Homework 3

PPHA 31002 | Statistics for Data Analysis I

Jay Ballesteros

November 17, 2025

STATISTICAL EXERCISES

1. Suppose a friend collects data from a random sample of college-educated, full-time workers: weekly.expenditures indicates a person's weekly expenditures in dollars, and weekly_income indicates that person's weekly income.

1.1. Explain what the value of 710.0 is, and how to properly interpret it. Is it substantively meaningful in this context?

710 is the coefficient of the dependent variable `weekly.expenditures` (intercept) and represents the predicted weekly expenditure when the weekly income (independendent variable) is 0. In this context, it means that even if a worker has no income, they are expected to spend $710 per week, it is meaningful as it provides information that even without income, spending persists (at $710)

1.2. Explain what the value of 0.6 is, and how to properly interpret it.

0.6 is the coefficient for the independent variable `weekly.income`. This 0.6 result indicates the expected change in weekly expenditures for each additional dollar of weekly income.

1.3. Interpret the R^2

The R^2 value of 0.13 indicates that the prediction errors are reduced by only 13%. In other words, only 13% of the variability in weekly expenditures (independent) can be explained by the weekly income (dependent). This suggests that while weekly income has some influence on expenditures, 87% of the variability is due to other factors not included in the model.

1.4. What is the predicted expenditure for a person with a weekly income of $200 based on the regression?

It is 830.0.$To get that result, I used the following equation: Y = \alpha + \beta * X$\$ Where Y is the predicted expenditure, α is the intercept (710.0), β is the slope (0.6), and X is the weekly income (200).

$$Y = 710.0 + 0.6 * 200 = 830.00$$

1.5. Suppose a worker in the sample reports weekly expenditures of $1,000 and has a weekly income of $200. Find the size of the residual for this worker.

The residual for this worker is 170.00.$To get that result, I used the following equation: \hat{\epsilon_i} = Y_i - \hat{Y_i}$\$ Where $\hat{\epsilon_i}$ is the residual, Y_i is the actual expenditure, and $\hat{Y}i$ is the predicted expenditure. As a result, the calculation is:

$$\hat{\epsilon_i} = 1000 - 830.00 = 170.00$$

1.6. Briefly explain why we have residuals.

Residuals are the differences between the actual values and the predicted values from a regression model Considering the last question, that worker spends 170.00 more that then model predicts. We have them to assess the accuracy of our model's predictions (a sort of prediction error), because probably not all factors influencing expenditures are captured in the model.

1.7. Calculate the predicted expenditure for a worker with a weekly income of $10,000. What disclaimer should we make of this predicted value?

$$Y = 710.0 + 0.6 * 10000 = 6710.00$$

The disclaimer for this predicted value is that it may not be reliable since, for instance, the maximum value of the `weekly_income` variable is 5000. So the prediction for a weekly income of $10,000 is beyond the range of the data used to fit the model, which can lead to inaccurate or misleading results.

2. A government official claims that 60% of residents favor expanding public transit. Based on survey data collected from a simple random sample of n = 100 residents, 51 indicate they are in favor of expanding public transit, and 49 respondents indicate that they do not favor expanding public transit. You plan to assess whether this survey provides sufficiently strong evidence to support the claim that the true proportion of residents who favor expansion is different from 60% (either higher or lower). Based on your conversation with the offical, you set a = 0.05 for your hypothesis test. Use the information above to develop a hypothesis test.

2.1. Write down the null and alternative hypotheses (using statistical notation) for a test that fits the situation detailed above.

For the null hypothesis,

$$H_0 : p = 0.60$$

Whereas for the alternative hypothesis,

$$H_A : p \neq 0.60$$

2.2. What is the appropriate test statistic to use for this analysis? Provide the correct formula and the value of the test statistic. Round to two decimal places.

