Econ 140 - Spring 2016
Section 9

GSI: Fenella Carpena

March 31, 2016

1 Assessing Studies Based on Multiple Regression

1.1 Internal Validity

Threat to Internal Validity	Examples/Cases	Implications for OLS Estimates	Possible Solutions
OVB	Example: $wages_i = \beta_0 + \beta_1 educ_i + u_i$ where $educ_i$ is educational attainment likely suffers from OVB, since factors like where parent's education and ability probably affect both levels of education and wages.	OLS estimators will be biased and inconsistent.	(1) If there is data on the omitted variable, we can include it. (2) If there is no data available, we can try to use proxy variables (e.g., ability is not observed so there will be no data available, but we can use GPA as a "proxy" for ability). (3) Use panel data (if possible). Note: Other solutions to be discussed in future chapters are: instrumental variables, or conduct a randomized controlled experiment.

Continued on next page

Continued from previous page

Threat to Internal Validity	**Examples/Cases**	**Implications for OLS Estimates**	**Possible Solutions**
Functional Form Misspecification	Example: True population regression function is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$, but the regression we run is $Y_i = \beta_0 + \beta_1 X_i + u_i$.	OLS estimators will be biased and inconsistent.	(1) Use economic intuition to choose the appropriate functional form, e.g., if you think there are decreasing returns to X, you might want to include a X^2 as a regressor or use $ln(X)$ instead of X (2) Look at scatterplots to visualize the relationship between X and Y, then select a functional form based on this scatter plot. (3) Check for possible non-linearities by adding polynomial terms & testing for significance.
Measurement Error	Cases: (A) Measurement error in X, (B) Measurement error in Y only	*Case (A):* If we have "classical measurement error" (i.e., the measurement error is uncorrelated with the true X and with the population regression error), OLS estimator will be biased and inconsistent. In particular, the estimator will be biased toward zero, this is known as "attenuation bias." *Case (B)*: Assuming LS assumptions still hold, OLS estimates will be unbiased and consistent. However, it will increase the SE of the estimates (i.e., less precise).	(1) Get a more accurate measure of the variables (2) For Case (A), develop a mathematical model of the measurement error, and use the resulting formula to adjust the estimate (e.g., if you have classical measurement error and you can make an informed guess about the value σ_w^2 (variance of the measurement error), you can use this σ_w^2 to correct for the bias). Note: Other solutions to be discussed in future chapters are: use instrumental variables.

Continued on next page

Continued from previous page

Threat to Internal Validity	**Examples/Cases**	**Implications for OLS Estimates**	**Possible Solutions**
Missing Data and Sample Selection	Cases: (A) Data is missing at random, (B) Missing data is selected based on X, (C) Missing data is selected based on Y. Examples: *Case (A):* You conducted a survey of a simple random sample of 100 individuals, but you lost a random subset of 20 completed questionnaires; *Case (B):* In a regression of test scores on student teacher ratio (STR), we use data only from districts with STR above 20; *Case (C):* Suppose we want to estimate the effect of GRE scores on the likelihood of success in grad school (grades). There is a sample selection problem here based on the outcome of interest, because grades are only available for those students who were admitted and enrolled.	*Case (A):* Assuming LS assumptions still hold, OLS estimates will be unbiased and consistent. However, we have a smaller sample size, which can make estimates less precise. *Case (B):* Assuming LS assumptions still hold, OLS estimates will be unbiased and consistent, but estimates can be less precise if there is not enough variation in the independent variables. *Case (C):* OLS estimates are biased and inconsistent.	Methods for estimating models with sample selection are beyond the scope of this course.
Simultaneity	There is a two-way relationship between the independent variable (X_i) and dependent variable (Y_i) in the regression model: $Y_i = \beta_0 + \beta_1 X_i + u_i$ and $X_i = \gamma_0 + \gamma_1 Y_i + v_i$. *Example from the textbook:* Suppose there is a govt policy that gives schools with poor test scores extra money to hire teachers and reduce class size. High STR might be related to low test scores, and low test scores would also lower STR because of the govt program.	OLS estimators will be biased and inconsistent.	Potential solutions will be discussed in future chapters: (1) Use instrumental variables, (2) Use randomized controlled experiments.
Heteroskedasticity	Occurs when the regression error is heteroskedastic, but SEs were calculated under homoskedasticity	Does not affect OLS coefficient estimates, but SEs are incorrect. Therefore, hypothesis tests/confidence intervals will be invalid.	Always use heteroskedastic-robust SEs.

