Regression Analysis
Final Project
(50 points)

This assignment is for candidates enrolled in our Programs in Analytics and Statistical Studies (PASS). Others are welcome to take it, and view model answers and compare their answers with it, but marks will be provided only to those registered in the above program.

Problem 5.5 on pages 237-240: This exercise uses the UBS dataset and works through a complete multiple linear regression analysis from beginning to end. The analysis follows the model building guidelines in Section 5.3. First, you'll view the data in a graph to make sure that there are no really extreme values in the data that might cause modeling problems later. Then, you'll fit a basic model with each predictor left untransformed and with no interaction terms; you'll find that this model is unsatisfactory. Next, you'll add twelve interactions (between the quantitative predictors and the indicator variables for region) to the model. This over-complicated model has some redundant predictor terms in it, so you'll next use a nested model F-test to remove three of the interactions you just added, then another nested model F-test to remove three more interactions. Next, you'll remove one more interaction using an individual t-test. The final model should include the three indicator variables for region, all four of the original quantitative predictors, and five interactions between indicator variables for region and quantitative predictors; this model should be satisfactory with respect to the model assumptions. The remaining parts of the exercise ask you to think about what the final estimated regression equation tells you about various economic ideas. The economic ideas in parts (j), (k), and (l) are best illustrated using the line graphs of Section 5.5; it can sometimes be difficult to get statistical software to construct these graphs correctly, so if you're having trouble here you can always answer the questions using just some simple algebra (or even by drawing the graphs roughly by hand). This exercise is long, but intended to be reasonably straightforward.

Problem 5.5: This problem is inspired by an example in Cook and Weisberg (1999). UBS AG Wealth Management Research conducts a regular survey of international prices and wages in major cities around the world (UBS, 2009). One variable measured is the price of a Big Mac hamburger, *Bigmac* (measured in the natural logarithm of minutes of labor required by an average worker to buy a Big Mac). The Big Mac is a common commodity that is essentially identical all over the world, and which therefore might be expected to have the same price everywhere. Of course it doesn't, and so economists use this so-called “Big Mac parity index” as a measure of inefficiency in currency exchange. The task is to build a multiple regression model to explain the variation in *Bigmac* for 73 cities in 2009 using the following predictor variables available in the **UBS** data file:

Wage = natural logarithm of average net wage, relative to New York = $\log_e(100)$.
Bread = natural logarithm of minutes of time required by average worker to buy 1 kg bread.
Rice = natural logarithm of minutes of time required by average worker to buy 1 kg rice.
Vac = average paid vacation days per year.
D_{As} = indicator variable for 14 cities in Asia.
D_{Em} = indicator variable for 17 cities in Eastern Europe or the Middle East.
D_{Sa} = indicator variable for 10 cities in South America or Africa.

The response variable and three of the predictor variables are expressed in natural logarithms to aid modeling since the original variables have highly skewed distributions (a few very high values relative to the rest). The reference region for the indicator variables is “North America, Western Europe, and Oceania” (32 cities).

a. Draw a scatterplot matrix of *Wage*, *Bread*, *Rice*, *Vac*, and *Bigmac*, and use different plotting symbols for each region [computer help #16 and #17]. Write a couple of sentences on anything of interest that you notice.

b. Fit a multiple regression model with D_{As}, D_{Em}, D_{Sa}, *Wage*, *Bread*, *Rice*, and *Vac* as predictors, and save the studentized residuals [computer help #31 and #35]. Draw a residual plot with these studentized residuals on the vertical axis and *Bread* on the horizontal axis. Write a couple of sentences about how to check three regression assumptions (zero mean, constant variance, independence) using residual plots like this. You should find that the zero mean assumption is most at risk of failing (why?), while the constant variance and independence assumptions probably pass.

c. To improve the model, consider interactions. In particular, it seems plausible that the effects of *Wage*, *Bread*, *Rice*, and *Vac* on *Bigmac* could vary according to region. So interactions between the indicator variables and quantitative predictors offer promise for improving the model. Create the 12 interactions: D_{As}*Wage*, D_{As}*Bread*, D_{As}*Rice*, D_{As}*Vac*, D_{Em}*Wage*, D_{Em}*Bread*, D_{Em}*Rice*, D_{Em}*Vac*, D_{Sa}*Wage*, D_{Sa}*Bread*, D_{Sa}*Rice*, and D_{Sa}*Vac* [computer help #6]. Next, fit the multiple regression model with D_{As}, D_{Em}, D_{Sa}, *Wage*, *Bread*, *Rice*, *Vac*, and these 12 interactions [computer help #31]. Which three interactions have the largest p-values in this model?