The appropriate test statistic for this analysis is the two proportion z-test The formula for this z-test is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where $\hat{p}$ is the sample proportion, p_0 is the hypothesized population proportion, and n is the sample size. Calculating the sample proportion:

$$\hat{p} = \frac{51}{100} = 0.51$$

Now, substituting the values into the formula:

$$z = \frac{0.51 - 0.60}{\sqrt{\frac{0.60(1-0.60)}{100}}} = \frac{-0.09}{\sqrt{\frac{0.60(0.40)}{100}}} = \frac{-0.09}{\sqrt{0.0024}} = \frac{-0.09}{0.049} \approx -1.84$$

2.3. Convert the test statistic you just found into a p-value. Round to three decimal places.

To convert the test statistic z = -1.84 into a p-value for a two-tailed test, we can use the standard normal distribution table or a statistical software. The p-value is calculated as:

$$p - value = 2 * P(Z < -1.84)$$

Using a standard normal distribution table or software, we find that:

$$P(Z < -1.84) \approx 0.0332$$

Thus, the p-value is:

$$p - value = 2 * 0.0332 = 0.0664$$

2.4. What is your conclusion (i.e., your decision) based on this test?

Based on the p-value of 0.0664 and the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. This means that there is not enough evidence to support the claim that the true proportion of residents who favor expanding public transit is different from 60%.

3. The organizers of a public health campaign claim that, after the implementation of the campaign, daily sugar intake among residents is now, on average, only 30 grams. Your friend is interested in investigating this claim, as she has a suspicion that the true average daily sugar consumption is greater than 30 grams for residents. You help her collect a simple random sample of n = 49 residents. Based on the sample, the mean daily sugar intake is 34 grams, and the standard deviation is 14 grams. She asks you to formally conduct a hypothesis test. Based on her tolerance for type I error, you set the significance level to be a = 0.05.

3.1. Write down the null and alternative hypotheses, using statistical notation, based on the situation detailed above.

3.2. What is the appropriate test statistic in this exercise? Provide the correct formula and the value of the test statistic. Round to two decimal places.

3.3. Convert the test statistic you just found into a p-value. Round to two decimal places.

3.4. What is your conclusion (i.e., your decision) based on this test? Would your conclusion be different if you had specified a two-sided alternative hypothesis instead? Briefly explain.

4. A survey conducted in June 2022 asked a random sample of U.S. adults about their preferences for electric vehicles versus gasoline-powered vehicles. The question stated: "If both options were available at a similar price, which type of car would you prefer to buy?" Among respondents under age 40, 100 out of 180 said they would prefer an electric vehicle, while among respondents age 40 and older, 155 out of 320 said they would prefer an electric vehicle. You are interested in testing whether the proportion of younger adults who state they prefer electric vehicles is greater than the proportion of older adults who state they prefer electric vehicles. The total sample size is n = 500.

4.1. Write down the null and alternative hypotheses, using statistical notation, for the situation detailed above.

4.2. What is the appropriate test statistic in this exercise? Provide the correct formula and the value of the test statistic. Round to two decimal places.

4.3. Convert the test statistic you just found into a p-value.

4.4. What is your conclusion (i.e., your decision) based on this test? Would your conclusion change if you had chosen a significance level of a = 0.01 instead? Briefly explain.

5. A local government implemented a job training program to help unemployed residents increase their earnings potential. Ten participants were randomly selected, and their monthly incomes (in dollars) were recorded before and after completing the program:

5.1. Formulate the appropriate null and alternative hypotheses, using statistical notation, based on the information provided above.

5.2.Calculate and report the sample mean and the sample variance of the differences in the participants monthly incomes. If needed, round to two decimal places.

5.3.What is the appropriate test statistic in this exercise? Provide the correct formula and the value of the test statistic. Round to two decimal places.

5.4.Convert the test statistic you just found into a p-value. Round to two decimal places.

5.5.Based on the data, can we conclude that participants' monthly incomes increased on average

6. Now, treat the data as two independent samples (Sample A and Sample B), instead of repeated observations on the same participants. Specifically, 10 randomly selected participants in Sample A did not complete the job training program (control group), whereas another 10 randomly selected participants in Sample B completed the job training program (treatment group). Results for these two independent samples are given in the table below (you can think of these as the monthly incomes for the participants in groups A and B). For the questions below, you should again use a significance level of a = 0.05.

6.1. Which hypothesis test should you use to investigate whether the group of participants who completed the job training program, on average, had higher incomes?

