Lecture 24: Inference for Linear Regression

- Hypothesis testing and confidence interval
- ANOVA
- CI v.s. Prediction interval

Introduction

– In last lecture we introduced residual analysis and types of outliers. We also talked about regression with categorical explanatory variables and log transformation on the response.

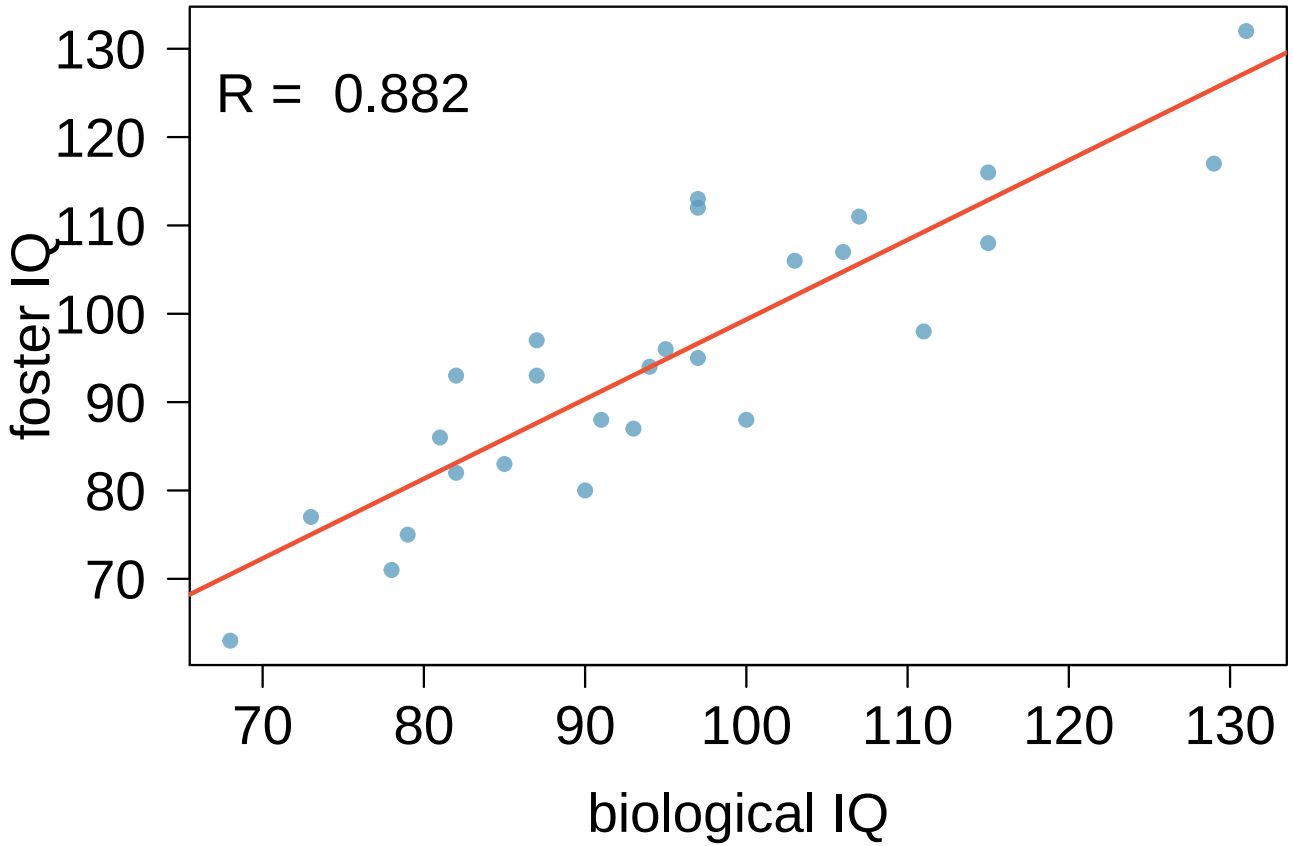
– Today, we will learn inference (e.g., hypothesis testing and confidence interval for the slope) for linear regression, how to understand regression output from software.

– We use t-test in inference for regression. We use F-test for ANOVA. There is an interesting relationship between the two.

– Lastly, we will compare confidence interval for the average values to the confidence interval for the predicted value.

Nature or 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.



Which of the following is false?

```
Coefficients:
             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
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.

(b) Roughly 78% of the foster twins' IQs can be accurately predicted by the model.

(c) The linear model is $\widehat{\text{fosterIQ}} = 9.2 + 0.9 \times \text{bioIQ}$.

(d) Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Which of the following is false?

