General pointers on writing an empirical social science paper

- Organize ideas before writing
- Define important terms and use terms consistently
- State hypothesis/goal/research question
- Describe methods
- Interpret results (in addition to stating them)
- Avoid excess verbiage
- Have others read and critique
- Edit, edit, edit

General pointers on writing an empirical social science paper

- Organize ideas before writing
- Define important terms and use terms consistently
- State hypothesis/goal/research question
- Describe methods
- Interpret results (in addition to stating them)
- Avoid excess verbiage
- Have others read and critique
- Edit, edit, edit

Will go through each of these sections in turn and tell you what elements each should contain. (Which sections you actually write will depend on your paper.)

- Introduction
- Data section
- Methods and Results
- (Conclusion)

Will illustrate elements of each of these sections with examples drawn from four papers (that I happen to know well):

- "Diversity, Social Goods Provision, and Performance in the Firm," *JEMS* 2014
- "Strategic Entry Deterrence and the Behavior of Pharmaceutical Incumbents Prior to Patent Expiration," *AEJMicro*, 2011
- "Countervailing Power in Wholesale Pharmaceuticals," *JIE*, 2010
- "Search, Obfuscation, and Price Elasticities on the Internet," *EMA*, 2009

Introduction

- Motivate why topic interesting/important
- Clearly state research question and describe (in non-technical terms) how you will analyze
- Describe what others have done and how what you're doing fits in
- Foreshadow results

Introduction

focus on this one

- Motivate why topic interesting/important
- Clearly state research question and describe (in non-technical terms) how you will analyze
- Describe what others have done and how what you're doing fits in
- Foreshadow results

Introduction

- Motivate why topic interesting/important

GALBRAITH [1952] SUGGESTED THAT large buyers have an advantage in extracting price concessions from suppliers. He called this effect the *countervailing power* of large buyers because he foresaw it as countervailing the market power of large suppliers. It has long been the conventional wisdom in the business press that such buyer-size effects exist.[1] Recently,

While we may have an academic interest in the sources of countervailing power and may have been fortunate that the pharmaceutical industry provides a good setting to explore the academic question, our findings about countervailing power in pharmaceuticals may have policy implications as well. In particular, they could shed light on the likely success of large healthcare procurement alliances.[4] If size alone is not sufficient for

Introduction

- Motivate why topic interesting/important

Legal and societal shifts in 20th Century America laid the groundwork for increased diversity in many settings. Schools ceased to be racially segregated. Barriers to female and minority employment diminished, leading to more diverse workplaces. Greater mobility resulted in neighborhoods fragmented in various dimensions. The consequent social benefits of this increased diversity, though difficult to quantify, may be quite important. With these broad social changes as a backdrop, the focus of this paper is

WHEN INTERNET COMMERCE first emerged, one heard a lot about the promise of "frictionless commerce." Search technologies would have a dramatic effect by making it easy for consumers to compare prices at online and offline merchants. This paper examines an environment where Internet price search plays a dominant role: small firms selling computer parts through Pricewatch.com. A primary observation is that the effect of the Internet on search frictions is not so clear-cut: advances in search technology are accompanied by investments by firms in obfuscation.

Introduction

- Clearly state research question and describe (in non-technical terms) how you will analyze

In this paper we test whether the implication from the theoretical papers cited in the previous paragraph–supplier competition is required for buyer-size discounts to emerge–holds in the pharmaceutical industry. Our data on

....

freely substitute among the competing generics. Using these sources of variation, we can identify instances in which buyers have good substitution opportunities, large size, both, or neither, and can therefore isolate the effect that each has on purchase price.

Introduction

- Clearly state research question and describe (in non-technical terms) how you will analyze

important. With these broad social changes as a backdrop, the focus of this paper is smaller but sharper. We are interested in how diversity in a group affects the provision of social goods in the group, and then, ultimately, performance of that group. In particular, we focus attention on diversity in a market environment, that created by a firm and its workforce. Regardless of the cause of the increased workplace diversity, it is the job of the managers to encourage the greatest productivity possible from their units, maximizing profits, perhaps, or some other quantifiable outcome. It is our goal, then, to shed light on how diversity is associated with those outcomes.

