Unit 6: Simple Linear Regression
Lecture 3: Confidence and prediction intervals for SLR

Statistics 101

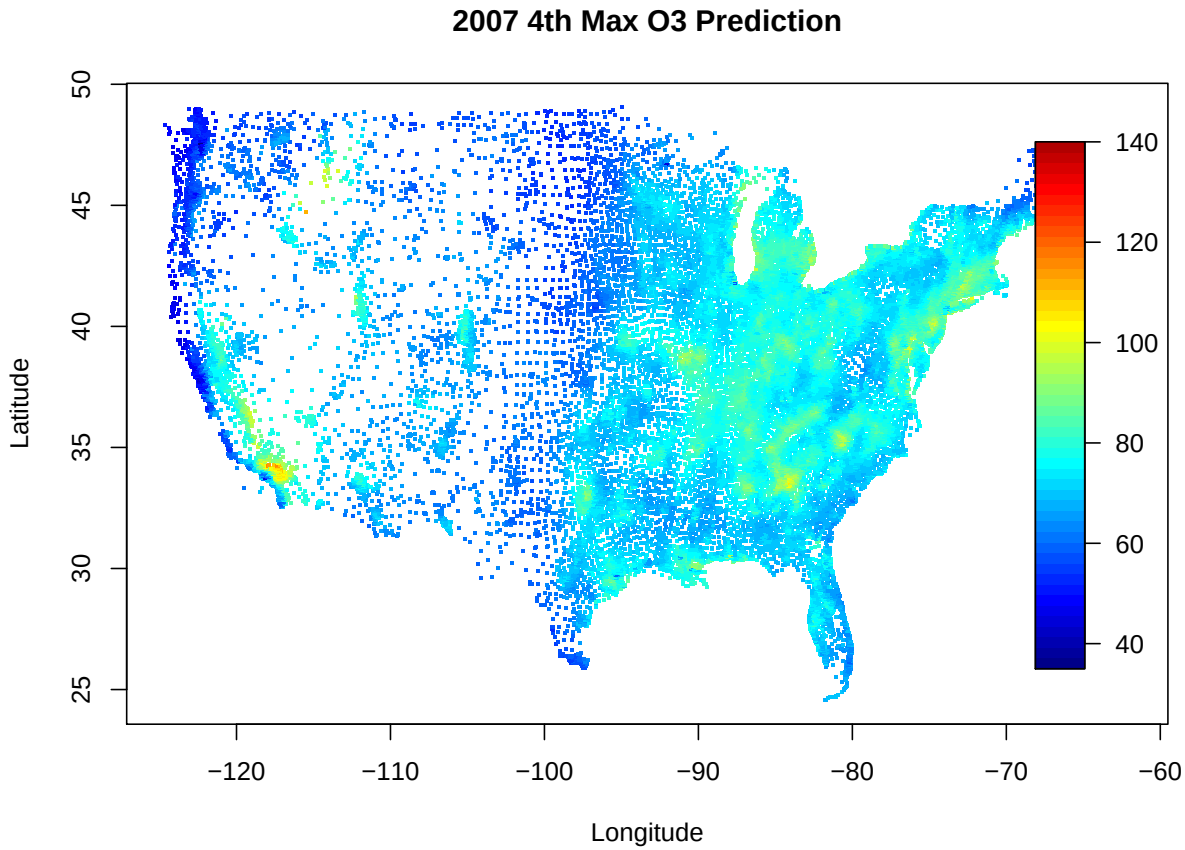
Thomas Leininger

June 19, 2013

Announcements

- Notes from HW: remember to check conditions and interpret findings in context when doing a CI/HT.
- Notes on project: link on schedule has example projects