```
Coefficients:
             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
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.

(b) *Roughly 78% of the foster twins' IQs can be accurately predicted by the model.*

(c) The linear model is $\widehat{\text{fosterIQ}} = 9.2 + 0.9 \times \text{bioIQ}$.

(d) Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Testing for the slope

The fitted least-squares regression line:

$$\widehat{\text{fosterIQ}} = 9.2 + 0.9 \times \text{bioIQ}$$

Does these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin?

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?

(a) $H_0 : b_0 = 0; H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0; H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0; H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$

Testing for the slope

The fitted least-squares regression line:

$$\widehat{\text{fosterIQ}} = 9.2 + 0.9 \times \text{bioIQ}$$

Does these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin?

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?

(a) $H_0 : b_0 = 0; H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0; H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0; H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$

Understanding regression output from software

- We will use a t-test in inference for regression.
- Recall: Test statistic, $T = \frac{point\ estimate - null\ value}{SE}$

	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

- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.
- Does the test statistics follow the t-distribution under the null hypothesis? Yes, under certain conditions.

Conditions for regression

Important for inference

- **Nearly normally distributed residuals** → check histogram or Q-Q plot of residuals
- **Constant variability of residuals** (homoscedasticity) → no fan shape in the residual plot
- **Independence of observations** (and hence residuals) → depends on data collection method, often violated for time-series data

Important regardless of doing inference

- **Linearity** → The data should show a linear trend

	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

Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

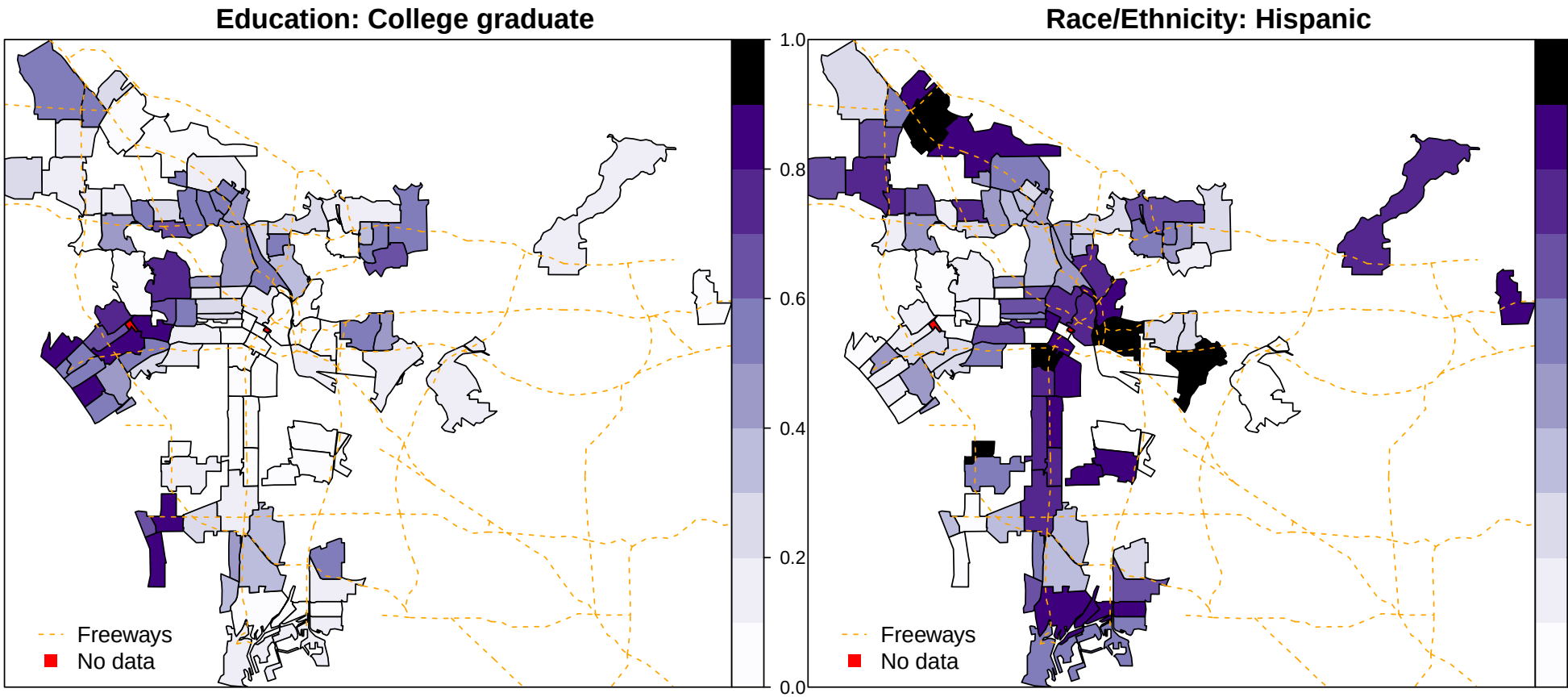
Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