Continued on next page

Continued from previous page

Threat to Internal Validity	Examples/Cases	Implications for OLS Estimates	Possible Solutions
Correlation of the error term across observations	Example: when the data are repeated observations of the same entity over time (e.g., panel or time series data)	Violates the i.i.d. assumption of least squares. Does not affect OLS coefficient estimates, but SEs are incorrect. Therefore, hypothesis tests/confidence intervals will be invalid.	Standard errors will need to be adjusted using clustered SEs.

1.2 External Validity

Threat to External Validity	Examples/Cases	Implications for OLS Estimates	Possible Solutions
Differences in population	Example: Lab studies on the toxic effects of chemicals are conducted using animal populations like mice, but the results are used to write health/safety regulations for humans.	Affects whether we can generalize the results to a different population/setting	If there are two or more studies on different but related populations/settings, external validity can be checked by comparing their results.
Differences in settings	Example: Examining the effect of student-teacher-ratio on test scores among elementary schools in California, but the results may not apply to elementary schools in Massachusetts.	Affects whether we can generalize the results to a different population/setting	If there are two or more studies on different but related populations/settings, external validity can be checked by comparing their results.

1.3 Review the Concepts Exercises, Chapter 9

1. What is the difference between external and internal validity? Between the population studied and the population of interest?

   Answer: Internal and external validity are different because the former deals with whether the statistical analysis about the causal effects are valid for the population being studied, while the latter deals with whether the result from the study can be generalized to other populations/settings.

   The difference between the *population studied* and the *population of interest* is that the former is the population of entities (e.g., people, companies, school districts, etc.) from which the sample was drawn, while the latter is the population to which the results of the study are generalized. For example, if we study the effect of class size on test scores using a random sample of elementary school districts in California, then the population studied is California elementary school districts. If a policy maker wants to generalize the results of the study to high schools, then the population of interest is high schools.

2. Key Concept 9.2 describes the problem of variable selection in terms of a trade-off between bias and variance. What is this trade-off? Why could including an additional regressor decrease bias? Increase variance?

   Answer: The trade-off is that we want to include as many variables as necessary in our multiple regression model, so as to avoid OVB. But at the same time, including more regressors can lead to less precise estimates (i.e. larger variance) of the coefficient of interest.

   Including an additional regressor can decrease bias if for example, without the regressor, our regression suffers from OVB. Including an additional regressor can increase variance, because as we add more and more variables, say X_2 to X_{100}, there is not enough variation left in X_1, so the variance of our estimate for β_1 will increase.

3. Economic variables are often measured with error. Does this mean that regression analysis is unreliable? Explain.

   Answer: The implications of measurement error on the regression analysis depends on whether the measurement error is in the dependent variable Y or the independent variable X. If the measurement error is in Y, then it will not affect the internal validity of the OLS regression, but our OLS estimates will be less precise. If the measurement error is in X, we would have biased and inconsistent estimates.

4. Suppose that a state offered voluntary standardized tests to all its third graders and that these data were used in a study of class size on student performance. Explain how sample selection bias might invalidate the results.

   Answer: Sample selection bias occurs here because the data consists of only those schools which volunteered to take the standardized test, so the sample is not representative of the population of schools. As an example, suppose we have four types of schools: (1) low class size and high performing students, (2) low class size and low performing students, (3) high class size and high performing students, (4) high class size and low performing students. If the data contains only schools of type (1) and type (3), then if we regressed student performance on class size, we would likely find a zero effect, because in this sample, student performance is high regardless of class size. However, this finding is not necessarily because class size has no effect on student performance, but rather because the sample is biased.

5. A researcher estimates the effect on crime rates of spending on police by using city-level data. Explain how simultaneous causality might invalidate the results.

   Answer: There is simultaneous causality because spending on police affects crime rates, and at the same time, crime rates affect spending on police. Spending on police affects crime rates because if we spend more resources on the police force, then there would be more police patrolling, etc. which would deter crime. Crime rates affect spending on police because in a given city where crime rates are high, the city officials might decide to do something about it and therefore spend more on police.

6. A researcher estimates a regression using two different software packages. The first uses the homoskedasticity-only formula for standard errors. The second uses the heterosekdasticity-robust formula. The standard errors are very different. Which should the researcher use? Why?

