EDUC 643 Assignment 02 Key

1. Diagnostics and assumption checks

1.1. Re-state your interpretation of the results of your hypothesis test from Assignment 1, Question 2.4. Then, write a paragraph in which you state the assumptions you have made in using an Ordinary Least Squares estimator to describe this relationship. (2 points)

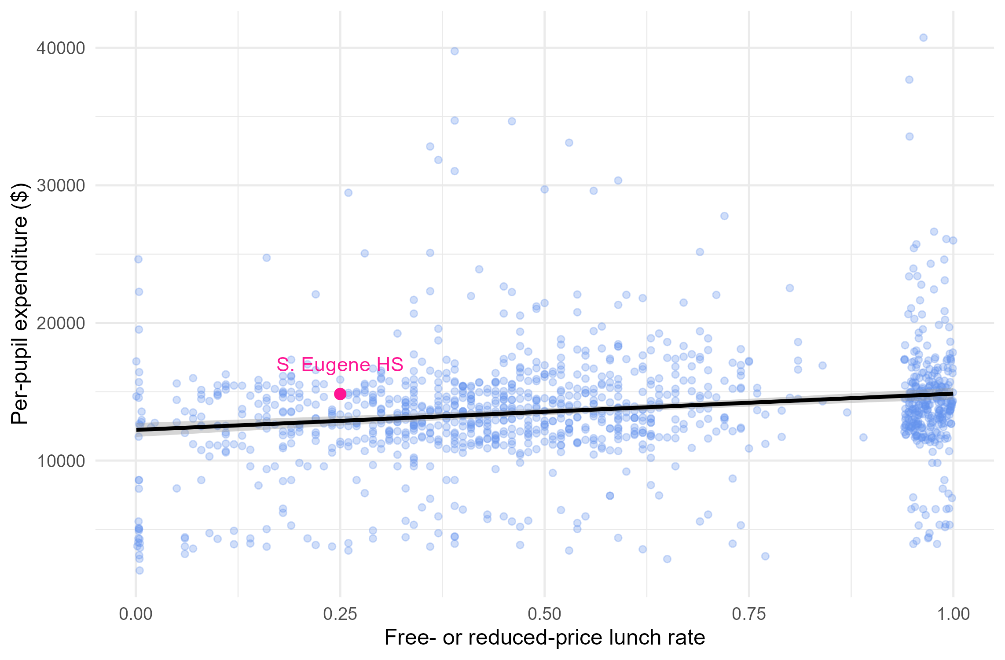
We estimate that an Oregon school in which all students (100 percent) receive free- or reduced-price lunch (*FRPL*) would spend, on average, $2,634.33 more per-student (*PPE*) than a school in which no students (0 percent) receive free- or reduced-price lunch. At an alpha-threshold of 0.05, we reject the null and conclude that, on average in the population of Oregon public schools, there is a relationship between FRPL rates and per-pupil expenditures.

In describing this relationship by means of an OLS estimator, we are making several assumptions about the nature of the underlying data and the relationship between *FRPL* and *PPE*. In particular, we are assuming that there is no measurement error in both variables, though this is particularly important for the values of *FRPL*. We are assuming that the relationship is linear, or that the mean of the distribution of *PPE* at each value of *FRPL* can be joined together by a straight line. We are assuming that the variance of the distribution of the residuals is identical for each value of *FRPL*. We additionally assume that our residuals are normally distributed. Finally, we assume that conditional on each value of *FRPL*, the values of *PPE* are independent from each other. Embedded into several of these assumptions is the assumption that there exist no unduly influential outliers in our data.

