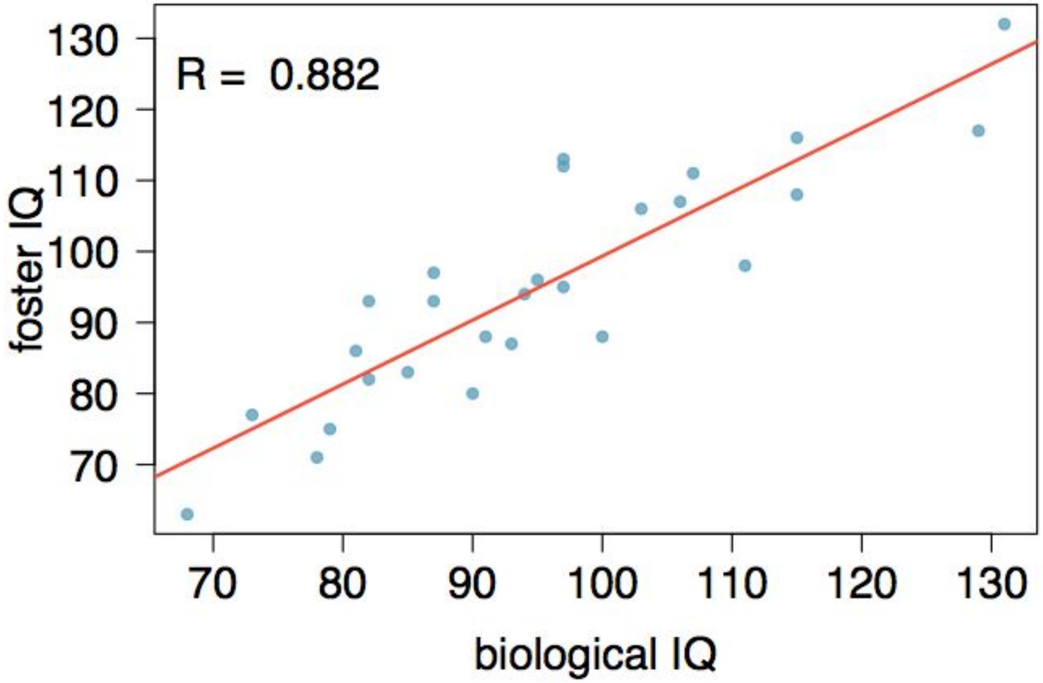
Section 8.4
Inference for Linear Regression

Stats 7 Summer Session II 2022

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.



Practice

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

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

Inference for the slope

- With a linear model our focuses are usually either:
  - Predicting values of the response for a given value of the explanatory/predictor variable
  - Studying the linear relationship between the response and the explanatory variables
    - Testing if the slope, β_1, is non zero (i.e. the two variables are linearly related)
    - Creating a confidence interval to estimate the slope
  - We could do the same investigation with the intercept, β_0, but the intercept is rarely of interest

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?

(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) $\boxed{H_0: \beta_1 = 0; H_A: \beta_1 \neq 0}$

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

- We always use a t-test in inference for regression.

  Remember: test statistic T = (point estimate - null value) / SE

- Point estimate = b_1 is the observed slope.
- SE_{b1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

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

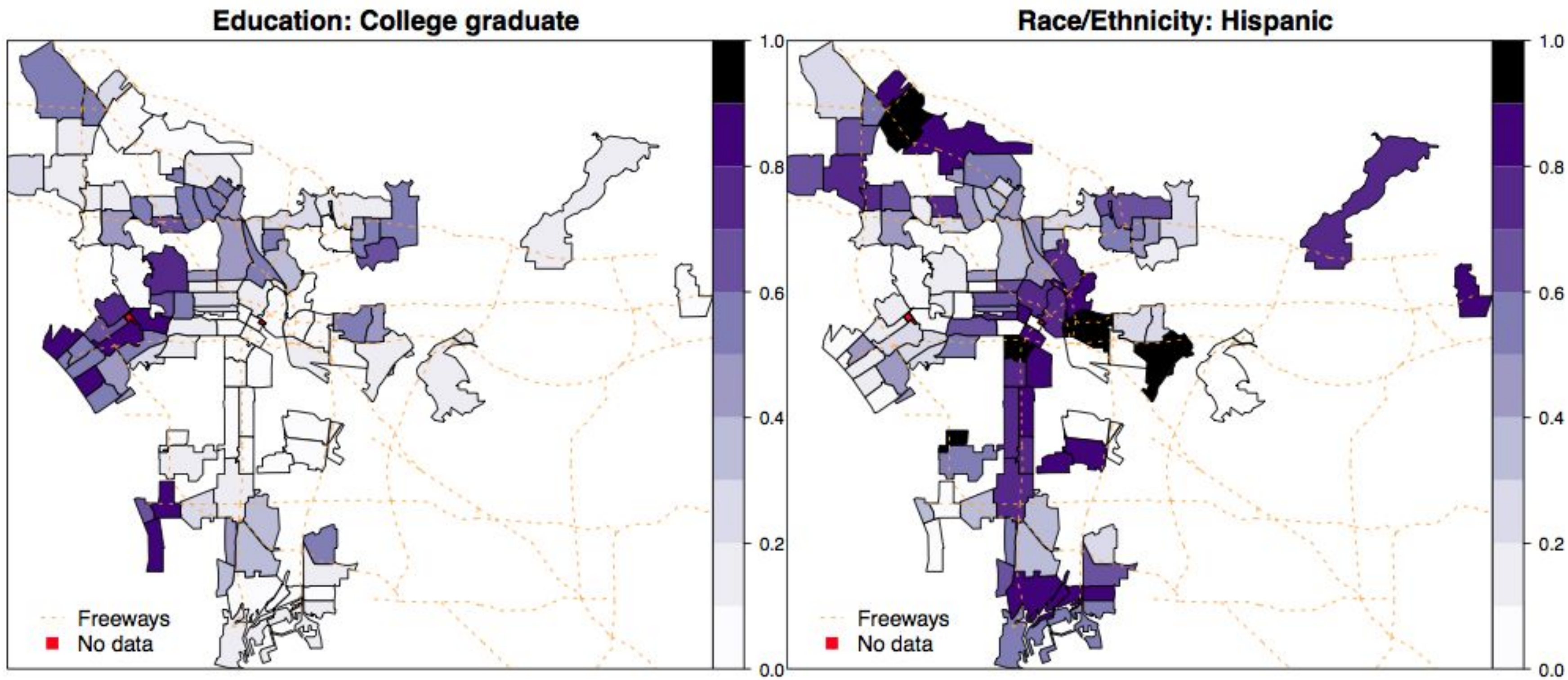
$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

