Unit 4: Regression and Prediction

3. Inference for Linear Regression

(Chapter 5.4)

11/21/2016

Quiz 8 - linear regression

Recap from last time

1. We can use the slope and intercept of a regression line to make predictions
2. We can also sometimes extrapolate, but this can be fraught
3. Just like t-tests and the other statistics we've explored so far, linear regression models are appropriate only when some conditions are met

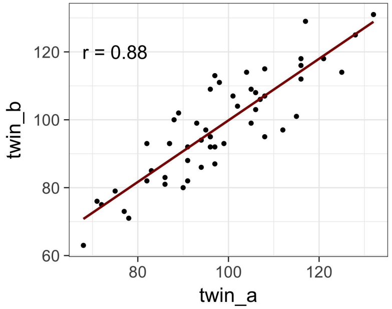
Key ideas

1. A regression's slope codes the relationship between the two measures
2. Correlation is equivalent to the slope of a regression for standardized values
3. Inference for regression parameters uses t-tests

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] 53 identical twins, separated within 6-months of birth and raised apart



Practice Question 1: Interpreting regression output

```
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  9.08670    6.92036    1.313    0.195
twin_a       0.90741    0.07004   12.957   <2e-16 ***
---
Residual standard error: 7.417 on 51 degrees of freedom
Multiple R-squared:  0.767,   Adjusted R-squared:  0.7624
F-statistic: 167.9 on 1 and 51 DF,  p-value: < 2.2e-16
```

Which of the following is false?

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

(b) Roughly 91% of the variance in twins' IQs can be predicted by the model.

(c) The linear model is `twin_b = 9.08 + 0.91 x twin_a`.

(d) Twins in group b with IQs higher than average IQs tend to have biological twins in group a with higher than average IQs as well.

Practice Question 1: Interpreting regression output

```
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  9.08670    6.92036    1.313    0.195
twin_a       0.90741    0.07004   12.957   <2e-16 ***
---
Residual standard error: 7.417 on 51 degrees of freedom
Multiple R-squared:  0.767,   Adjusted R-squared:  0.7624
F-statistic: 167.9 on 1 and 51 DF,  p-value: < 2.2e-16
```

Which of the following is false?

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

(b) Roughly 91% of the variance in twins' IQs can be predicted by the model.

(c) The linear model is `twin_b = 9.08 + 0.91 x twin_a`.

(d) Twins in group b with IQs higher than average IQs tend to have biological twins in group a with higher than average IQs as well.

Practice Question 2: Testing the relationship

Assuming that these 53 pairs of twins are a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of a biological twin is a significant predictor of IQ of the other 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) $H_0: \beta_1 = 0; H_A: \beta_1 \neq 0$

Practice Question 2: Testing the relationship

Assuming that these 53 pairs of twins are a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of a biological twin is a significant predictor of IQ of the other 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) $H_0: \beta_1 = 0; H_A: \beta_1 \neq 0$

Analyzing the slope of the regression line

	estimate	std.error	t-value	p-value
(Intercept)	9.0867	6.9203	1.3130	0.1950
twin_a	0.9074	0.0700	12.956	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 of the slope.

Degrees of freedom of 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.

Analyzing the slope of the regression line

	estimate	std.error	t-value	p-value
(Intercept)	9.0867	6.9203	1.3130	0.1950
twin_a	**0.9074**	**0.0700**	**12.956**	**0.0000**