   Answer: The researcher should use the heteroskedastic-robust SEs. This is because if the regression error is homoskedastic, then both homoskedastic and heteroskedastic SEs will be consistent. However, if the regression error is heteroskedastic, then the homoskedastic SEs are inconsistent, but the heteroskedastic errors will be consistent. In other words, homoskedastic SEs are consistent only if the regression errors are homoskedastic, while heteroskedastic SEs are consistent regardless of whether the regression errors are homoskedastic or heteroskedastic.

2 Panel Data

2.1 Notation: Quick Explanation

Example 2.1.1. Suppose you have the following data set. Fill in the columns for i, t, DItaly, DYear2001 in the following table, where DItaly is a dummy variable equal to 1 if the country is Italy, and DYear2001 is a dummy variable equal to 1 if the year is 2001. What are n and T in this data?

> **Answer:** n denotes the total number of entities (here, countries) so $n = 3$. T denotes the number of time periods we observe, so $T = 4$. Answers to the table are below in blue font color.

i	t	Country	Year	GDP	Interest rate	Europe	DItaly	DYear2001
1	1	Brazil	2000	100	22	0	0	0
1	2	Brazil	2001	150	18	0	0	1
1	3	Brazil	2002	170	15	0	0	0
1	4	Brazil	2003	220	12	0	0	0
2	1	USA	2000	1000	4	0	0	0
2	2	USA	2001	1150	3	0	0	1
2	3	USA	2002	1400	2	0	0	0
2	4	USA	2003	1900	2	0	0	0
3	1	Italy	2000	60	6	1	1	0
3	2	Italy	2001	130	5	1	1	1
3	3	Italy	2002	150	5	1	1	0
3	4	Italy	2003	190	3	1	1	0

2.2 Regression with Panel Data

Example 2.2.1. Some states in the US have enacted laws that allow citizens to carry concealed weapons. Proponents of argue that if more people carry concealed weapons, crime will decline because criminals will be deterred from attacking other people. Opponents argue that crime will increase because of accidental or spontaneous use of weapons. As a researcher, you would like to examine the effect of concealed weapons laws on violent crime. Suppose that you have a data on 50 US states for the years 1977 to 1999. The data contains violent crime rate (the variable name is *vio*) for the 50 states across these years, and a variable called *guns* which is dummy variable equal to 1 if the state has a law in effect that year which allows citizens to cary concealed weapons.

Type # 1: Regression with Only Entity Fixed Effects. For a regression with only entity fixed effects, we write the population regression as follows

$$vio_{it} = \beta_1 guns_{it} + \alpha_i + u_{it}$$

where, as we've discussed in the previous section of these notes, i denotes the entity, and t denotes time.

(a) What is α_i?

Answer: Here α_i is state fixed effect (there is one for each state i). Note that it does not have a subscript t because it does not vary over time.

(b) What does α_i capture?

Answer: α_i contains **ALL** state-specific characteristics that affect vio_{it} but are **not changing over time**. In our example, the state fixed effect α_i would contain all state-specific characteristics that are constant over time that affect violent crime. Examples include cultural views and political ideology (assuming these do not change over time). All of these non-time varying characteristics will be contained in α_i.

(c) In practice, how do we estimate the above regression?

Answer: In practice, one way we can estimate the above regression is by adding a dummy variable for each state. In this case, there are 50 states in the data. However, note because of the **dummy variable trap,** if we were to estimate the above regression with an intercept, we should include only 49 state dummies in the regression (instead of all 50 state dummies). (Exercise: Explain why.)

Type # 2: Regression with Only Time Fixed Effects. For a regression with only time fixed effects, we write the population regression as follows

$$vio_{it} = \beta_1 guns_{it} + \lambda_t + u_{it}$$

where, as we've discussed in the previous section of these notes, i denotes the entity, and t denotes time.

(a) What is λ_t?

Answer: Here λ_t is the time fixed effect. Note that it does not have a subscript i because it is the same for all entities (here, states).

(b) What does λ_t capture?

Answer: λ_t captures the effect of year t on vio_{it}; it **varies over time but not across states**. In our example, the time fixed effect would contain time trends that influence the violent crime, but do not vary across states. Examples might include the national unemployment rate, national GDP, and other national, US-wide economic trends.

(c) In practice, how do we estimate the above regression?

Answer: As before, in practice, one way we can estimate the above regression is by adding a dummy variable for each time period. In this case, the data goes from 1977 to 1999, so there are 23 years in the data. Again, because of the **dummy variable trap,** if we were to estimate the above regression with an intercept, we should include only 22 year dummies in the regression.

