Unit 7: Multiple linear regression

3. Confidence and prediction intervals

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

April 13, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bitly.com/sta101sp15

Outline

1. Uncertainty of predictions
    1. Confidence intervals for average values
    2. Prediction intervals for specific predicted values
    3. Recap - CI vs. PI
    4. Confidence and prediction intervals for MLR

- Regression models are useful for making predictions for new observations not include in the original dataset.

- Regression models are useful for making predictions for new observations not include in the original dataset.
- If the model is good, the predictions should be close to the true value of the response variable for this observation, however it may not be exact, i.e. $\hat{y}$ might be different than y.

- Regression models are useful for making predictions for new observations not include in the original dataset.
- If the model is good, the predictions should be close to the true value of the response variable for this observation, however it may not be exact, i.e. $\hat{y}$ might be different than y.
- With any prediction we can (and should) also report a measure of uncertainty of the prediction:
  - Use a *confidence interval* for the uncertainty around the expected value of predictions (average of a group of predictions) -- e.g. predict the average final exam score of a group of students who scored the same on the midterm.
  - Use a *prediction interval* for the uncertainty around a single prediction -- e.g. predict the final exam score of one student with a given midterm score.

1. Uncertainty of predictions
    1. Confidence intervals for average values
    2. Prediction intervals for specific predicted values
    3. Recap - CI vs. PI
    4. Confidence and prediction intervals for MLR

Confidence intervals for average values

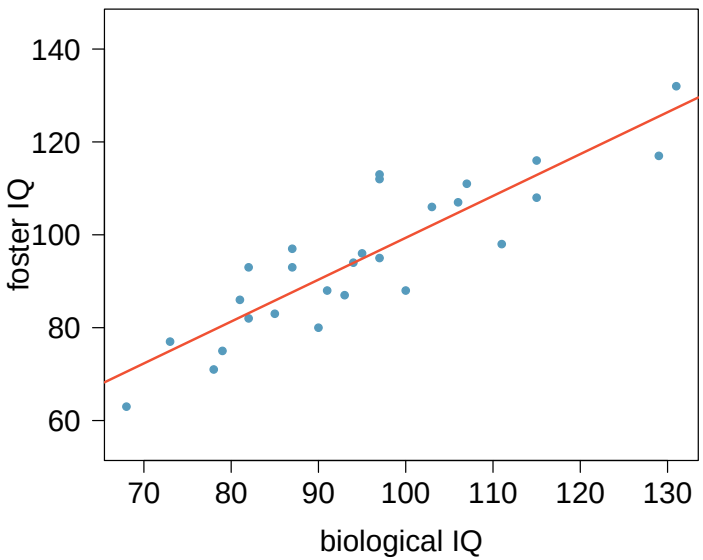
A confidence interval for the average (expected) value of y, $E(y)$, for a given $x^\star$, is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s is the standard deviation of the residuals, calculated as $\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$.

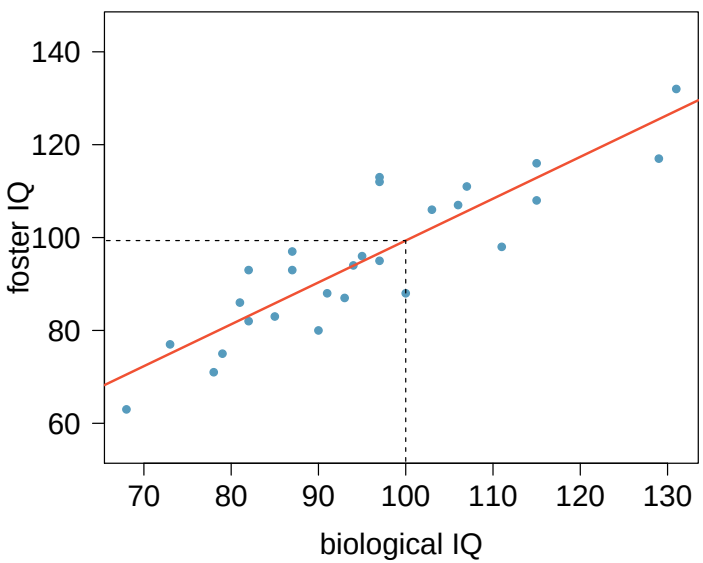
Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```



Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

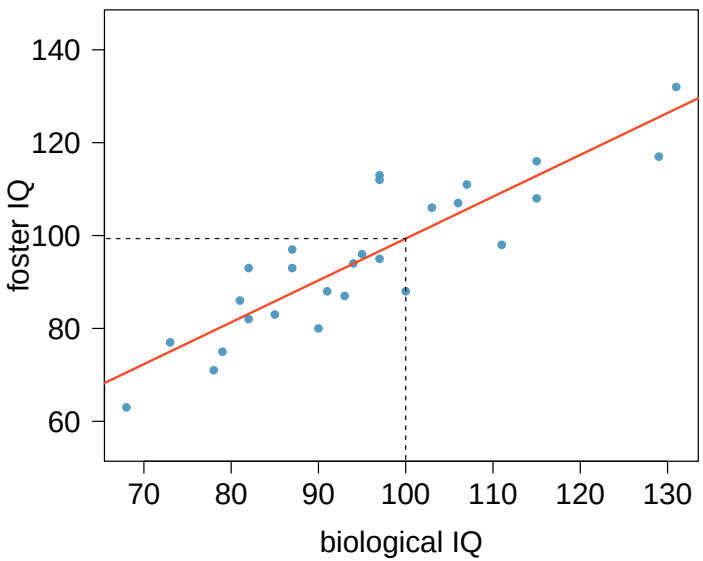
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```



$$\hat{y} \quad = \quad 9.2076 + 0.90144 \times 100 \approx 99.35$$

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```

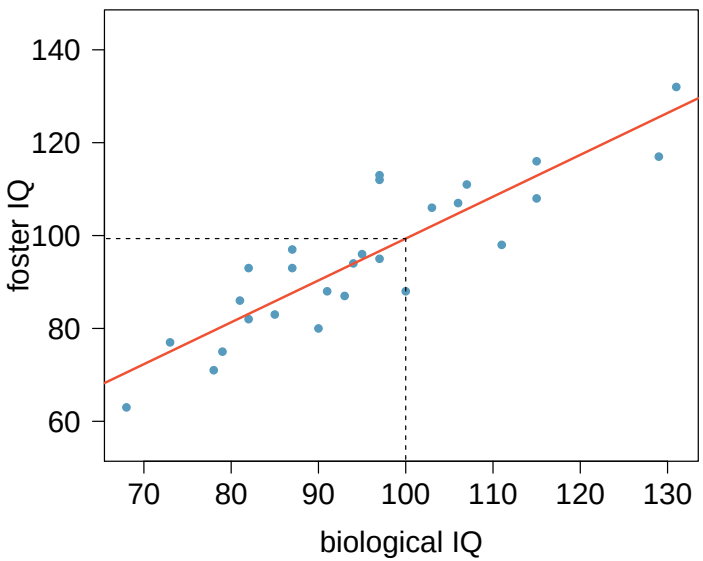


$$\hat{y} = 9.2076 + 0.90144 \times 100 \approx 99.35$$

$$df = n - 2$$

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```

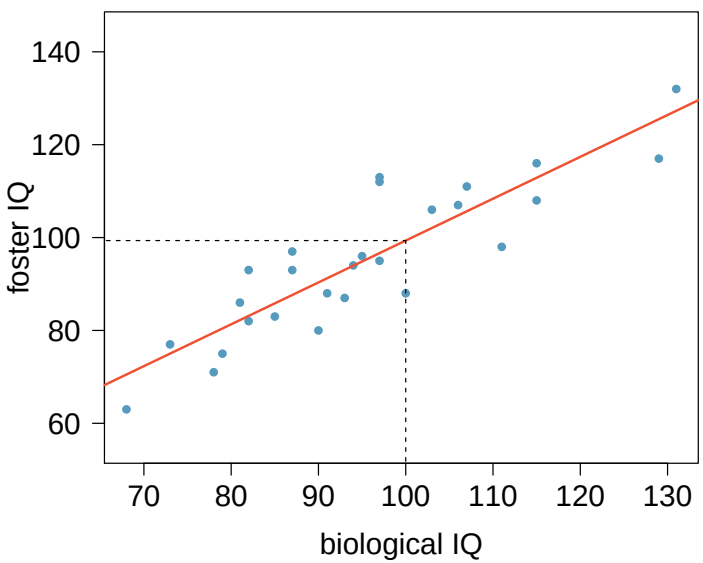


$$\hat{y} = 9.2076 + 0.90144 \times 100 \approx 99.35$$

$$df = n - 2 \qquad t^\star = 2.06$$

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```



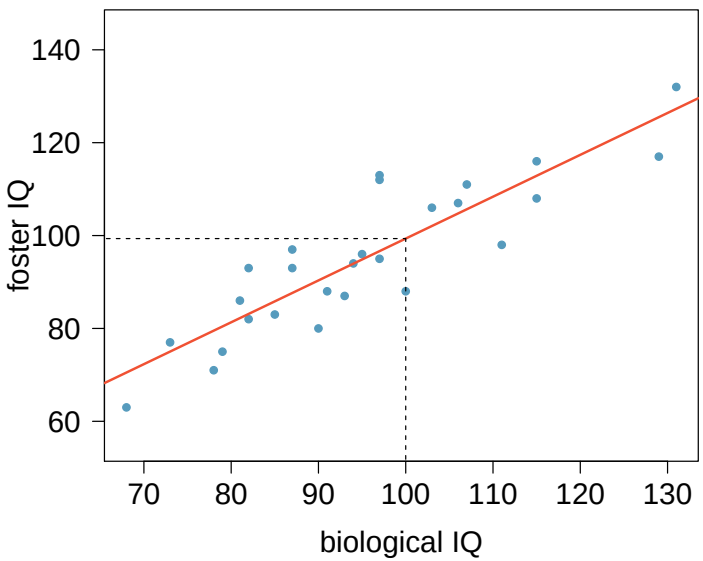
$$\hat{y} = 9.2076 + 0.90144 \times 100 \approx 99.35$$