Introduction

- Clearly state research question and describe (in non-technical terms) how you will analyze

partial answer

We have a unique data set from a firm that operates numerous small offices in the United States and abroad. They have provided us with eight years of individual-level employee survey data as well as office-level measures of diversity and performance. The survey data furnish us with several indicators of firm social capital, such the level of cooperation among employees. The data allow us to address two distinct questions. First, broadly speaking, do we find lower levels of social goods provision in more diverse offices? Such a finding has been made in the economics literature cited above, but our results provide an interesting complement to those: economists have previously focused on the effects of diversity in communities instead of workplaces,[3] and we measure diversity on two dimensions not explored in this literature, gender and tenure.

Introduction

- Describe what others have done and how what you're doing fits in

Economists' interest in the effect of diversity on the provision of social goods is ongoing. Studies have found evidence that social goods are provided at a lower level in communities exhibiting fragmentation on various dimensions. Vigdor (2004) finds that census response rates are lower in census tracts with higher ethnic fragmentation. Costa and Kahn (2003) find that desertion rates are higher in Civil War military companies with higher age and occupational fragmentation. Glaeser et al. (2000) find that trust is lower among Harvard undergraduates when race and nationality fragmentation is higher. Several studies have documented that school funding is higher in more homogenous communities (see, e.g., Goldin and Katz, 1999; Poterba, 1997; Miguel and Gugerty, 2005).[1] These results may be quite important in contexts where social goods provision is the output of interest. However, in some contexts, the social good may be an "intermediate good." In the workplace, cooperation, trust, and other social goods may be important elements of the functioning of an office, but firm owners ultimately care about an office's performance, as reflected in revenues, costs, and profits. We would also like to address this additional question, so often missing in the economics literature.

Introduction

- Foreshadow results

Our results show that large buyers (chain drugstores) receive no discount relative to small buyers (independent drugstores) on antibiotics with unexpired patents–antibiotics for which drugstores have no substitution opportunities and thus effectively face monopoly suppliers. For off-patent antibiotics–antibiotics for which drugstores have some substitution opportunites–chain drugstores receive a statistically significant but small discount relative to independents, at most 2%.

Introduction

- Foreshadow results

Our first empirical result is a striking confirmation that price search technologies can dramatically reduce search frictions. We estimate that the firm faces a demand elasticity of −20 or more for its lowest quality memory modules!

Our second main empirical result is a contribution to the empirics of loss leaders. We show that charging a low price for a low-quality product increases our retailer's sales of medium- and high-quality products. Intuitively, this happens because one cannot ask a search engine to find "decent-quality memory module sold with reasonable shipping, return, warranty, and other terms." Hence, many consumers use Pricewatch to do what it is good at—finding websites that offer the lowest prices for *any* memory module—and then search within a few of these websites to find products that better fit their preferences.

Data

- Cite sources for data
- Describe any special data treatments you've performed
- Describe structure of data set (# obs, level of observation, period of time covered, etc.)
- Present variable definitions
- Present summary statistics

Probably need tables

- Discuss interesting facts, observations, shortcomings

Data

- Cite sources for data

Our data were provided by a professional services firm that operates over sixty offices in the United States and abroad. The offices range in size from just a few employees to nearly 100 at their headquarters. They administered anonymous employee satisfaction surveys approximately annually from 1995 to 2002. Table I contains summary statistics on the variables we created with these data, which we describe below.

Our dataset, collected by the pharmaceutical-marketing-research firm IMS America, covers virtually all prescription antibiotics sold in the United States from January 1992 to August 1996.[16] It includes nationwide quantities and revenues from wholesale transactions between manufacturers/distributors and retailers each month.

Data

- Describe any special data treatments you've performed

Our primary measure of the degree to which an incumbent has engaged in presentation proliferation, *PresHerf*, is a Herfindahl-style measure that is also constructed from the presentation-level revenue data. Specifically, we define $PresHerf_{it} = w_i \sum_k z_{idkt}^2 + (1 - w_i) \sum_k z_{ihkt}^2$, where w_i is the fraction of the sales of drug i which are made through drugstores and z_{idkt} and z_{ihkt} are the fractions of drug i's revenues in year t in the drugstore and hospital markets, respectively, which are accounted for by presentation k.[37] *PresHerf* will be large in markets where a small number of

For tenure diversity, we calculated the standard deviation of tenure for each office, and then divided by the number of employees in the office. Finally, we scaled the expression linearly so that the measure takes on values of 0 for offices where everyone has worked for the firm the same amount of time and positive values for offices with some variance in the amount of time the employees have worked there, 1 being an upper bound in our data set. Note that the scaling for both of these diversity variables is arbitrary, but choosing a scale on the unit interval aided interpretation.