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

Note that here **p-value** = P(observing a slope at least as different from 0 as the one observed if in fact there is no relationship between x and y)

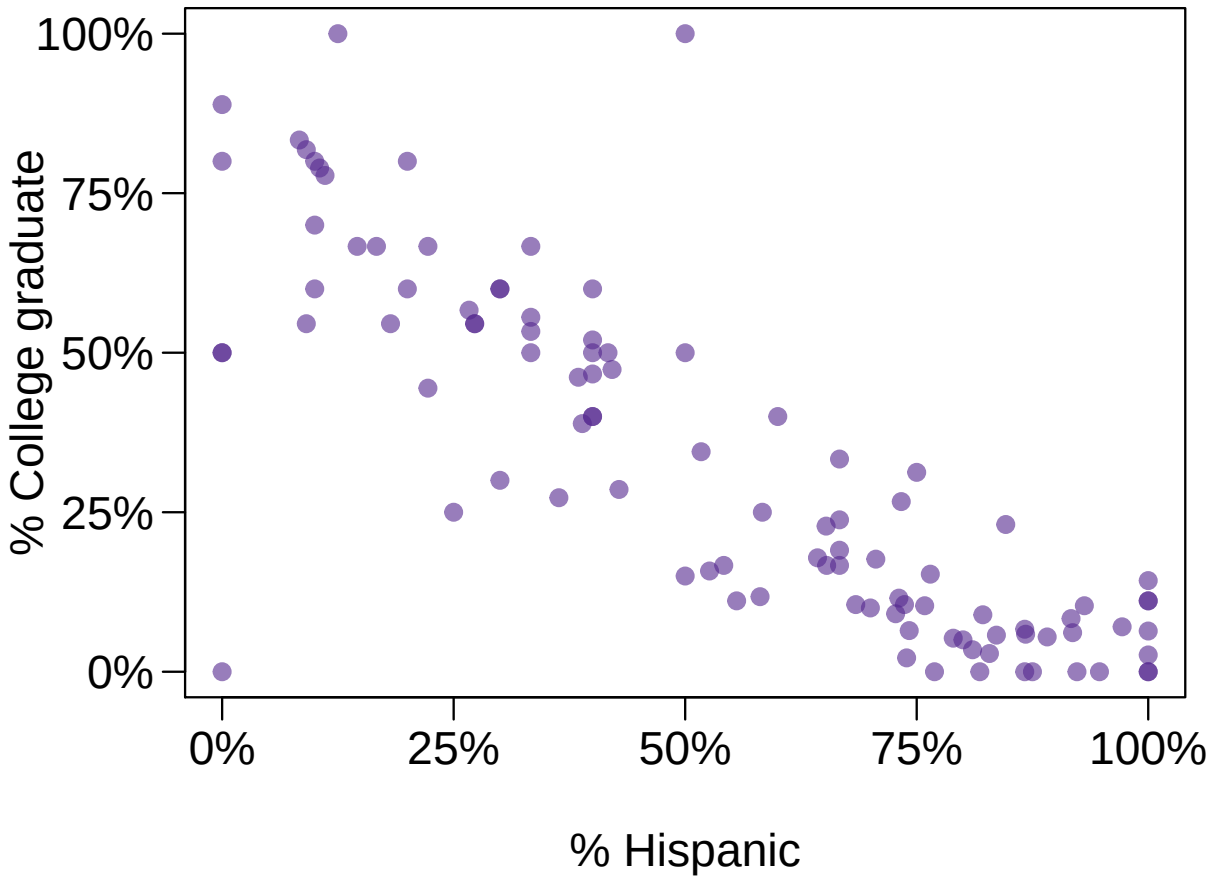
% College graduate vs. % Hispanic in LA

What can you say about the relationship between % college graduate and % Hispanic in a sample of 100 zip code areas in LA?