$$df = n - 2 \qquad t^\star = 2.06$$

$$ME = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}}$$

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```



$$\hat{y} = 9.2076 + 0.90144 \times 100 \approx 99.35$$

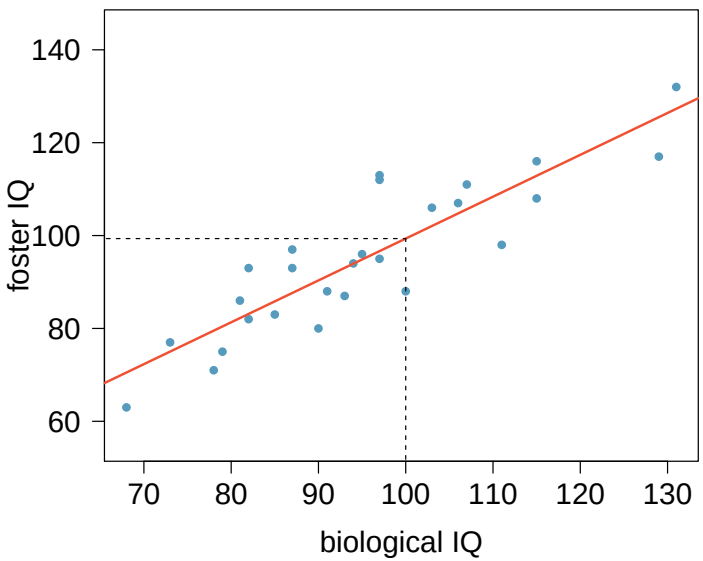
$$df = n - 2 \qquad t^\star = 2.06$$

$$ME = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}}$$

$$\approx 3.2$$

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

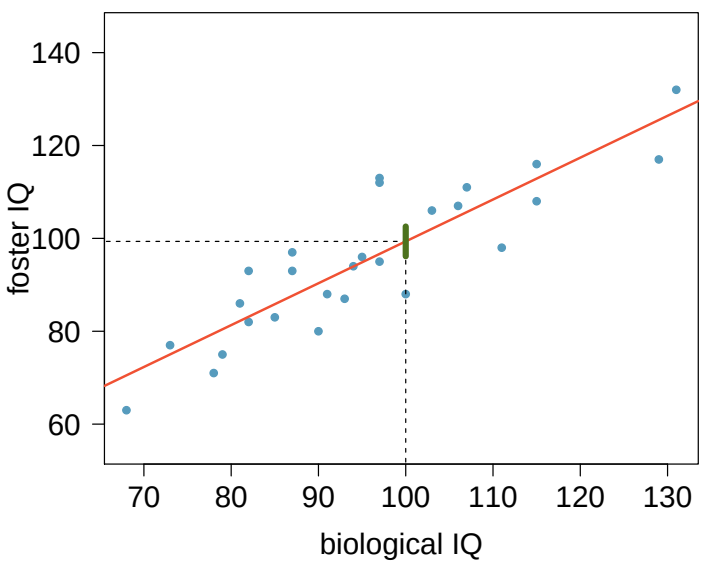
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```



$$\begin{aligned} \hat{y} &= 9.2076 + 0.90144 \times 100 \approx 99.35 \\ df &= n - 2 \qquad t^\star = 2.06 \\ ME &= 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} \\ &\approx 3.2 \\ CI &= 99.35 \pm 3.2 \end{aligned}$$

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```



$$
\begin{aligned}
\hat{y} &= 9.2076 + 0.90144 \times 100 \approx 99.35 \\
df &= n - 2 \qquad t^\star = 2.06 \\
ME &= 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} \\
&\approx 3.2 \\
CI &= 99.35 \pm 3.2 \\
&= (96.15, 102.55)
\end{aligned}
$$

