Lecture 20 - Regression: Inference, Outliers, and Intervals

Sta102 / BME102

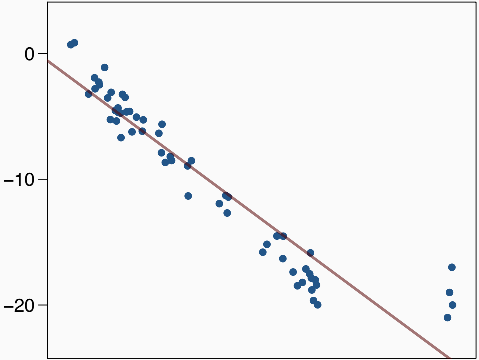
April 13, 2016

Colin Rundel

Types of outliers in linear regression

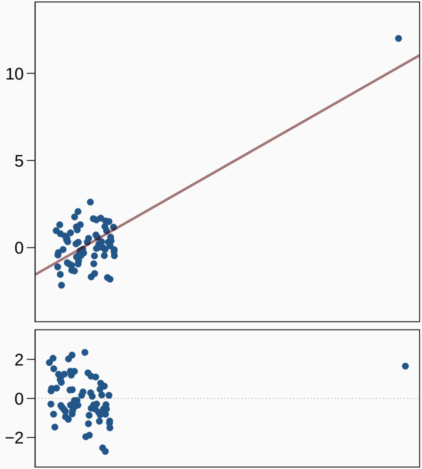
Types of outliers

Is regression robust? Think about how the regression line would change with and without “outlier(s)”.



Types of outliers

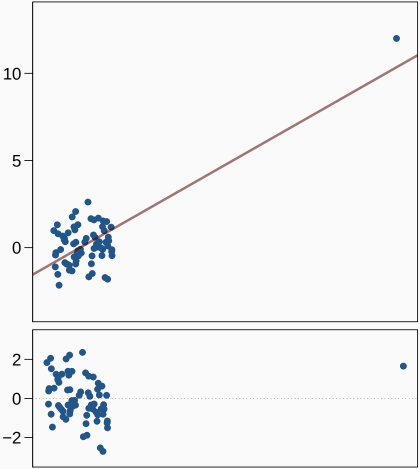
How does the following point influence the least squares line?



Types of outliers

How does the following point influence the least squares line?

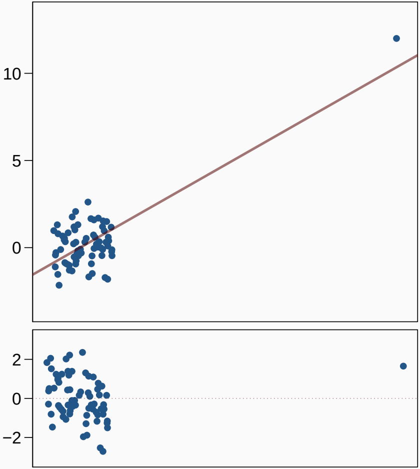
Without the outlier there is *no* relationship between X and Y ($\mathrm{Cor}(X, Y) = 0$).



Types of outliers

How does the following point influence the least squares line?

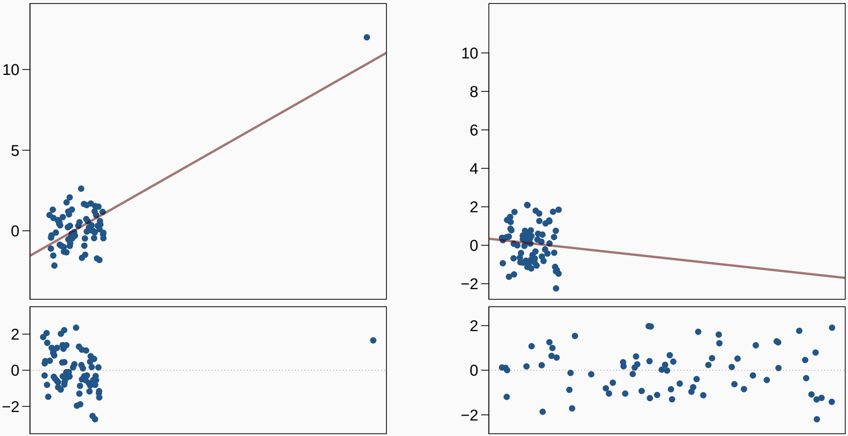
What would have happened if the outlier was directly above the other points?



