Sta 111 - Summer II 2017
Probability and Statistical Inference

20. Inference for linear regression

Lu Wang

Duke University, Department of Statistical Science

August 2, 2017

Outline

1. Hypothesis testing for the slope
    1. Understanding regression output from software
    2. Use a *t*-test in inference for regression

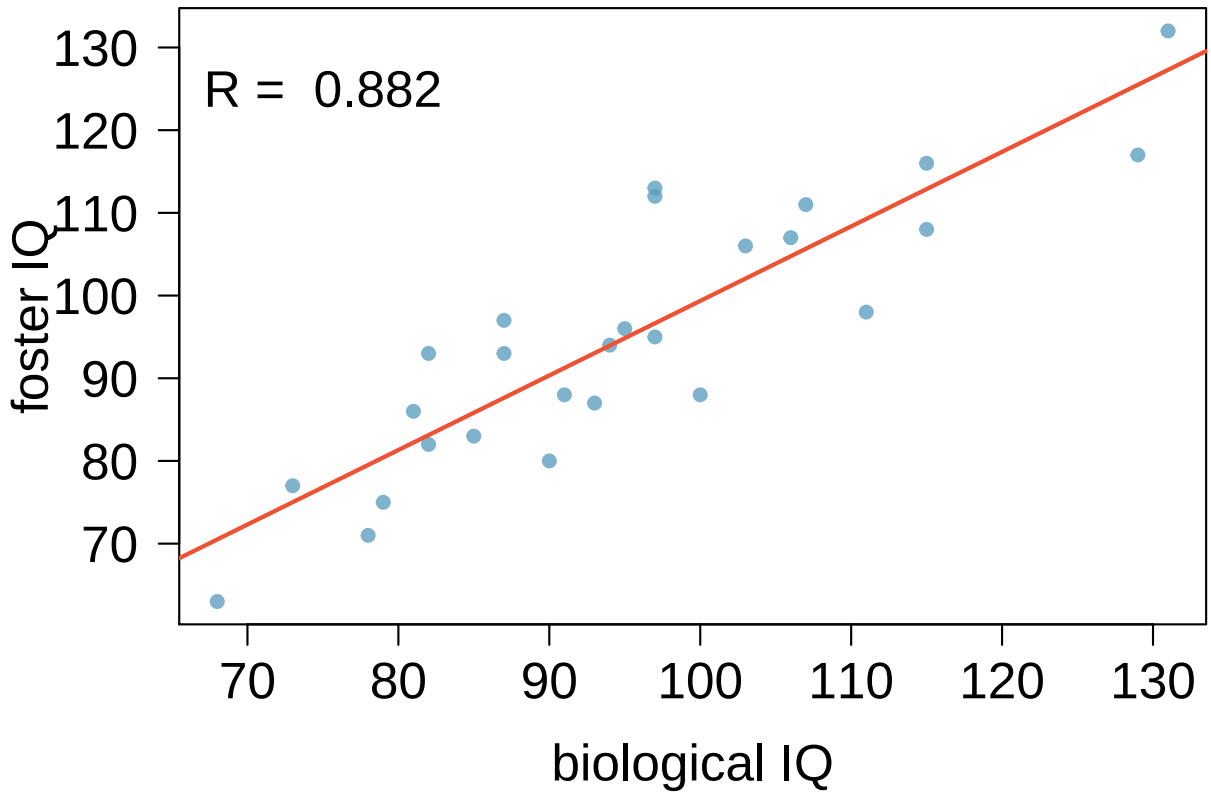
2. CI for the slope

3. Summary

4. Caution

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 27 identical twins, one raised by foster parents, the other by the biological parents.



