Problem Set 1
Predictive Analytics for Business Strategy
McDermott

*Be sure to include all group member names on submitted assignment file. **Josh Andreychuk, Dustin Boler, Thomas Kappler, Malcolm Moody, Nick Seefeld**

*Use this file and add your solution after each part of each question. Put your solution in **bold**.

1. We are interested in estimating demand for beer. We observe data on prices and quantities of 8 different beer brands i at 1000 different stores s. To start, suppose the true causal/exogenous model for sales is

$$\log Quantity_{is} = b_0 + b_1 \log price_{is} + b_2 Quality_i + U_{is}$$

We don't directly observe quality. Use the first worksheet in PS1.xlsx. You will also need to make some adjustments since the columns related to brand are strings (Stata will read them as letters and not as numbers). There are multiple ways of fixing this, but here is one method:

Step 1: Create a new column in Excel called something like brandNum. In the cells, use **=LEFT(B2,1)**. If you've learned Python, where you can take the first element in an array, you may recognize this as taking the first character in the B2 cell. This will still be read as a string so we will need another step.

Step 2: Import the spreadsheet into Stata. You can verify the type of variable brandNum is by typing des (or describe) brandNum. You will see that it is a string. In order to convert it, you can use the following code: destring brandNum, gen(brandNum2)

*Note when we regress log(Y) on log(X), the coefficient on log(X) tells us the percentage change in Y for a 1% increase in X (also known as an elasticity).

   a. Estimate by OLS a model of log quantity on log price (and nothing else). Interpret coefficient on log price. Is it statistically significant? Does its sign seem reasonable?

```
. reg log_quantity log_price
```

Source	SS	df	MS	Number of obs	=	8,000
				F(1, 7998)	=	20.51
Model	106.91905	1	106.91905	Prob > F	=	0.0000
Residual	41686.0621	7,998	5.21206078	R-squared	=	0.0026
				Adj R-squared	=	0.0024
Total	41792.9812	7,999	5.22477575	Root MSE	=	2.283

log_quantity	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
log_price	.6288727	.1388481	4.53	0.000	.3566943	.9010512
_cons	.6477426	.4203228	1.54	0.123	-.1761996	1.471685

By regressing log quantity on log price and evaluating the coefficient on log price, we estimate that a 1% increase in price will result in a .63% increase in quantity demanded. Log price is statistically significant because P = 0. This seems unintuitive because an increase in price normally results in a decrease in demand.

b. We don't observe quality, but we do observe brand. Would a set of brand dummy variables be a valid way to control for quality? Explain why or why not.

Yes, brand as a proxy for quality is a valid way to control for quality. Brand reputation is often closely tied to perceived quality. Consumers may associate certain brands with higher quality and might be willing to pay more for them, or buy them more frequently. By including brand dummy variables, you can account for the different levels of perceived quality across brands.

However, brand is not a direct measure of quality. While brand can be a proxy for quality, it's not a direct measure. There can be omitted variable bias. If there are other factors that influence both the brand and quantity sold that aren't included in your model, this could lead to omitted variable bias. For example, if certain brands are more likely to be sold in certain locations or during certain seasons, and you don't control for location or season, this could bias your results.

c. Regress log quantity on log price and the 8 brand dummy variables (use the "i." on the brand in Stata). Interpret the coefficient on log price. Is it statistically significant? Is it causal? Explain

```
. reg log_quantity log_price i.brandNum2
```

Source	SS	df	MS	Number of obs	=	8,000
				F(8, 7991)	=	36.22
Model	1462.51619	8	182.814524	Prob > F	=	0.0000
Residual	40330.465	7,991	5.04698598	R-squared	=	0.0350
				Adj R-squared	=	0.0340
Total	41792.9812	7,999	5.22477575	Root MSE	=	2.2465

log_quantity	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
log_price	-2.867728	.2536451	-11.31	0.000	-3.364939	-2.370517
brandNum2						
2	.0693167	.1004775	0.69	0.490	-.1276455	.2662789
3	.5572959	.1039594	5.36	0.000	.3535083	.7610834
4	.7647594	.107875	7.09	0.000	.5532962	.9762226
5	.0511954	.10047	0.51	0.610	-.1457519	.2481428
6	-.4990644	.1033812	-4.83	0.000	-.7017186	-.2964102
7	-1.193684	.1181543	-10.10	0.000	-1.425297	-.9620709
8	1.510911	.1271171	11.89	0.000	1.261728	1.760093
_cons	11.05557	.7623452	14.50	0.000	9.561176	12.54997