Leave clues in the text---other crucial information will be contained in the tables.

Data

- Describe structure of data set (# obs, level of observation, period of time covered, etc.)

Our data were provided by a professional services firm that operates over sixty offices in the United States and abroad. The offices range in size from just a few employees to nearly 100 at their headquarters. They administered anonymous employee satisfaction surveys approximately annually from 1995 to 2002. Table I contains summary statistics on the variables we created with these data, which we describe below.

From the survey responses, we can identify the office, gender, and tenure of the individual employees, enabling us to create office-level measures of diversity in those two dimensions. For gender, we calculated the standard deviation of a dummy variable for male for each office and scaled it linearly to fall into [0, 1], where 0 indicates an all-male or all-female office and 1 is an office evenly divided. This variable is called *GendDiversity*. In our data, the minimum value is 0 and the maximum is 1. Note that this firm employs more women than men, and that we have both male-dominated and female-dominated offices among our observations where *GendDiversity* is near 0.

Data

- Present variable definitions
- Present summary statistics

Probably need tables

TABLE II.
SUMMARY STATISTICS, CITY-LEVEL VARIABLES

Variable	Obs.	Mean	Std. Dev.	Min	Max
At the city level:					
CAvgAge	61	33.9	2.7	29.6	41.7
CPolitics	47	75.3	58.8	1	227
CPercMinority	64	43.6	20.5	2.0	89.5
CPercMale	64	48.8	1.2	46.5	51.4
COfficeRent in annual dollars per ft^2	59	42.15	37.10	15.60	197.80
CPopulation in thousands	67	1,462	1,818	81	8,008

Pretty self-explanatory names

Report all of these

Note that there's a lot of information in these two tables about structure of the data set

TABLE I.
SUMMARY STATISTICS

Variable	Obs.	Mean	Std. Dev.	Min	Max
At the employee level					
Satisfaction	1,707	3.943	0.990	1	5
Morale	1,683	3.592	1.017	1	5
Cooperate	1,541	4.038	1.036	1	5
Male	1,648	0.329	0.470	0	1
TenureYears	1,665	2.570	2.087	0.25	7
At the office-year level					
Unemploy	269	4.77	1.84	1.4	12.2
Number	269	4.94	3.12	2	19
AvgSatisfaction	269	4.06	0.58	2	5
AvgDPerception	269	4.73	0.36	3	5
AvgMorale	269	3.74	0.66	1	5
AvgCooperate	248	4.14	0.64	2	5
AvgGender	269	0.29	0.25	0	1
AvgTYears	269	2.32	1.14	0.25	6.25
GendDiversity	269	0.58	0.41	0	1
TenureDiversity	269	0.11	0.11	0	1
Revenues in thousands	269	3,794	3,660	3	23,900

Also note that these are not computer output!!!

Note that there's a lot of information in these two tables about structure of the data set

TABLE I.
SUMMARY STATISTICS

Variable	Obs.	Mean	Std. Dev.	Min	Max
At the employee level					
Satisfaction	1,707	3.943	0.990	1	5
Morale	1,683	3.592	1.017	1	5
Cooperate	1,541	4.038	1.036	1	5
Male	1,648	0.329	0.470	0	1
TenureYears	1,665	2.570	2.087	0.25	7
At the office-year level					
Unemploy	269	4.77	1.84	1.4	12.2
Number	269	4.94	3.12	2	19
AvgSatisfaction	269	4.06	0.58	2	5
AvgDPerception	269	4.73	0.36	3	5
AvgMorale	269	3.74	0.66	1	5
AvgCooperate	248	4.14	0.64	2	5
AvgGender	269	0.29	0.25	0	1
AvgTYears	269	2.32	1.14	0.25	6.25
GendDiversity	269	0.58	0.41	0	1
TenureDiversity	269	0.11	0.11	0	1
Revenues in thousands	269	3,794	3,660	3	23,900

And they do not include sci notation and a million significant digits.