d. Let's see if we can remove these three interactions without significantly reducing the ability of the model to explain *Bigmac*. Do a "nested model F-test" by fitting a reduced multiple regression model with D_{As}, D_{Em}, D_{Sa}, *Wage*, *Bread*, *Rice*, *Vac*, D_{As}*Wage*, D_{As}*Bread*, D_{As}*Rice*, D_{As}*Vac*, D_{Em}*Bread*, D_{Em}*Vac*, D_{Sa}*Bread*, D_{Sa}*Rice*, and D_{Sa}*Vac*, and adding D_{Em}*Wage*, D_{Em}*Rice*, and D_{Sa}*Wage*, to make a complete model [computer help #34]. What are the values of the F-statistic and the p-value for this test? Does this mean that we can remove these three interactions without significantly reducing the ability of the model to explain *Bigmac*?

Hint: See Section 3.3.4.

e. Let's see if we can remove three more interactions without significantly reducing the ability of the model to explain *Bigmac*. Do a "nested model F-test" by fitting a reduced multiple regression model with D_{As}, D_{Em}, D_{Sa}, *Wage*, *Bread*, *Rice*, *Vac*, D_{As}*Bread*, D_{As}*Rice*, D_{As}*Vac*, D_{Em}*Bread*, D_{Em}*Vac*, and D_{Sa}*Bread*, and adding D_{As}*Wage*, D_{Sa}*Rice*, and D_{Sa}*Vac*, to make a complete model [computer help #34]. What are the values of the F-statistic and the p-value for this test? Does this mean that we can remove these three interactions without significantly reducing the ability of the model to explain *Bigmac*?

f. Now fit a multiple regression model with D_{As}, D_{Em}, D_{Sa}, *Wage*, *Bread*, *Rice*, *Vac*, D_{As}*Bread*, D_{As}*Rice*, D_{As}*Vac*, D_{Em}*Bread*, D_{Em}*Vac*, and D_{Sa}*Bread* [computer help #31]. If we want to have a more parsimonious model that preserves hierarchy, which predictor term can we now consider removing? Do an individual t-test to show this formally.

Hint: See Section 3.3.5.

g. Our final model has the following predictors: D_{As}, D_{Em}, D_{Sa}, *Wage*, *Bread*, *Rice*, *Vac*, D_{As}*Bread*, D_{As}*Rice*, D_{As}*Vac*, D_{Em}*Vac*, and D_{Sa}*Bread*. Fit this model and save the studentized residuals [computer help #31 and #35]. Draw a residual plot with these studentized residuals on the vertical axis and *Bread* on the horizontal axis [computer help #15 and #36]. Does it appear that the zero mean regression assumption that appeared to be violated in part (b) has now been corrected?

Note: Although the values of the regression standard error, s, and adjusted R^2 suggest that the model from part (f) may be preferable to this model, the individual t-test from part (f) suggests otherwise. For the purposes of this exercise, both models give very similar results and conclusions, and we will proceed with the model from this part as our final model for the remainder of the questions.

h. Write out the least squares regression equation for this final model [computer help #31]; that is, replace the $\hat{b}$'s with numbers in:

$$\widehat{Bigmac} = \hat{b}_0 + \hat{b}_1 D_{As} + \hat{b}_2 D_{Em} + \hat{b}_3 D_{Sa} + \hat{b}_4 Wage + \hat{b}_5 Bread + \hat{b}_6 Rice + \hat{b}_7 Vac + \hat{b}_8 D_{As} Bread + \hat{b}_9 D_{As} Rice + \hat{b}_{10} D_{As} Vac + \hat{b}_{11} D_{Em} Vac + \hat{b}_{12} D_{Sa} Bread.$$

i. An economist reasons that as net wages increase, the cost of a Big Mac goes down, all else being equal (fewer minutes of labor would be needed to buy a Big Mac since the average wage is higher). According to our final model, is the economist correct?

j. The economist also reasons that as the cost of bread increases, the cost of Big Macs goes up, all else being equal (food prices tend to fall and rise together). According to our final model, is the economist correct?