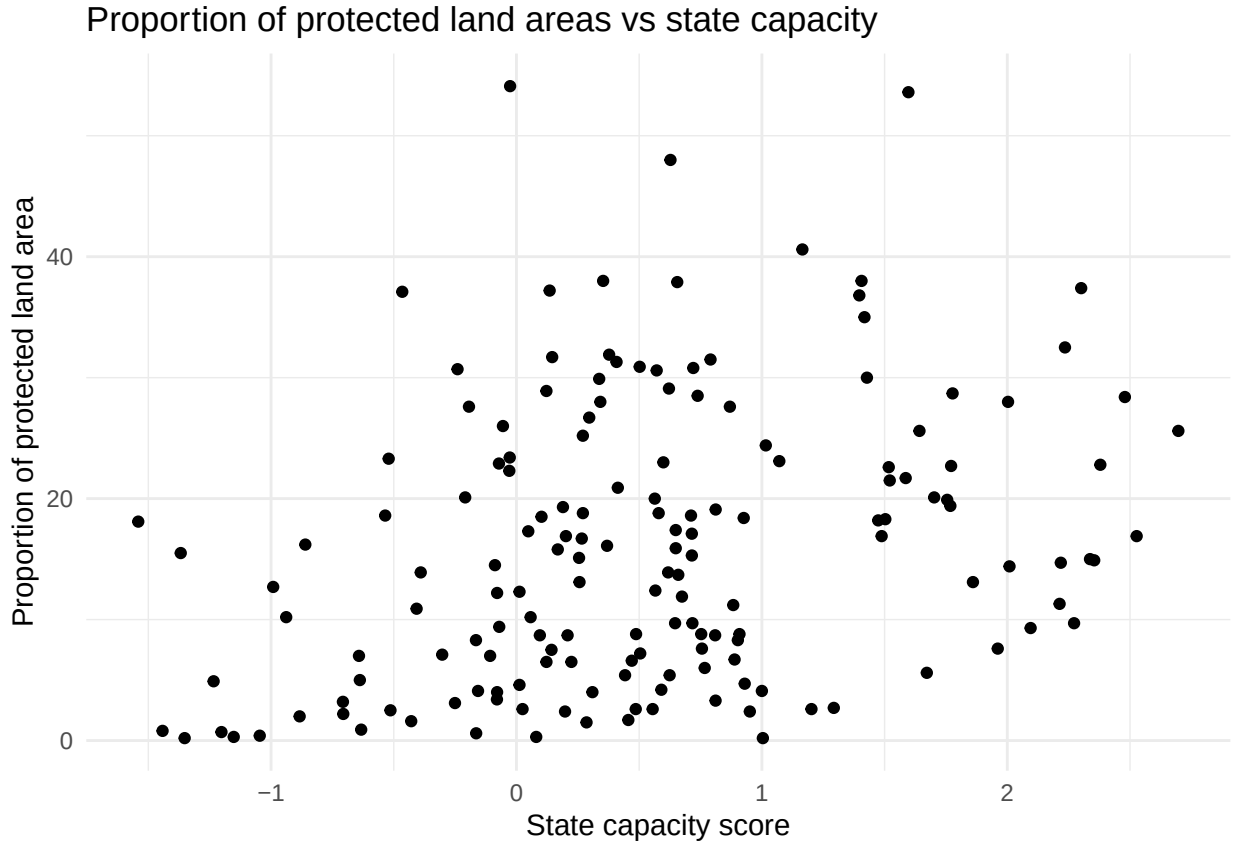
6.2. State your null and alternative hypotheses using statistical notation.

6.3. What is the appropriate test statistic in this exercise? Provide the correct formula and the value of the test statistic. Round to two decimal places.

6.4. What is the outcome of the hypothesis? Briefly compare the outcome of this hypothesis test with the conclusion of your analysis from question (5v). In other words, why do you arrive at the same conclusion or a different conclusion?

