How do sociologists use linear regression?

Linear regression is one of the most powerful and commonly used statistical techniques in sociology. We learned how ordinary least squares regression is used to model the relationship between an interval or ratio dependent variable and one or more independent variables. We discussed how the regression coefficient is equal to the change in the dependent variable associated with a one-unit change in the independent variable. The t-statistic indicates whether this change in the dependent variable is significantly different from zero or if it may be due to chance. We also discussed how any variables that are correlated with both the independent and dependent variables should be “controlled” for - that is, included in the regression to net out its effect on the dependent variable.

So how do sociologists use linear regression?

Example 1: Is there a gender gap in lawyers’ incomes?

Professional women face the largest gender gap in income. Dinovitzer and colleagues (2009) sought to determine the mechanisms (i.e., the reasons) behind the gender gap for women lawyers. They tackled several explanations that might rationally account for why the gender gap in pay exists, including gender differences in human capital (i.e., credentials, like law school rank), gender differences in networking with senior lawyers, and gender differences in firm size. They said that if they considered all of these mechanisms and a gender gap in pay still existed, this would indicate discrimination, likely due to unconscious bias. In more statistical terms, if they controlled for every possible non-discriminatory reason there might be differences in pay between men and women lawyers and there was still an association between gender and income, this would indicate discrimination.

What is the main independent variable of interest? Is it nominal, ordinal, interval, or ratio?

What is the dependent variable? Is it nominal, ordinal, interval, or ratio?

What variables are Dinovitzer and colleagues controlling for?

Dinovitzer and colleagues wanted to see if gender (the independent variable, nominal, specifically binary) had an effect on income (the dependent variable, ratio) controlling for human capital, networking, and firm size (the control variables). Because the dependent variable is continuous, the authors used ordinary least squares regression. Their results are shown in table 1.

Table 1. Ordinary least squares regression on logged income.[1]

Variable	Coefficient
Male	0.052***
Law School Tier (Excluded = Tier 4)	
Ranked 1-10	0.129***
Ranked 11-20	0.114***
Ranked 21-40	0.063***
Ranked 41-100	0.023
Tier 3 (Ranked 101-137)	0.007
Firm Size (Excluded = 2-20)	
21-100	0.221***
101-250	0.363***
251+	0.467***
Networking	
Recruitment committee	0.065**
Meals with partners/associates	0.027
Recreation with partners	-0.014
$\boldsymbol{R^2}$	0.720

Source: After the JD.
Note: *p < 0.050, **p < 0.010, ***p < 0.001. N = 3,590. Controls for race, marital status, years since taking the bar, GPA, compensation practices of the firm (e.g., bonuses), area of law practiced, and legal market (i.e., place) are included.

[1] The authors use the natural log of income, instead of modelling income directly, to reduce heteroskedasticity - the fact that the variability of the residuals are unequal across the values of one of the independent variables.

How would you interpret the coefficient and significance for “Ranked 1-10”?

How would you interpret the coefficient and significance for “251+” firm size?

How would you interpret the coefficient and significance for “Recruitment committee”?

Many of the mechanisms Dinovitzer and colleagues proposed are significant - that is, they do explain to some extent why certain lawyers are paid more while others are paid less. The results for law school tier indicate that attending a higher tier law school is associated with being paid more. For instance, going to a top 10 law school is expected to increase logged income by 0.129 compared to going to a tier 4 law school. Because p < 0.001, the difference between the logged incomes of lawyers who attended a top 10 school and those that attended a tier 4 school are not due to chance. Though going to a tier 3 school is associated with a 0.007 increase in logged income compared to going to a tier 4 school, this increase is not significant and may be due to chance (p > 0.050).

How would you interpret the coefficient and significance for “Ranked 11-20”?

How would you interpret the coefficient and significance for “Ranked 21-40”?

How would you interpret the coefficient and significance for “Ranked 41-100”?

The results also indicate that lawyers at large firms are paid more. For instance, lawyers at firms with more than 250 lawyers are expected to have a logged income that is 0.467 higher than lawyers at firms with 2-20 lawyers. This difference is significant (p < 0.001).

How would you interpret the coefficient and significance for “21-100” firm size?

How would you interpret the coefficient and significance for “101-250” firm size?

Finally, only one of the networking variables was significant. Serving on a recruiting committee was associated with a 0.065 increase in logged income (p < 0.010).

How would you interpret the coefficient and significance for “Meals with partners/associates”?

How would you interpret the coefficient and significance for “Recreation with partners”?

What does the coefficient and significance for “Male” mean for Dinovitzer and colleagues’ hypothesis?

