Konstantin Kucheryavyy
Graduate School of Public Policy
University of Tokyo

Data Science for Practical Economic Research

Assignment 2, 90 points

Due Date: Wednesday, June 16th, end of day

INSTRUCTIONS:

- For this assignment, use the data set that you described in Assignment 1.
- You can submit this assignment later than the deadline, but each day of a delay would cost you 5 points.

THE ASSIGNMENT: In this assignment you are asked to play with the linear (multiple) regression model. To review the linear regression model, see the slides Linear Regression. Also, see the demo code.

The steps below are typical steps you would perform when working with a linear model.

1. **(5 points)** Choose one variable as your outcome and several other variables as your predictors. Explain the logic behind your choice.
   Comment: When you are choosing predictor variables for your outcome, you should typically have in mind some story or, more formally, a theory. You *cannot* just throw in into your regression all the variables that you have. Doing so is problematic for several reasons. One of them is chance (or spurious) correlations. Take a look at this article
   `https://www.bbc.com/news/magazine-27537142`,
   and check this hillarious web resource:
   `http://www.tylervigen.com/spurious-correlations`.

2. Fit a linear regression (*Ordinary Least Squares*, OLS) model.

   (a) **(5 points)** How well does your model fit the data? Use R^2 as well as plots to answer this question. You can plot outcomes in the data (Y) versus fitted

outcomes ($\hat{Y}$).

Comment: See slides 20-22 for an overview of R^2. The plot of Y versus $\hat{Y}$ is called the **residual scatter**. A perfect fit $\left(R^2 = 1\right)$ corresponds to all points on the 45 degree line, no fit ($R^2 = 0$) corresponds to all points on the vertical line corresonding to the average value of Y in the sample ($\overline{Y}$).

(b) **(5 points)** Provide interpretation of your estimated coefficients. Pay special attention to the signs. Do they make sense?

Comment: See slides 9 and 27 for the interpretation of estimated coefficients of a linear model.

(c) **(5 points)** Is there evidence of error terms heteroskedasticity? Use the Breusch-Pagan and/or White test for heteroskedasticity to check this formally.

Comment: See slide 53. Also, see the demo code on how to perform the the Breusch-Pagan and White tests — it is very simple.

(d) **(5 points)** Calculate heteroskedasticity-robust standard errors. Compare them with regular standard errors.

Comment 1: See the demo code — it is very simple.

Comment 2: Even if heteroskedacticity of error terms is present, the estimated coefficients are still unbiased and consistent (however they are not asymptotically efficient, but often this is not a big deal if your sample is large). Also, the goodness-of-fit measure R^2 is unchanged. However, the usual OLS t-statistics do not have t distributions in the presence of heteroskedasticity. The consequences of this is that we cannot test the hypothesis $\mathrm{H}_0 : \beta_j = 0$ using the t-statistics. Similarly, F-statistics are no longer F-distributed, and we cannot test the hypothesis $\mathrm{H}_0 : \beta_1 = \cdots = \beta_p = 0$ using the F-statistics. Econometricians (White, 1980) have developed results that allow to adjust standard errors and t- and F-statistics so that they are *valid* in the presence of heteroskedasticity of *unknown* form. Validity here means that, for large samples, t-statistics calculated using robust standard errors has a distribution that is close to the t-distribution. However, this is not true for small samples (at the same time, if standard erorrs are homogenous, then t-statistics has the t-distribution for any sample size). *Practically, for any large sample you are working with, you should just calculate heteroskedasticity-robust standard errors instead of the regular standard errors.* Alternatively, you can calculate both types of errors and compare them. If the difference is not large, you can argue that your conclusions are not sensitive

to the standard error is use.

(e) **(5 points)** Provide 95% confidence intervals for the estimated coefficients based on the heteroskedasticity-robust standard errors. For which of the coefficients can you reject the null hypothesis $H_0 : \beta_j = 0$? Comment on the coefficients (if any) for which you *cannot* reject the null hypothesis $H_0 : \beta_j = 0$.

(f) **(5 points)** Is there evidence of high leverage observations? Look at observations with the highest leverage statistics. Do they make sense? Can they be caused by incorrect data entry process?
Comment: See slides 54-55. Do not discard such boservations unless you strongly believe they are the result of an error in the data entry process! They can be a source of valuable information.

(g) **(5 points)** Is there evidence of outliers? Look at observations with the largest studentized residuals. Do they make sense? Can they be caused by incorrect data entry process?
Comment: See slides 56-57. Do not discard such boservations unless you strongly believe they are the result of an error in the data entry process! They can be a source of valuable information.

(h) **(5 points)** Is there evidence of collinearity between regressors? Look at the variane inflation factors (VIF). If you find some regressors "highly" correlated with other regressors, think about the reasons for that. Does it make sense to combine some of the regressors into one regressor? Does it make sense to exclude some of the regressors due to collinearity?
Comment: See slides 58-63. Collinearity in itself does not violate any assumptions of OLS. So, theoretically, there is no reason to discard highly collinear regressors. In fact, it is impossible to even formally define "highly collinear". But looking at regressors with, say, high VIF, may give you a clue that you are duplicating information by including some of the regressosrs. In this case it might make sense to drop some of them or combine them into one (for example, by adding them up).

(i) **(5 points)** The whole purpose of Machine Learning methods is prediction. So, let's do it. Use your estimated model to predict some outcome for a (hypothetical) observation that is not in your sample.
Comment: Once you made a prediction for your outcome, you might wonder

how precise is your prediction. To address this question, we use confidence and/or prediction intervals. If error terms are homoskedastic, then we can use the formulas for confidence/prediction intervals provided in the slides. If, on the other hand, there is evidence of heteroskedasticity, then things blow up. While construction of confidence/prediction intervals can be worked out analytically under some additional assumptions, it is quite painful (those students who are interested in this question can read Section 8-4d of Wooldrige "Introductory Econometrics. A Modern Approach.").

3. **(40 points)** Repeat the previous exercise (2a-i), but this time fit a log-linear model. That is, take logs of your outcome and of the regressors for which it makes sense, work with the transofrmed data the same way you worked with the orginal data in the previous exercise.
   Comment: Log-tranformation of the data is one of the cheap ways to get a "better" model that is still very simple. Sometimes it helps, sometimes it is not. In any case, often, it is worth trying. If you believe that in your case a log-linear model does not make sense, provide arguments for this.