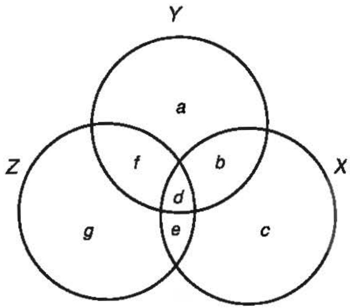
Quantitative Methods in Political Science: Linear Regression: Statistical Inference, Dummies and Interactions

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova | David M. Grundmanns

Week 6 - 13 October 2021

Quiz

Consider a linear model: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$. The figure represents the total variation of Y, X and Z each with a circle. Which of the following statements are true?



1. X and Z are correlated. Their covariation is represented by $d + e$.
2. The graph makes transparent that Z is related to both, Y and X.
3. The relationship between Y and X statistically controlling for Z is accounted for by the area of b.
4. Not controlling for Z in the model would attribute $b + d$ to the variation of X that is shared with Y.
5. If we fail to control for Z we will end up with a biased estimate of X's effect on Y.

The Course

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using software
  - R
  - Basic programming skills

Overview: Week 6

Significance Testing

- Significance Test for One Coefficient: CI, t-Test and p-value

Categorical Variables in Regression

Interactions

Significance Testing

Statistical Inference for Linear Models

- This lecture: Classical statistical regression inference including
  - confidence intervals for estimated coefficients,
  - significance tests for estimated coefficients using confidence intervals, t-test, p-values
- Next lecture: Interpretation of regression inference including
  - how to make results accessible to non-technical readers,
  - how to learn about quantities of interest,
  - how to display uncertainty of own results, and
  - which tools to use (predicted probabilities, expected values, and first differences).

Confidence Intervals for Regression Coefficients

- To assess the uncertainty around our estimates, we construct confidence intervals, such that this interval contains the true population parameter in, e.g., 95% of the hypothetically repeated samples.
- More formally, let α $(0 < \alpha < 1)$ be the level of significance and δ be a positive number. Then, the confidence interval around β_j is defined as

$$Pr(\hat{\beta}_j - \delta \leq \beta_j \leq \hat{\beta}_j + \delta) = 1 - \alpha.$$

- One strategy: Assume normal sampling distribution (i.e., *normal approximation*) and given that we know the standard errors of the coefficients we can construct confidence intervals (i.e., $\delta \approx 1.96 \cdot SE(\hat{\beta}_j)$)

Confidence Intervals for Regression Coefficients

- Other strategy: We analytically construct a confidence interval using a normalized test statistic. The test statistic t^* for our hypothesized value of β_j can be calculated as

$$t^* = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}}} \sim t_{(n-k-1)}.$$

- Since we use $\hat{\sigma}^2$ instead of the true population variance, σ^2, the test statistic is no longer normally distributed, but *t*-distributed with $n - k - 1$ degrees of freedom, where k is the number of independent variables.

Confidence Intervals for Regression Coefficients

- With such a normalized test statistic, t^*, and equal probability density at the lower and upper tails, a confidence interval for the true value β_j is given as

$$Pr(t_{(\frac{\alpha}{2})} \leq t^* \leq t_{(1-\frac{\alpha}{2})}) = 1 - \alpha.$$

- Substituting in our explicit expression for t^* and relying on the symmetry of the t-distribution, yields

$$Pr(\hat{\beta}_j - t_{(1-\frac{\alpha}{2})} \cdot SE(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{(1-\frac{\alpha}{2})} \cdot SE(\hat{\beta}_j)) = 1 - \alpha.$$

- Or more simply, we have the known expression:

$$\hat{\beta}_j \pm t_{(1-\frac{\alpha}{2})} \cdot SE(\hat{\beta}_j)$$

- When $n - k - 1 > 120$, then one can use the 97.5 percentile of the standard normal (i.e., 1.96) rather than the t-distribution (in fact, use 2 as a rule-of-thumb (!), i.e. $\delta \approx 2 \cdot SE(\hat{\beta}_j)$) to construct a 95% confidence interval around the true value β_j.

Is 95% Good Enough? Type I and Type II Errors

In general, though, tests are flawed. Tests detect things that don't exist (*false positive*), and miss things that do exist (*false negative*).

