only valid for the range of the x values, i.e. the explanatory values.

§10.1 Simple Linear Regression

A few warnings and tips:

- The regression line and the inferential result is only valid for the range of the x values, i.e. the explanatory values.
- β_0 and β_1 are the y-intercept and the slope of the population regression line respectively.

§10.1 Simple Linear Regression

A few warnings and tips:

- The regression line and the inferential result is only valid for the range of the x values, i.e. the explanatory values.
- β_0 and β_1 are the y-intercept and the slope of the population regression line respectively.
- We learned how to compute the statistics corresponding to β_0 and β_1 in Chapter 2.

Recall:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

Recall:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

$$b_1 = r \frac{s_y}{s_x},$$

Recall:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

$$b_1 = r \frac{s_y}{s_x},$$

and

$$b_0 = \bar{y} - b_1 \bar{x}.$$

We have not yet discussed the parameter residuals ϵ_i. The observed residuals e_i can be calculated using the definition below.

§10.1 Simple Linear Regression

We have not yet discussed the parameter residuals ϵ_i. The observed residuals e_i can be calculated using the definition below.

Definition (residual)

$e_i =$ observed response$-$predicted response$= y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$.

§10.1 Simple Linear Regression

We have not yet discussed the parameter residuals ϵ_i. The observed residuals e_i can be calculated using the definition below.

Definition (residual)

$e_i =$ observed response$-$predicted response$= y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$.

Example ($\hat{y} = 12.5 - 1.5x$)

What's the residual for the point (5,6)?

§10.1 Simple Linear Regression

We have not yet discussed the parameter residuals ϵ_i. The observed residuals e_i can be calculated using the definition below.

Definition (residual)

e_i = observed response−predicted response= $y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$.

Example ($\hat{y} = 12.5 - 1.5x$)

What's the residual for the point (5,6)? It's 1.

For simple linear regression, we can also estimate the standard deviation of the residuals σ:

For simple linear regression, we can also estimate the standard deviation of the residuals σ:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}.$$

§10.1 Simple Linear Regression

For simple linear regression, we can also estimate the standard deviation of the residuals σ:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}.$$

In this calculation of a sample variance, we divide by $n - 2$ to make s^2 an unbiased estimator or σ^2.

§10.1 Simple Linear Regression

For simple linear regression, we can also estimate the standard deviation of the residuals σ:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}.$$

In this calculation of a sample variance, we divide by $n-2$ to make s^2 an unbiased estimator or σ^2. We usually use technology for most of these calculations.

Definition (CI & significance test for regression slope)

The **level C CI for** β_1 is $b_1 \pm t^* \mathrm{SE}_{b_1}$. Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Definition (CI & significance test for regression slope)

The **level C CI for** β_1 is $b_1 \pm t^* \mathrm{SE}_{b_1}$. Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.
To test the hypothesis $H_0 : \beta_1 = 0$, compute the test statistic

$$t = \frac{b_1}{\mathrm{SE}_{b_1}}.$$

The degrees of freedom are $n-2$, and the P-value for the test against

$$\begin{array}{lcr} H_a : \beta_1 > 0 & \text{is} & P(T > t) \\ H_a : \beta_1 < 0 & \text{is} & P(T < t) \\ H_a : \beta_1 \neq 0 & \text{is} & 2P(T > |t|) \end{array}$$

§10.1 Simple Linear Regression

In the previous slide,

$$\mathrm{SE}_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}},$$

In the previous slide,

$$\mathrm{SE}_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}},$$

and the corresponding test for b_0 is exactly the same, using

$$\mathrm{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}.$$

§10.1 Simple Linear Regression

We may not want a confidence interval for the slope or intercept, however. In general, we would probably be more concerned with a confidence interval for the mean response!

We may not want a confidence interval for the slope or intercept, however. In general, we would probably be more concerned with a confidence interval for the mean response!

Definition (confidence interval for mean response)

The **level C CI for the mean response** μ_y when x takes the particular value x^* is

$$\hat{\mu_y} \pm t^* \mathrm{SE}_{\hat{\mu}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

§10.1 Simple Linear Regression

We may not want a confidence interval for the slope or intercept, however. In general, we would probably be more concerned with a confidence interval for the mean response!

Definition (confidence interval for mean response)

The **level C CI for the mean response** μ_y when x takes the particular value x^* is

$$\hat{\mu_y} \pm t^* \mathrm{SE}_{\hat{\mu}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Here,

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$

§10.1 Simple Linear Regression

We may not want a confidence interval for the slope or intercept, however. In general, we would probably be more concerned with a confidence interval for the mean response!