By regressing log quantity on log price and the 8 brand dummy variables (brand 1 is the omitted baseline) and evaluating the coefficient on log price, we estimate that a 1% increase in price will result in a 2.87% decrease in quantity demanded on average. This model is statistically significant because P = 0. Price is probably causal because we've controlled for differences in price due to brand, but we can't dismiss the possibility that other confounding factors have not been controlled for. Ultimately, this is a better model than controlling for log price alone because we've reduced the confounding factors.

2. In class, we looked at estimating a difference-in-differences model in which there was a clear before and after. In this problem set, you will also need to use fixed effects for states so that you can estimate a difference between states of the differences across time within states.
   You are interested in the impact of SundayHours (the number of hours a store is open on Sunday) on Profits, so that should always be included in your models. You also do not need to generate any new variables for this assignment. Also, when asked to "write down a model" you should only include variables that you have data on. Use the second worksheet in PS1.xlsx.

   a. Write down a fixed effects model that controls for state fixed effects but does not include any controls for time.

      reg Profits SundayHours i.State

b. Display and interpret the results of this first regression.

```
. reg Profits SundayHours i.State
```

Source	SS	df	MS	Number of obs	=	120
				F(12, 107)	=	2.37
Model	220901.317	12	18408.4431	Prob > F	=	0.0094
Residual	829354.224	107	7750.97406	R-squared	=	0.2103
				Adj R-squared	=	0.1218
Total	1050255.54	119	8825.67682	Root MSE	=	88.04

Profits	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
SundayHours	-33.68284	11.12633	-3.03	0.003	-55.73949	-11.62618
State						
2	17.72576	41.66468	0.43	0.671	-64.86962	100.3211
3	13.72473	39.46373	0.35	0.729	-64.50752	91.95698
4	32.66298	41.88333	0.78	0.437	-50.36583	115.6918
5	19.47548	40.1596	0.48	0.629	-60.13624	99.0872
6	78.80361	40.79637	1.93	0.056	-2.07044	159.6777
7	15.61714	40.81277	0.38	0.703	-65.28941	96.52369
8	12.05547	40.29256	0.30	0.765	-67.81984	91.93078
9	19.36114	40.3187	0.48	0.632	-60.56598	99.28825
10	-10.09591	39.53204	-0.26	0.799	-88.46357	68.27174
11	.1462017	40.196	0.00	0.997	-79.53769	79.83009
12	-39.08646	39.5536	-0.99	0.325	-117.4969	39.32393
_cons	928.495	64.70769	14.35	0.000	800.2196	1056.77

By regressing Profits on Sunday Hours and the 12 states (state 1 is the omitted baseline) and evaluating the coefficient on Sunday Hours, we estimate that an increase of 1 Sunday Hour of operation will result in an average $33.68 decrease in profits.

c. Can you think of any possible confounding factors in answer 2a.? If yes, explain why it is/they are (a) confounding factor(s).

If you only control for state fixed effects but not time fixed effects, you would be controlling for factors that are constant within each state over time, but not for factors that change over time. This could leave your estimates susceptible to several additional confounding factors:

1. National Trends: Any national trends that affect all states equally and change over time could confound your results. For example, if there is a national trend towards longer opening hours or changes in shopping behavior over time, this could be correlated with both `SundayHours` and `Profits`.
2. Economic Cycles: Broader economic cycles, such as recessions or periods of economic growth, can affect all states and change over time. These economic conditions could influence both `SundayHours` and `Profits`.