- Statistical inference is basically a decision problem between two alternatives:
  - H_0: Null hypothesis.
  - H_A: Alternative hypothesis.
- A 95% confidence interval means that under repeated experiments the given interval includes the true parameter, β, 95 out of 100 times. Hence, with H_0 being true, we falsely reject H_0 5 times out of 100 even though we should not have done so (type I error).

Is 95% Good Enough? Type I and Type II Errors

	H_0 is true	H_0 is false
Not reject H_0	correct decision	Type II error (false negative)
Reject H_0	Type I error (false positive)	correct decision

- Consider the errors for a case in which the hypotheses are "H_0: No disease" and "H_A: Disease". Which error would you "prefer"?

Type I and Type II Errors

- Assume that H_0 is that a patient has no disease.

	H_0: No disease	H_A: Disease
Not reject H_0	correct decision	Type II error (false negative)
Reject H_0	Type I error (false positive)	correct decision

- Then, for a type I error, the patient is told that s/he has the disease even though s/he does not. The test to diagnose the patient is positive ("Yes, you have the disease"), but falsely so.
- For a type II error, however, the patient is not diagnosed of having a disease even though s/he does have it. The test is negative ("No worries, you do not have the disease"), but falsely so.

Hypothesis Test for Coefficient Using t-Test

- In testing statistical significance of a regression coefficient, we usually want to know if our estimated coefficient $\hat{\beta}_j$ is different from zero, i.e.: $\beta_j^* \neq 0$.
- We test the null hypothesis about the true population parameter, β_j, against the (two-sided) alternative hypothesis:
  - $H_0 : \beta_j^* = 0$
  - $H_A : \beta_j^* \neq 0$
- Thus, we construct a test statistic, t^*, given our hypotheses about β_j^*

$$t^* = \frac{\hat{\beta}_j - \beta_j^*}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim t_{(n-k-1)}.$$

- We reject the null hypothesis, H_0, at the $\alpha-\%$ significance level if

$$\mid t^* \mid > t_{(1-\frac{\alpha}{2}, n-k-1)} \text{ ("critical value")},$$

where $n - k - 1$ are the degrees of freedom with k independent variables.

Another way to say the same thing: Computing p-Values

- So far we used a classical approach to hypothesis testing:
    - Specifying alternative (and null) hypothesis
    - Choose significance level (α)
    - Get the respective critical value ($t_{(1-\frac{\alpha}{2},n-k-1)}$) and compare it to test statistic (t^*)
    - H_0 is either rejected or not rejected at a chosen significance level
- Different scholars might prefer different significance levels (and the null might be not rejected at the 5% but at the 10% level. Which level is correct?)

Another way to say the same thing: Computing p-Values

- Alternative strategy: Given the observed t^*, what is the smallest significance level at which the null hypothesis would be rejected? This is called the p-value ($p \in (0,1)$).

$$p = Pr(|t_{(n-k-1)}| > |t^*|) = 2Pr(t_{(n-k-1)} > |t^*|)$$

  where $Pr(t > t^*)$ is the area to the right of t^* (given $(n-k-1)df$)
- Small $p-$values are evidence against the null, large p-values provide little evidence against the null.
- Say $p = .03$, then we would observe a value of the t statistic as extreme as we did in only 3% of all random samples if the H_0 is true. Thus, this is pretty strong evidence against the null. Hence, H_0 is not likely to be true.

Categorical Variables in Regression

Categorical Variables in Regression: Introduction

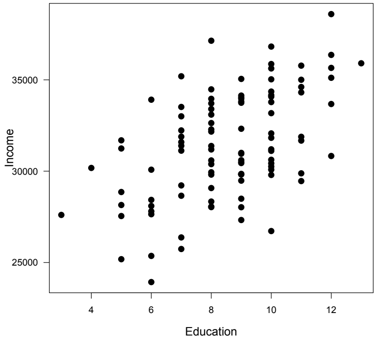
- In political science, variables are often qualitative or categorical.
- We can easily include qualitative information as independent variables in our regression model.
- Examples for qualitative data are:
  - Vote choice (Did vote or did not vote).
  - Gender (Is male or female).
  - Regime type (Is a democracy or an autocracy).
  - Membership status (Is a EU member state or not).

Dummy Variables

- Qualitative information often comes in the form of binary information. These zero-one variables are called dummies or dummy variables.
- These variables come with a trade-off:
  - Downside: Loss in information.
  - Upside: Dummy variables are easy to interpret.
- Good coding practice: Name your variable after the "1" category, e.g., it should be "female" and not "gender". This helps to avoid confusion!
- For further notes on "Coding style and Good Computing Practice", see Jonathan Nagler's website and, more recently a very interesting and helpful article by Nick Eubank (2016) in *The Political Methodologist*.

