inference for linear regression

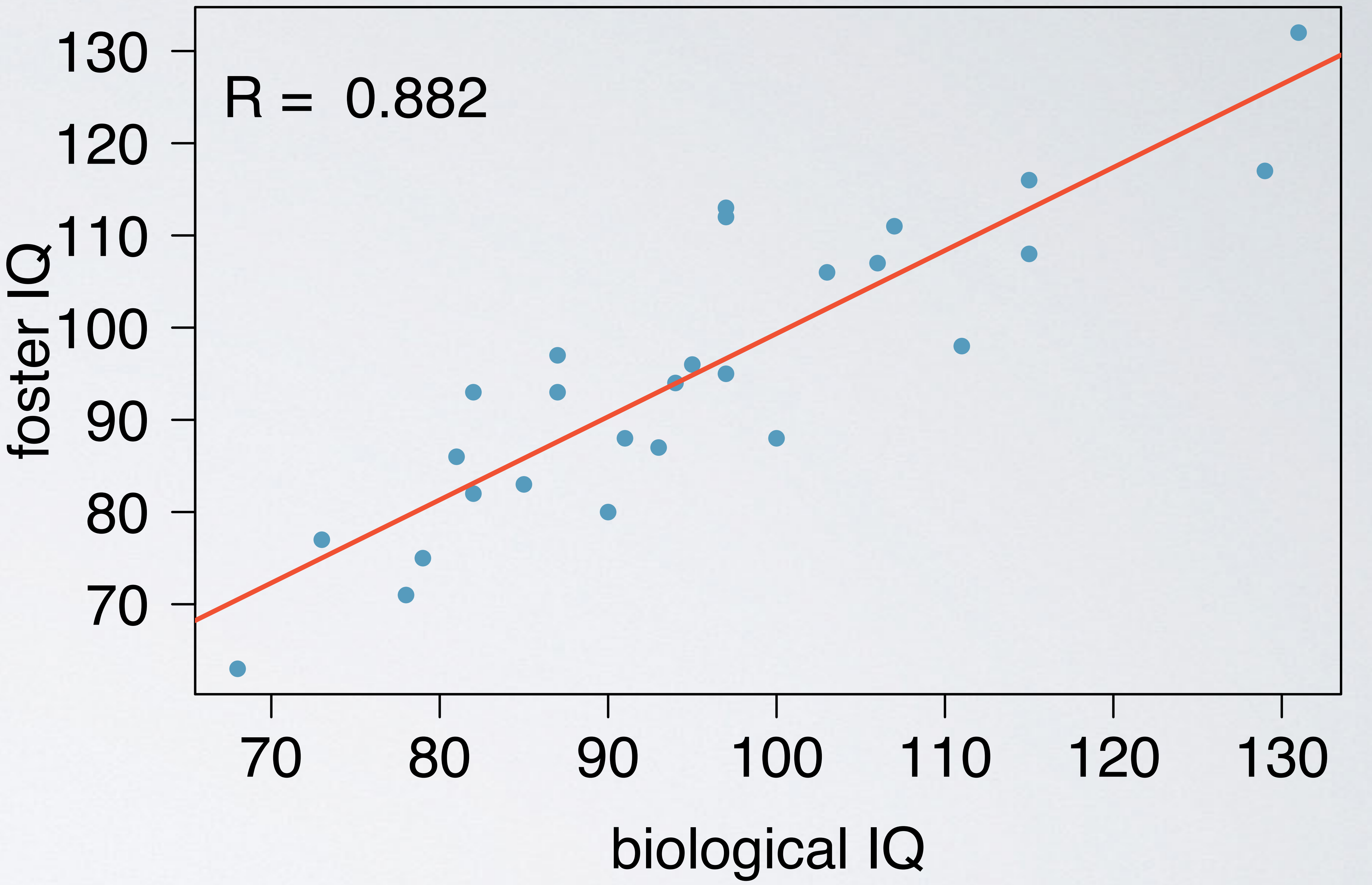
- hypothesis testing for significance of predictor
- confidence interval for slope
- conditions for inference

Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

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.



results

regression output:

	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

linear model:

$$\widehat{fosterIQ} = 9.2076 + 0.9014\ bioIQ$$

R^2:

$$R^2 = 0.78$$

testing for the slope - hypotheses

Is the explanatory variable a significant predictor of the response variable?