1.2. Examine the school called “South Eugene High School” (ncesid == 410474000573) in the NERD$ data. Characterize its observed value of per-pupil expenditure and the proportion of students at the school classified as low-family-income. How do the observed values differ from the predicted values for this school? (2 points)

**Figure 1**

*Relationship between school-level per-pupil expenditure in U.S. dollars and the proportion of students receiving free- or reduced-price lunch for Oregon public schools 2018-19.*

**

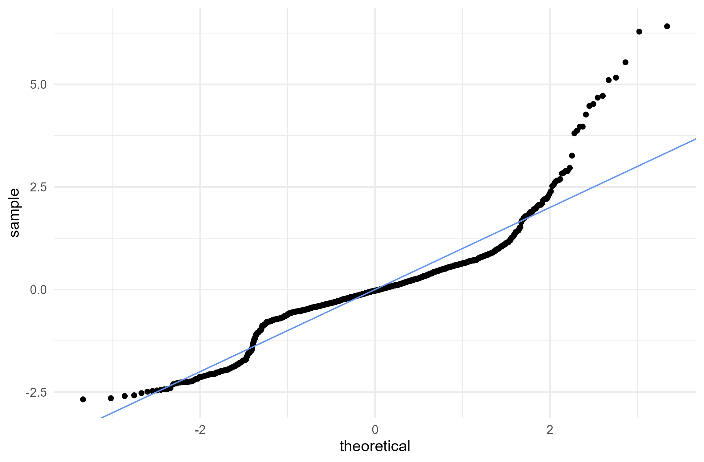
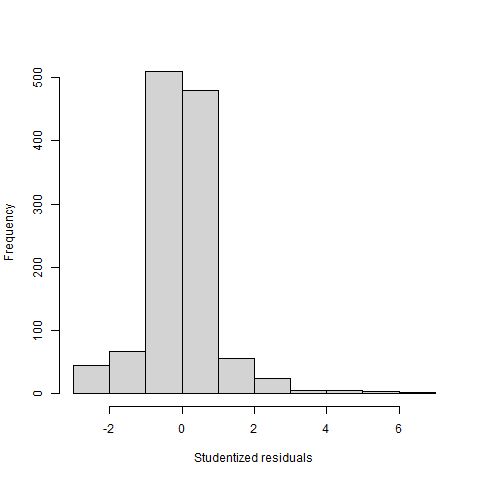
South Eugene High School spends $14,836 per-student and 25 percent of its student body receives free- or reduced price lunch. As we highlight in Figure 1, its observed *PPE* is slightly above what we would predict. In fact, based on the model we fit in Assignment 1, we would predict that S. Eugene HS would spend $12,885 per-student (). Corresponding to the vertical distance that S. Eugene HS is from the line of best fit in Figure 1, the school spends approximately $1,991 per-pupil more than expected.

1.3. Assess the residuals from the linear regression you conducted in Assignment 1 for evidence on the fitted model’s linearity, the normality and homoscedasticity of the residuals, and for the presence of outliers. In these assessments you should minimally present three different figures. (3 points)

In fitting our estimates of the bivariate relationship between FRPL and PPE, we made several assumptions about the linearity of the relationship, and the normality and homoscedasticity of the residuals. As we document in Figure 2, the studentized residuals are mound-shaped with a single peak and roughly normal, though there is a long positive tail (Panel A), which has a corresponding deviation from the quantile-quantile rank line at higher values of the studentized residuals in Panel B. Thus, our normality assumption is generally met, though with some concern about the upper tail. Both Figure 1 and Figure 3 provide strong evidence that the relationship is best characterized as a linear one given that there is no clear pattern of under- or over-predicting that would be suggestive of non-linearity. Finally, there is no strong evidence of heteroscedasticity of the residuals. In Figure 3, the scatter of the residuals is roughly uniform throughout the fitted values, though there does appear to be a cluster of fitted values at both positive and negative extremes.