Visualization of the Day

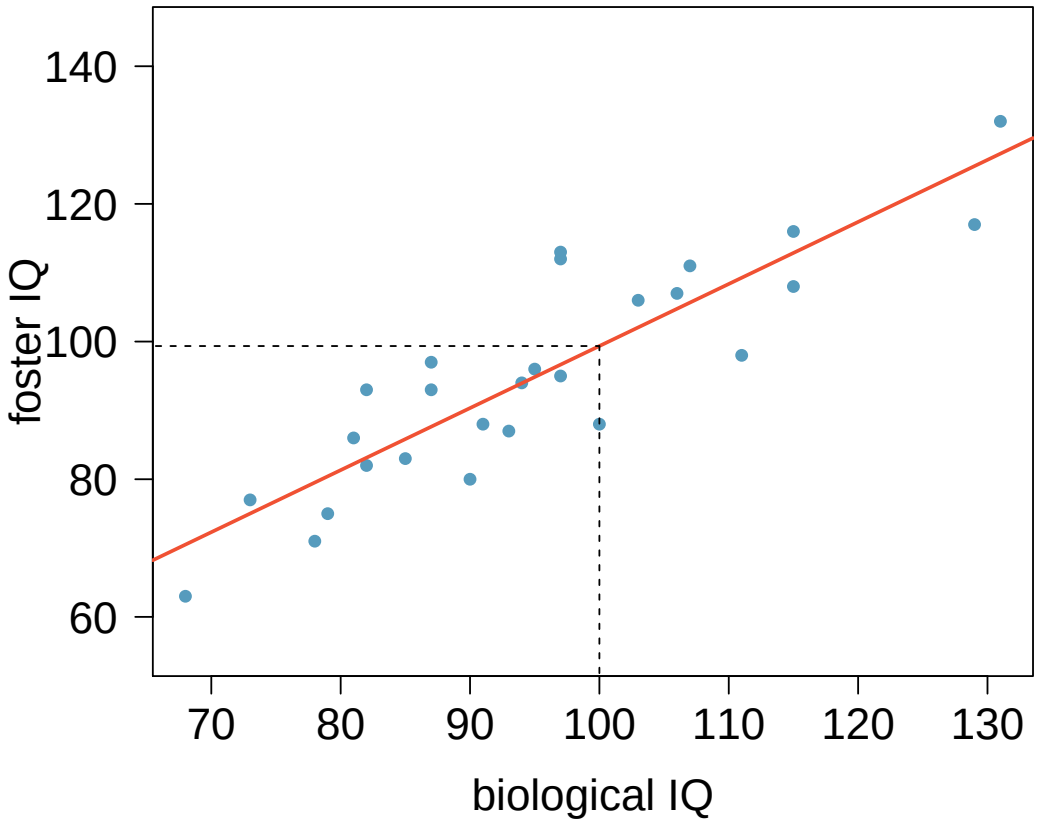


http://stat.duke.edu/~tjl13/s101/DailyO3_2007_160.gif

Can we make CIs for predicting a foster twin's IQ?

Two type of intervals available:

- Confidence interval for the average foster twin's IQ
- **Prediction** interval for a single foster twin's IQ



Confidence intervals for average values

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

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

where s_y is the standard deviation of the residuals, calculated as

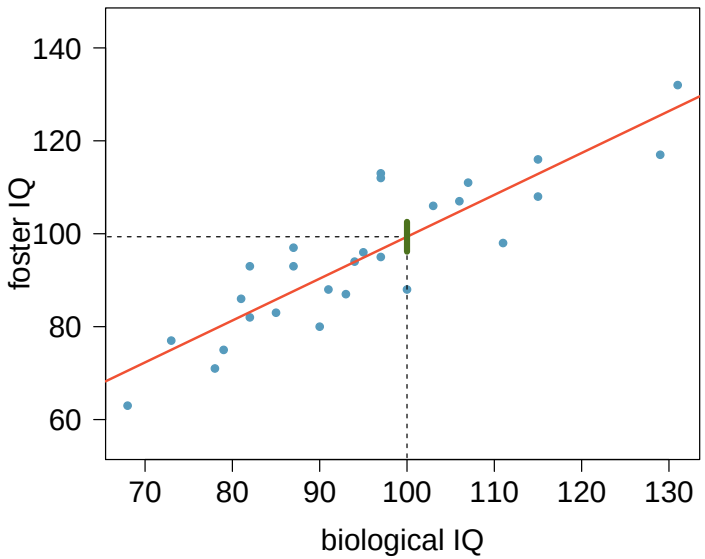
$$s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}.$$

s_y is called `residual standard error` in R regression output.

