EDUC 643 Assignment 01 Key

1. Descriptive statistics

1.1. Summarize the dataset. Specifically, create a table to show the means and standard deviations of the continuous variables (i.e., exclude the geographical identifiers and the *level* and *locale* categorical variables). Write 2-3 sentences to report and interpret these statistics. (1 point)

**Table 1**

***Summary statistics for per-pupil expenditures and school and community characteristics for Oregon public schools, 2018-19.***

|  | Mean | SD | Min | Median | Max |
| --- | --- | --- | --- | --- | --- |
| Per-pupil expenditure ($) | 13701.44 | 4273.52 | 2011.41 | 13586.98 | 40734.41 |
| Enrollment | 467.28 | 355.60 | 20.70 | 405.30 | 2866.09 |
| SES index for all families | 0.16 | 0.56 | -1.29 | 0.07 | 1.83 |
| Median family income ($) | 53856.90 | 11364.84 | 28570.00 | 50569.00 | 114467.00 |
| BA+ rate | 0.27 | 0.13 | 0.08 | 0.24 | 0.67 |
| Unemployment rate | 0.07 | 0.01 | 0.02 | 0.07 | 0.11 |
| SNAP receipt rate | 0.14 | 0.05 | 0.02 | 0.14 | 0.28 |
| Free- or reduced-price lunch rate | 0.56 | 0.30 | 0.00 | 0.51 | 1.00 |
| Source: National Education Resource Database on Schools (NERD$) |

There are 1,193 school-level observations in the data; none of the variables have any missing values. As we document in Table 1, with respect to our variables of interest, the mean per-pupil expenditure was $13,701, with a standard deviation of $4,274, suggesting substantial variability in expenditures. The school-level mean of the proportion of students receiving free- or reduced-price lunch (FRPL) was 0.56, with a standard deviation of 0.30, again suggesting a high degree of variability around the mean FRPL rate.

1.2 Create a plot (make sure to label the x and y axis) to visualize the relationship between the variables *ppe* and *frpl*. Include a line of best fit on this display. (1 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.*



1.3 Write 1-2 sentences to interpret this visualized relationship, relying on the five features of bivariate relationships we introduced in class. (1 point)

In Figure 1, on average, it appears that there is a positive relationship between the proportion of a school’s student body who receive free- or reduced-price lunch, and the school’s per-pupil expenditures. This relationship is roughly linear and fairly weak in both magnitude and strength with seemingly substantial variation away from the line of best fit throughout the distribution. There are a few potentially outlying values of schools that spend considerably more per-student than others.

#### 2. Bivariate analysis

2.1. Write a formal linear model that describes the simple, bivariate relationship between school-level per-pupil expenditure and the school-level average receipt of free- or reduced-price lunch and interpret each of the terms in this model. (1 point)

We postulate the following linear model to characterize the relationship between school-level per pupil expenditure (*PPE*) as a function of the average receipt of free- or reduced-price lunch (*FRPL*) at the school:

$$PPE\_{i}= β\_{0}+β\_{1}\left(FRPL\_{i}\right)+ ε\_{i}$$

where *PPE* for the *i*th school is a function of an intercept (*β0*), which will represent the value of *PPE* when *FRPL* is equal to 0. Our parameter of interest is *β1*, which represents the slope of the relationship between *FRPL* and *PPE*. Finally, *εi* is an idiosyncratic school-level error term, which represents the variance in PPE unexplained by our model.

2.2. State your null hypothesis about the relationship between per-pupil expenditure and the proportion of students receiving free- or reduced-price lunch. (1 point)

There is no relationship between per-pupil expenditure and the proportion of students receiving free- or reduced-price lunch in school, on average in the population of Oregon public schools.