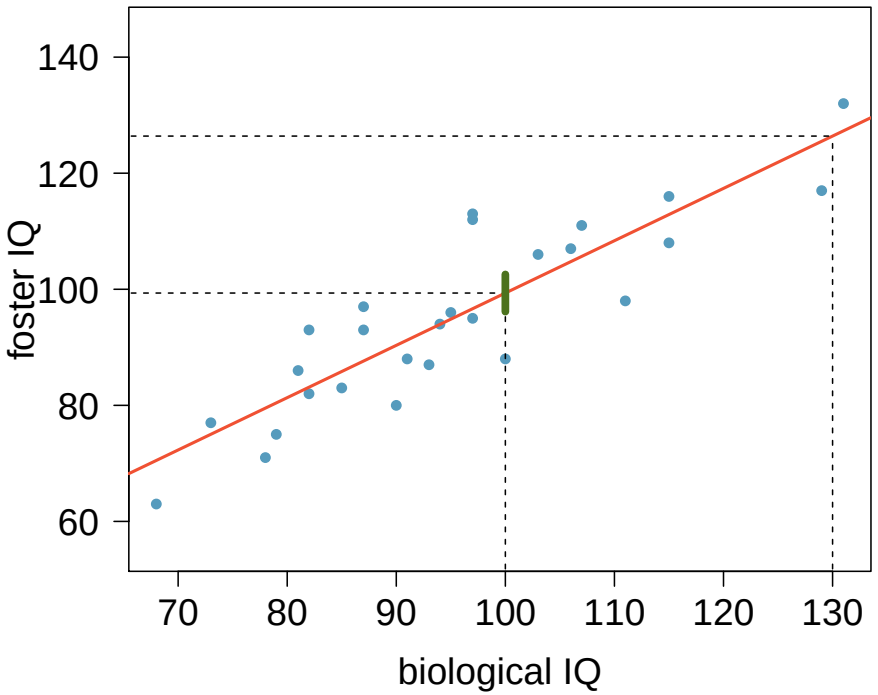
```
# load data
install.packages("faraway") # dataset is in this package
library(faraway)
data(twins)

# fit model
m = lm(Foster ~ Biological, data = twins)

# create a new data frame for the new observation
newdata = data.frame(Biological = 100)

# calculate a prediction
# and a confidence interval for the prediction
predict(m , newdata, interval = "confidence")
```

```
# load data
install.packages("faraway") # dataset is in this package
library(faraway)
data(twins)

# fit model
m = lm(Foster ~ Biological, data = twins)

# create a new data frame for the new observation
newdata = data.frame(Biological = 100)

# calculate a prediction
# and a confidence interval for the prediction
predict(m , newdata, interval = "confidence")
```

```
     fit       lwr       upr
 99.3512 96.14866 102.5537
```

Clicker question

How would you expect the width of the 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 130 points ($x^\star = 130$) to compare to the previous confidence interval (where $x^\star = 100$)?



(a) wider

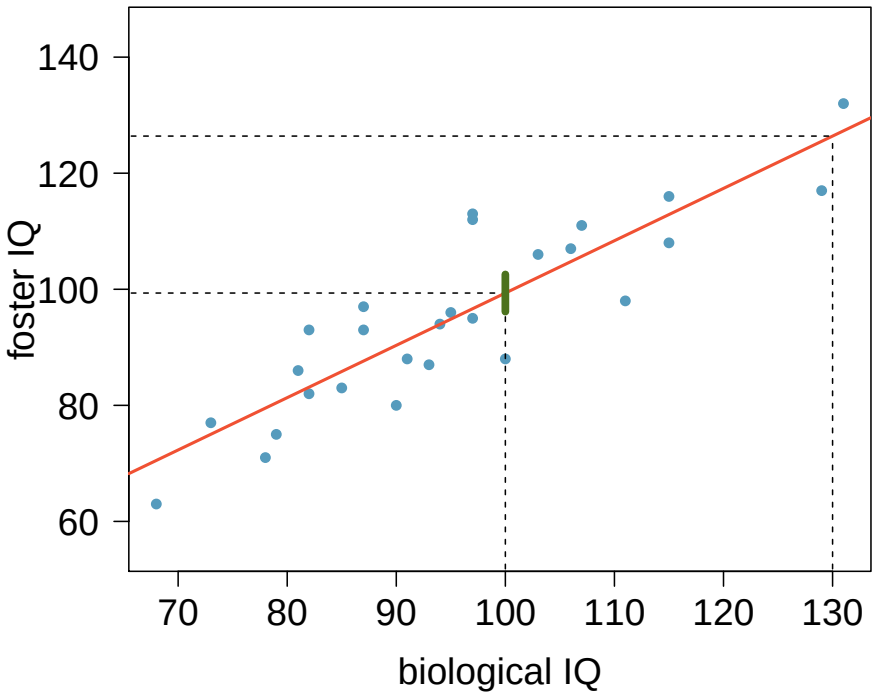
(b) narrower

(c) same width

(d) cannot tell

Clicker question

How would you expect the width of the 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 130 points ($x^\star = 130$) to compare to the previous confidence interval (where $x^\star = 100$)?