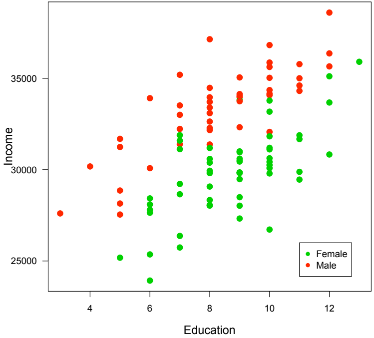
Example: Income and Education – A Gender Gap

- Suppose we want to examine the relationship between education and income among women and men.
- We collected the following fake data:



Example: Income and Education – A Gender Gap

- Suppose we want to examine the relationship between education and income among women and men.
- We collected the following fake data:



Example: Income and Education – A Gender Gap

- Our model: $Income = \beta_0 + \beta_1 * Education + \beta_2 * Female + \epsilon$
- Suppose we find the following estimates:

$$\widehat{Income} = 25934 + 894 \cdot Education - 3876 \cdot Female$$

- Using the *Female*-dummy, we get two regression lines. One for males and one for females:
- For females (if $Female = 1$) we obtain:

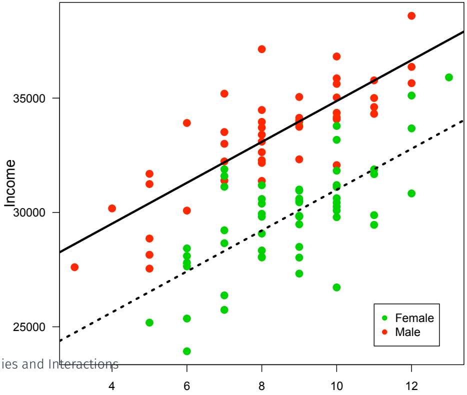
$$\widehat{Income} = (25934 - 3876 \cdot 1) + 894 \cdot Education = 22058 + 894 \cdot Education.$$

- For men (if $Female = 0$) we obtain:

$$\widehat{Income} = (25934 - 3876 \cdot 0) + 894 \cdot Education = 25934 + 894 \cdot Education.$$

Example: Income and Education – A Gender Gap

- Solid line for males: $\widehat{Income} = 25934 + 894 \cdot Education$
- Dashed line for females: $\widehat{Income} = 22058 + 894 \cdot Education$
- This illustrates that dummy variables shift the intercept up or down.



Using Dummy Variables for Multiple Categories

- Dummy variable trap.
  - Base group is represented by the intercept.
  - If we were to add a dummy variable for each group, we would introduce perfect multi-collinearity.
  - Statistical software usually warns you of this.
- Solution: Split a k-category variable into $k-1$ binary dummies.
- Interpretation is always relative to the baseline category.
- Suppose you analyze the effect of different social classes (lower, middle upper) on income ($\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2$):

	Dummy Variables		
Social Class	D_1	D_2	
lower	0	0	$\hat{Y} = \hat{\beta}_0$
middle	1	0	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$
upper	0	1	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2$

Cleverly using Dummy Variables for Multiple Categories

- What if we want to test the difference between middle and upper class?
- Cleverly construct dummy variables such that an estimated coefficient identifies this difference.

	Dummy Variables		
Social Class	$\tilde{D}_1$	$\tilde{D}_2$	
lower	0	0	$\hat{Y} = \hat{\beta}_0$
middle	1	0	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$
upper	1	1	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$

- When estimating $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{D}_1 + \hat{\beta}_2 \tilde{D}_2$ then the estimated coefficient of the second dummy, $\hat{\beta}_2$, represents (by design!) the difference between middle and upper class.