DATA EXERCISE

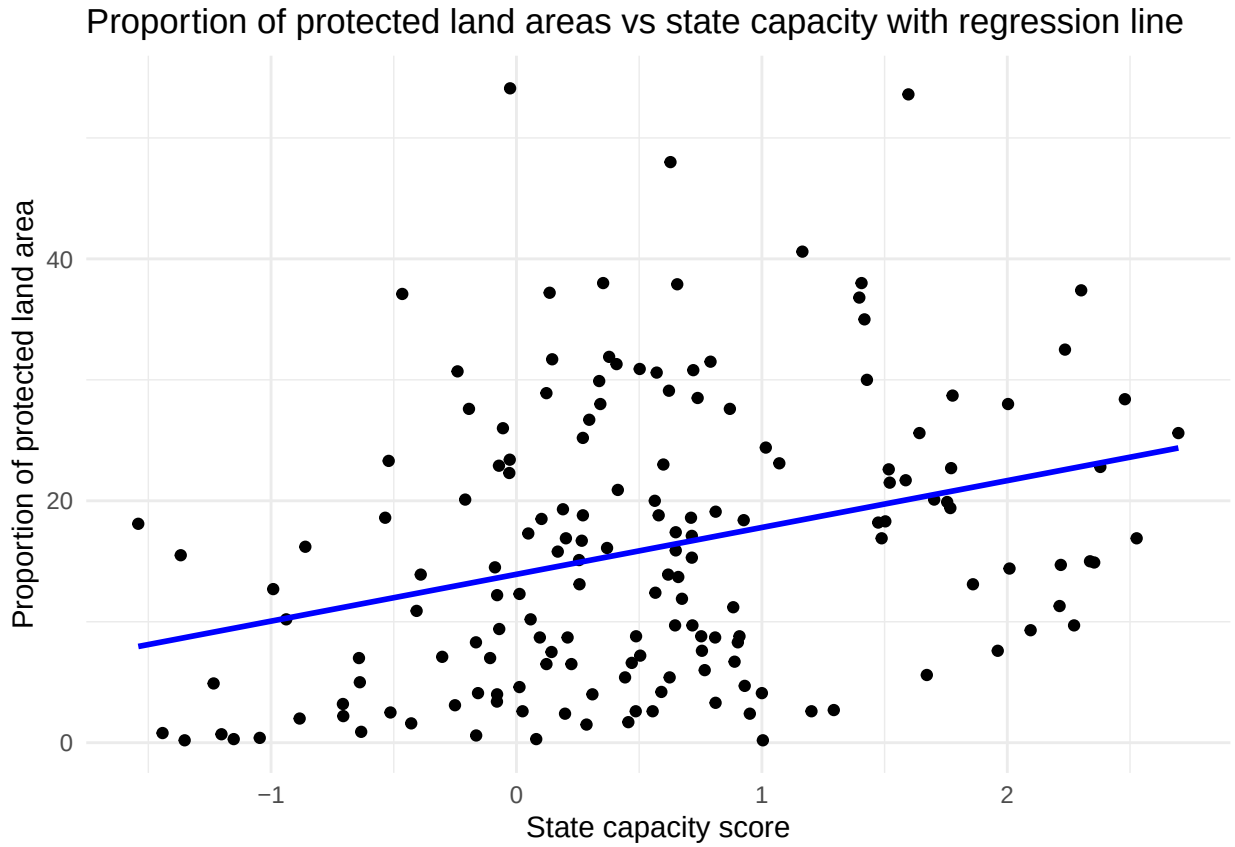
7. Construct a scatterplot showing the relationship between the proportion of protected land areas and state capacity score. The proportion of protected land areas should be measured on the vertical axis (y-axis), and state capacity on the horizontal axis (x-axis). Make sure to label both axes and include a title as well. Provide a brief description of the relationship between these two variables based on a visual inspection. [1.5pts]



Based on the scatterplot, there appears to be a positive relationship between the state capacity score and the percentage of protected land area. As the state capacity score increases, the percentage of protected land area also tends to increase. This suggests that countries with higher state capacity scores may be more effective in protecting their land areas.

8. Run a regression of percent of protected land area (dependent variable) on the state capacity score (inde- pendent variable), and plot the regression line atop scatterplot you just created. [1pt]

```
## `geom_smooth()` using formula = 'y ~ x'
```



9. Report the estimated intercept and coefficient, interpret them, and explain if the estimated intercept is substantively meaningful in this context.