(a) *wider*

(b) narrower

(c) same width

(d) cannot tell

How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

$$x^\star = 100 \qquad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$$

How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

$x^\star = 100$ $\qquad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$

$x^\star = 130$

How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

$$x^\star = 100 \qquad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$$

$$x^\star = 130 \qquad ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(130 - 95.3)^2}{26 \times 15.74^2}} =$$

How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

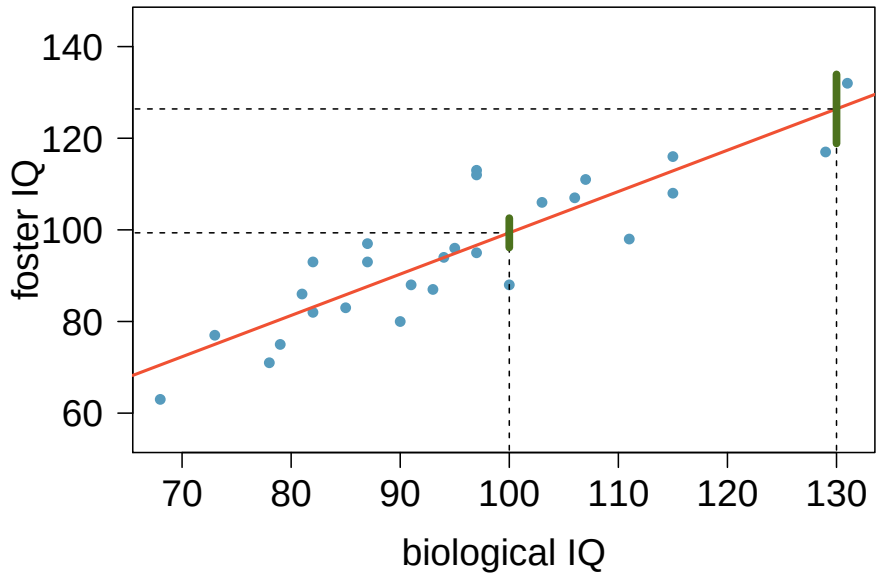
$$x^\star = 100 \qquad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$$

$$x^\star = 130 \qquad ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(130 - 95.3)^2}{26 \times 15.74^2}} = 7.53$$

How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

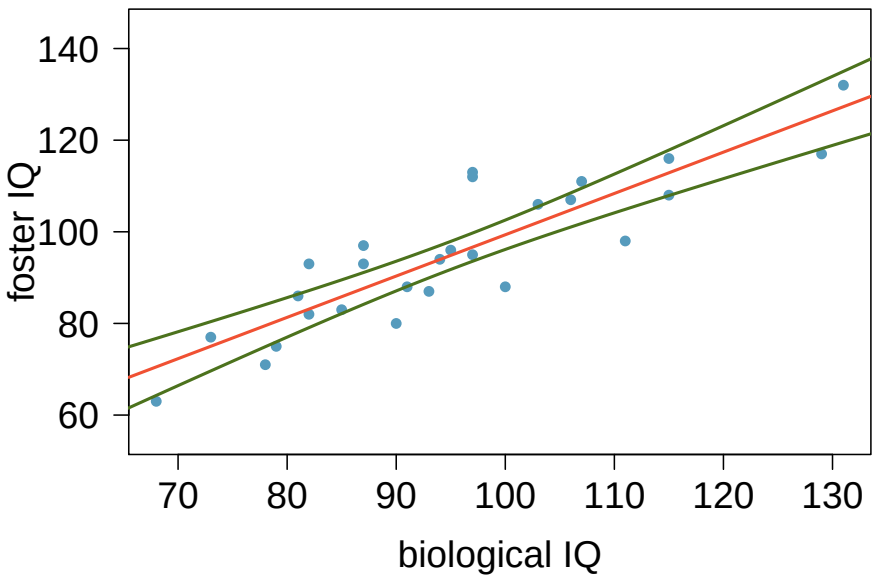
$$x^\star = 100 \qquad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$$

$$x^\star = 130 \qquad ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(130 - 95.3)^2}{26 \times 15.74^2}} = 7.53$$



The width of the confidence interval for $E(y)$ increases as $x^\star$ moves away from the center.

- Conceptually: We are much more certain of our predictions at the center of the data than at the edges (and our level of certainty decreases even further when predicting outside the range of the data -- extrapolation).
- Mathematically: As $(x^\star - \bar{x})^2$ term increases, the margin of error of the confidence interval increases as well.



1. Uncertainty of predictions
    1. Confidence intervals for average values
    2. Prediction intervals for specific predicted values
    3. Recap - CI vs. PI
    4. Confidence and prediction intervals for MLR

Clicker question

Earlier we learned how to calculate a confidence interval for average y, $E(y)$, for a given $x^\star$.
Suppose we're not interested in the average, but instead we want to predict a future value of y for a given $x^\star$.
Would you expect there to be more uncertainty around an average or a specific predicted value?

(a) more uncertainty around an average
(b) more uncertainty around a specific predicted value
(c) equal uncertainty around both values
(d) cannot tell

Clicker question

Earlier we learned how to calculate a confidence interval for average y, $E(y)$, for a given $x^\star$.
Suppose we're not interested in the average, but instead we want to predict a future value of y for a given $x^\star$.
Would you expect there to be more uncertainty around an average or a specific predicted value?

(a) more uncertainty around an average

(b) *more uncertainty around a specific predicted value*

(c) equal uncertainty around both values

(d) cannot tell

Prediction intervals for specific predicted values

A *prediction interval* for y for a given $x^\star$ is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s is the standard deviation of the residuals.

Prediction intervals for specific predicted values

A *prediction interval* for y for a given $x^\star$ is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s is the standard deviation of the residuals.