Interactions

Modeling Interactions

- So far, we have only been adding variables in an additive manner, e.g.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon.$$

- Suppose, however, we want to test a hypothesis that the relation between an independent variable X_i and dependent variable Y depends on the value of another dummy variable D.
- Think of: $Income = \beta_0 + \beta_1 Education + \beta_2 Female + \beta_3 Education \cdot Female + \epsilon$
- The effect of X_i on Y is also called conditional because the hypothesized effect is conditional on D.
- In other words, if D is 1, the relation between X_i and Y is different than when D is zero.
- This is what we also call an interaction effect.
- Interaction model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 X_1 \cdot D + \ldots + \epsilon$

Modeling Interactions: Interpretation

- An interaction effect conditions the effect of an independent variable (e.g., *Education*) on the dependent variable.
- Interaction model if $D = 0$ (condition is absent):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_3 X_1 0 + \epsilon = \beta_0 + \beta_1 X_1 + \epsilon$$

- Interaction model if $D = 1$ (condition is present):

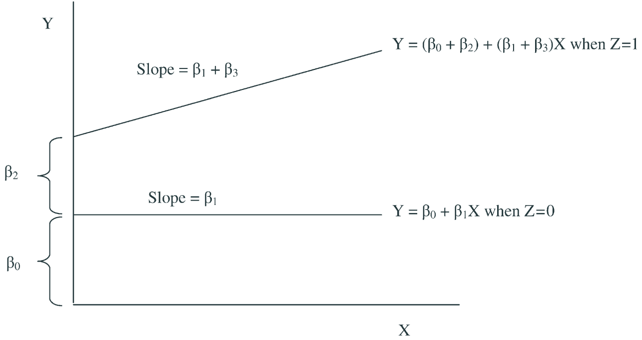
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_3 X_1 1 + \epsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon$$

- In other words, we get an intercept shift and a change in slopes.
- Do not interpret constitutive terms (i.e, $\hat{\beta}_1$ and $\hat{\beta}_2$) as if they are unconditional effects!

Modeling Interactions: Interpretation

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

Hypothesis H_1: An increase in X is associated with an increase in Y when condition Z is met, but not when condition Z is absent.



Modeling Interactions with Continuous Variables

- Interactions between dummy variables and continuous variables are the easiest to understand.
- But, we can interact continuous variables as well.
- Assume instead of a dummy D, X_2 to be continuous.
- Example: *Temporally-proximate presidential elections will reduce the effective number of electoral parties if and only if the number of presidential candidates is sufficiently low.*
- Thus,

$$\begin{aligned} ElectoralParties = \beta_0 &+ \beta_1 Proximity + \beta_2 PresidentialCandidates \\ &+ \beta_3 Proximity \cdot PresidentialCandidates + \epsilon \end{aligned}$$

Modeling Interactions with Continuous Variables

- In this case, the effect of the independent variable on the dependent variable gradually changes as another variable changes.

$$
\begin{aligned}
Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \ldots + \epsilon \\
Y =& \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) \cdot X_1 + \ldots + \epsilon
\end{aligned}
$$

- Marginal effect of X_1 on Y (i.e., $\frac{\delta Y}{\delta X_1} = \beta_1 + \beta_3 X_2$) represents the effect of change in X_1 on the expected change in Y, especially when the change in the independent variable (X_1) is infinitely small (marginal).
- The standard error of this marginal effect is (next week you will understand how to get variances and covariances):

$$
\hat{\sigma}_{\frac{\delta Y}{\delta X_1}} = \sqrt{var(\hat{\beta}_1) + X_2^2 var(\hat{\beta}_3) + 2X_2 cov(\hat{\beta}_1, \hat{\beta}_3)}
$$

- Of course, you may also interpret the marginal effect of X_2 on Y analogously.

Modeling Interactions with Continuous Variables

Table 1 The impact of presidential elections on the effective number of electoral parties. Dependent variable: Effective number of electoral parties

Regressor	*Model*
Proximity	−3.44** (0.49)
PresidentialCandidates	0.29* (0.07)
Proximity*PresidentialCandidates	0.82** (0.22)
Controls	——
Constant	3.01** (0.33)
R^2	0.34
N	522

*$p < 0.05$; **$p < 0.01$ (two-tailed). Control variables not shown here. Robust standard errors clustered by country in parentheses.

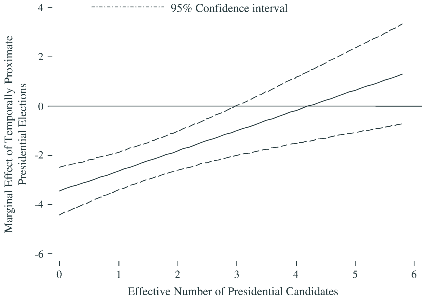


Fig. 3 The marginal effect of temporally proximate presidential elections on the effective number of electoral parties.