Type # 3: Regression with Both Entity and Time Fixed Effects. For a regression with both entity and time fixed effects, we write the population regression as follows

$$vio_{it} = \beta_1 guns_{it} + \alpha_i + \lambda_t + u_{it}$$

where, as we've discussed in the previous section of these notes, i denotes the entity, and t denotes time.

(a) What are state and time fixed effects able to control for? What are they **not** able to control for?

Answer: In the above regression where we have both state and time fixed effects, we are controlling for both **state-specific characteristics that do not change over time** (these are the α_i) and **factors that vary over time but not across entities** (these are the λ_t). However, these fixed effects do not control for **state-specific characteristics** that **vary over time**, which can still be a source of omitted variable bias.

(b) What are possible sources of OVB in this regression?

Answer: An example of an omitted variable could be state GDP per capita (a measure of wealth, which varies over time across states). Note that state GDP per capita would be contained in u_{it}. For state GDP per capita to be cause omitted variable bias, we would need it to be correlated with both violent crime (which is possible if during a recession, people don't have jobs and turn to crime to get money) and gun ownership laws (which is possible if as people become richer, they want to buy more guns).

(c) What does serial correlation mean and how does it arise in the above regression? Is serial correlation a threat to internal validity? If so, explain how we can address this issue.

Answer: In panel data, the regression error u_{it} can be correlated over time within a particular entity. This is called serial correlation. The intuition is that what happens in one year tends to be correlated with what happens the next year, within an entity. In our example, u_{it} can be serially correlated; i.e., the u in one year is correlated with the u in another year. An example of why serial correlation might occur here is because u_{it} contains state GDP per capita (as explained above), and if there is a recession, state GDP per capita this year would be correlated with next year's state GDP per capita. Serial correlation is a problem because it is a threat to internal validity; to address this issue, we can use **clustered standard errors**. Here, our "clusters" would be states.

2.3 Review the Concepts Exercises, Chapter 10

1. Why is it necessary to use two subscripts, i and t, to describe panel data? What does i refer to? What does t refer to?

   Answer: Panel data consists of entities (e.g., states) over a period of time (e.g., every year from year 2000 to 2012). Therefore, we need two subscripts, one to denote a particular entity (e.g., the state of California) and another to denote the particular point in time (e.g., year 2004). Typically, i refers to entities and t refers to time.

2. A researcher is using a panel data set on $n = 1000$ workers over $T = 10$ years (from 2001 to 2010) that contains the workers' earnings, gender, education, and age. Assume that the education variable is changing over time for at least some individuals in the sample (for example, because some individuals return to school). The researcher is interested in the effect of education on earnings. Give some examples of unobserved person-specific variable that are correlated with both education and earnings. Can you think of examples of time-specific variables that might be correlated with education and earnings? How would you control for these person-specific and time-specific effects in a panel data regression?

   Answer: An example of a person-specific variables that is correlated with both education and earnings is "ability". Specifically, it may be the case that people who are of high ability choose to get obtain more years of schooling (e.g., get a master's degree). At the same time, ability would also be correlated with wages "ability" is a characteristic that employers value. An example of a time-specific variable that might be correlated with education and earnings is GDP (or other indicator of nation-wide economic conditions). For example, during a recession, economic conditions are bad so earnings are also low. Furthermore, economic conditions and education are likely to be correlated since when economic conditions are good, college enrollment goes down, for example because high school graduates can easily find good jobs (so in turn they choose to work instead of going to college). We can control for person-specific and time-specific effects in a panel data regression by adding both person fixed effects and time fixed effects.

3. Can the regression that you suggested in response to Question 2 above be used to estimate the effect of gender on an individual's earnings? Can that regression be used to estimate the effect of the national unemployment rate on an individual's earnings? Explain.

   Answer: No for both questions. Note that if a regression includes person fixed effects, it absorbs all individual characteristics that are not changing over time. One of these characteristics is gender, so its effects on earnings cannot be determined separately from the person fixed effect. Similarly, if the regression includes time fixed effects, it absorbs all trends over time that that do not vary across individuals. The national unemployment rate is the same for all individuals at any point in time, so it is absorbed in the time fixed effect, and cannot be determine desperately from the time fixed effect.

4. In the context of the regression you suggested in Question 2, explain why the regression error for a given individual might be serially correlated.

   Answer: Serial correlation means that for a given person, the error terms are correlated over time, e.g. u_{it} is correlated with $u_{i,t+1}$ or u_{it} is correlated with $u_{i,t-5}$. Further, note that some of the factors contained in u are those characteristics specific to the individual that affect wages and are changing over time. (Exercise: Can you give examples of factors that are contained in u?)