- The formula is very similar, except the variability is higher since there is an added 1 in the formula.

Prediction intervals for specific predicted values

A *prediction interval* for y for a given $x^\star$ is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s is the standard deviation of the residuals.

- The formula is very similar, except the variability is higher since there is an added 1 in the formula.
- Prediction level: If we repeat the study of obtaining a regression data set many times, each time forming a XX% prediction interval at $x^\star$, and wait to see what the future value of y is at $x^\star$, then roughly XX% of the prediction intervals will contain the corresponding actual value of y.

Application exercise: 7.3 Prediction interval

See course website for details

Application exercise: 7.3 Prediction interval

See course website for details

We already found that $\hat{y} \approx 99.35$ *and* $t^\star_{25} = 2.06$.

$$ME = 2.06 \times 7.729 \times \sqrt{1 + \frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} \approx 16.24$$

Application exercise: 7.3 Prediction interval

See course website for details

We already found that $\hat{y} \approx 99.35$ *and* $t^\star_{25} = 2.06$.

$$\begin{aligned} ME &= 2.06 \times 7.729 \times \sqrt{1 + \frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} \approx 16.24 \\ CI &= 99.35 \pm 16.24 \end{aligned}$$

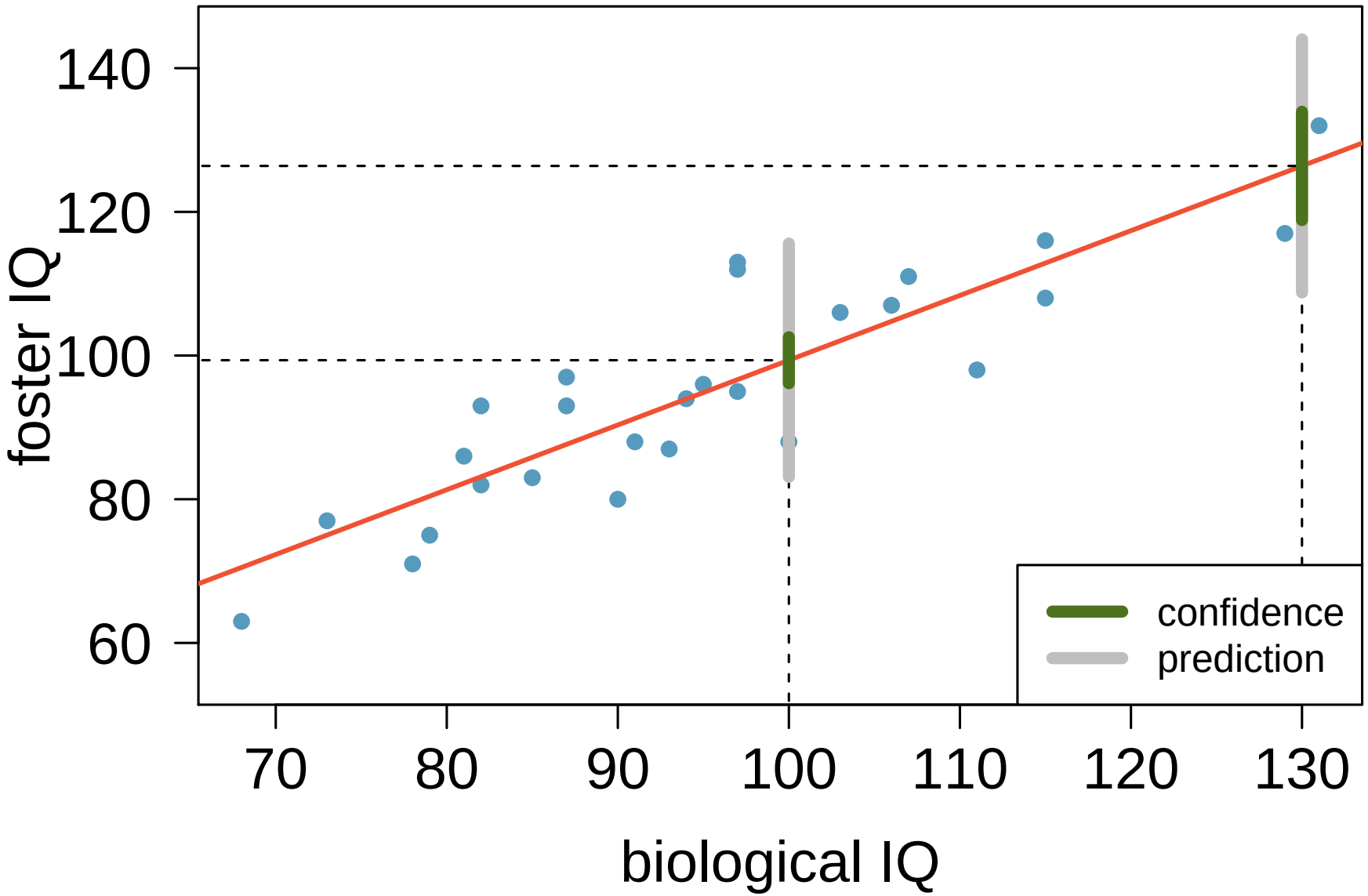
Application exercise: 7.3 Prediction interval