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
```



$$\hat{y} =$$

$$df = n - 2 = \quad , \; t^\star =$$

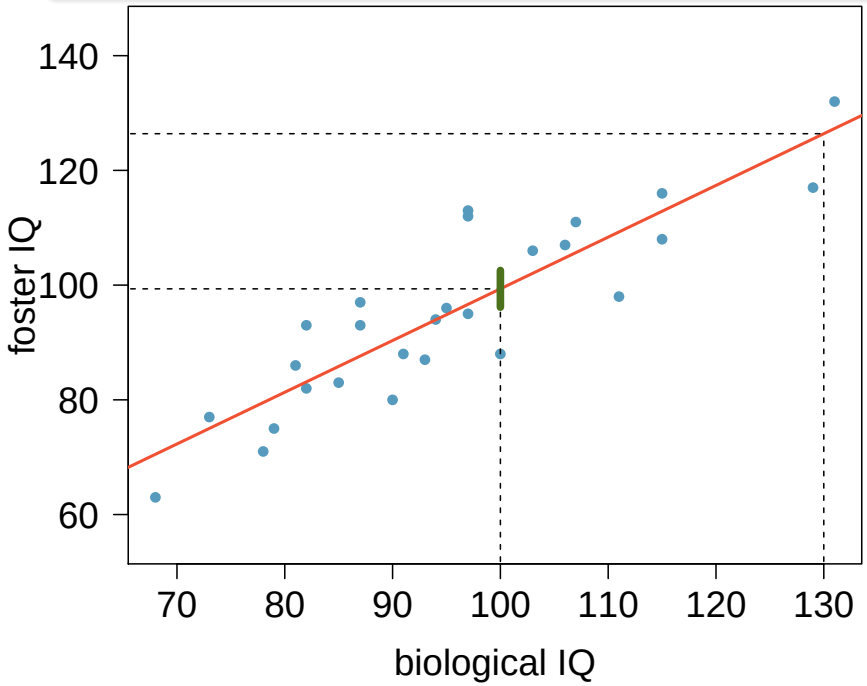
$$ME =$$

$$CI = 99.35 \pm 3.2$$

$$= (96.15, 102.55)$$

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 $x^\star = 100$)?



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

(a) wider

(b) narrower

(c) same width

(d) cannot tell

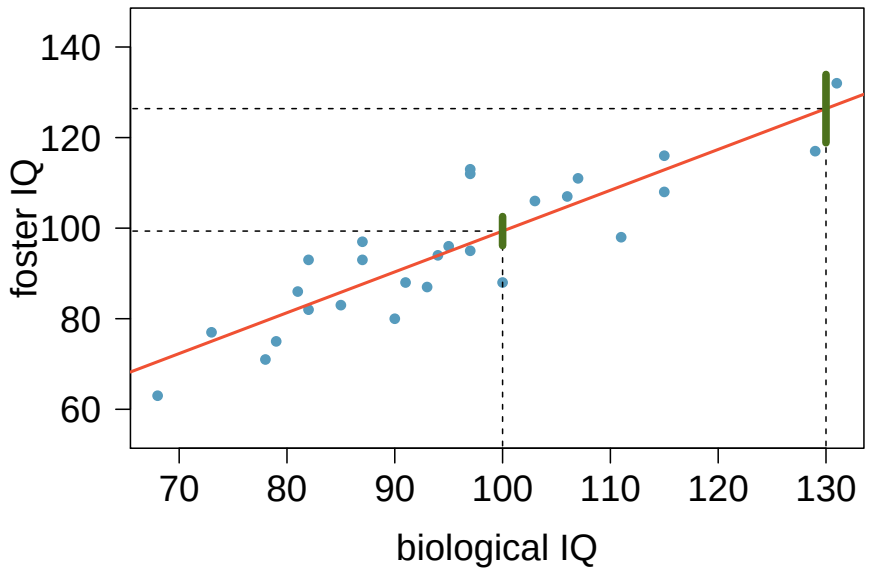
How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

$x^\star = 100$

$$ME_{100} = 2.06 \times 7.729 \times$$

$x^\star = 130$

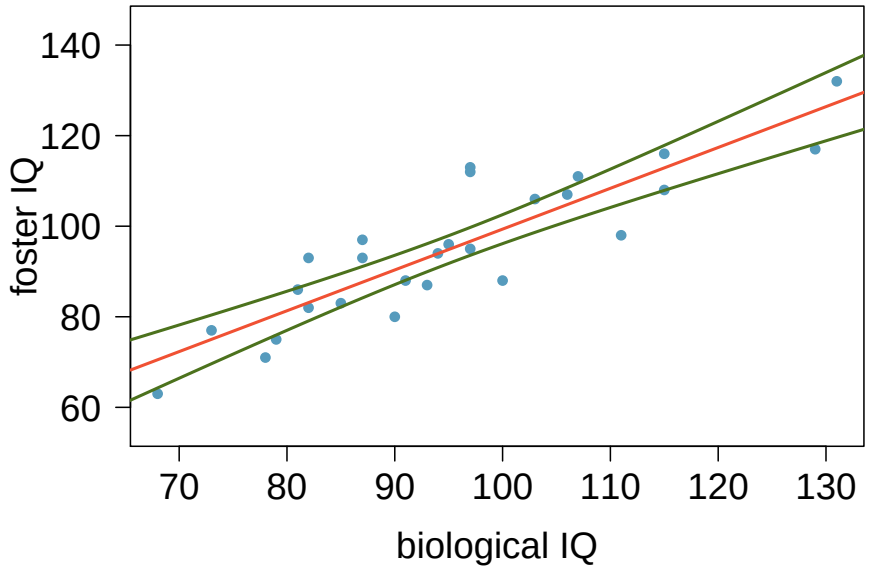
$$ME_{130} = 2.06 \times 7.729 \times$$