$$df = 27 - 2 = 25$$

$$p - value = P(|T| > 9.36) < 0.01$$

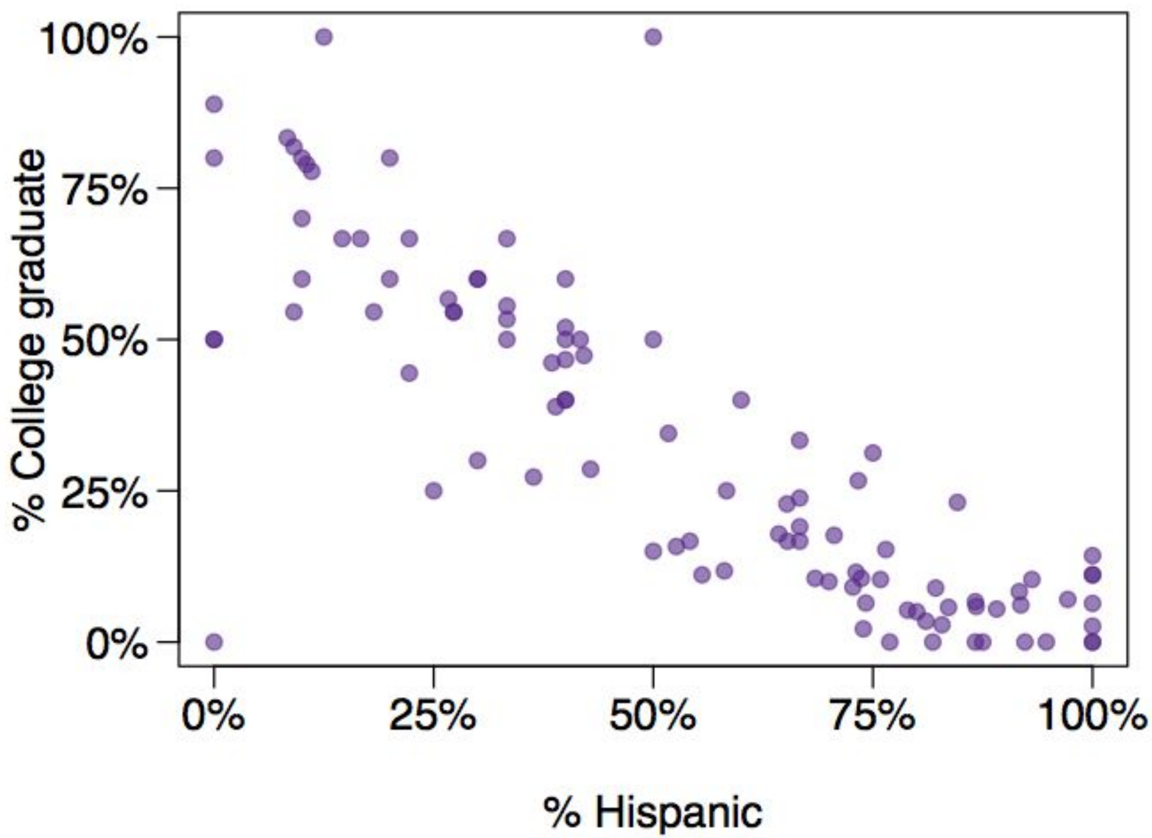
Percent college graduate vs. percent Hispanic in LA

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

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

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	$\boxed{0.9014}$	$\boxed{0.0963}$	9.36	0.0000

(a) 9.2076 ± 1.65 x 9.2999

$\boxed{\text{(b) } 0.9014 \pm 2.06 \times 0.0963}$

(c) 0.9014 ± 1.96 x 0.0963

(d) 9.2076 ± 1.96 x 0.0963

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

```
> qt((1 - 0.95) / 2, df = 25)
[1] -2.059539
```

(0.7, 1.1)

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:
$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$
  - Confidence interval:
$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$
- 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_{b1}, 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

- 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.
- The ultimate goal is to have independent observations.