Statistical Inference

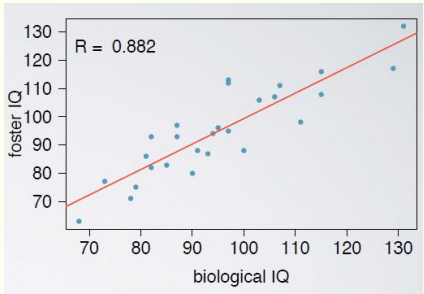
Introduction to Linear Regression

Behnam Bahrak

Spring 2020

Inference for Linear Regression

- 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.



Behnam Bahrak
bahrak@ut.ac.ir

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):

$$H_0 : \beta_1 = 0$$

The explanatory variable is not a significant predictor of the response variable, i.e. no relationship → slope of the relationship is 0.

H_A (something going on):

$$H_A : \beta_1 \neq 0$$

The explanatory variable is a significant predictor of the response variable, i.e. relationship → slope of the relationship is different than 0.

Testing for the Slope - Mechanics

- Use a t-statistic in inference for regression

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

b_1 is the point estimate; SE_{b_1} is the SE

t-statistic for the slope:	$T = \frac{b_1 - 0}{SE_{b_1}} \qquad df = n - 2$

Focus on degrees of freedom

- Degrees of freedom for linear regression:
  - $df = n - 2$
- Lose 1 df for each parameter estimated
- In linear regression we estimate 2 parameters:

β_0 and β_1

Standard Errors

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}$$

$$SE_{b_0} = \frac{s}{\sqrt{n}} \times \sqrt{1 + \frac{(\bar{x})^2}{s_x^2}} \qquad SE_{b_1} = \frac{s}{\sqrt{n}} \times \frac{1}{s_x}$$

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

p-value = P(|T| > 9.36) ≈ 0

Confidence Interval for the Slope

point estimate ± margin of error

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

Example

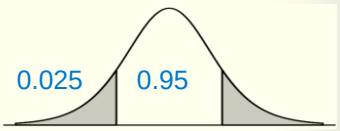
- Calculate the 95% confidence interval for the slope of the relationship between biological and foster twins' IQs?

	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

$$df = 27 - 2 = 25$$

$$t^{*}_{25} = 2.06$$

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



R

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

Example

- 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:

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

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_{b_1}, 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.

Variability Partitioning

- So far: t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.
- Alternative: 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

	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:

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

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

explained variability in y: $SS_{Reg} = 6724.66 - 1493.53 = 5231.13$

ANOVA Output

	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:

total degrees of freedom: $df_{Tot} = 27 - 1 = 26$

regression degrees of freedom: $df_{Reg} = 1$ only 1 predictor

residual degrees of freedom: $df_{Res} = 26 - 1 = 25$

ANOVA Output

	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

MS regression: $MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$

MS residual: $MS_{Res} = \frac{SS_{Res}}{df_{Res}} = \frac{1493.53}{25} = 59.74$

F statistic — ratio of explained to unexplained variability

$$F_{(1,25)} = \frac{MS_{Reg}}{MS_{Res}} = 87.56$$

ANOVA

	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			

small p-value → reject H_0

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

- The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

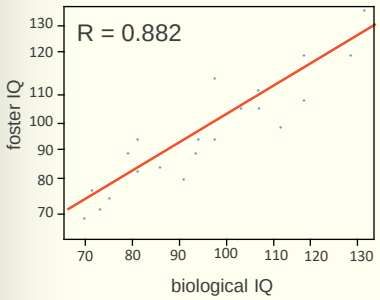
Revisiting R^2

- R^2 is the proportion of variability in y explained by the model:
  - large → linear relationship between x and y exists
  - small → evidence provided by the data may not be convincing
- Two ways to calculate R^2:

  (1) using correlation: square of the correlation coefficient

  (2) from the definition: proportion of explained to total variability

Revisiting R^2



	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			

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

(2) $R^2 = \dfrac{\text{explained variability}}{\text{total variability}} = \dfrac{SS_{Reg}}{SS_{Tot}} = \dfrac{5231.13}{6724.66} \approx 0.78$