H_0 (nothing going on): The explanatory variable is not a significant predictor of the response variable, i.e. no relationship $\rightarrow$ slope of the relationship is 0.

$$H_0 : \beta_1 = 0$$

H_A (something going on): The explanatory variable is a significant predictor of the response variable, i.e. relationship $\rightarrow$ slope of the relationship is different than 0.

$$H_A : \beta_1 \neq 0$$

testing for the slope - mechanics

use a t-statistic in inference for regression

b_1

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

SE_{b_1}

t-statistic for the slope:

$$T = \frac{b_1 - 0}{SE_{b_1}} \qquad df = n - 2$$

focus on:
df = n - 2

Lose 1 df for each parameter estimated, and in linear regression we estimate 2 parameters: β_0 and β_1

calculating the test statistic

	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

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

$$df = 27 - 2 = 25$$

$$\text{p-value} = P(|T| > 9.36) \approx 0$$

confidence interval for the slope

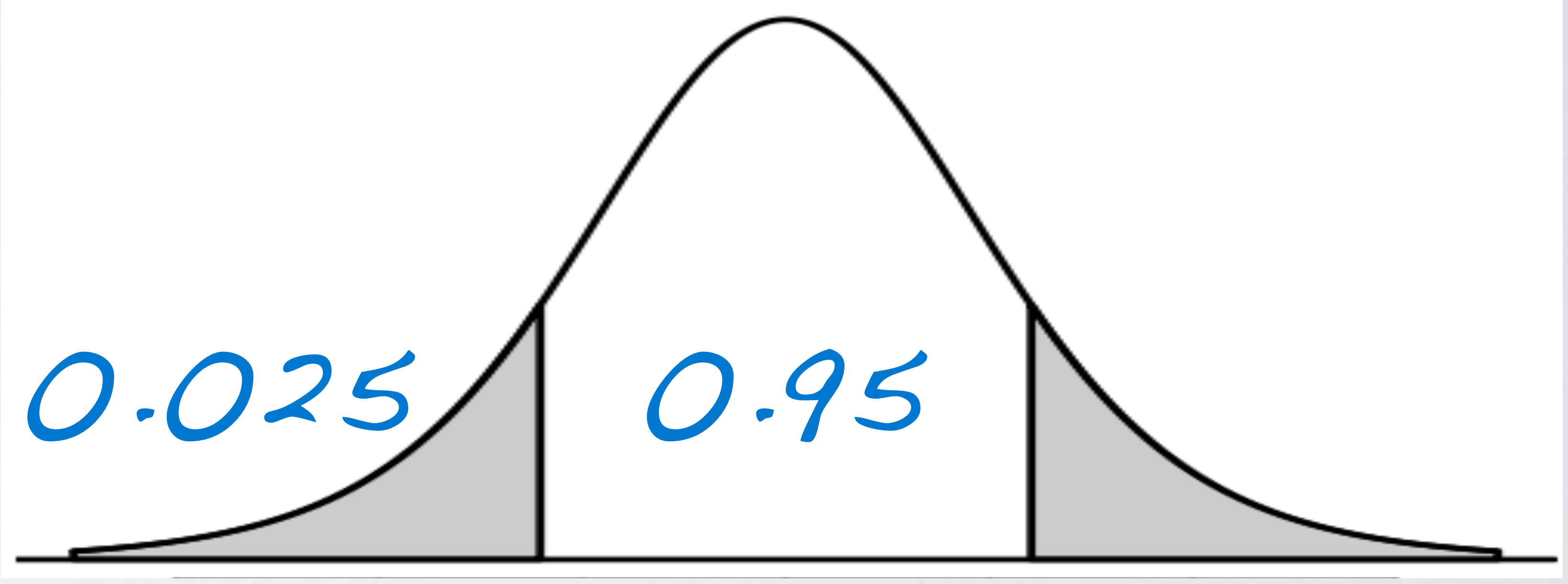
point estimate $\pm$ margin of error

$$b_1 \pm t^{\star}_{df} SE_{b_1}$$

practice

Calculate the 95% confidence interval for the slope of the relationship between biological and foster twins' IQs?

	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



R

```
> qt(0.025, df = 25)
[1] -2.059539
```

$df = 27 - 2 = 25$

$t^*_{25} = 2.06$

$0.9014 \pm 2.06 \times 0.0963 = (0.7, 1.1)$

practice

Interpret the 95% confidence interval for the slope of the relationship between biological and foster twins' IQs: (0.7, 1.1)

We are 95% confident that for each additional point on the biological twins' IQs, the foster twins' IQs are expected on average to be higher by 0.7 to 1.1 points.

recap - inference for regression

hypothesis test:

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

confidence interval:

$$b_1 \pm t^\star_{df} SE_{b_1}$$

- Null value is often 0, since we usually check for **any** relationship between the explanatory and the response variables.
- Regression output gives b_1, SE_{b1}, and two-tailed p-value for the t-test for the slope where the null value is 0.
- Inference on the intercept is rarely done.

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 already 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.

Image source: http://openclipart.org/detail/1695/warning-sign-by-zeimusu