The estimated intercept for the model is 13.923514 and the coefficient for the state capacity score is 3.872412. In this context, the results can be interpreted as follows: - The intercept (13.923514) represents the predicted percentage of protected land area when the state capacity score is 0. However, this value may not be substantively meaningful because a state capacity score of 0 may not be realistic or applicable in real-world scenarios. - The coefficient (3.872412) indicates that for each one-unit increase in the state capacity score, the percentage of protected land area is expected to increase by approximately 3.87 percentage points, holding all else constant. This suggests a positive relationship between state capacity and the extent of land protection.

10. Calculate the predicted percent of land under protection in a country with a state capacity of 3. To what extent should we assign

substantive meaning to this predicted value?

Using the regression equation:

$$Y = \alpha + \beta * X$$

Where Y is the predicted percentage of protected land area, α is the intercept (13.923514), β is the coefficient (3.872412), and X is the state capacity score (3). Calculating the predicted value:

$$Y = 13.923514 + 3.872412 * 3 = 25.54075$$

The predicted percentage of land under protection in a country with a state capacity of 3 is approximately 25.54%. This predicted value can be assigned substantive meaning as it provides an estimate of the extent of land protection based on the state capacity score. However, it is important to consider the context and

limitations of the model, such as the range of state capacity scores in the data and potential confounding factors that may influence land protection.

11. One of your friends suggests that assuming a linear relationship between the percentage of protected land area and the state capacity score might be imposing a very strong and unrealistic assumption about the structure of the data. They suggest you split the state capacity score into deciles (i.e. 10 equal-sized groups), and regress the outcome on indicators for a country's status in the 2nd, 3rd,. . . , 10th decile. Use the ntile() command from the dplyr library to create the deciles. Then, use the factor() function to store the decile variable you just created to be recognized as a factor variable in R. This way, you do not need to create separate indicator variables for each decile category. Simply include the factor variable as the explanatory variable in the lm() function and R implictly assumes that you would like to include various binary indicators variables for each category in your regression. Note that R will automatically omit the first category to use as the reference group. Report the estimates from your regression.

```
##
## Call:
## lm(formula = percent_of_protected_land_area ~ capacity_deciles,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.138  -7.706  -2.535   7.765  39.265
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.900      2.685   2.198 0.029469 *
## capacity_deciles2     7.688      3.797   2.025 0.044598 *
## capacity_deciles3     8.935      3.797   2.354 0.019866 *
## capacity_deciles4    10.994      3.855   2.852 0.004952 **
## capacity_deciles5    11.869      3.855   3.078 0.002466 **
## capacity_deciles6    12.606      3.855   3.270 0.001330 **
## capacity_deciles7    10.238      3.855   2.655 0.008761 **
## capacity_deciles8     8.662      3.855   2.247 0.026080 *
## capacity_deciles9    17.837      3.855   4.627 7.87e-06 ***
## capacity_deciles10   12.950      3.855   3.359 0.000988 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.07 on 153 degrees of freedom
## Multiple R-squared:  0.1435, Adjusted R-squared:  0.09312
## F-statistic: 2.848 on 9 and 153 DF,  p-value: 0.003948
```

```
library(jtools)
summ(decile_model)
```

Observations	163
Dependent variable	percent_of_protected_land_area
Type	OLS linear regression

F(9,153)	2.85
R^2	0.14
Adj. R^2	0.09

	Est.	S.E.	t val.	p
(Intercept)	5.90	2.68	2.20	0.03
capacity_deciles2	7.69	3.80	2.03	0.04
capacity_deciles3	8.94	3.80	2.35	0.02
capacity_deciles4	10.99	3.86	2.85	0.00
capacity_deciles5	11.87	3.86	3.08	0.00
capacity_deciles6	12.61	3.86	3.27	0.00
capacity_deciles7	10.24	3.86	2.66	0.01
capacity_deciles8	8.66	3.86	2.25	0.03
capacity_deciles9	17.84	3.86	4.63	0.00
capacity_deciles10	12.95	3.86	3.36	0.00

Standard errors: OLS

The regression results from the decile model indicate the following estimated coefficients for each decile category (with the first decile as the reference group):

12. Which regression—the one based on the single variable for state capacity or its transformation to deciles— seems to provide better predictions of annual earnings within the sample? Explain your answer.