Definition (confidence interval for mean response)

The **level C CI for the mean response** μ_y when x takes the particular value x^* is

$$\hat{\mu_y} \pm t^* \mathrm{SE}_{\hat{\mu}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Here,

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}.$$

Note: $\mathrm{SE}_{\hat{\mu}}$ only takes into account the variability due to the model — it does not include the variability due to the deviations.

If we want a confidence interval which includes the variability due to deviations for a future observation, then we're actually creating a confidence interval for y and not μ_y.

If we want a confidence interval which includes the variability due to deviations for a future observation, then we're actually creating a confidence interval for y and not μ_y.

Definition (prediction interval for future observation)

The **level C CI for a future observation** y from the subpopulation corresponding to x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

§10.1 Simple Linear Regression

If we want a confidence interval which includes the variability due to deviations for a future observation, then we're actually creating a confidence interval for y and not μ_y.

Definition (prediction interval for future observation)

The **level C CI for a future observation** y from the subpopulation corresponding to x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Here,

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}.$$

§10.1 Simple Linear Regression

If we want a confidence interval which includes the variability due to deviations for a future observation, then we're actually creating a confidence interval for y and not μ_y.

Definition (prediction interval for future observation)

The **level C CI for a future observation** y from the subpopulation corresponding to x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Here,

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$

Note: Prediction intervals are wider than confidence intervals for mean responses.

§10.1 Simple Linear Regression

If we want a confidence interval which includes the variability due to deviations for a future observation, then we're actually creating a confidence interval for y and not μ_y.

Definition (prediction interval for future observation)

The **level C CI for a future observation** y from the subpopulation corresponding to x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}.$$

Here, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Here,

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}.$$

Note: Prediction intervals are wider than confidence intervals for mean responses. Why?

§10.1 Simple Linear Regression

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8

The regression line (where age is the explanatory variable) is

$$\hat{y} = -0.153125 + 0.90625x.$$

§10.1 Simple Linear Regression

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8

The regression line (where age is the explanatory variable) is

$$\hat{y} = -0.153125 + 0.90625x.$$

What's the test statistic to test the claim that the slope is not zero?

§10.1 Simple Linear Regression

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8

The regression line (where age is the explanatory variable) is

$$\hat{y} = -0.153125 + 0.90625x.$$

What's the test statistic to test the claim that the slope is not zero? To calculate this, we need s and then SE_{b_1}.

§10.1 Simple Linear Regression

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8
$\hat{y}$	1.2063	1.6594	1.3875	1.6594	0.9344	0.7531
e_i^2	0.0113	0.0035	0.0127	0.0017	0.0012	0.0022
$x - \bar{x}$	-0.0667	0.4333	0.1333	0.4333	-0.3667	-0.5667

Thus, $s = \sqrt{\frac{0.0113+0.0035+0.0127+0.0017+0.0012+0.0022}{6-2}} \approx 0.0901.$

§10.1 Simple Linear Regression

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8
$\hat{y}$	1.2063	1.6594	1.3875	1.6594	0.9344	0.7531
e_i^2	0.0113	0.0035	0.0127	0.0017	0.0012	0.0022
$x - \bar{x}$	-0.0667	0.4333	0.1333	0.4333	-0.3667	-0.5667

Thus, $s = \sqrt{\frac{0.0113+0.0035+0.0127+0.0017+0.0012+0.0022}{6-2}} \approx 0.0901$. And so $\text{SE}_{b_1} = \frac{s}{\sqrt{\sum(x-\bar{x})^2}} = \frac{0.0901}{\sqrt{0.8533}} = 0.0975$.

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8
$\hat{y}$	1.2063	1.6594	1.3875	1.6594	0.9344	0.7531
e_i^2	0.0113	0.0035	0.0127	0.0017	0.0012	0.0022
$x - \bar{x}$	-0.0667	0.4333	0.1333	0.4333	-0.3667	-0.5667

Thus, our t test statistic is $t = \frac{b_1}{\text{SE}_{b_1}} = \frac{0.90625}{0.0975} = 9.29$.

§10.1 Simple Linear Regression

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8

Now, we have $t = 9.29$, let's get the P-value for our test of the claim that $b_1 = 0$, using $\alpha = 0.01$.

We're going to work with the same basic example:

Example (tree heights & ages)

Below is the height and ages of several trees on a holiday tree farm. Ages are in years. Heights are in meters.

Age	1.5	2.0	1.7	2.0	1.2	1.0
Height	1.1	1.6	1.5	1.7	0.9	0.8

Now, we have $t = 9.29$, let's get the P-value for our test of the claim that $b_1 = 0$, using $\alpha = 0.01$. We calculate $2P(T > 9.29) = 0.0007$. Thus, we reject H_0 and do not support our claim.