% College educated vs. % Hispanic in LA - another look

What can you say about the relationship between of % college graduate and % Hispanic in a sample of 100 zip code areas in LA?



% College educated vs. % Hispanic in LA - linear model

Which of the below is the best interpretation of the slope?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
%Hispanic	-0.7527	0.0501	-15.01	0.0000

(a) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.

(b) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.

(c) An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.

(d) In zip code areas with no Hispanic residents, % of college graduates is expected to be 75%.

% College educated vs. % Hispanic in LA - linear model

Which of the below is the best interpretation of the slope?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
%Hispanic	-0.7527	0.0501	-15.01	0.0000

(a) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.

(b) *A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.*

(c) An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.

(d) In zip code areas with no Hispanic residents, % of college graduates is expected to be 75%.

% College educated vs. % Hispanic in LA - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % Hispanic and % college graduates in zip code areas in LA?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
hispanic	-0.7527	0.0501	-15.01	0.0000

How reliable is this p-value if these zip code areas are not randomly selected?

% College educated vs. % Hispanic in LA - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % Hispanic and % college graduates in zip code areas in LA?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
hispanic	-0.7527	0.0501	-15.01	0.0000

Yes, the p-value for % Hispanic is low, indicating that the data provide convincing evidence that the slope parameter is different than 0.

How reliable is this p-value if these zip code areas are not randomly selected?

% College educated vs. % Hispanic in LA - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % Hispanic and % college graduates in zip code areas in LA?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7290	0.0308	23.68	0.0000
hispanic	-0.7527	0.0501	-15.01	0.0000

Yes, the p-value for % Hispanic is low, indicating that the data provide convincing evidence that the slope parameter is different than 0.

How reliable is this p-value if these zip code areas are not randomly selected?

Not very...

Confidence interval for the slope

1. Recall that a **confidence interval** is calculated as **point estimate** ± **ME**
2. Use a t-distribution to create confidence intervals for the slope
3. In the case of a simple linear regression, the degrees of freedom associated with the slope is $n - 2$
4. A CI for a slope: $b_1 \pm T^{\star}_{n-2} \times \text{SE}_{b_1}$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n - 2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	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

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n-2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	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

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

$$n = 27 \qquad df = 27 - 2 = 25$$

$$95\%: \; t^{\star}_{25} = 2.06$$

$$0.9014 \pm 2.06 \times 0.0963$$

$$(0.7 \;\; , \;\; 1.1)$$

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{\text{SE}_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2}\text{SE}_{b_1}$$

- The null value is often 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 *two-tailed* p-value for the t-test for the slope where the null value is 0.
- We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope.

Caution

- If conditions for fitting the regression line do not hold, then the inference presented here should not be applied.
  - The standard error or distribution assumption of the point estimates may not be valid.
- 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 already have population data.
- If you have a sample that is non-random (biased), inference on the results will be unreliable.

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 requires *analysis of variance (ANOVA)*.

ANOVA output - Sum of squares

	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:

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

$$SS_{Err} = \sum (y - \hat{y})^2 = \sum e_i^2 = 1493.53 \quad \text{(unexplained variability in residuals)}$$

$$SS_{Reg} = \sum (\hat{y} - \bar{y})^2 = SS_{Tot} - SS_{Err} \quad \text{(explained variability in y)}$$
$$= 6724.66 - 1493.53 = 5231.13$$

	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			

Degrees of freedom:

$$\text{df}_{\text{Tot}} = n - 1 = 27 - 1 = 26$$

$$\text{df}_{\text{Reg}} = 1 \quad \text{(there is only 1 predictor)}$$

$$\text{df}_{\text{Res}} = \text{df}_{\text{Tot}} - \text{df}_{\text{Reg}} = 26 - 1 = 25$$

	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 squares:

$$\text{MS}_{\text{Reg}} = \frac{\text{SS}_{\text{Reg}}}{\text{df}_{\text{Reg}}} = \frac{5231.13}{1} = 5231.13$$

$$\text{MS}_{\text{Err}} = \frac{\text{SS}_{\text{Err}}}{\text{df}_{\text{Err}}} = \frac{1493.53}{1} = 59.74$$

F-statistic:

$$F_{(1,25)} = \frac{\text{MS}_{\text{Reg}}}{\text{MS}_{\text{Err}}} = 87.56 \quad \text{(ratio of explained to unexplained variability)}$$

	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			

F-statistic:

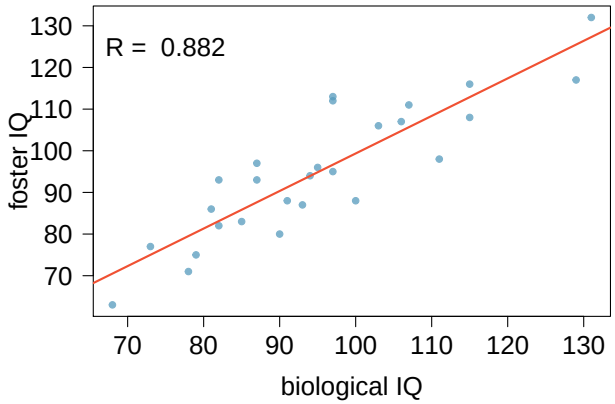
$$F_{(1,25)} = \frac{\text{MS}_{\text{Reg}}}{\text{MS}_{\text{Err}}} = 87.56 \quad \text{(ratio of explained to unexplained variability)}$$

The null hypothesis is $\beta_1 = 0$ and the alternative is $\beta_1 \neq 0$. With a large F-statistic, and a small p-value, we reject H_0 and conclude that the slope is significantly different than 0, i.e., the explanatory variable is a significant predictor of the response variable.

Revisit R^2

- Remember, R^2 is the proportion of variability in y explained by the model:
  - A large R^2 suggests a linear relationship between x and y exists
  - A small R^2 suggests the evidence provided by the data may not be convincing
- There are actually two ways to calculate R^2:
  1. From the definition: proportion of explained to total variability
  2. Using correlation: square of the correlation coefficient

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			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{\text{SS}_{\text{Reg}}}{\text{SS}_{\text{Tot}}} = \frac{5231.13}{6724.66} \approx 0.78$$

$$R^2 = \text{square of correlation coefficient} = 0.882^2 \approx 0.78$$

Confidence interval for average values

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

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

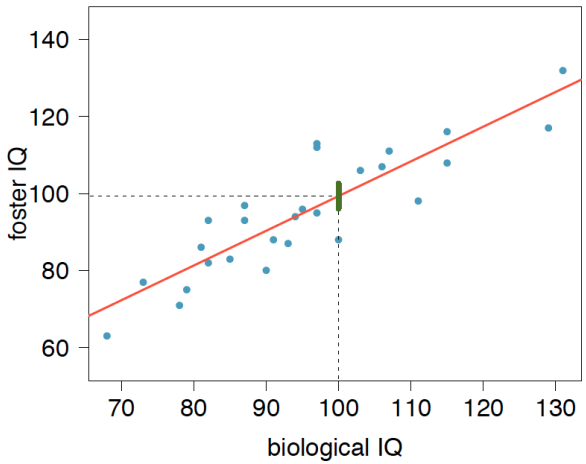
where s is standard deviation of the residuals, calculated as