Testing for the slope

The fitted least-squares regression line:

$$\widehat{\text{fosIQ}} = 9.21 + 0.90 \times \text{bioIQ}.$$

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

We can frame this question into a hypothesis test. 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

Recall: Test statistic $= \frac{\textit{point estimate} - \textit{null value}}{SE}$

In the case of regression:

- point estimate = b_1 is the observed slope.
- SE is the standard error associated with the slope.
- rely on statistical software to compute the point estimates and SE.

```
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,^^IAdjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

Does the test statistic follow the t-distribution under the null hypothesis?

Yes, under certain conditions ...

Recall: conditions for regression

Important for inference

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

Important regardless of doing inference

- *Linearity* $\rightarrow$ The data should show a linear trend.

Testing for the slope: $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

	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

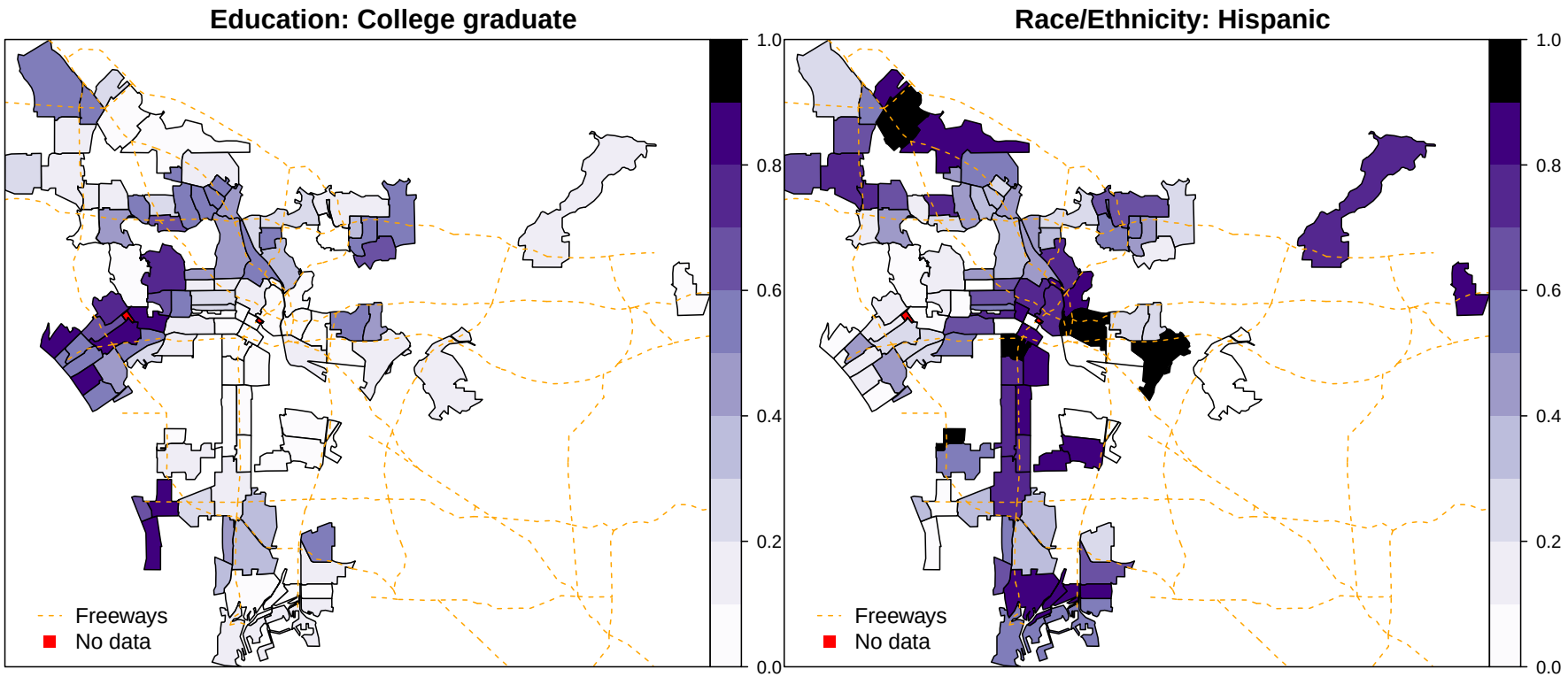
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.
  - We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

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