Data---do this for the data-driven version

- Discuss interesting facts, observations, shortcomings.
  - Show us correlation tables
  - Plot histograms of variables
  - Create XY plots of pairs of variables
  - Test whether two sets of observations are from the same distribution
  - Exhibit the truncation or top-coding of a variable
  - Be creative, but selective---include what is interesting or surprising or important and tell us why

Data---do this for the data-driven version

- Discuss interesting facts, observations, shortcomings.
  - Show us correlation tables
  - Plot histograms of variables
  - Create XY plots of pairs of variables
  - Test whether two sets of observations are from the same distribution
  - Exhibit the truncation or top-coding of a variable
  - Be creative, but selective---include what is interesting or surprising or important and tell us why

Don't forget this part!!

Methods and Results

- Reiterate objective of paper
- Describe method
- Present results in table
- Describe and interpret results in text

Methods and Results

- Reiterate objective of paper

Our theory discussion provided guidance on an empirical strategy, and, in particular, suggested two estimating equations, one where cooperative effort is a function of diversity and a second where output is a function of diversity. We now address the first of these and turn to our results on social capital within the office. We use our employee-level data and focus on explaining perceived levels of cooperation. Most particularly, we will be interested in measures of diversity as explanatory variables.

Note how this language is drawing a clear connection between the paper's objective and the regressions we are about to describe.
A good results section should draw that connection repeatedly.

Methods and Results

- Describe method

The dependent variable in all our regressions is a difference in log price. For example, in the regression comparing chain versus independent drugstores discussed in Section IV(i), the dependent variable is

$$(1) \qquad \Delta^{CI}_{i,j,m,t} = \ln\left(PRICE^{C}_{i,j,m,t}\right) - \ln\left(PRICE^{I}_{i,j,m,t}\right),$$

where $PRICE^{C}_{i,j,m,t}$ is the average wholesale price in month t paid by chain drugstores for drug i in presentation j produced by manufacturer m and $PRICE^{I}_{i,j,m,t}$ is that price paid by independent drugstores. In additional regressions using hospital and HMO data discussed in Section IV(ii), we introduce analogous dependent variables, Δ^{HD}, Δ^{OD}, Δ^{HO}, where D denotes all drugstores, H denotes hospitals, and O denotes HMO's. This differenced specification has several advantages, providing readily interpretable coefficients and accounting for drug, presentation, manufacturer and time fixed effects, as well as their interactions.

Different specifications arranged in columns

No scientific notation!!!

Methods and Results

- Present results in table

Significance indicated by both SEs and asterisks

TABLE IV
WEIGHTED LEAST SQUARES REGRESSIONS OF THE DIFFERENCE IN LOG PRICE

	Δ^{CI}	Δ^{HD}	Δ^{OD}	Δ^{HO}	Δ^{HC}
ONPAT	0.002	−0.077***	−0.079***	0.015	−0.071***
	(0.001)	(0.017)	(0.018)	(0.016)	(0.015)
OFFPAT × *BRANDED*	−0.003**	−0.328***	−0.205***	−0.043	−0.288***
	(0.002)	(0.059)	(0.055)	(0.040)	(0.053)
OFFPAT × *GENERIC* × *ONEGEN*	−0.017*	−0.151***	−0.121**	−0.005	−0.124***
	(0.010)	(0.057)	(0.051)	(0.018)	(0.044)
OFFPAT × *GENERIC* × *MULTGEN*	−0.003	−0.145***	−0.043**	−0.054***	−0.117***
	(0.002)	(0.020)	(0.017)	(0.015)	(0.017)
R^2	0.0003	0.0355	0.0209	0.0038	0.0351
Observations	107,164	107,287	71,463	69,644	93,181
Manufacturer-Drug Clusters	791	740	630	588	694

Notes: For each observation, the weight in the weighted least squares estimation procedure is the natural logarithm of the sum of revenue in the two relevant channels. An exhaustive set of dummies is included in each regression and the constant term omitted. White [1980] heteroskedasticity-robust standard errors reported in parentheses below coefficient estimates. Standard errors are adjusted for non-independence within manufacturer-drug clusters. Significantly different from zero in a two-tailed t-test with degrees of freedom equal to the number of unique manufacturer-drug clusters minus one at the *10% level; **5% level; ***1% level.

