Linear Regression with R/Stata

Day 3: Categorical Predictors & Inference

Constantin Manuel Bosancianu

February 27, 2019

Wissenschaftszentrum Berlin
Institutions and Political Inequality
bosancianu@icloud.com

Preamble

Recap from yesterday

- ✓ Model fit: residual standard error & R^2 (coefficient of determination);
- ✓ (adjusted) R^2 is most used measure of fit;
- ✓ Interpretation of coefficients in the simple and multiple case is identical;
- ✓ Multiple regression coefficients are *partial* coefficients—effect of X while keeping all other predictors constant.

NLSY example

	Model 1
(Intercept)	81.285***
	(2.303)
Mother's IQ	0.563***
	(0.061)
Mother graduated HS	5.647*
	(2.258)
Mother's age	0.225
	(0.331)
Num. obs.	434

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
DV (outcome) is child's IQ measured at age 3.

Predicting children's IQ (IQ rescaled by 100, age by 18)

Categorical predictors

Categorical predictors

Linear regression can accommodate them without a problem.

Indicator (“dummy”) variables: can take only two values—female (yes/no), country in Europe (yes/no), post-crisis (yes/no) etc.

The coefficient for such variables is interpreted as a difference in the levels of Y for the two categories.

Multi-category variables can be transformed into a series of indicator variables.

Example: Boston house prices

	Cont. only	Cont. and cat.
(Intercept)	−50.878***	−47.981***
	(6.121)	(5.755)
Average num. rooms	11.842***	11.327***
	(0.948)	(0.894)
Next to Charles river (dich.)		9.575***
		(2.506)
Num. obs.	92	92

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
DV (outcome) is average house price in Boston townships (in 1,000s of USD).

Comparison of regressions

The model is $Price = a + b_1 Rooms + b_2 River + e$.

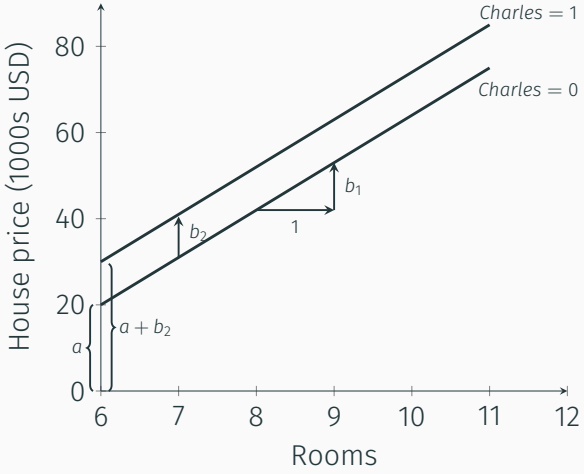
Example: Boston house prices

	Cont. only	Cont. and cat.
(Intercept)	20.175***	19.980***
	(0.679)	(0.635)
Average num. rooms	11.842***	11.327***
	(0.948)	(0.894)
Next to Charles river (dich.)		9.575***
		(2.506)
Num. obs.	92	92

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
DV (outcome) is average house price in Boston townships (in 1,000s of USD).

Number of rooms rescaled by 6

Visualizing the model



Dummy variable regression

More than 2 categories

Turn into a set of indicator variables—28 EU countries turned into 28 variables.

For n categories, only $n - 1$ indicator variables can be in the regression model. One category must be used as a reference category.

Works the same for an indicator variable: 2 categories (male/female) result in only 1 variable included in model.

For the Boston price data, take the example of crime rate: low (1–4.99 crimes per 1,000 residents), middle (5–14.99 crimes/1,000), and high (15 crimes and above/1,000).

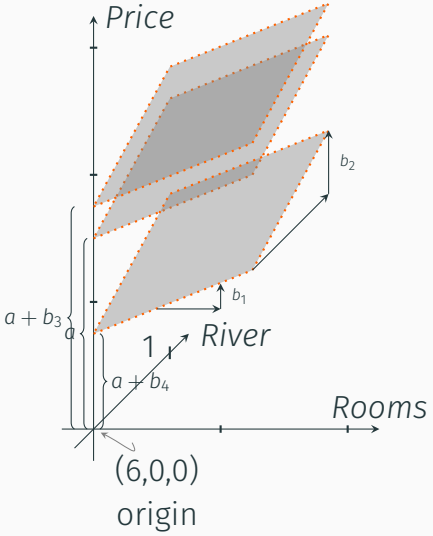
Example: Boston house prices

	Model 1	Model 2
(Intercept)	19.980***	20.062***
	(0.635)	(0.682)
Average num. rooms	11.327***	11.428***
	(0.894)	(0.901)
Next to Charles river (dich.)	9.575***	8.764***
	(2.506)	(2.376)
Moderate crime (5 to 15/1000)		2.579
		(1.714)
High crime (over 15/1000)		−7.876**
		(2.353)
Num. obs.	92	92

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
DV (outcome) is average house price in Boston townships (in 1,000s of USD).