   One example of a factor contained in u that leads to serial correlation is the following. Note that u contains information about the individual's employer; these characteristics may be changing over time, e.g., if a person switches jobs, so it is not absorbed in the individual fixed effect. Now, suppose that between 2001-2005 the individual works at firm A, then from the 2006-2010, the individual switched to a new job and worked for firm B. Then, note that for this individual, the error term for years 2001-2005 (which contains employer characteristics) will be correlated with each other, and the error term from 2006-2010 (which again contains employer characteristics) will also be correlated with each other.

3 Midterm 2 Practice Problems/Review

MT2, Fall 2013, Question 3. You are someone who loves wine seeking to use your knowledge of econometrics to understand how wine prices are determined. You hypothesize that the current market price of fancy wine is affected by the weather during the growing season just before the grapes are harvested. Accordingly, you obtain data on bottles of wine harvested between 1970 and 2000 and sold in auction *this year*, and hope to estimate the following equation:

$$Price_i = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + \beta_3 RainfallHarv_i + \beta_5 Vintage_i + u_i,$$

where $i = 1, \ldots, n$ indexes bottles of wine sold this year and the variables are defined as follows:

Variable	Description
Age	The age of the bottle of wine (this year minus harvest year)
$Temp$	The average temperature during the growing season in the area where the grapes were grown
$RainfallHarv$	The amount of rainfall that August through September (when the grapes were harvested)
$Vintage$	The year that the grapes were harvested

(a) Are any of the OLS assumptions going to be violated with certainty? If so, which one? Explain.

Answer: Violates no perfect multicollinearity assumption, since $Age = 2016 - Vintage$ where 2016 is the current year.

(b) Unfortunately, the rainfall data you want were not collected in the 1970s for some of the vineyards. You suspect that vineyards producing low-end wines were the last to start collecting weather data. You must decide whether to: (i) run the full regression on a subset of the wines for which rainfall data is available even for the 1970s, or (ii) run a regression without the rainfall variables, but using all of the wines in your sample. Briefly discuss the trade-offs to adopting strategy (i) versus strategy (ii) [Hint: you may want to list the cost(s) and benefit(s) of each strategy.]

Answer: If we choose option (i) instead of option (ii), the pros and cons are as follows (note that answers may vary depending on your argument).

- Pro: Excluding rainfall might cause OVB if rainfall affects prices and is also correlated with the regressors (e.g., if rainfall is correlated with temperature)

- Pro: Excluding rainfall might decrease the R^2 of the model. We might care about having a high R^2 if we want to use the model for forecasting.
- Con: The regression will be estimated on a subsample that is not randomly selected, because low-end wines were the last to start collecting weather data according to the question prompt. This is a problem if we want to get estimates that are meaningful for the average wine, but low-end wines are priced differently from the rest.
- Con: If we loose too much data, we would have less precise estimates.

(c) You decide to estimate the following, shorter regression on your entire sample:

$$Price_i = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + u_i,$$

but you would like to know by what *percent* the price changes when age changes by a year, all else equal. Select the statement that is *most* correct:

(i) You should regress $Price_i = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + u_i$.

(ii) You should regress $\log(Price_i) = \beta_0 + \beta_1 Age_i + \beta_2 Temp_i + u_i$.

(iii) You should regress $Price_i = \beta_0 + \beta_1 \log(Age_i) + \beta_2 Temp_i + u_i$.

(iv) You should regress $\log(Price_i) = \beta_0 + \beta_1 \log(Age_i) + \beta_2 Temp_i + u_i$.

(v) You should run a different regression that is not specified in another option.

Answer: The answer is (ii). Option (i) is incorrect because here, a 1 unit change in age is associated with a β_1 change in price, holding all else constant. Option (iii) is incorrect because here, a 1% change in age is associated with a $0.01\beta_1$ change in price, holding all else constant. Option (iv) is incorrect because here, a 1% change in age is associated with a 1% change in price, holding all else constant.

(d) Discuss the external validity of your study with respect to California wines.

Answer: Note that answers will vary depending on your argument/assumptions.

We first notice that the question does not specify the geographic location of the vineyards for which we have data, so whether the study is externally valid for CA is uncertain.

Next, recall that the two main threats to external validity are differences in population and differences in setting, so we need to consider these differences for external validity. For example:

- Differences in population: The variety of grapes used is a factor that affects the types of wine in our sample, and it might differ in CA and other regions. Also, the data are from wines sold at auction this year; are these wines comparable to CA wines?
- Differences in setting: The barrel in which the wine is stored, the type of soil, weather conditions, etc. are all factors about the setting that might be different in our sample vs. CA and other regions.

