

Final Monday: 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 first column one suffers from omitted variable bias. One of the omitted variables is median family income. Since median income is likely to be negatively correlated with the proportion of black, and income itself is positively correlated with the price of French fries (as we can see from the positive coefficient of income in the second column) there is a negative bias in the first column.
- 3) The significance of the coefficient depends on the t-stat, which we obtain by dividing the coefficient estimate by the standard error. In the second column the coefficient estimate of the proportion of black is higher, which increases the t-stat, given a more or less constant standard error. The standard error does not in fact differ that much between the 2 columns, reflecting 2 off-setting effects: by including a variable that is correlated with the proportion of black we decrease precision (~multicollinearity), but since the income variable reduces the noise (σ), there is also an offsetting positive effect on precision.
- 4) There are only 4 type of restaurants in the data, and including a fourth dummy would result in perfect collinearity.

$$5) \frac{\partial \log(\text{price})}{\partial NJ} = 0.078 + 0.312 * \text{propblack}$$

The price difference between NJ and PE is higher the larger the proportion of black.
E.g. a 10% point increase in the proportion of black increases the price difference with an extra 3%, *ceteris paribus*.

$$\frac{\partial \log(\text{price})}{\partial \text{propblack}} = -0.327 + 0.312 * NJ$$

The price of fries decreases with the proportion of black in Pennsylvania,
but there is almost no change in NJ (sum = -0.015).

- 6) In choosing between the different models, we need to evaluate which specification is least likely to violate the Gauss Markov assumptions, and in particular the zero conditional mean assumption. Since every column includes more variables, we need to consider whether these variables are significant. If they are significant and if there is reasons to believe they might be correlated with some of the other variables it is better to leave them in, since we might be violating the zero conditional mean assumption otherwise. (An additional benefit is that adding significant variables reduces the noise in the model and therefore increases precision). An F-

test of joint significance (comparing the restricted model in column 3 (or 2) with the unrestricted model in column 4 indicates that the added variables are jointly significant. This is not surprising given that t-tests indicate that many of the variables individually are significant (population density, NJ dummy, BK and RR dummies). 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).

On the other hand the t-test on the coefficient of the interaction term in column 5 is low. Hence including the interaction effect does not add to the model, and we are left with the specification in column 4.

- 7) To show the robustness (or in this case the sensitivity) of the results. By comparing the columns, we can see that the significance of the income variable is robust across the different specifications, while this is clearly not the case for the coefficients on the proportion of blacks.
- 8) a. If we know in which county the restaurants are located, we can control for any county-level unobservables that might be correlated with both the dependent and the independent variables and therefore reduce the possibility of omitted variable bias.
b. One would need to create a dummy variable for each county, and include 28 out the 29 dummy variables in the model.

B. Fertility in Botswana

- 1) False:
Based on the regressions, it is not really possible to evaluate this statement, as one would need to do a t-test testing $\beta_{educ} = \beta_{heduc}$. (in this particular case, since the 95% confidence intervals don't overlap, it is likely that we can reject that null-hypothesis).
In addition, one would need to evaluate whether the Gauss-Markov assumptions hold to know what we can learn from this regression about the impact of education. Some of the assumptions (e.g. the zero conditional mean assumption (because of inclusion of endogeneous variables such as *usemeth*), as well as the random sampling assumption (because many observations are missing) are likely to be violated in this regression (see further).
- 2) The trade-off is mainly between violations of the random sampling assumption and the zero conditional mean assumption. The information of husband's education is missing for many observations. Including this variable might possible lead to a non-random sample (since the observations on husband's might be missing for non-random subset of the sample). Violation of the random sampling assumptions might be solved by dropping this variable. On the other hand, given that the variable has a significant impact, and is likely correlated with other independent variables in the model (e.g. education of the respondent), leaving the variable out would lead to a violation of the zero conditional mean assumption. In addition to the impact on the bias, the impact on precision of the estimates is also ambiguous (more observations versus reducing noise).

- 3) Using robust standard errors was probably a good idea, since the homoscedasticity assumption is likely to be violated in this regression. E.g. because the variance of the number of children is likely to depend on age of the women (the older she is, the more children she could have had). A more efficient way of dealing with this heteroscedasticity would possibly be using Weighted Least Squares, which is BLUE if one knows the form of the heteroscedasticity. E.g. if one could correctly assume the effect of age on variance is linear, one could use age as weight, and obtain an efficient estimate.
- 4) The second regression is a linear probability model in which the homoscedasticity assumption is violated by construction (because $Var(y|x) = p(x)[1 - p(x)]$). Since the LPM might predict variances that are negative, we can't use WLS (in most cases) and use robust standard errors, which gives consistent estimates.
- 5) The use of contraceptives (at least once in the past) increases the predicted probability of having more than 2 children with approximately 24 percentage point. The precise estimate is 27 percentage points. This might not seem very intuitive since one might expect the use of contraceptives to decrease the probability of having more than 2 kids. We might however expect reverse causality (women might use contraceptives once they have more than 2 kids) or spurious correlation (certain types of women are more likely to use contraceptives and to have more kids).
- 6) Information about the (randomly selected) locations where contraceptives were distributed could potentially serve as a good instrumental variable for the use of contraceptives. The problem with the regressions as they stand is that the use of contraceptives is clearly endogenous. To improve the regressions we can now address this problem by using the 2 stage Least Squares method.
 - a. In a first stage one would estimate: $contraceptives = f(\text{location with free distribution})$
 $x = \pi_0 + \pi_1 z + v$ (with $x = usemeth$ and z a dummy indicating whether the location was randomly selected to get the free distribution:
 - b. In the second stage : $y = \beta_0 + \beta_1 \hat{x} + otherfactors + u$ with \hat{x} the prediction based on the first stage
- 7) In addition to the problems related to the violations of the assumptions, a problem with predictions would be the fact that the LPM can predict outcomes that are smaller than 0 or bigger than 1, which don't make sense. If the main purpose is to obtain sensible predictions, one would definitely want to use a probit or a logit model .

C. True or False

- 1) True: Because of zero conditional mean assumption violations (omitted variables, reverse causality, measurement error, ...) one cannot correctly identify the effect that x has on y .
- 2) True. In a time series model, we might want to add lags of the independent variables, if we believe that certain independent variables might have a long(er) run effect. If such effects indeed exists, omitting these lagged variables would result in a violation of the zero conditional mean assumption, and therefore in bias. On the other hand, many independent variables are likely to be correlated over time. Adding lags therefore introduces multicollinearity and might decrease the precision of the estimate.