Excluded category: low crime (1–4.99 per 1,000 residents)

Visualizing the model



Inference in simple regression

Inference: from sample to population

How do we know that what we find in a sample is also valid in the population?

In the NLSY example, 434 tested children are a sample of the population (all US children aged 3).

How much would a and b vary if different samples of size n would be selected?

Ways of presenting this:

- ✓ Standard errors (based on variance)
- ✓ Confidence intervals

Sampling variance

We can compute *how much* the coefficients are expected to change, on average.

$$V(a) = \frac{\sigma_e^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

$$V(b) = \frac{\sigma_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma_e^2}{(n-1)\sigma_x^2} \tag{2}$$

Small sampling variances are desirable—a and b would not be very different if we randomly selected another sample.

Sampling variance: behavior

$$V(b) = \frac{\sigma_e^2}{(n-1)\sigma_X^2} \tag{3}$$

- ✓ larger n means smaller $V(a)$ and $V(b)$;
- ✓ as σ_e^2 increases, so do $V(a)$ and $V(b)$;
- ✓ as $\sum_{i=1}^{n}(x_i - \bar{x})^2$ increases, $V(a)$ and $V(b)$ get smaller.

The formulas, in fact, use σ_ϵ^2, but we have to rely on σ_e^2 as an approximation (e_i=sample residuals; ϵ_i=(hypothetical) population residuals)

Null Hypothesis Significance Testing (NHST)

A standard t test of the hypothesis that $b \neq 0$, with the null hypothesis $b_0 = 0$.

$$t_0 = \frac{b - b_0}{\sigma_b} \tag{4}$$

Because σ_b is the standard deviation of an estimated quantity (b), it's technically a *standard error*.

Null Hypothesis Significance Testing (NHST)

$$t_0 = \frac{b - b_0}{\sigma_b}$$

t_0: t distribution with $n - k - 1 = n - 2$ degrees of freedom.

If t_0 is larger than the critical value at that level of significance, then the H_0 is rejected and H_1 is (indirectly) supported.

⏭ More NHST

Confidence intervals

An alternative way of presenting uncertainty. Once we have $V(a)$ and $V(b)$, they are easy.

A $100(1-\alpha)\%$ confidence interval for b is:

$$[b - t_{\frac{\alpha}{2}}\sigma_b; b + t_{\frac{\alpha}{2}}\sigma_b] \tag{5}$$

Critical t values for $\alpha = 0.05$ (two-tailed):

- ✓ $t_{0.025} \approx 1.96$ for $n \geq 500$;
- ✓ $t_{0.025} \approx 2$ for $n \approx 60$;
- ✓ $t_{0.025} \approx 2.1$ for $n \approx 20$;

Confidence intervals

With b, t and $V(b)$ we have all the "ingredients" to construct the confidence interval (CI).

Two aspects are of relevance:

- ✓ the width of the interval—the wider, the more uncertainty we have about β (population value);
- ✓ whether it intersects 0—if it does, we can't be sure that β is not, in fact, 0.

Inference in multiple regression

The case of multiple regression

Constructing CIs and conducting NHST is done in the same way as for simple regression.

$$V(b_j) = \underbrace{\frac{1}{1 - R_j^2}}_{\text{VIF}} \times \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_j - \bar{x}_j)^2} \tag{6}$$

The second part is the same as for simple regression. The first part is called the *variance inflation factor* (VIF).

R_j^2 is the model fit from a regression of X_j on all the other Xs (predictors) in the model.

VIF and multicollinearity

VIF is the reason why we didn't add all 3 indicator variables for crime rate.

Knowing the values on 2 of the indicators gives us the value for the third indicator as well.

Adding all 3 to the regression means that R_j^2 is 1, and that the variance for the indicator variables is ∞.

This logic also works when correlations between predictors are too high (e.g., above 0.80–0.85).

Inference for multiple slopes: F test

It can show you whether a model with k predictors fits the data better than a model with no predictors.

The null hypothesis for such a test would be:

$$H_0 : b_1 = b_2 = \cdots = b_k = 0 \tag{7}$$

This is called a “global” test, or an “omnibus” test, based on the F distribution.