$$\frac{1}{n-2} \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

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
```



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

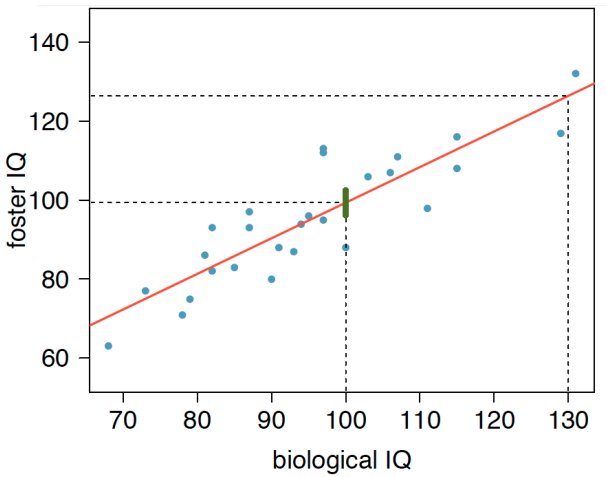
$$df = n - 2, \quad t^\star = 2.06$$

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

$$\approx 3.2$$

$$CI = 99.35 \pm 3.2 = (96.15, 102.55)$$

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 $x^\star = 100$)?



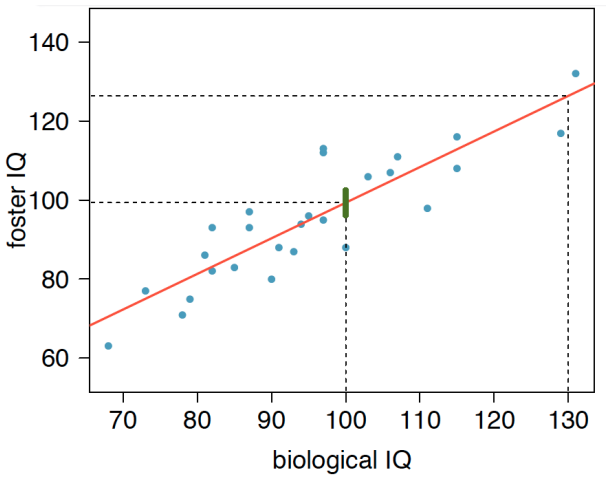
(a) wider

(b) narrower

(c) same width

(d) cannot tell

How 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

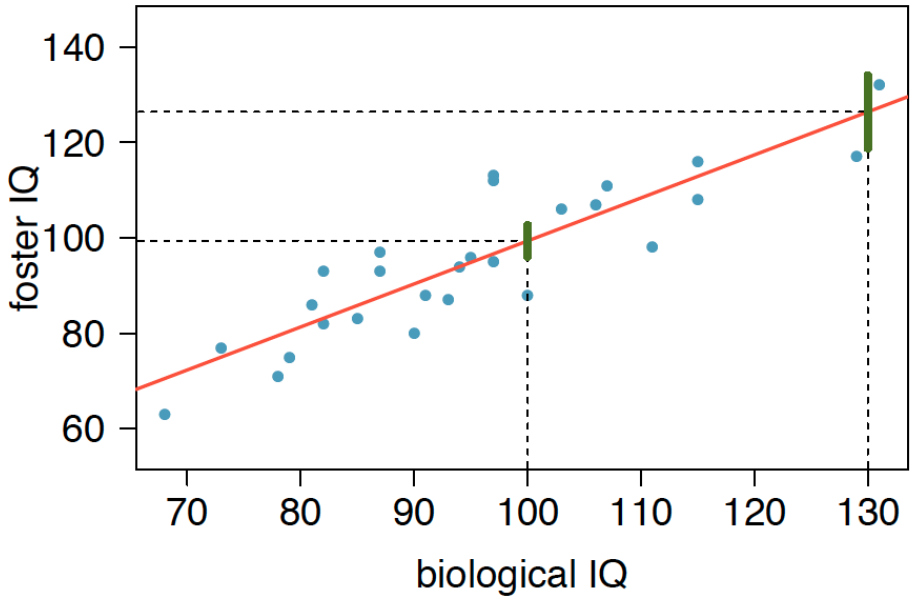
(c) same width

(d) cannot tell

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