§10.1 Simple Linear Regression

Obviously, these hypothesis tests involve a lot do calculations! Minitab, TI calculators, and technology **is your friend** for these!

Obviously, these hypothesis tests involve a lot do calculations! Minitab, TI calculators, and technology **is your friend** for these! Let's do a confidence interval for the mean response.

Obviously, these hypothesis tests involve a lot do calculations! Minitab, TI calculators, and technology **is your friend** for these! Let's do a confidence interval for the mean response.

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

§10.1 Simple Linear Regression

Obviously, these hypothesis tests involve a lot do calculations! Minitab, TI calculators, and technology **is your friend** for these! Let's do a confidence interval for the mean response.

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

Well,

$$\hat{\mu}_y = -0.153125 + 0.90625(1.5) = 1.2063,$$

Obviously, these hypothesis tests involve a lot do calculations! Minitab, TI calculators, and technology **is your friend** for these! Let's do a confidence interval for the mean response.

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

Well,

$$\hat{\mu}_y = -0.153125 + 0.90625(1.5) = 1.2063,$$

and we have $\text{SE}_{\hat{\mu}} =$

$$s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 0.0901\sqrt{\frac{1}{6} + \frac{(1.5 - 1.5667)^2}{0.8533}} = 0.0374$$

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

Further, we have $t^* = 2.131$.

§10.1 Simple Linear Regression

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

Further, we have $t^* = 2.131$. Thus, our CI is

§10.1 Simple Linear Regression

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

Further, we have $t^* = 2.131$. Thus, our CI is

$$\hat{\mu}_y \pm t^* \mathrm{SE}_{\hat{\mu}} =$$

Example (CI for mean response)

Using the same example, compute a 90% CI for the mean response when the age of the tree is 1.5 years.

Further, we have $t^* = 2.131$. Thus, our CI is

$$\hat{\mu}_y \pm t^*\mathrm{SE}_{\hat{\mu}} = 1.2063 \pm 2.132(0.0374) = (1.127, 1.286).$$

Example (prediction interval)

Using the same example, compute a 90% prediction interval for $\hat{y}$ when the age of the tree is 1.5 years.

Example (prediction interval)

Using the same example, compute a 90% prediction interval for $\hat{y}$ when the age of the tree is 1.5 years.

The only thing that changes in this problem is that $\hat{y} = \hat{\mu}_y$, and we use $\mathrm{SE}_y =$

Example (prediction interval)

Using the same example, compute a 90% prediction interval for $\hat{y}$ when the age of the tree is 1.5 years.

The only thing that changes in this problem is that $\hat{y} = \hat{\mu}_y$, and we use $\mathrm{SE}_y =$

$$s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 0.0901\sqrt{1 + \frac{1}{6} + \frac{(1.5 - 1.5667)^2}{0.8533}}$$

$$= 0.0975.$$

Example (prediction interval)

Using the same example, compute a 90% prediction interval for $\hat{y}$ when the age of the tree is 1.5 years.

The only thing that changes in this problem is that $\hat{y} = \hat{\mu}_y$, and we use $\mathrm{SE}_y =$

$$s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 0.0901\sqrt{1 + \frac{1}{6} + \frac{(1.5 - 1.5667)^2}{0.8533}}$$

$$= 0.0975.$$

Thus, our prediction interval is

Example (prediction interval)

Using the same example, compute a 90% prediction interval for $\hat{y}$ when the age of the tree is 1.5 years.

The only thing that changes in this problem is that $\hat{y} = \hat{\mu}_y$, and we use $\mathrm{SE}_y =$

$$s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 0.0901\sqrt{1 + \frac{1}{6} + \frac{(1.5 - 1.5667)^2}{0.8533}}$$

$$= 0.0975.$$

Thus, our prediction interval is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}} =$$

§10.1 Simple Linear Regression

Example (prediction interval)

Using the same example, compute a 90% prediction interval for $\hat{y}$ when the age of the tree is 1.5 years.

The only thing that changes in this problem is that $\hat{y} = \hat{\mu}_y$, and we use $\text{SE}_y =$

$$s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 0.0901\sqrt{1 + \frac{1}{6} + \frac{(1.5 - 1.5667)^2}{0.8533}}$$

$$= 0.0975.$$

Thus, our prediction interval is

$$\hat{y} \pm t^* \text{SE}_{\hat{y}} = 1.2063 \pm 2.132(0.0975) = (0.998, 1.414).$$

Notice: Predictions intervals are wider than CI for mean response.

Notice: Predictions intervals are wider than CI for mean response. This makes sense: means vary less than individual observations!

§10.1 Simple Linear Regression

Notice: Predictions intervals are wider than CI for mean response. This makes sense: means vary less than individual observations!

I highly recommend using Minitab as much as possible for these problems!