Final Exam, Fall 2014, Question 4. You are interested in learning how the value of cars evolves as they grow older. You obtain cross sectional data on a large number of cars, and for each vehicle, i, you observe the dollar value (Y_i), the age in years (X_{1i}) and the mileage in miles (X_{2i}). You also observe the make and model of each vehicle, (e.g. "Ford Focus"), and you generate a set of dummy variables, $D_i^1, \ldots, D_i^m$, such that D_i^1 equals 1 if vehicle i is of make and model #1 (e.g. "Ford Focus") and 0 otherwise, and similarly for all makes and models $1, \ldots, m$.

(a) You estimate that $\hat{Y}_i = 30,000 - 3,000 \cdot X_{1i} + 80 \cdot X_{1i}^2$. Based on the estimate, what is the average marginal effect of vehicle age on dollar value when a car is brand new? When it is 10 years old?

> **Answer:** Since the model is characterized by a quadratic specification with respect to age, the average marginal effect of age on car value depends on age itself. Specifically, it is given by the function $AME(X_{1i}) = \frac{\partial Y_i}{\partial X_{1i}} = -3,000 + 160 \cdot X_{1i}$. Hence, it will be $AME(0) = -3,000$ for a brand new car and $AME(10) = -1,400$ for a ten years old car.

(b) Next, you run the regression $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{2i} + u_i$. Interpret the coefficient β_3.

> **Answer:** It is the marginal effect on car value of an additional mile of mileage, conditional on age staying constant. Formally: $\beta_3 = E[\frac{\partial Y_i}{\partial X_{2i}}|X_{1i}]$.

(c) Next, you consider the regression

$$Y_i = \sum_{j=1}^{m} \alpha_j D_i^j + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{2i} + u_i. \tag{1}$$

What is the interpretation of the coefficient α_1?

> **Answer:** Coefficient α_1 is the expected value of a vehicle that: (1) belongs to model #1 (e.g. "Ford Focus"), (2) is brand new ($X_{1i} = 0$) and (3) has zero mileage ($X_{2i} = 0$).

(d) Assume that for each make and model you observe a large number of vehicles. Write a specification such that the value of cars can evolve *differently* with age for each make and model. Using the regression equation you specified what will be the value of a 5 year old car with 60,000 miles be if it is of make and model #1? And if it is of make and model #2? State your answers as functions of the regression coefficients.

> **Answer:** An appropriate model would read as follows:
>
> $$Y_i = \sum_{j=1}^{m} \left(\alpha_j D_i^j + \beta_{1j} X_{1i} \cdot D_i^j + \beta_{2j} X_{1i}^2 \cdot D_i^j \right) + \beta_3 X_{2i} + u_i. \tag{2}$$
>
> where is included a full set of $2m$ parameters to represent the interaction effect of car age with every group's dummy (so to fit a separate quadratic effect of age for each model of car). Each pair of parameters is denoted as (β_{1j}, β_{2j}) for each model j, with $j = 1, \ldots, m$.
>
> The expected "value" (Y_i) of a 5 year old car with 60,000 miles belonging to group 1 is given by $\alpha_1 + 5\beta_{11} + 25\beta_{21} + 60,000\beta_3$, while a car belonging to group 2 has expected value $\alpha_2 + 5\beta_{12} + 25\beta_{22} + 60,000\beta_3$.

(e) Suppose you obtain panel data, so that you observe each vehicle, its age and its mileage at multiple times, $t = 1, \ldots, T$. With panel data, regression (1) becomes

$$Y_{it} = \sum_{j=1}^{m} \alpha_j D_i^j + \beta_1 X_{1it} + \beta_2 X_{1it}^2 + \beta_3 X_{2it} + u_{it}, \tag{3}$$

where t indexes the time of observation. If you add *vehicle fixed effects* to regression (3), can you still include mileage (X_{2it}) as a regressor? Can you still include $D_i^1, \ldots, D_i^m$ as regressors? Why or why not?

Answer: One can still include mileage in the regression, given that it is a variable that has both a cross-sectional and a time variation (that is, it varies over time for the same vehicle as well as across vehicles at the same time). However, it is impossible to also include model-specific dummy variables, given that each of them would be perfectly collinear with the set of vehicle-fixed effects of that model's vehicles.