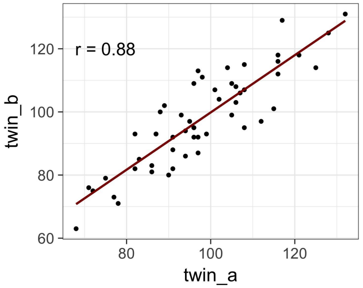
$$T = \frac{.9074 - 0}{.0700} = 12.956$$

$$df = 53 - 2 = 51$$

$$p - value = P(|T| > 12.956) < .001$$

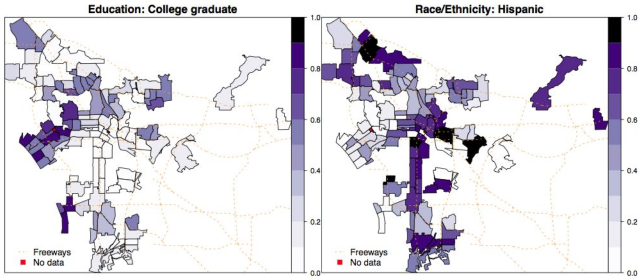
What is the relationship between slope and correlation?

	estimate	std.error	t-value	p-value
(Intercept)	9.0867	6.9203	1.3130	0.1950
twin_a	**0.9074**	**0.0700**	**12.956**	**0.0000**



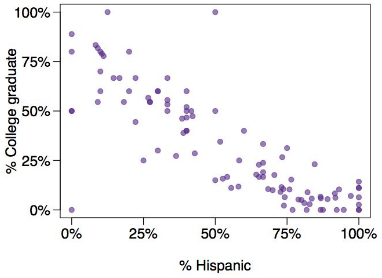
Let’s try another example

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



Let’s try another example

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



Practice Question 3: Inference from a logistic regression

	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

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

(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 areas with no Hispanic residents, % college graduates is expected to be 75%.

Practice Question 3: Inference from a logistic regression

	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

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

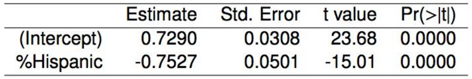
(a) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.

(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 areas with no Hispanic residents, % college graduates is expected to be 75%.

So what should we conclude?



	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

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

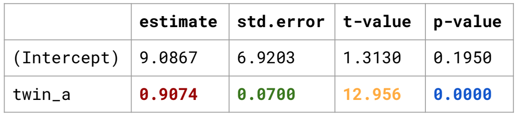
Yes, the p-value for % Hispanic is low, indicating that the data provide convincing evidence that the slope parameter is different than 0.

How reliable is this evidence if the zip codes are not randomly selected?

Not very...

Practice Question 4: Confidence intervals for regression estimates

Remember that a confidence interval is calculated as point estimate ± 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 53 twins).**



	estimate	std.error	t-value	p-value
(Intercept)	9.0867	6.9203	1.3130	0.1950
twin_a	0.9074	0.0700	12.956	0.0000

(a) 9.0867 ± 1.65 x 6.9203

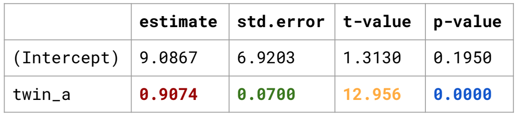
(b) .9074 ± 2.01 x .0700

(c) .9074 ± 1.96 x .0700

(d) 9.0867 ± 1.96 x .0700

Practice Question 4: Confidence intervals for regression estimates

Remember that a confidence interval is calculated as point estimate ± 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 53 twins).**



	estimate	std.error	t-value	p-value
(Intercept)	9.0867	6.9203	1.3130	0.1950
twin_a	0.9074	0.0700	12.956	0.0000

(a) 9.0867 ± 1.65 x 6.9203

(b) .9074 ± 2.01 x .0700

(c) .9074 ± 1.96 x .0700

(d) 9.0867 ± 1.96 x .0700

n = 53   df = 53 - 2 = 51

95%: t_{51}^{*} = 2.06

0.9074 ± 2.01 x 0.0700

(0.767, 1.05)

Inference for linear regression

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

Hypothesis test: $T = \frac{b_1 - null\ value}{SE_{b_1}}$ $df = n - 2$

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 focusing on the slope.

Key ideas

1. A regression’s slope codes the relationship between the two measures
2. Correlation is equivalent to the slope of a regression for standardized values
3. Inference for regression parameters uses t-tests