With and without

$R = 0.72, R^2 = 0.522$

$R = -0.091, R^2 = 0.0083$



Some terminology

Outliers are points that fall away from the main body of other points - but not all outliers are created equal.

Some terminology

Outliers are points that fall away from the main body of other points - but not all outliers are created equal.

- Outliers that are horizontally distant from the body of other points are called *leverage* points.

Some terminology

Outliers are points that fall away from the main body of other points - but not all outliers are created equal.

- Outliers that are horizontally distant from the body of other points are called *leverage* points.
- Points with high leverage that actually influence the *slope* of the regression line are called *influential* points.

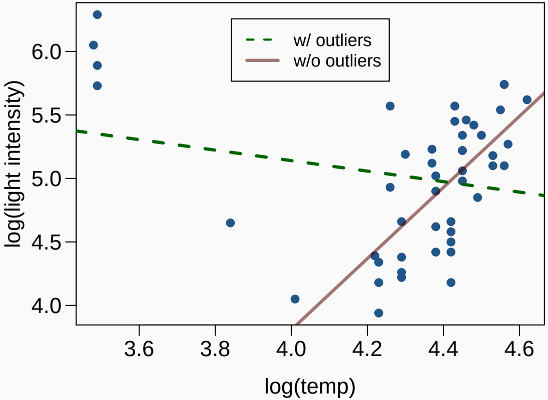
Some terminology

Outliers are points that fall away from the main body of other points - but not all outliers are created equal.

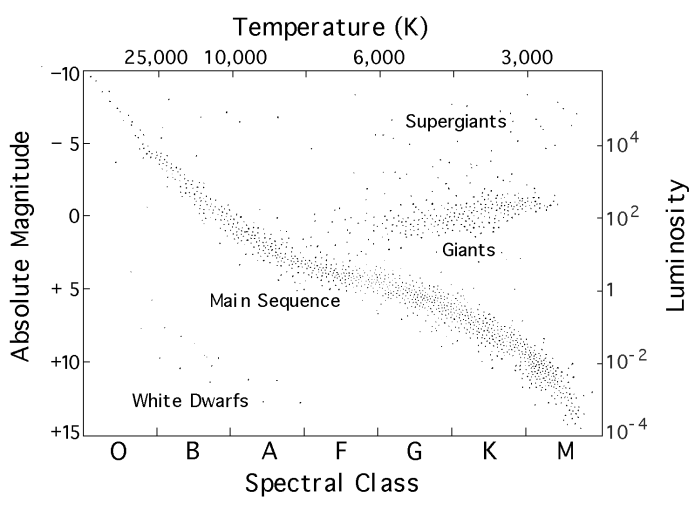
- Outliers that are horizontally distant from the body of other points are called *leverage* points.
- Points with high leverage that actually influence the *slope* of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential.

Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.

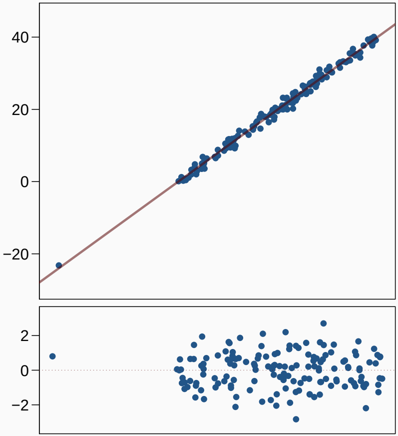


Hertzsprung-Russell Diagram



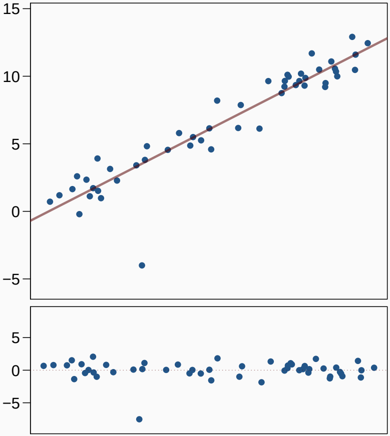
Types of outliers

Which type of outlier is displayed below?