2.3 Formally test your hypothesis using an Ordinary Least Squares estimation strategy. Report your results in a formatted table. (2 points)

**Table 2**

***Ordinary Least Squares estimates of the relationship between school-level per-pupil expenditures and free- and reduced price lunch rates for Oregon public schools, 2018-19***

|  | Model 1 |
| --- | --- |
| (Intercept) | 12226.90\*\*\* |
|  | (258.38) |
| Free- or reduced-price lunch (0-1) | 2634.33\*\*\* |
|  | (407.23) |
| Num.Obs. | 1193 |
| R2 | 0.03 |
| RMSE | 4198.60 |
| *Notes*: + *p* < 0.1, \* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001. Cells report coefficients and standard errors in parentheses. |

2.4. Interpret the results of your test in 1-2 sentences. (2 points)

We estimate that an Oregon school in which all students (100 percent) receive free- or reduced-price would spend, on average, $2,634.33 more per-student than a school in which no students (0 percent) receive free- or reduced-price lunch (Table 2). 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.

2.5. Assess the quality of your model fit using the *R2* statistic. (1 point)

The *R2* for our model is 0.03, which means that our model explains 3 percent of the overall variation in per-pupil expenditure. This seems like a relatively small proportion of the overall variance and suggests that many other factors contribute to variation in *PPE*.

2.6. Select an alpha-threshold and describe the corresponding confidence interval for your estimates of the relationship between *frpl* and *ppe*. (1 point)

With an alpha-threshold of 0.05, our corresponding 95 percent confidence interval for the relationship between *frpl* and *ppe* is $1,835.36 - $3,433.30. This implies that were we to repeat the exercise of measuring the proportion of students receiving FRPL and school PPE—as well as the relationship between the two—for the population of Oregon schools 100 times, that in 95 cases we would expect our estimate of this relationship to fall between $1,835 and $3,433.

3. Regression assumptions

3.1. What assumptions have you made in using an Ordinary Least Squares estimator in your analysis in Section 2? (1 point)

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.

3.2. 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? (1 point)

**Figure 2**

*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 2, its observed *PPE* is slightly above what we would predict. In fact, based on the model we fit, we would predict that S. Eugene HS would spend $12,885 per-student ($\hat{PPE}=12226.90+\left(0.25\right)\left(2634.33\right)=\$12,885.48$). Corresponding to the vertical distance that S. Eugene HS is from the line of best fit in Figure 2, the school spends approximately $1,991 per-pupil more than expected.

3.3. Assess the residuals from the linear regression you conducted in Section 2 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. (2 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 3, 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 2 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 4, 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 3**

*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 4**

*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.*



3.4. 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? (1 point)

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.

3.5. Given what you have found in 3.3 and 3.4, 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? (2 points)

Generally, we find our OLS assumptions to be roughly—though far from 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.

#### 4. Putting it together

4.1. Imagine you are an analyst working for Charlene Williams, Director of the Oregon Department of Education. You have been tasked with describing to her whether schools with greater levels of student financial need receive more money. Write a short paragraph reporting the results of your analysis, while introducing appropriate caveats and nuances as needed. In particular, you should think about how you want to introduce the ideas of relationship magnitude, model fit, omitted variables, and correlation vs. causation for a lay audience. (2 points)

Using publicly available data from the National Education Resource Database on Schools, we found that public schools in Oregon who educate more children from low-income families spend *more* per-child than schools educating fewer children from low-income families. In particular, we estimate that schools have 10 percentage points more free- and reduced-price lunch recipients spend, on average, $263 more per child per year. Given that the median school in Oregon enrolls around 405 students, this represents an annual budgetary difference of more than $106,700 for school educating 10 percentage points more FRPL-receiving students. While this seems like a meaningful difference, there is little concrete evidence suggesting just how much more compensatory funds would be required to promote a more equitable schooling system for students living in poverty. Further, there is substantial variation in per-pupil expenditure that is not explained by our model—some schools with very few low family-income students spend close to $25,000 per-student, per-year while some schools that educate nearly exclusively low family-income students spend less than $5,000 per-student, per-year. It may be valuable to investigate further the role that school size, school location, school grade-level, community resources, and local governance play in determining school funding. While there is some evidence of socio-economic-based equity funding in our analysis, if the goal of the Oregon Department of Education is to systematically provide more resources to schools that educate more students living in poverty, our analysis also shows that there is substantial scope to strengthen and increase the magnitude of this relationship.