Dinovitzer and colleagues found that controlling for human capital, firm size, and networking did not eliminate the effect of gender on income. They found that male lawyers were expected to make logged income 0.052 greater than female lawyers. This effect is significant ($p < 0.001$), meaning that it is not due to chance. As they hypothesized, the typical rationales for the gender gap in income do not fully explain the observed differences in the incomes of male and female lawyers. They conclude that there are likely discriminatory processes, like unconscious bias, at play.

What limitations does Dinovitzer et al’s data or data analysis have?

One common complaint about studies that tackle the gender wage gap is that they do not control for productivity. This is particularly important, critics argue, because women may be expected to do more unpaid work at home than men, which may restrict their time to complete paid work. Employers may therefore be responding to the amount or quality of work that women do, rather than the sex of the employee doing it. Dinovitzer and colleagues’ study could be improved by including a measure of productivity, like number of hours billed, which is commonly used in studies of law.

Example 2: How do multiracial kids fit into the American system of race?

Several hypotheses exist to explain how the contours of the American system of racial oppression and privilege are changing to account for multiracial individuals. One is the “Latin Americanization” hypothesis, which states that the United States is moving toward a racial system based on skin color used in many Latin American countries. Another is that the United States is moving toward a new binary system where multiracial individuals will be categorized as black or white and treated accordingly. Campbell (2009) sought to adjudicate between these two hypotheses by analyzing the GPA of multiracial students. She predicted that if the Latin Americanization hypothesis was correct, skin color (on a scale from 1 to 5) would be related to GPA. On the other hand, if the binary hypothesis was correct, she predicted that being perceived as black by others would be related to GPA. Because social class has been shown to predict GPA, Campbell decided to net out the effect of having a parent with a Bachelor’s degree.

What are the main independent variables of interest? Are they nominal, ordinal, interval, or ratio?

What is the dependent variable? Is it nominal, ordinal, interval, or ratio?

What variable is Campbell controlling for? Is it nominal, ordinal, interval, or ratio?

Campbell wanted to determine if skin color (the independent variable, interval) or being perceived as black by others (the independent variable, nominal, specifically binary) was associated with a student's GPA (the dependent variable, ratio). She controlled for having a parent with a Bachelor's degree (the control variable, nominal, specifically binary). Because the dependent variable is continuous, Campbell used ordinary least squares regression. The results are shown in table 2.

Table 2. Ordinary least squares regression on GPA.

	Model 1	**Model 2**
Perceived as "Black" by interviewer	-0.030	
Skin color		0.030
Parent with a Bachelor's degree	0.285**	0.286**

Source: ADHealth, wave 3.
Note: *p < 0.050, **p < 0.010, ***p < 0.001. N = 510. Skin color is measured on a scale of 1 (= black) to 5 (= white). Controls for gender, age, having two parents at home, log of family income, and foreign born are included.

How do you interpret the coefficient and significance for "Parent with a Bachelor's degree" in both models?

Campbell found that socioeconomic status was associated with GPA. Having a parent with a bachelor's degree was expected to increase a student's GPA by 0.285 or 0.286 points (depending on the model). Because p < 0.010, this difference is not due to chance.

How do you interpret the coefficient and significance for “Perceived as ‘Black’ by interviewer”?

How do you interpret the coefficient and significance for “Skin color”?

What does this mean for the Latin Americanization and binary hypotheses?

Though being perceived as Black by the interviewer was associated with a 0.030 decrease in GPA, this difference may be due to chance. Likewise, though a one unit increase in the lightness of a student’s skin color is associated with a 0.030 increase in GPA, this difference is not significant. Because neither of these coefficients are significant, Campbell’s results cannot adjudicate between the Latin Americanization and binary hypotheses. On the other hand, the results do show the enduring impact of socioeconomic status on student outcomes.

What limitations does Campbell’s data or data analysis have?

Research shows that perception of an individual's race is dependent on their social status and class (Penner & Saperstein, 2008). In her study, Campbell uses whether the *interviewer* perceived of the child as black. *Teachers and principals* - ostensibly the people who might discriminate against children they perceive as black - may have more or less or different information than the interviewer about the child’s class background upon which they base their evaluation of the child’s race. Similarly, perception of a child’s skin color may be affected by the same processes or may be dependent on the racial makeup of the school environment. Including some school demographics or direct measures of how school employees identify the child’s race would improve the study. (Of course, the latter may be very difficult to obtain.)

References

Campbell, M. E. (2009). Multiracial Groups and Educational Inequality: A Rainbow or a Divide? *Social Problems*, *56*(3), 425–446.

Dinovitzer, R., Reichman, N., & Sterling, J. (2009). The differential valuation of women’s work: A new look at the gender gap in lawyers' incomes. *Social Forces*, *88*(2), 819–864.

Penner, A. M., & Saperstein, A. (2008). How social status shapes race. *Proceedings of the National Academy of Sciences*, *105*(50), 19628–19630.