F test

$$F_0 = \frac{\frac{RegSS}{k}}{\frac{RSS}{n-k-1}} = \frac{n-k-1}{k} \times \frac{R^2}{1-R^2} \tag{8}$$

RegSS and RSS are the same quantities we discussed in the model fit section, n is the sample size, and k is the number of predictors.

F-statistic has an F-distribution with n and $n-k-1$ degrees of freedom.

If F_0 surpasses the critical value for the test, then H_0 is rejected, and you may conclude that at least one of the $b_1, b_2, \ldots, b_k$ slopes is different from 0.

F test (cont.)

It can be used in the same way to see if a model with k predictors fits the data better than a model with fewer predictors.

Two models are compared: the "null" (predictors we're not interested in), and the "full" (all predictors).

$$Null: Y = a + b_4X_4 + b_5X_5 + e_1 \tag{9}$$
$$Full: Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_2 \tag{10}$$

This this case $H_0 : b_1 = b_2 = b_3 = 0$.

F test (cont.)

$$F_0 = \frac{\frac{RegSS_1 - RegSS_0}{q}}{\frac{RSS_1}{n-k-1}} = \frac{n-k-1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \tag{11}$$

$RegSS_0$, RSS_0, and R_0^2 refer to the null model. $RegSS_1$, RSS_1, and R_1^2 refer to the null model.

n is the sample size, k is the number of predictors in the full model (5), and q is the number of predictors in the full model that are not in the null model (3).

If the test is significant, then at least one of b_1, b_2 or b_3 is $\neq 0$.

Regression on population data

What do SEs mean in our California 1992 example, where our regression contains *all* the counties in California?

Two strategies:

1. Only focus on a and b and ignore the *SE*s, as they are meaningless.
2. Adopt the "superpopulation" assumption, e.g. the counties in California in 1992 are a sample out of all the possible ways in which counties in California might have developed historically.

The second strategy is accepted in the discipline, particularly if the goal is to make predictions outside of the sample.

More interpretation

Predicting emancipative values

In South Africa (from WVS 6):

- ✓ Age: in years (centered at 37 years);
- ✓ Gender: Female = 1; Male = 0;
- ✓ Education: 3 dummies (primary, secondary, tertiary);
- ✓ Marital status: single vs. everyone else;
- ✓ Income: 10 deciles;
- ✓ Supervisor: R. is a supervisor at work.

Emancipative values are measured on a constructed scale, ranging from 0 to 1.

Results

	Model 1	Model 2
(Intercept)	0.405 (0.009)***	0.408 (0.010)***
Age (centered)	0.000 (0.000)	0.000 (0.000)
Gender (female)	0.021 (0.004)***	0.022 (0.005)***
Education (secondary)	0.011 (0.008)	0.013 (0.008)
Education (tertiary)	0.028 (0.010)**	0.030 (0.011)**
Marital status (single)	0.009 (0.006)	0.011 (0.006)*
Income (decile)	0.006 (0.001)***	0.005 (0.001)***
Supervisor at work (yes)		−0.013 (0.008)
R^2	0.021	0.020
Adj. R^2	0.019	0.018
Num. obs.	3423	3083

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. Standard errors in brackets.
DV (outcome) is index of emancipative values at individual level (0-1 range).

For education, reference category is primary education completed.

Thank you for the kind attention!

NHST

Step 1: Under the null hypothesis (H_0) we expect $\beta = 0$ in the population.

Step 2: Our sample regression, however, produced a $\beta = b$, with a standard error of $SE = s$.

Step 3: A t-test for whether b is statistically significant is $t = \frac{b-0}{s}$.

NHST: final steps

Step 4: Compare this t value with a critical value of the t-test, for the # of d.f. in your regression, at a 5% significance level (probability of rejecting H_0 when it is true).

Step 5: If our t result is larger, β is statistically significantly different from 0 in the population. If t is smaller, it is not.

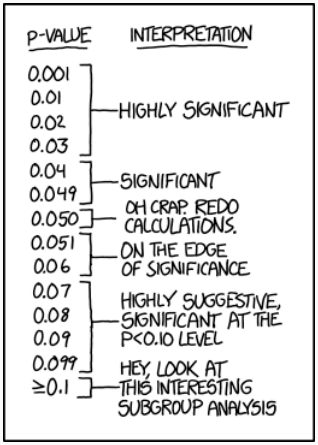
NHST

Thankfully, the software does this automatically. Given the size of the coefficient and SE, it can compute very precisely the probability that we would see an effect as strong as β *if* H_0 *were true*.

All we need to do as researchers is check if $p < .05$.

If $p > .05$, then we cannot be sure that the effect in the population is not, in fact, 0.

NHST (reality is messy and painful)



Go Back