**Figure 2**

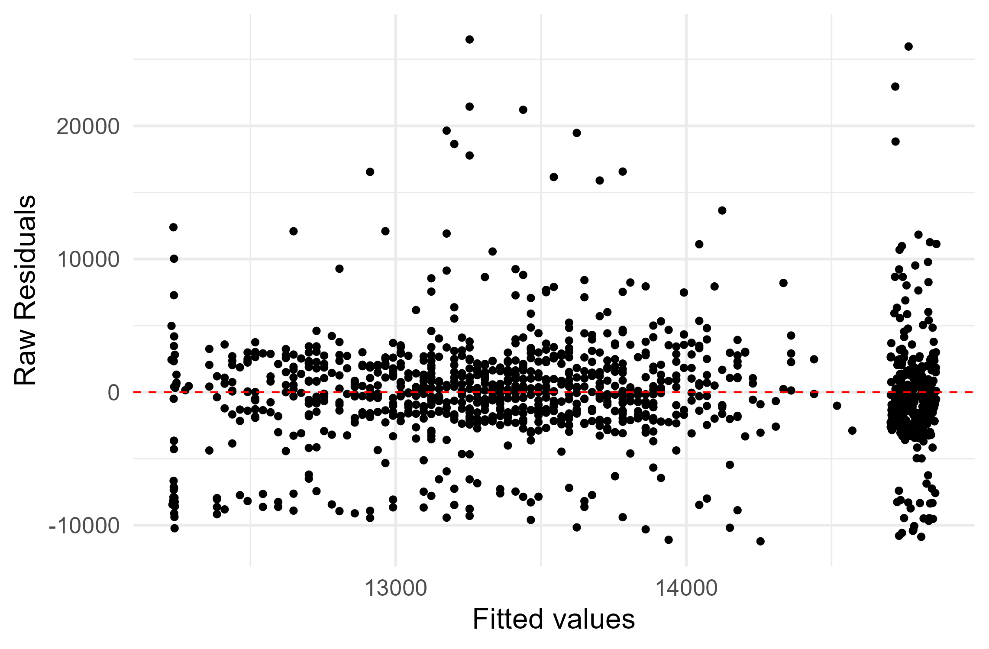
*Residual distribution from a linear fit of the relationship between school-level per-pupil expenditure in U.S. dollars and the proportion of students receiving free- or reduced-price lunch for Oregon public schools 2018-19.*



(a) Histogram of studentized residuals (b) Q-Q plot of studentized residuals

**Figure 3**

*Residuals vs. fitted values plot from a linear fit of the relationship between school-level per-pupil expenditure in U.S. dollars and the proportion of students receiving free- or reduced-price lunch for Oregon public schools 2018-19.*



1.4. What are some ways in which you could imagine the residuals in this dataset not being independently distributed? In other words, what sort of clustering might be present in these data and how would this affect your inference? (2 points)

This is a school-level dataset of public schools in the state of Oregon. It is likely that the values of per-pupil expenditure are strongly correlated within school districts, which would be an example of non-independence. Other such ways in which the residuals might be correlated across values of FRPL could be by larger geography such as county or rural/remote locations, by schools that educate large proportions of English learners and/or students with disabilities, or by schools that are part of particular targeted educational investments from the state or federal governments.

1.5. Given what you have found in 1.3 and 1.4, to what extent do you feel like your OLS regression assumptions have been met in this analysis? Are there solutions you would consider implementing if any of the regression assumptions are not met? If so, what are they (note: you do not need to actually implement any of the solutions, just describe what they might be)? What are some of the reasons not to implement any “fixes” to violations to your assumptions? (3 points)

Generally, we find our OLS assumptions to be roughly—though not perfectly—met in this context. We cannot test whether measurement error is present given these data alone. However, if we are worried about this (particularly with respect to rates of family poverty), we could possibly use prior-year values of free- and reduced-price lunch and create a composite school-level measure. We could also use Census- or alternative-source measures of child poverty. The relationship is largely linear based on our inspection of the plots. If we were concerned about this, we could implement a transformation (which we will learn to do in Unit 5). There is minimal evidence of heteroscedasticity, but if we were concerned about this, we could adjust our inference by using heteroscedastic-robust standard errors. There is some evidence that the residuals are not normally distributed. Further, there are several substantive reasons why we might worry that our residual independence assumption is not met. To address this concern, we could again adjust our inference by either fitting a multi-level model or adjusting our standard errors for group-level clustering.