Methods and Results

- Describe and interpret results in text

The first column of Table IV presents results from our main regression. It examines the difference in prices paid by chain drugstores and independent drugstores in different supply regimes and, as such, provides a fairly clean test of the importance of size in those different supply regimes. Chain drugstores and independent drugstores should not differ in their substitution opportunities: they cannot be restrictive against on-patent brand-name drugs and can only be slightly restrictive against off-patent brand-name drugs or single-source generics, but both can be restrictive against multiple source generics. The only difference is that chains will tend to be larger-volume buyers than independents. Recall that the dependent variable is the log of the difference between prices paid by chain and independent drugstores. Therefore, the interpretation of the 0.002 coefficient estimate on *ONPAT* is that chain drugstores pay 0.2% more for branded on-patent drugs than independent drugstores do. The coefficient is not significantly different from zero, and given its small standard error, one should conclude that the effect is precisely estimated to be about zero.

Methods and Results

- Describe and interpret results in text

The first column of Table IV presents results from our main regression. It examines the difference in prices paid by chain drugstores and independent drugstores in different supply regimes and, as such, provides a fairly clean test of the importance of size in those different supply regimes. Chain drugstores and independent drugstores should not differ in their substitution opportunities: they cannot be restrictive against on-patent brand-name drugs and can only be slightly restrictive against off-patent brand-name drugs or single-source generics, but both can be restrictive against multiple source generics. The only difference is that chains will tend to be larger-volume buyers than independents. Recall that the dependent variable is the log of the difference between prices paid by chain and independent drugstores. Therefore, the interpretation of the 0.002 coefficient estimate on *ONPAT* is that chain drugstores pay 0.2% more for branded on-patent drugs than independent drugstores do. The coefficient is not significantly different from zero, and given its small standard error, one should conclude that the effect is precisely estimated to be about zero.

Methods and Results

- Describe and interpret results in text

This result taken on its own suggests that the variants of theoretical models in which buyer-size discounts emerge with a monopoly supplier may not be relevant in our market. In the variant of symmetric-information bargaining models (see Horn and Wolinsky [1988b]; Stole and Zwiebel [1996]; Chipty and Snyder [1999]; Raskovich [2003]; Segal [2003]; Adilov and Alexander [2006]; Inderst and Wey [2007]; and Normann, Ruffle and Snyder [2007]), large-buyer discounts emerge with a monopoly supplier if the bargaining-surplus function is concave. Lott and Roberts [1991] and Levy [1999] suggest that large-buyer discounts may be a pass-through of cost savings from lower per-unit warehousing and distribution costs. Such cost savings would also be passed through by a monopoly supplier. We find that large buyers are not able to extract any discount at all in our market in the face of a monopoly supplier.

Again, this language is drawing a clear connection between the paper's objective and the regression results---it is not enough to just repeat the numbers from the table.

Conclusion---probably not necessary for you

- Varies greatly by writer. If you do include one, it should be viewed as an opportunity to make your final case---this is the last thing that the reader reads.
  - —Reiterate results
  - —Discuss policy implications
  - —Discuss future research
  - —Tie loose ends together

Additional comments

- Structure and style of papers can and do deviate from what I have presented. If you have good reasons, deviating is fine, but essential elements should not be omitted.
- My example papers were 20-35 pages long---yours will be shorter.
- Clarity is always our primary goal---nothing else can be achieved without clarity.

Lecture 16: Running regressions–Practical issues

Prof. Esther Duflo

14.310x

Practical issues with regression

- Reading a regression output
- Dummy Variables
- Other Functional Form issues

Reading a regression output in R

```
> mod1 <-lm(ftvoteshare~fncandidates)
> summary(mod1)

Call:
lm(formula = ftvoteshare ~ fncandidates)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46077 -0.05541 -0.01163 -0.01163  0.85447

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.011631   0.009869   1.179                0.239
fncandidates 0.133901   0.006533  20.496 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1528 on 370 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.5317,    Adjusted R-squared:  0.5304
F-statistic: 420.1 on 1 and 370 DF,  p-value: < 0.00000000000000022
```

What R gives you when you type summary(...)