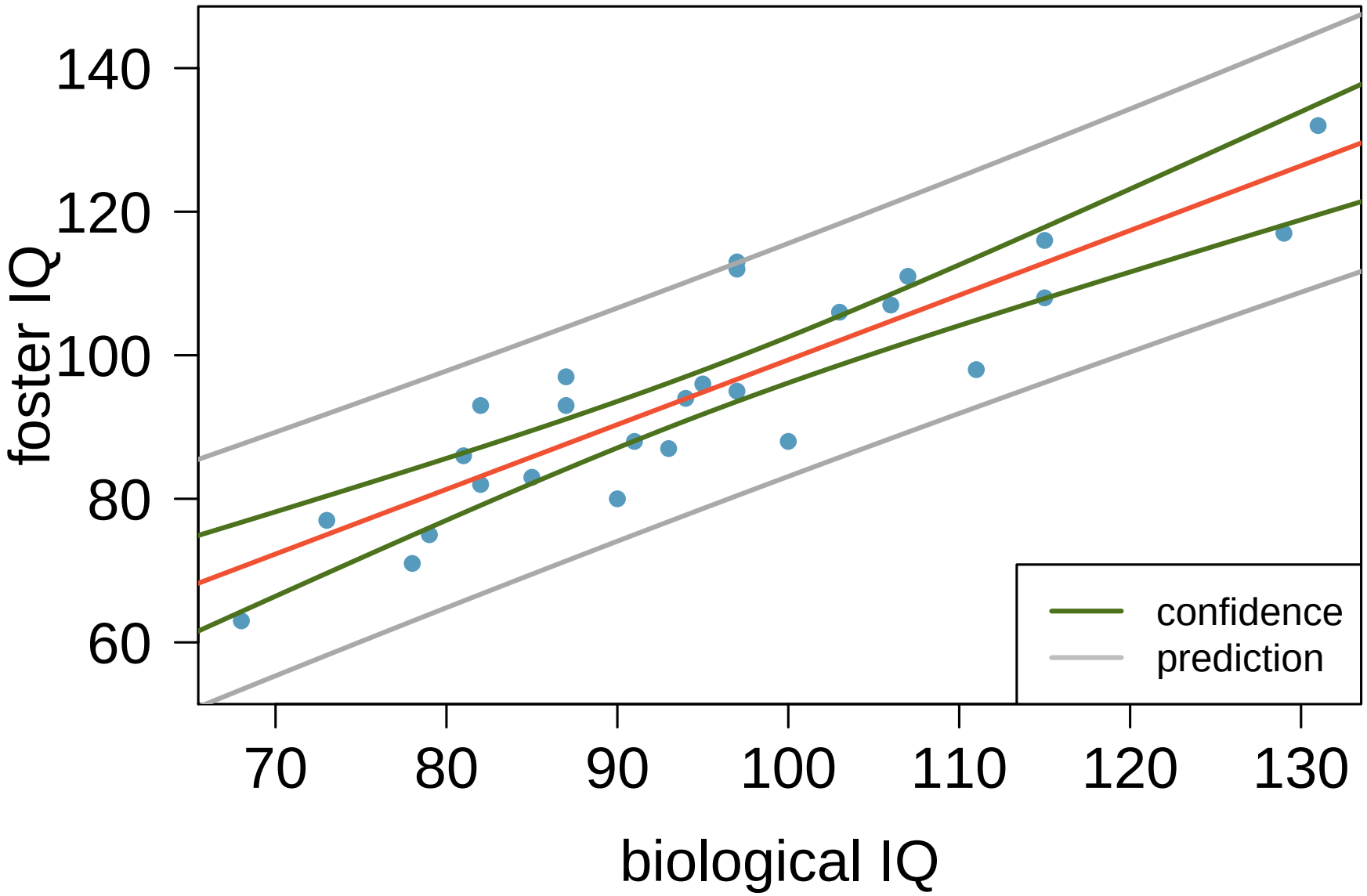
See course website for details

We already found that $\hat{y} \approx 99.35$ *and* $t^\star_{25} = 2.06$.

$$
\begin{aligned}
ME &= 2.06 \times 7.729 \times \sqrt{1 + \frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} \approx 16.24 \\
CI &= 99.35 \pm 16.24 \\
&= (83.11, 115.59)
\end{aligned}
$$

1. Uncertainty of predictions
    1. Confidence intervals for average values
    2. Prediction intervals for specific predicted values
    3. Recap - CI vs. PI
    4. Confidence and prediction intervals for MLR


foster IQ
140
120
100
80
60
70
80
90
100
110
120
130
biological IQ
confidence
prediction


foster IQ
140
120
100
80
60
70
80
90
100
110
120
130
biological IQ
confidence
prediction

- A prediction interval is similar in spirit to a confidence interval, except that

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a ``moving target", the random future value of y, while

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a ``moving target", the random future value of y, while
  - the confidence interval is designed to cover the ``fixed target", the average (expected) value of y, $E(y)$,

  for a given $x^\star$.

