

**Final Wednesday: Answer Key****A. Discrimination and (French) fries**

- 1) A one percentage point increase in the proportion of blacks in the neighborhood (defined by zip code) is predicted to increase the price for French fries with 0.138 %, ceteris paribus  
A 1 percent increase in the median family income of the neighborhood is predicted to increase the price of French fries with 0.13%, ceteris paribus.
- 2) The estimation in the second column suffers from omitted variable bias. There are many omitted variables, included some of the ones in column 3. On average, the variables that are added in column 3 seem to reduce the positive bias in column 2. Indeed most (all) of them are likely to be positively correlated with the proportion of black population, and have themselves a positive effect on y, resulting in a likely positive bias in column 2 (although the direction of the bias also depends on the correlation between the different variables, given that this is multivariate case).
- 3) 90% confidence interval:  $\hat{\beta} \pm 1.654 * se(\hat{\beta}) = [-0.013, 0.091]$
- 4) False: Since the dependent variable is in the log-form, we need to take the  $\exp(-0.092)$  to interpret its meaning. We find that the predicted price for a restaurant in a neighborhood where the proportion of black is 0, is 0.91 cents.
- 5) Keeping everything else constant, the price of French fries in New Jersey is predicted to be 8.654 % ( $=100 * (\exp(0.083) - 1)$ ) higher than the price of French fries in Pennsylvania. The coefficient is significant at the 1% level.
- 6) False. The models in column 4 and 5 control for possible price differences across chains, which might be important to control for possible unobservables that are different across chains. Moreover, by controlling for chains, we can reduce the amount of noise in the model and increase the precision of the estimates. To see whether omitted variable bias problem would be a possible problem we need to evaluate 1) whether there is a correlation between any of the other independent variables and the chain dummy (e.g. because a certain chain might be more likely to locate in a certain type of neighborhood); 2) whether there is a correlation between the chain dummies and the independent variables. In column 4 (and 5) the t-tests indicate the fast food dummies are indeed significant. We could also do a joint F-test to confirm this result. Using the R-squared form of the F-test to compare column (3) with (4) we find:

$$F \equiv \frac{(R_{ur}^2 - R_r^2) / q}{(1 - R_{ur}^2) / (n - k - 1)} = \frac{(0.43 - 0.25) / 3}{(1 - 0.43) / (391 - 10)} = 40.1$$

This is clearly bigger than the critical value at 1% significance level which leads us to reject the null-hypothesis that the fast food dummies jointly don't have an effect on the log(price).

7) False. While the coefficient for the proportion is black is significantly positive in column 2, it turns negative and becomes insignificant once we control for more variables (and reduce the potential omitted variable bias problem). We hence don't find evidence of price discrimination against African Americans. The results, as they stand, do however suggest there might be price discrimination against high-income families, as the coefficient is significantly positive in all the specifications (though we need to worry whether there might be other omitted variables that would bias this result).

8) One would first need to create a dummy variable that equals 1 for Wendy's restaurant and zero for all others. Including that variable and the interaction effect of that variable with proportion of black, and taking out the other fast food restaurant dummies out of the model, gives us the correct model. Hence:

$$\log(\text{price}) = \beta_1 \text{black} + \beta_2 \log(\text{income}) + \beta_3 \text{crimerate} + \beta_4 \text{density} + \beta_5 \text{wcar} + \beta_6 \text{NJ} + \beta_7 \text{Wendy} + \beta_8 (\text{Wendy} * \text{black}) + u$$

If  $\beta_8$  is found to be significantly positive, there is evidence of Wendy's discriminating against African-Americans.

9) Many of you argued that local wages should be included to account for differences in costs of producing French fries. This could indeed be included if you had sufficient evidence that fast food restaurants are wage-takers, and hence local wages can be taken exogenously (e.g. because they'll pay only the minimum possible). In that case, we might indeed want to include that, especially since it might be correlated with income, and as such cause a violation of the zero conditional mean assumption. Hence the model would be:

$$\log(\text{price}) = \beta_1 \text{black} + \beta_2 \log(\text{income}) + \beta_3 \text{crimerate} + \beta_4 \text{density} + \beta_5 \text{wcar} + \beta_6 \text{NJ} + \beta_7 \text{BK} + \beta_8 \text{KFC} + \beta_9 \text{RR} + \beta_{10} \text{wages} + u$$

(see also Monday's exam on inclusion of county dummies).

## B. Fertility in Botswana

1) Coefficient of education in the second regression:

Keeping everything else constant, an additional year of education of the woman is predicted to decrease the probability of having more than 2 kids with 1.87 percentage points. The coefficient is significant at the 1% level.

2) No. Correlation between independent variables does not violate any assumption, unless 2 variables are perfectly collinear, which is obviously not the case here. Collinearity between 2 variables does cause multicollinearity, which reduces the precision of the estimates. In this case however, the precision of the estimates is still very high (all coefficients are significant at the 1% level), so it is not a reason for concern.

3) In this regression, there is some clear reasons to suspect heteroscedasticity, i.e. the variation of the independent variables (nr of kids) is likely to depend e.g. on the women's age (since among very young women the variation of the number of kids will be low, but among older women there can be much more variation). Weighted Least Squares is the efficient estimator in case of heteroscedasticity, and given that we have a good idea about the form of the heteroscedasticity here, we can use this to construct weights.

- 4) a: No: adding the 2 samples together does not cause a violation of any assumption. Given that the sum of 2 random variables is still random, and given that this is a case of sampling on an exogenous variable (if we include the rural (or urban) dummy), we don't violate the random sampling assumption. Adding the 2 samples together also does not affect any of the other assumptions. One might only wonder whether it makes sense to assume that the independent variables affect the dependent variables in the same way in the urban and rural areas, which is what we look at in point b.
- b. One would need to do a Chow-test. This can be done in 2 ways. The first method is to interact all the independent variables with the rural (or urban) dummies, and testing for joint significance of all the interaction effects and the urban (or rural) dummy. Or one could estimate the regression separately for the rural sample, for the urban sample, and for all observations together and use the following form of the F-test: 
$$F = \frac{[SSR_p - (SSR_1 + SSR_2)] [n - 2(k + 1)]}{SSR_1 + SSR_2} \cdot \frac{1}{k + 1}$$
- 5) By creating dummy variables for 2 of the 3 religions, and including these in the model. If we suspect that religion is correlated with any of the other independent variables (e.g. *usemeth*), and itself likely affects the number of children, this would clearly improve the regression by reducing omitted variable bias problem.
- 6) If the women have not given the correct information, we have a problem of measurement error on one of the independent variables, which causes a violation of the zero conditional mean assumption. The implications depend on the type of measurement error. If the measurement error is random, we know that the coefficients will be biased towards zero (attenuation bias). However, if the measurement error is systematic (i.e. correlated with the true value), the problem is even larger as we now will have biased and inconsistent results.

### C. True or False

- 1) False: In order to make accurate inferences, we also need the normality assumption.
- 2) False: Autocorrelation can occur in time-series regressions, even when a fixed time trend is controlled for (e.g. because of a shock that has effects in more than one time period).