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

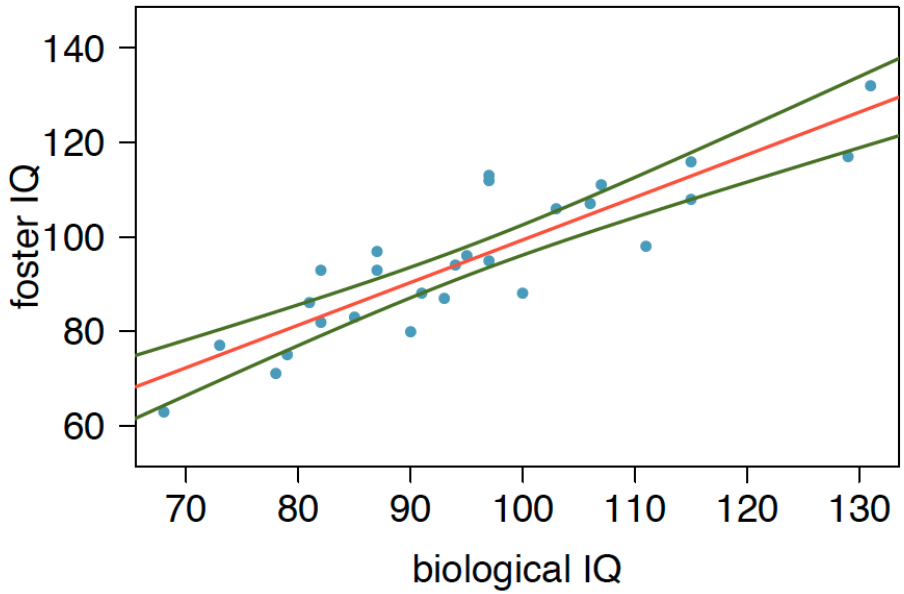
$$x^\star = 130, \quad ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(\mathbf{130} - 95.3)^2}{26 \times 15.74^2}} = 7.53$$



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.



Recap

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 $x^\star$.

We would expect there to be more uncertainty around a specific predicted value.

Prediction interval for specific predicted values

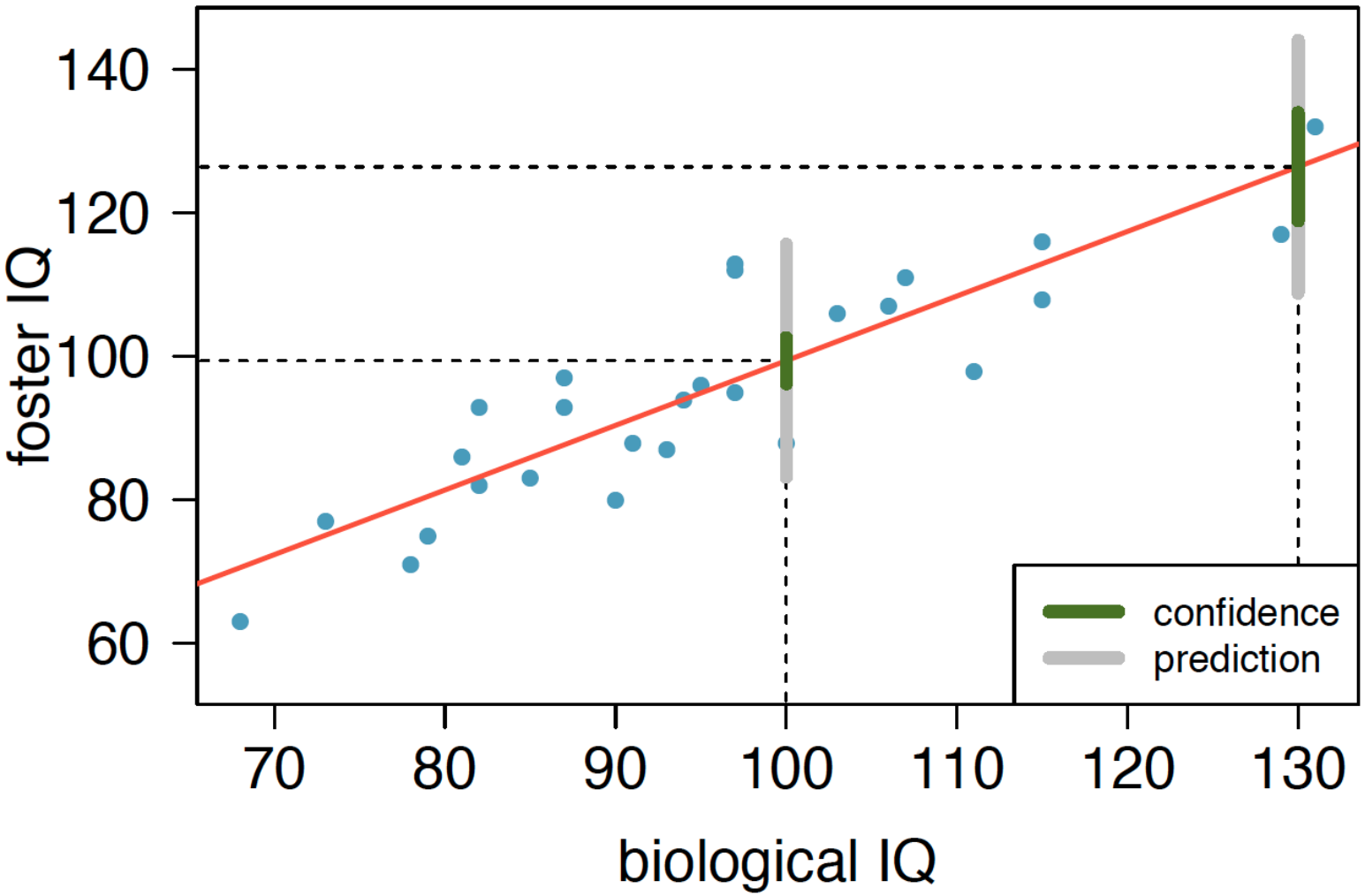
A **prediction interval** for y for a given $x^\star$ is

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

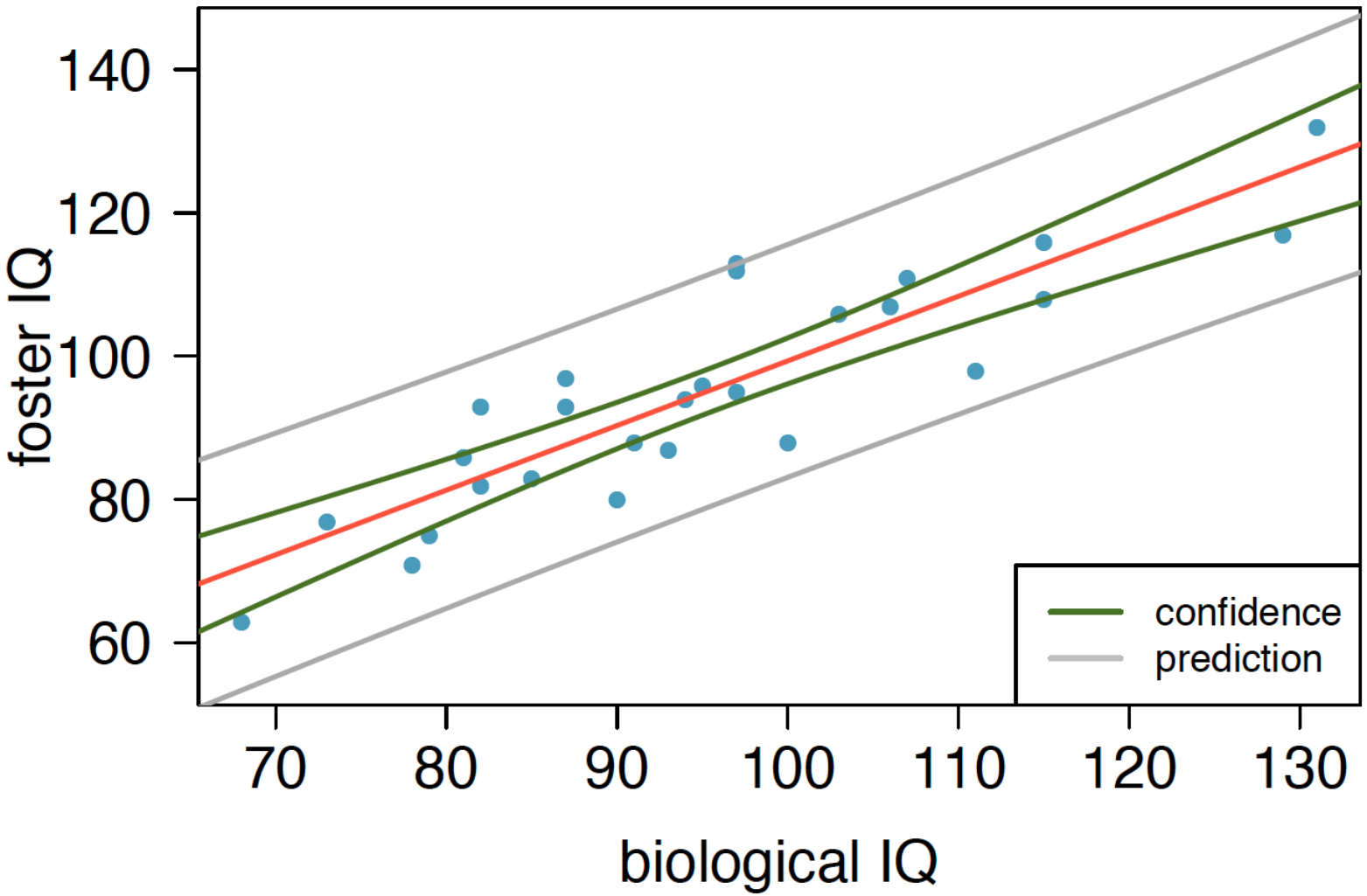
where s is the standard deviation of the residuals.

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

CI for $E(Y)$ v.s. PI for y



CI for $E(y)$ v.s. PI for y



CI for $E(y)$ v.s. 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 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 $E(y)$ v.s. 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 future $x^\star$ 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.

Recap

Many of the inference procedures we introduced that were used for samples from a normal distribution can be extended to the simple linear regression model. The theorems that allowed us to conclude that various statistics had t distribution continues to apply in the regression case.

In the next lecture, we will introduce multiple linear regression.

Suggested reading:

- D.S. Sec. 11.2, 11.3
- OpenIntro3: 7.4