Types of outliers

Which type of outlier is displayed below?



Recap

Are following statements true or false?

(1) Influential points always change the intercept of the regression line.

(2) Influential points always reduce R^2.

(3) It is much more likely for a high leverage point to be influential, than a low leverage point.

(4) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

Recap

Are following statements true or false?

(1) Influential points always change the intercept of the regression line. *False*

(2) Influential points always reduce R^2.

(3) It is much more likely for a high leverage point to be influential, than a low leverage point.

(4) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

Recap

Are following statements true or false?

(1) Influential points always change the intercept of the regression line. *False*

(2) Influential points always reduce R^2. *False*

(3) It is much more likely for a high leverage point to be influential, than a low leverage point.

(4) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

Recap

Are following statements true or false?

(1) Influential points always change the intercept of the regression line. *False*

(2) Influential points always reduce R^2. *False*

(3) It is much more likely for a high leverage point to be influential, than a low leverage point. *True*

(4) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

Recap

Are following statements true or false?

(1) Influential points always change the intercept of the regression line. *False*

(2) Influential points always reduce R^2. *False*

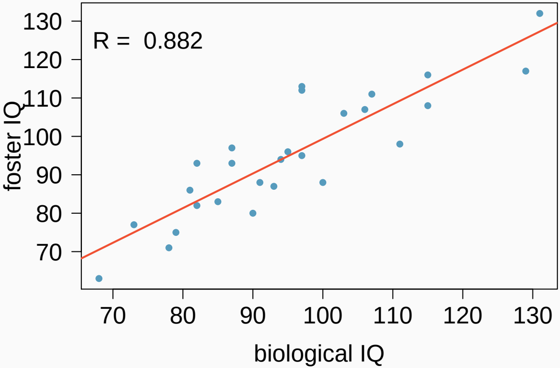
(3) It is much more likely for a high leverage point to be influential, than a low leverage point. *True*

(4) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear. *False*

Inference for linear regression

Nature vs. nurture?

In 1966 Cyril Burt published a paper called “The genetic determination of differences in intelligence: A study of monozygotic twins reared apart” The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Finding the regression line

	Foster IQ (y)	Biological IQ (x)
mean	$\bar{y} = 95.11$	$\bar{x} = 95.30$
sd	$s_y = 16.08$	$s_x = 15.73$
correlation	$R = 0.8819$	

Finding the regression line

	Foster IQ (y)	Biological IQ (x)
mean	$\bar{y} = 95.11$	$\bar{x} = 95.30$
sd	$s_y = 16.08$	$s_x = 15.73$
correlation	$R = 0.8819$	

$$b_1 = \frac{s_y}{s_x} R = \frac{16.08}{15.73} 0.8819 = 0.90$$

Finding the regression line

	Foster IQ (y)	Biological IQ (x)
mean	$\bar{y} = 95.11$	$\bar{x} = 95.30$
sd	$s_y = 16.08$	$s_x = 15.73$
correlation	$R = 0.8819$	