Recap

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.



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?

Prediction intervals for specific predicted values

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

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

- 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: Prediction interval

Calculate a 95% prediction 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.

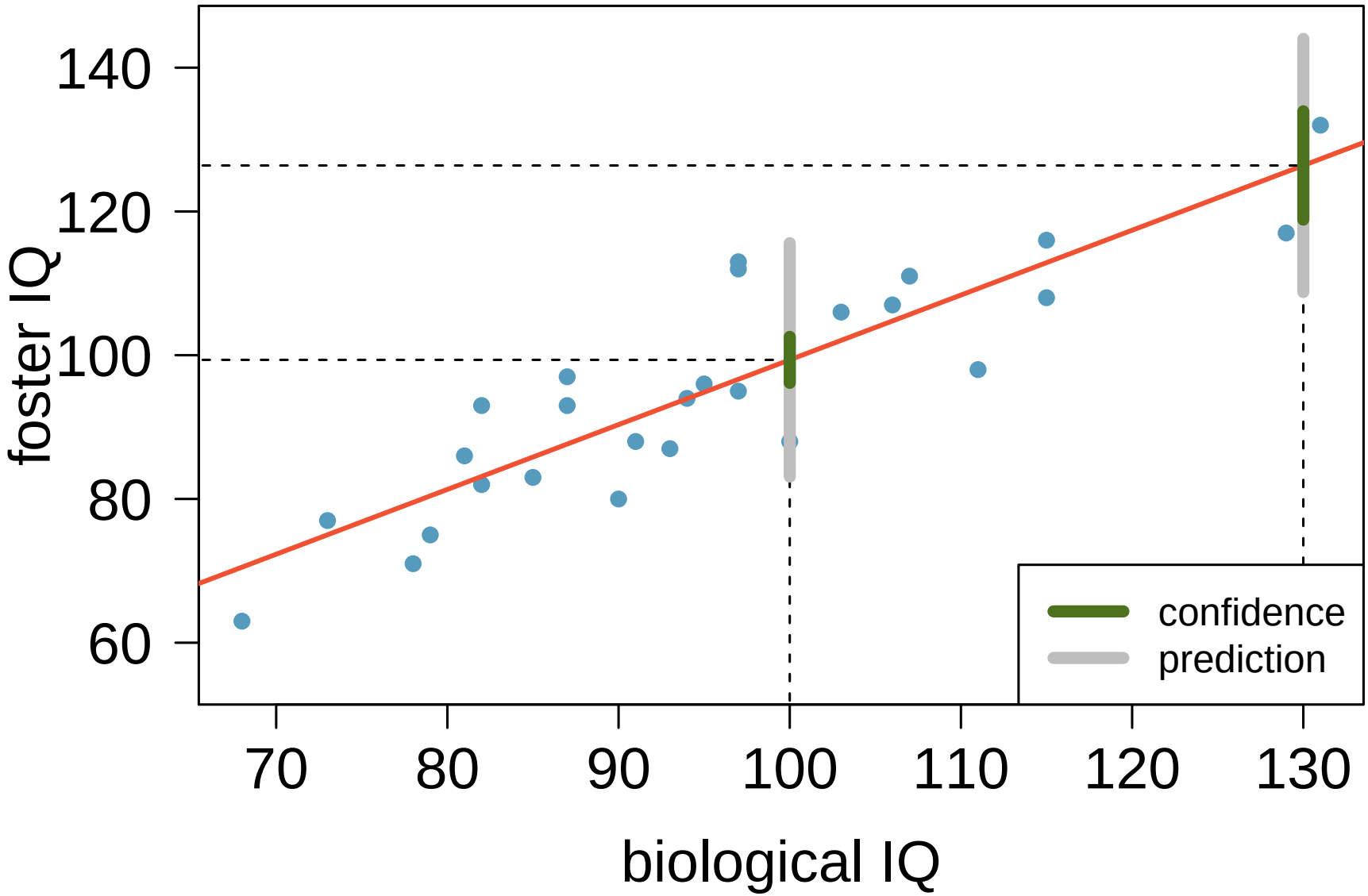
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

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

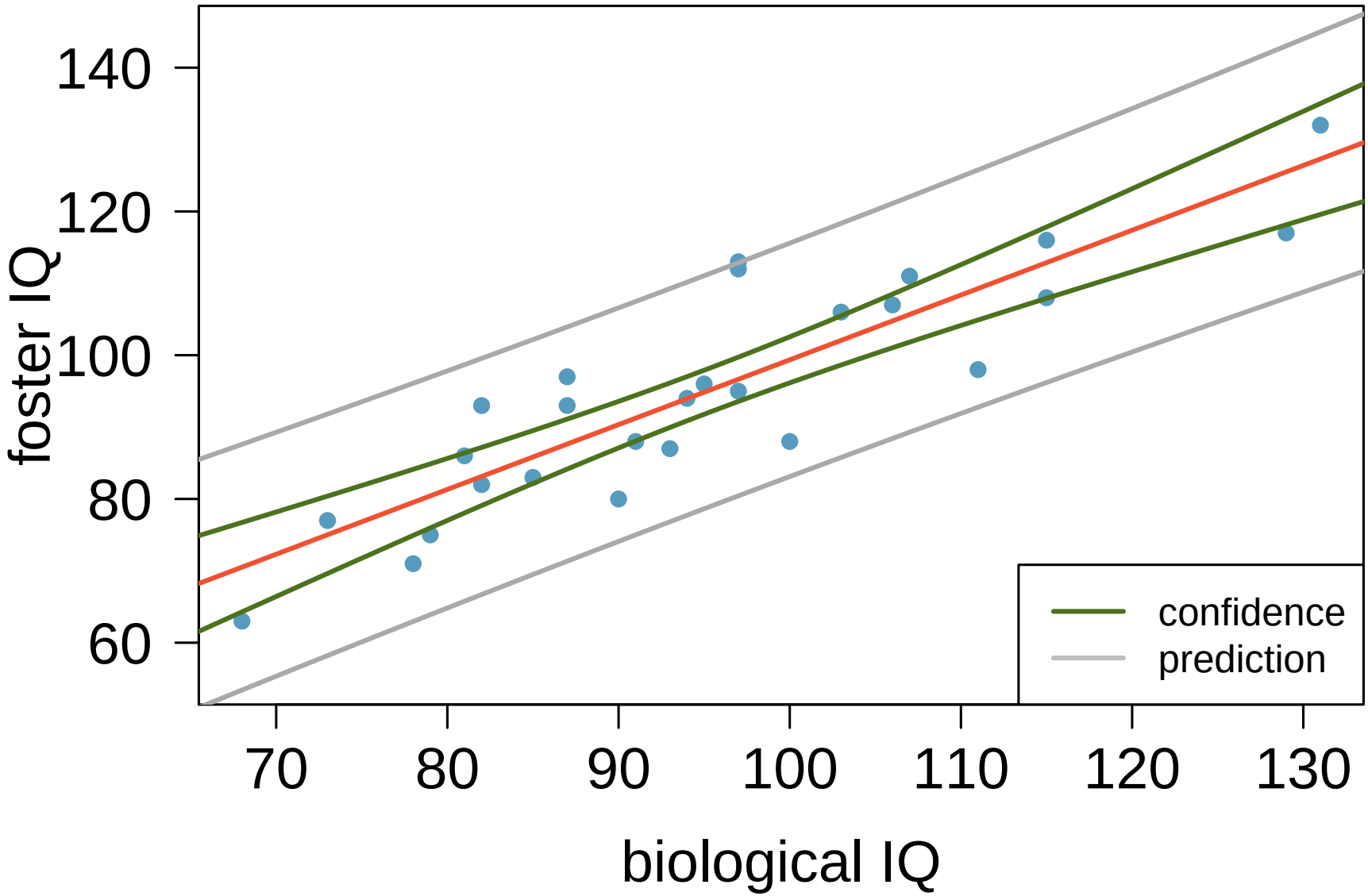
We already found that $\hat{y} \approx 99.35$ and $t^{\star}_{25} = 2.06$.

$$\hat{y} \pm t^{\star}_{n-2}\, s_y \sqrt{1 + \frac{1}{n} + \frac{(x^{\star} - \bar{x})^2}{(n-1)s_x^2}} =$$

CI for $E(y)$ vs. PI for y (1)



CI for $E(y)$ vs. PI for y (2)



CI for $E(y)$ vs. PI for y - differences

- 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.

CI for $E(y)$ vs. PI for y - similarities

- 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.

For further discussion of confidence intervals and predictions intervals for y given a specific level of x, see the video below:

http://www.youtube.com/watch?feature=player_embedded&v=qVCQi0KPR0s