- Intercept
- Coefficients and standard errors of estimated coefficients
- The T statistics: testing the hypothesis that each coefficient is zero.
- And the associated p value, and even stars if you cannot read p value!
- The R squared
- The F stat of the regression: test the hypothesis that all coefficients are zero

Doing more with the model

- The coefficients are stored in a vector called "coef"
- The variance covariance matrix of the coefficient is a matrix called "vcov"
- You can use the "hypothesistesting" in the library "car" to test any linear hypothesis of the form $R\beta = 0$ as seen in last lecture
- you can export the coefficients and the standard errors in a data frame (or use the "stargazer" package to do much better than that... we will see it later hopefully)
- you can visualize prediction and residuals (see R code)

Dummy Variables

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

D_i is a dummy variable , or an indicator variable, if it takes the value 1 if the observation is in group A, and 0 if in group B. Example:

- RCT: 1 if in treatment group , 0 otherwise
- 1 if male, 0 if female
- 1 before great depression, 0 after
- 1 before generic substitution act passed, 0 otherwise,
- 1 if the house has a deck in the backyard, 0 otherwise,

Interpretation

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

Without any control variable, it is easy to verify that $\widehat{\beta} = \overline{Y_A} - \overline{Y_B}$. So you can always estimate the difference between treatment and control group for an RCT using an OLS regression framework.

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity.
- So what do we do?

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity.
- So what do we do?
- We typically omit ONE group (if we don't do it, R may do it for us), and then what is the interpretation of each coefficient?

From a categorical variable to dummy variables

- What if you don't have two groups, but, say, 50 (e.g. 50 states): Your original variable is takes discrete values 1 to 50.
- It usually does not make much sense to include it directly as a regressor
- Transform it into 50 dummy variables: for each state, the dummy $= 1$ if the observation is from that state, and 0 otherwise.
- Careful, what happens if you introduce all of them and the constant?
- R will complain about multi-colinearity.
- So what do we do?
- We typically omit ONE group (if we don't do it, R may do it for us), and then what is the interpretation of each coefficient?
- It is the difference between the value of this group and the value for the omitted (reference) group.

with other variables in the regression

With other variables in the regression

$$Y_i = \alpha + \beta D_i + X_i \gamma + \epsilon_i$$

In that case β is the difference in intercept between group A and group B. This is the most frequent way that RCT are analyzed: the matrix X are "control" variables: things that did not affect the assignment but may have been different at baseline.

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$:

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$ An estimate of the difference between the treatment effect for males and for female. [we call this the interaction effect]

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$ An estimate of the difference between the treatment effect for males and for female. [we call this the interaction effect]

How do you obtain, for example, an estimate of the mean for males?

Dummy variables and Interactions

Now imagine you have two sets of dummy variables, say, Treatment and control, and Male and Female.
You can run:

$$Y_i = \alpha + \beta D_i + \gamma M_i + \delta M_i * D_i + \epsilon_i$$

How do we interpret these coefficients:

- $\widehat{\alpha}$: An estimate of mean for women in the control group
- $\widehat{\beta}$: An estimate of the difference between the treatment and control group means for women [we call this the treatment main effect]
- $\widehat{\gamma}$: An estimate of the difference between Males and Females. [we call this the gender main effect]
- $\widehat{\delta}$ An estimate of the difference between the treatment effect for males and for female. [we call this the interaction effect]

How do you obtain, for example, an estimate of the mean for males?
How do you obtain an estimate of the treatment effect for males?

Interaction of a dummy variable and a continuous variable

Practical issues with regression

- Dummy Variables
- Other Functional Form issues

Other functional form issues

- Transforming the dependent variable
- Non linear transformations of the independent variables

Transformations of the dependent variable

- Suppose $Y_i = AX_{1i}^{\beta_1} X_{2i}^{\beta_2} e^{\epsilon_i}$ then run linear regression

$$log(Y_i) = \beta_0 + \beta_1 logX_{1i} + \beta_2 logX_{12} + \epsilon_i$$

  to estimate β_1 and β_2. Note that β_1 and β_2 are *elasticities*: when X_1 changes by 1%, Y changes by β_1% .

- Returns to education formulation

$$logY_i = \beta_0 + \beta_1 S_i + \epsilon_i$$

  When education increases by 1 year, wages increases by $\beta_1 \times$ %.

Transformations of the dependent variable

- Box Cox Transformation
  Suppose $Y_i = \frac{1}{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}$
  then run regression

$$\frac{1}{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- Discrete choice model
  Suppose

$$P_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\epsilon_i}}$$

  P_i is the percentage of individuals choosing a particular option (e.g. buying a particular car)
  then run regression:

$$Y_i = log(\frac{Pi}{1-P_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Non linear transformation of the independent variables

- When running a kernel regression as exploratory analysis we may realize that the relationship between two variables does not appear to be linear.
- Does it mean we cannot run OLS?