$$b_1 = \frac{s_y}{s_x} R = \frac{16.08}{15.73} 0.8819 = 0.90$$

$$b_0 = \bar{y} - b_1 \, \bar{x} = 95.11 - 0.90 \, 95.30 = 9.2$$

Regression Output

```
summary(lm(twins$Foster ~ twins$Biological))
```

```
## Call:
## lm(formula = twins$Foster ~ twins$Biological)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>♦t♦)
## (Intercept)       9.20760    9.29990   0.990    0.332
## twins$Biological  0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

Conditions for inference

In order to conduct *inference*, the following conditions must be met:

1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

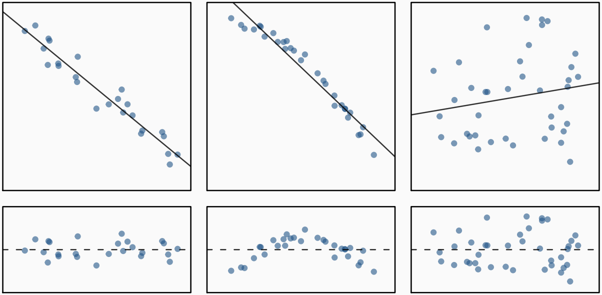
- The relationship between the explanatory and the response variable should be linear.

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class.

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class.
- Check using a *scatterplot* (x vs y) or a *residual plot* (x vs resid) .



Conditions: (2) Nearly normal residuals

- The residuals should follow a nearly normal distribution.

Conditions: (2) Nearly normal residuals

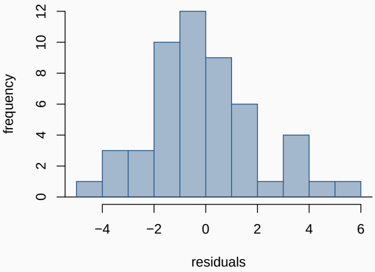
- The residuals should follow a nearly normal distribution.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

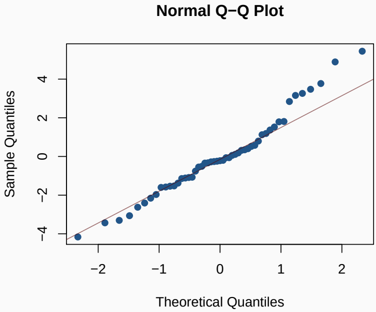
Conditions: (2) Nearly normal residuals

- The residuals should follow a nearly normal distribution.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Checked using a histogram or normal probability plot of residuals.

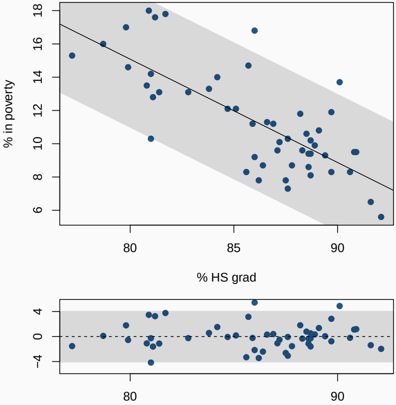
Conditions: (2) Nearly normal residuals

- The residuals should follow a nearly normal distribution.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Checked using a histogram or normal probability plot of residuals.



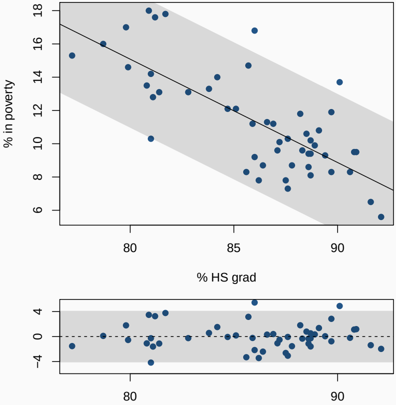


Conditions: (3) Constant variability



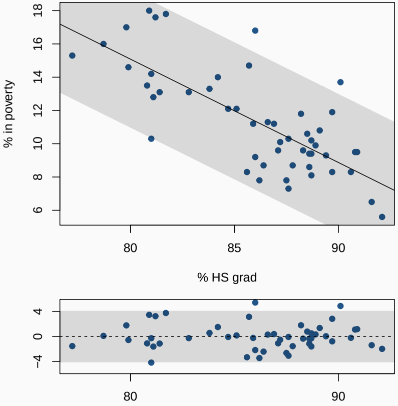
- The variability of the residuals from the least squares line should be constant.

Conditions: (3) Constant variability



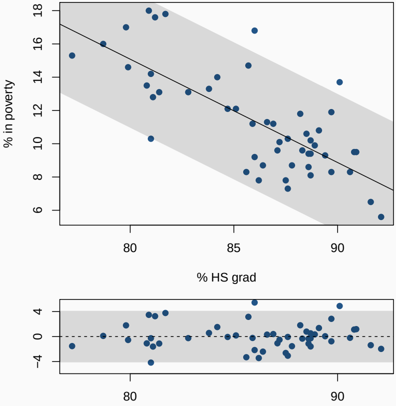
- The variability of the residuals from the least squares line should be constant.
- This implies that the variability of *any* region of the residual plot should be the same as any other region.

Conditions: (3) Constant variability



- The variability of the residuals from the least squares line should be constant.
- This implies that the variability of *any* region of the residual plot should be the same as any other region.
- Also referred to as *homoscedasticity* / *heteroscedasticity*.

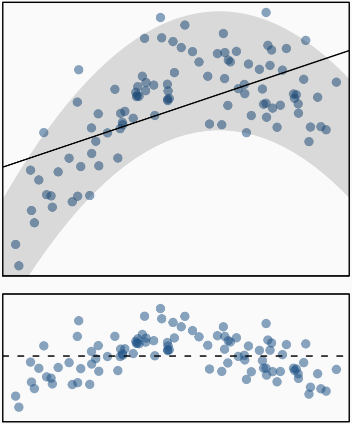
Conditions: (3) Constant variability



- The variability of the residuals from the least squares line should be constant.
- This implies that the variability of *any* region of the residual plot should be the same as any other region.
- Also referred to as *homoscedasticity* / *heteroscedasticity*.
- Check using a residuals plot.

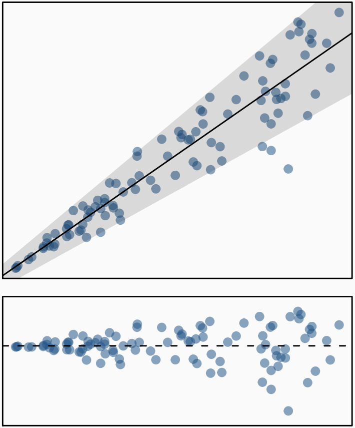
Checking conditions

What condition is this linear model violating?

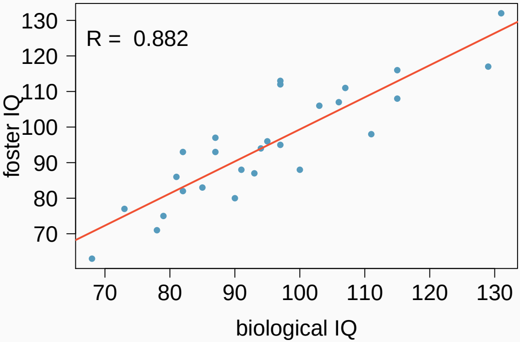


Checking conditions (II)

What condition is this linear model obviously violating?



Back to Nature vs nurture



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin.

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin.