Hint: This is trickier to answer than part (i) since the "Bread effect" depends on D_{As} and D_{Sa}. Write this effect out as:

$$\hat{b}_0 + \hat{b}_1 D_{As} + \hat{b}_2 D_{Em} + \hat{b}_3 D_{Sa} + \hat{b}_4 m_{Wage} + \hat{b}_5 Bread + \hat{b}_6 m_{Rice} + \hat{b}_7 m_{Vac} + \hat{b}_8 D_{As} Bread + \hat{b}_9 D_{As} m_{Rice} + \hat{b}_{10} D_{As} m_{Vac} + \hat{b}_{11} D_{Em} m_{Vac} + \hat{b}_{12} D_{Sa} Bread,$$

replacing the $\hat{b}$'s with numbers, m_{Wage} with the sample mean of Wage, m_{Rice} with the sample mean of Rice, and m_{Vac} with the sample mean of Vac [computer help #10]. Then create this as a variable in the dataset [computer help #6], and draw a predictor effect line graph with the "Bread effect" variable on the vertical axis, Bread on the horizontal axis, and "Region" to mark four separate lines [computer help #42]. This should produce one line for each region; the economist may be correct for some, all, or none of the regions! To ensure that your line graphs are correct, make sure that you can recreate the predictor effect plots in Section 5.5 first, particularly Figure 5.15 on page 230.

k. The economist also reasons that as the cost of rice increases the cost of Big Macs goes up, all else being equal (food prices tend to fall and rise together). According to our final model, is the economist correct?

Hint: This is similar to part (j) since the "Rice effect" depends on D_{As}. Write this effect out as:

$$\hat{b}_0 + \hat{b}_1 D_{As} + \hat{b}_2 D_{Em} + \hat{b}_3 D_{Sa} + \hat{b}_4 m_{Wage} + \hat{b}_5 m_{Bread} + \hat{b}_6 Rice + \hat{b}_7 m_{Vac} + \hat{b}_8 D_{As} m_{Bread} + \hat{b}_9 D_{As} Rice + \hat{b}_{10} D_{As} m_{Vac} + \hat{b}_{11} D_{Em} m_{Vac} + \hat{b}_{12} D_{Sa} m_{Bread},$$

replacing the $\hat{b}$'s with numbers, m_{Wage} with the sample mean of Wage, m_{Bread} with the sample mean of Bread, and m_{Vac} with the sample mean of Vac [computer help #10]. Then create this as a variable in the dataset [computer help #6], and draw a predictor effect line graph with the "Rice effect" variable on the vertical axis, Rice on the horizontal axis, and "Region" to mark four separate lines [computer help #42]. Again, the economist may be correct for some, all, or none of the regions.

l. The economist also reasons that as vacation days increase the cost of a Big Mac goes up, all else being equal (productivity decreases, leading to an increase in overall production costs). According to our final model, is the economist correct?

Hint: This is similar to part (j) since the "Vac effect" depends on $D_{As}$and D_{Em}. Write this effect out as:

$$\hat{b}_0 + \hat{b}_1 D_{As} + \hat{b}_2 D_{Em} + \hat{b}_3 D_{Sa} + \hat{b}_4 m_{Wage} + \hat{b}_5 m_{Bread} +$$
$$\hat{b}_6 m_{Rice} + \hat{b}_7 Vac + \hat{b}_8 D_{As} m_{Bread} + \hat{b}_9 D_{As} m_{Rice} + \hat{b}_{10} D_{As} Vac + \hat{b}_{11} D_{Em} Vac +$$
$$\hat{b}_{12} D_{Sa} m_{Bread},$$

replacing the $\hat{b}$'s with numbers, m_{Wage} with the sample mean of Wage, m_{Bread} with the sample mean of Bread, and m_{Rice} with the sample mean of Rice [computer help #10]. Then create this as a variable in the dataset [computer help #6], and draw a predictor effect line graph with the "Vac effect" variable on the vertical axis, Vac on the horizontal axis, and "Region" to mark four separate lines [computer help #42]. Again, the economist may be correct for some, all, or none of the regions.