- 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*).

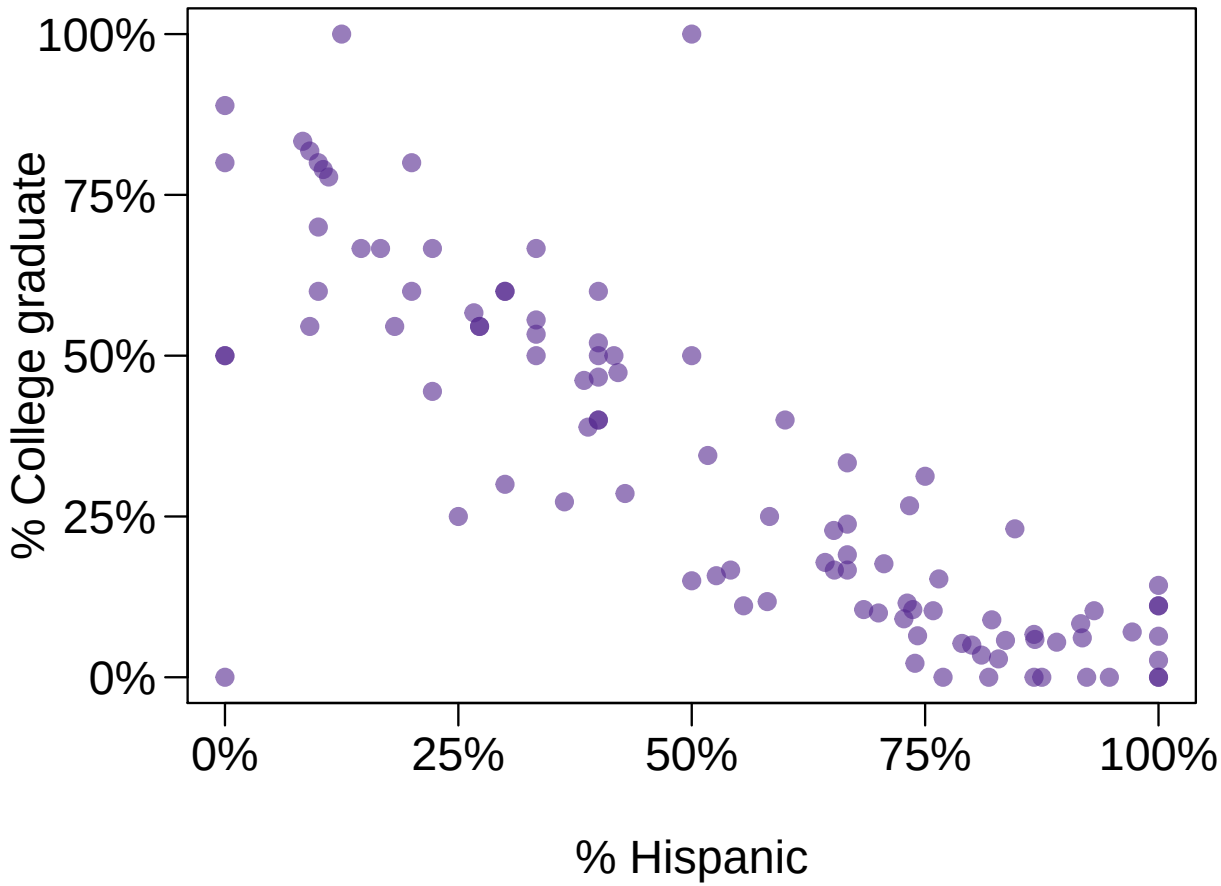
% College graduate vs. % Hispanic in LA

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

Do these data provide 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?

Confidence interval for the slope

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

What is the 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

$$
\begin{aligned}
n &= 27 \qquad df = 27 - 2 = 25 \\
95\%: \; t^\star_{25} &= 2.06 \\
0.9014 &\pm 2.06 \times 0.0963 = (0.7, 1.1)
\end{aligned}
$$

Recap: Inference for regression uses the *t*-distribution

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

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

  - 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_{b_1}, and *two-tailed* p-value for the *t*-test for the slope where the null value is 0.
  - If your test is one-sided and the point estimate is in the direction of H_A, then you can halve the software's p-value to get the one-tail area.
- 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 estimate may not be valid.
- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- If you have a sample that is non-random, inference on the results will be unreliable.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.