What are the appropriate hypotheses?

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin.

What are the appropriate hypotheses?

First consider what the null hypothesis should be, if there is no relationship between the two variables what value of the slope would we expect to see?

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin.

What are the appropriate hypotheses?

First consider what the null hypothesis should be, if there is no relationship between the two variables what value of the slope would we expect to see?

H_0: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

We are interested in inference on β_1 which we estimate using the point estimate b_1.

It turns out that after normalizing our point estimate has a T distribution with $n-2$ degrees of freedom.

$$T_{df=n-2} = \frac{b_1 - \beta_1}{SE}$$

where,

$$SE_{b_1} = \frac{1}{\sqrt{n-1}} \frac{s_e}{s_x} = \frac{1}{\sqrt{n-1}} \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \epsilon_i^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Data + Regression Output

	Foster IQ (y)	Biological IQ (x)
mean	$\bar{y} = 95.11$	$\bar{x} = 95.30$
sd	$s_y = 16.08$	$s_x = 15.73$
n = 27	$R = 0.8819$	

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.20760    9.29990   0.990    0.332
## twins$Biological  0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

$$
\begin{aligned}
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25
\end{aligned}
$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

$$\begin{aligned} T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\ df &= 27 - 2 = 25 \\ \text{p-value} &= P(|t| > 9.36) < 0.01 \end{aligned}$$

Confidence interval for the slope

Since we know the sampling distribution we can also construct a confidence interval: *point estimate* $\pm$ *CV* $\times$ *SE*.

What is the correct 95% confidence interval for the slope parameter?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Confidence interval for the slope

Since we know the sampling distribution we can also construct a confidence interval: *point estimate* $\pm$ *CV* $\times$ *SE*.

What is the correct 95% confidence interval for the slope parameter?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$n = 27 \qquad df = 27 - 2 = 25 \qquad t^{\star}_{25} = 2.06$$

Confidence interval for the slope

Since we know the sampling distribution we can also construct a confidence interval: *point estimate* $\pm$ *CV* $\times$ *SE*.

What is the correct 95% confidence interval for the slope parameter?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$n = 27 \qquad df = 27 - 2 = 25 \qquad t^\star_{25} = 2.06$$

$$\begin{aligned} 95\%\ CI &= PE \pm CV \times SE \\ &= 0.9014 \pm 2.06 \times 0.0963 \\ &= (0.7,\ 1.1) \end{aligned}$$

Recap

Inference for the slope for a SLR model (only one explanatory variable):

Recap

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test for β_1:

$$T = \frac{b_1 - \beta_1}{SE_{b_1}} \qquad df = n - 2$$

Recap

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test for β_1:

$$T = \frac{b_1 - \beta_1}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval for β_1:

$$b_1 \pm t^{\star}_{df=n-2} \times SE_{b_1}$$

Recap

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test for β_1:

$$T = \frac{b_1 - \beta_1}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval for β_1:

