Due: Monday, May 1, 2023, 11:49 p.m.

Data Analytics and Modeling: Quantitative Analysis for Economic Strategy (ECON453)
Spring, 2023
Problem Set 4 (and 5)

Instructions:

- Submit your well labeled R code as an addendum along with your write-up. Start your R code with a comment with your name, Econ453, pset 4&5.
- I will look at your handwritten (hopefully typed) answers first. Your handwritten answers should be self contained. I will go over your code to see your technique and to see that your code matches with our answers. Please do not just dump output from R code and expect me to hunt down the answers.
- Where needed, data tables are available on d2L in an EXCEL file.
- Upload your completed homework as a single PDF file to the d2L. Please do not e-mail or submit hard copies.

1. The following table gives household level survey data on health care expenses, income, and education for households in the US, Canada, and Mexico. Only a few observations are shown here; complete data are on d2L in pset1.xlsx, sheet "medical_expenses." This is the same dataset used for pset1.

COUNTRY	LOCATION	MEDICALEXPN (100$)	INCOME (1000$)	EDUCATION (Years)
USA	RURAL	22.265	19	12
USA	RURAL	8.98	42	10
.	.	.	.	.
USA	URBAN	48.061	64	17
.	.	.	.	.
CANADA	RURAL	2.287	23	5
.	.	.	.	.
CANADA	URBAN	20.722	29	13
CANADA	URBAN	38.404	70	15
.	.	.	.	.
MEXICO	RURAL	13.065	7	3
MEXICO	RURAL	15.457	28	12
.	.	.	.	.
MEXICO	URBAN	49.915	76	18
MEXICO	URBAN	9.104	9	5
.	.	.	.	.

a) Using the data for all countries estimate a regression model that explains medical expenses as a function of income, education, country, and location (rural vs. urban). Note: You need to create dummy variables both country and location (urban vs. rural). Hint: Create dummy variables for USA, Canada and Mexico and use any two of them. Also, create a dummy variable for urban location and use it. In all, the regression model will have six beta coefficients (including an intercept).

b) Test the null hypothesis that country and location (urban vs rural) jointly have no effect on medical expenses. Interpret your results using complete sentences. Hint: Use `linearHypothesis` function in R.

c) Using the estimated model, construct a 95% confidence interval for expected medical expenses for a rural Mexican household with an income of $50,000 and 12 years of education. Compare this to that of an urban US household with the same level of income and education. Do both intervals overlap?

2. The following table gives standardized test scores for randomly selected fourth graders from two schools (Lincoln School and Kennedy School) and for two years (2014 and 2016). Only a few observations are shown here; complete data are on d2L in pset1.xlsx, sheet "scores." This is the same dataset used for pset1.

year	school	score
2014	Lincoln	40
2014	Lincoln	12
.	.	.
.	.	.
2014	Lincoln	59
2014	Kennedy	66
2014	Kennedy	93
.	.	.
.	.	.
2014	Kennedy	73
2016	Lincoln	65
2016	Lincoln	95
.	.	.
.	.	.
2016	Lincoln	82
2016	Kennedy	68
2016	Kennedy	85
.	.	.
.	.	.
2016	Kennedy	99
2016	Kennedy	51

.
We will redo some of the hypotheses testing you did in pset2 using a regression model. Regression analysis provides a convenient tool for accommodating various hypothesis testing.

For answering the following questions, you need to create Kennedy School Dummy and 2016 dummy as follows: dken = 1 if observation belongs to Kennedy School and 0 if not. d2016 = 1 if the observation is for year 2016 and 0 if the observation is for year 2014.

a) Recall that Question (1c) in pset 2 is, "*The School District, which encompasses both schools, claims that average test scores increased for both schools between 2014 and 2016. Does the data support the School District's claim? Verify the claim one school at a time.*"

   Redo this test using regression analysis. This can be done with regression model, $score_i = \beta_0 + \beta_1 d2016_i + \varepsilon_i$

Hint: Using data for Kennedy School, regress scores on d2016 and estimate betas. The hypothesis of interest is: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 > 0$. Then, repeat this, using data for Lincoln School.

b) Recall that Question (1d) in pset 2 is, "*A county official claims that performance of fourth graders does not differ between both schools. Does the data support the claim? Verify the claim for 2014 and 2016 separately.*"

Redo this test using regression analysis. This can be done with regression model, $score_i = \beta_0 + \beta_1 dken_i + \varepsilon_i$

Hint: Using data for 2014, regress scores on d2016 and estimate betas. The hypothesis of interest is: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Then, repeat this, using data for 2016.

c) Using data for both schools for both years, estimate the model $score_i = \beta_0 + \beta_1 dken_i + \beta_2 d2016 + \varepsilon_i$.

Using estimates from this model test the hypothesis that average test scores are the same for both schools and for both years against the alternative that they are not. Hint: Here, $H_0 : \beta_1 = 0, \beta_2 = 0$, H_1 : At least one of β_1 *and* $\beta_2 \neq 0$

3. Use Wages data from Chapter 8 of the text for this question. Data are saved on d2L (Look for jaggia_ba_2e_ch08_data.xlsx file on d2L). This data set has cross sectional data on individual wages, a dummy for graduate degree and age of the individual. Hint: Parts (a) – (d) of this question is based on Example 8.5 (pages 289 – 291) from the text.

a) Draw scatter plot for the data with age on x-axis and wages on y-axis. What can you infer from this chart? You may use EXCEL for this chart.

b) There are 160 observations in the dataset. Partition the sample into two sub-samples: the first 140 observations (training set) and the last 20 observations (validation set). Using the first 140 observations, fit two regression models. In model 1, regress *wages* on *graduate* and *age*. In model 2, regress *wages* on *graduate*, *age*, and age^2. Summarize estimated coefficients, standard errors, p-values, R-square, and Adjusted R-square from both models in a well formatted table. Which model would you choose? Justify your choice. Note: You may use EXCEL/WORD to format this table.

c) Using estimated models, predict expected wages for a 30-year-old individual with a graduate degree.

d) According to estimated model 2, at what age are wages at the maximum? Do these calculations in R.

e) Cross validation. Calculate, MSE, RMSE, MAPE, MAE, MAPE for both models over the training set and over the validation set. Present these results in a table similar to the following:

Cross Validation Results for Wage Data

	Model 1	Model 2
Training Set (Observations 1 – 140):		
R-square		
Adjusted R-Square		
$\left[\text{correlation}\left(wages, \hat{wages}\right)\right]^2$		
MSE		
RMSE		
MAE		
MAPE		
Validation Set (Observations 141 – 160):		
$\left[\text{correlation}\left(wages, \hat{wages}\right)\right]^2$		
MSE		
RMSE		
MAE		
MAPE		

Comment on these cross-validation results. Based on these results, which model would you choose?