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a ``moving target", the random future value of y, while
  - the confidence interval is designed to cover the ``fixed target", the average (expected) value of y, $E(y)$,

  for a given $x^\star$.
- Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for a given $x^\star$ and confidence level. This makes sense, since

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a ``moving target", the random future value of y, while
  - the confidence interval is designed to cover the ``fixed target", the average (expected) value of y, $E(y)$,

  for a given $x^\star$.
- Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for a given $x^\star$ and confidence level. This makes sense, since
  - the prediction interval must take account of the tendency of y to fluctuate from its mean value, while

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a ``moving target", the random future value of y, while
  - the confidence interval is designed to cover the ``fixed target", the average (expected) value of y, $E(y)$,

  for a given $x^\star$.
- Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for a given $x^\star$ and confidence level. This makes sense, since
  - the prediction interval must take account of the tendency of y to fluctuate from its mean value, while
  - the confidence interval simply needs to account for the uncertainty in estimating the mean value.

- For a given data set, the error in estimating $E(y)$ and $\hat{y}$ grows as $x^\star$ moves away from $\bar{x}$. Thus, the further $x^\star$ is from $\bar{x}$, the wider the confidence and prediction intervals will be.

- For a given data set, the error in estimating $E(y)$ and $\hat{y}$ grows as $x^\star$ moves away from $\bar{x}$. Thus, the further $x^\star$ is from $\bar{x}$, the wider the confidence and prediction intervals will be.
- If any of the conditions underlying the model are violated, then the confidence intervals and prediction intervals may be invalid as well. This is why it's so important to check the conditions by examining the residuals, etc.

1. Uncertainty of predictions
    1. Confidence intervals for average values
    2. Prediction intervals for specific predicted values
    3. Recap - CI vs. PI
    4. Confidence and prediction intervals for MLR

- In the case of multiple linear regression (regression with many predictors), confidence and prediction intervals for a new prediction works exactly the same way.
- However the formulas are much more complicated since we no longer have just one x, but instead many xs.
- For confidence and prediction intervals for MLR we will focus on the concepts and leave the calculations up to R.

```
load(url("http://www.openintro.org/stat/data/evals.RData"))

# fit a model
m = lm(score ~ rank + gender + language +
    cls_perc_eval + cls_students, data = evals)

# create a data frame with the new observation (mine)
prof = data.frame(rank = "teaching", gender = "female",
    language = "english", cls_perc_eval = 90,
    cls_students = 120)
```

Prediction of evaluation my evaluation score

```
load(url("http://www.openintro.org/stat/data/evals.RData"))

# fit a model
m = lm(score ~ rank + gender + language +
    cls_perc_eval + cls_students, data = evals)

# create a data frame with the new observation (mine)
prof = data.frame(rank = "teaching", gender = "female",
    language = "english", cls_perc_eval = 90,
    cls_students = 120)
```

```
predict(m , prof , interval = "prediction")
```

Prediction of evaluation my evaluation score

```
load(url("http://www.openintro.org/stat/data/evals.RData"))

# fit a model
m = lm(score ~ rank + gender + language +
    cls_perc_eval + cls_students, data = evals)

# create a data frame with the new observation (mine)
prof = data.frame(rank = "teaching", gender = "female",
    language = "english", cls_perc_eval = 90,
    cls_students = 120)
```

```
predict(m , prof , interval = "prediction")
```

```
       fit      lwr      upr
1 4.352334 3.315313 5.389355
```

Prediction of evaluation my evaluation score

```
load(url("http://www.openintro.org/stat/data/evals.RData"))

# fit a model
m = lm(score ~ rank + gender + language +
    cls_perc_eval + cls_students, data = evals)

# create a data frame with the new observation (mine)
prof = data.frame(rank = "teaching", gender = "female",
    language = "english", cls_perc_eval = 90,
    cls_students = 120)
```

```
predict(m , prof , interval = "prediction")
```

```
       fit      lwr      upr
1 4.352334 3.315313 5.389355
```

```
predict(m , prof , interval = "confidence")
```

Prediction of evaluation my evaluation score

```
load(url("http://www.openintro.org/stat/data/evals.RData"))

# fit a model
m = lm(score ~ rank + gender + language +
    cls_perc_eval + cls_students, data = evals)

# create a data frame with the new observation (mine)
prof = data.frame(rank = "teaching", gender = "female",
    language = "english", cls_perc_eval = 90,
    cls_students = 120)
```

```
predict(m , prof , interval = "prediction")
```

```
       fit      lwr      upr
1 4.352334 3.315313 5.389355
```

```
predict(m , prof , interval = "confidence")
```

```
       fit      lwr      upr
1 4.352334 4.210043 4.494626
```