$$b_1 \pm t^{\star}_{df=n-2} \times SE_{b_1}$$

- The null value is almost always 0, since we are usually checking for *any* relationship between the explanatory and the response variable.

Recap

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test for β_1:

$$T = \frac{b_1 - \beta_1}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval for β_1:

$$b_1 \pm t^\star_{df=n-2} \times SE_{b_1}$$

- The null value is almost always 0, since we are usually checking for *any* relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and the *two-tailed* p-value for the t-test of the slope when the null hypothesis is $\beta_1 = 0$.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you have population data.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you have population data.
- If you have a sample that is non-random (biased), the results will be unreliable.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you have population data.
- If you have a sample that is non-random (biased), the results will be unreliable.
- The ultimate goal is to have independent observations – and you know how to check for those by now.

Variability partitioning

- We considered the t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.

Variability partitioning

- We considered the t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.

- However, we can also consider the variability in y explained by x, compared to the unexplained variability.

Variability partitioning

- We considered the t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.
- However, we can also consider the variability in y explained by x, compared to the unexplained variability.
- *Partitioning* the variability in y to explained and unexplained variability is something we have already done *(ANOVA)*.

Sums of Squares

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

ANOVA Model:

$$\hat{y}_{ij} = \bar{y}_i$$

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2$$

$$SSG = \sum_i \sum_j (\bar{y}_i - \bar{y})^2$$

$$SSE = \sum_i \sum_j (y_i - \bar{y}_i)^2$$

Regression Model:

$$\hat{y}_i = b_0 + b_1 x_i$$

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2$$

$$SSG = \sum_i \sum_j (b_0 + b_1 x_i - \bar{y})^2$$

$$SSE = \sum_i \sum_j (y_i - b_0 + b_1 x_i)^2$$

ANOVA Table - Linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of Squares: $SS_{Tot} \quad = \quad \sum_i (y_i - \bar{y})^2 = 6724.66$ *(total variability in y)*

ANOVA Table - Linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of Squares:

$$SS_{Tot} = \sum_i (y_i - \bar{y})^2 = 6724.66 \textit{ (total variability in y)}$$

$$\begin{aligned} SS_{Err} &= \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 \\ &= 1493.53 \textit{ (unexplained variability in residuals)} \end{aligned}$$

ANOVA Table - Linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of Squares:

$$SS_{Tot} = \sum_i (y_i - \bar{y})^2 = 6724.66 \ \textit{(total variability in y)}$$

$$SS_{Err} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$$= 1493.53 \ \textit{(unexplained variability in residuals)}$$

$$SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2 = SS_{Tot} - SS_{Err}$$

$$= 5231.13 \ \textit{(explained variability in y)}$$

ANOVA Table - Linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of Squares:

$$SS_{Tot} = \sum_i (y_i - \bar{y})^2 = 6724.66 \text{ (total variability in y)}$$

$$SS_{Err} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$$= 1493.53 \text{ (unexplained variability in residuals)}$$

$$SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2 = SS_{Tot} - SS_{Err}$$

$$= 5231.13 \text{ (explained variability in y)}$$

Degrees of freedom:

$$df_{Tot} = n - 1 = 27 - 1 = 26$$

ANOVA Table - Linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of Squares: $SS_{Tot} = \sum_i (y_i - \bar{y})^2 = 6724.66$ *(total variability in y)*

$$SS_{Err} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$= 1493.53$ *(unexplained variability in residuals)*

$$SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2 = SS_{Tot} - SS_{Err}$$

$= 5231.13$ *(explained variability in y)*

Degrees of freedom: $df_{Tot} = n - 1 = 27 - 1 = 26$

$$df_{Reg} = 2 - 1 = 1$$

ANOVA Table - Linear Regression

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of Squares: $SS_{Tot} = \sum_i (y_i - \bar{y})^2 = 6724.66$ *(total variability in y)*

$$SS_{Err} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$= 1493.53$ *(unexplained variability in residuals)*

$$SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2 = SS_{Tot} - SS_{Err}$$

$= 5231.13$ *(explained variability in y)*

Degrees of freedom: $df_{Tot} = n - 1 = 27 - 1 = 26$

$$df_{Reg} = 2 - 1 = 1$$

$$df_{Res} = df_{Tot} - df_{Reg} = 26 - 1 = 25$$

ANOVA output - F-test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Mean sq.: $MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$

ANOVA output - F-test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Mean sq.:
$$MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$$
$$MS_{Err} = \frac{SS_{Err}}{df_{Err}} = \frac{1493.53}{25} = 59.74$$

ANOVA output - F-test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

$$\textit{Mean sq.: } MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$$

$$MS_{Err} = \frac{SS_{Err}}{df_{Err}} = \frac{1493.53}{25} = 59.74$$

$$\textit{F-statistic: } F_{(1,25)} = \frac{MS_{Reg}}{MS_{Err}} = 87.56 \quad \textit{(ratio of explained to unexplained variability)}$$

ANOVA output - F-test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Mean sq.: $$MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$$

$$MS_{Err} = \frac{SS_{Err}}{df_{Err}} = \frac{1493.53}{25} = 59.74$$

F-statistic: $$F_{(1,25)} = \frac{MS_{Reg}}{MS_{Err}} = 87.56$$ *(ratio of explained to unexplained variability)*

This test compares our regression model to an intercept *only* model - which is equivalent to a null hypothesis of $\beta_1 = 0$ and the alternative of $\beta_1 \neq 0$.

Regression Output

```
summary(lm(twins$Foster ~ twins$Biological))

## Call:
## lm(formula = twins$Foster ~ twins$Biological)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.20760    9.29990   0.990    0.332
## twins$Biological  0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

ANOVA output - R^2 calculation

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

ANOVA output - R^2 calculation

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

$$
\begin{aligned}
R^2 = \frac{\text{explained variability}}{\text{total variability}} &= \frac{SS_{Reg}}{SS_{Tot}} = \frac{5231.13}{6724.66} = 0.7779 \\
&= 1 - \frac{SS_{Err}}{SS_{Tot}} = 1 - \frac{1493.53}{6724.66} = 0.7779
\end{aligned}
$$