Non linear transformation of the independent variables

- When running a kernel regression as exploratory analysis we may realize that the relationship between two variables does not appear to be linear.
- Does it mean we cannot run OLS?
- No!
- We can use polynomial or other transformations of the data to represent non linearities
- or partition the range of X.

Polynomial models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_k X_{1i}^k + \epsilon_i$$

- You can chose straight polynomial, or series expansion, or orthogonal polynomials or whatever.
- If you assume that the model is known, this is just standard OLS. You may want to plot the curve, compute the derivative with respect to X at key points, etc.
- If you assume that the model is now known, this is a non-parametric method: you realize there is bias (because the shape is never quite perfect) and variance (as you add more Xs) and you promise to add more terms as the number of observation increases. This is called *series* regression.

Other non linear transformations

- Take log of X
- Interact the X, such as the slope of one depends on the level of another.
- Potentially lots of variables and their transformations... How to chose? This is where machine learning tools can become handy (more on that later!)

Using dummies for approximation

- Partition the range of X is interval, $X^0, \ldots X^J$
- Define the dummies as:
  $D_{1i} = I_{[X^0 \leq X_{1i} < X^1]}$
  $D_{2i} = I_{[X^1 \leq X_{1i} < X^2]} \cdots$

Using dummies for approximation

- Partition the range of X is interval, $X^0, \ldots X^J$
- Define the dummies as:
  $D_{1i} = I_{[X^0 \leq X_{1i} < X^1]}$
  $D_{2i} = I_{[X^1 \leq X_{1i} < X^2]} \cdots$
- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + \epsilon_i$$

(note no intercept. why?)

Using dummies for approximation

- Partition the range of X is interval, $X^0, \ldots X^J$
- Define the dummies as:
  $D_{1i} = I_{[X^0 \leq X_{1i} < X^1]}$
  $D_{2i} = I_{[X^1 \leq X_{1i} < X^2]} \cdots$
- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + \epsilon_i$$

  (note no intercept. why?)
- Define Piece wise linear variables as:
  $S_{1i} = I_{[X^0 \leq X_{1i} < X^1]}(X_{1i} - X^1)$ $S_{2i} = I_{[X^1 \leq X_{1i} < X^2]}(X_{1i} - X^2)$

Using dummies for approximation

- Partition the range of X is interval, $X^0, \ldots X^J$
- Define the dummies as:
  $D_{1i} = I_{[X^0 \leq X_{1i} < X^1]}$
  $D_{2i} = I_{[X^1 \leq X_{1i} < X^2]} \cdots$
- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + \epsilon_i$$

  (note no intercept. why?)
- Define Piece wise linear variables as:
  $S_{1i} = I_{[X^0 \leq X_{1i} < X^1]}(X_{1i} - X^1)$ $S_{2i} = I_{[X^1 \leq X_{1i} < X^2]}(X_{1i} - X^2)$
- Run regression

$$Y_i = \beta_1 X_{1i} + \beta_2 S_{1i} + \cdots + \beta_J S_{j-1i} + \epsilon_i$$

Locally Linear Regression

- What size of interval should we chose?
- This should by now sound very familiar: either you are willing to assume that you *know* the shape of the function: Then, just cut it as you know it is relevant.
- Or.... we are trying to guess the shape of the function
- And then we have the familiar bias/variance trade off: we are now in fact performing a non parametric regression technique knows a locally linear regression: around each point where we are interested in evaluating the function, we run a weighted regression of Y_i on X_i, where the weights will be given by a Kernel, for observations in a bandwidth. We take the predicted value from the regression as best predictor for Y_i. So it is exactly like a Kernel regression, but we use a linear regression in each little interval instead!
- Why on earth?

Locally Linear Regression

- What size of interval should we chose?
- This should by now sound very familiar: either you are willing to assume that you *know* the shape of the function: Then, just cut it as you know it is relevant.
- Or.... we are trying to guess the shape of the function
- And then we have the familiar bias/variance trade off: we are now in fact performing a non parametric regression technique knows a locally linear regression: around each point where we are interested in evaluating the function, we run a weighted regression of Y_i on X_i, where the weights will be given by a Kernel, for observations in a bandwidth. We take the predicted value from the regression as best predictor for Y_i. So it is exactly like a Kernel regression, but we use a linear regression in each little interval instead!
- Why on earth?
  - It has better properties (especially at the boundaries)
  - And the slope is often of interest
  - This is what R do when you specify the option "LOESS" (or