3. Policy Changes: If there are national-level policy changes over time, these could also confound your results. For instance, changes in national labor laws could affect both `SundayHours` and `Profits`.
4. Seasonality: Seasonal factors or specific events that change over time and affect all states, such as holiday shopping seasons, could also confound your results.
By including time fixed effects in your model (along with state fixed effects), you can control for all of these time-varying factors that affect all states equally. However, without time fixed effects, these could be potential sources of confounding.

d. Write down the model that you will estimate that would determine the effect of the number of hours a store is open on Sundays (you should include controls for entity and time).

reg Profits SundayHours i.State i.Period

e. Can you think of any possible confounding factors in answer 2d.? If yes, explain why it is/they are (a) confounding factor(s).

In a Difference-in-Differences (DiD) model with state and time fixed effects, you're controlling for factors that are constant within each state over time (state fixed effects) and factors that are constant across states but vary over time (time fixed effects). This helps to control for many potential confounding factors.

However, there are still potential confounding factors that this model does not control for:

1. Time-Varying State Factors: These are factors that vary both across states and over time. Examples could include changes in state-level policies, economic conditions, or population changes. If these factors are correlated with both `SundayHours` and `Profits`, they could confound your estimates.
2. Simultaneous Events: If other events coincide with the changes in `SundayHours` and also affect `Profits`, this could confound your results. For instance, if a change in business strategy leads to both longer `SundayHours` and higher `Profits`, it could be this strategy change, rather than the longer hours, that is driving the increase in profits.
3. Non-Parallel Trends: The DiD model assumes that in the absence of treatment (changes in `SundayHours`), the trends in the outcome (the `Profits`) would have been the same in the treatment and control groups. This is known as the parallel trends assumption. If this assumption is violated - for example, if profits were already trending upwards in states where `SundayHours` increased, even before the increase - this could bias your results.

4. Spillover Effects: If there are spillover effects between states - for instance, if changes in `SundayHours` in one state affect profits in neighboring states - this could bias your results.

f. Display and interpret the results of this second regression.

```
. reg Profits SundayHours i.State i.Period
```

Source	SS	df	MS	Number of obs	=	120
				F(21, 98)	=	6.57
Model	614107.537	21	29243.216	Prob > F	=	0.0000
Residual	436148.005	98	4450.48984	R-squared	=	0.5847
				Adj R-squared	=	0.4957
Total	1050255.54	119	8825.67682	Root MSE	=	66.712

Profits	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
SundayHours	-68.41689	9.633715	-7.10	0.000	-87.53469	-49.2991
State						
2	-24.82126	32.08355	-0.77	0.441	-88.49002	38.84751
3	5.353015	29.92473	0.18	0.858	-54.03164	64.73767
4	-11.92708	32.29627	-0.37	0.713	-76.01798	52.16382
5	-5.223887	30.61091	-0.17	0.865	-65.97024	55.52246
6	45.45023	31.23579	1.46	0.149	-16.53618	107.4367
7	-17.93112	31.25184	-0.57	0.567	-79.94938	44.08715
8	-14.67112	30.74162	-0.48	0.634	-75.67687	46.33463
9	-7.746785	30.7673	-0.25	0.802	-68.80349	53.30992
10	-21.17153	29.99224	-0.71	0.482	-80.69016	38.34709
11	-25.12368	30.64671	-0.82	0.414	-85.94107	35.69372
12	-50.88836	30.01354	-1.70	0.093	-110.4493	8.672536
Period						
2	35.65862	27.3016	1.31	0.195	-18.52053	89.83777
3	17.60023	27.38124	0.64	0.522	-36.73695	71.93741
4	39.34142	27.46151	1.43	0.155	-15.15505	93.83789
5	93.4151	27.42311	3.41	0.001	38.99484	147.8354
6	113.284	27.77513	4.08	0.000	58.16518	168.4029
7	134.4217	27.35709	4.91	0.000	80.13244	188.711
8	123.7879	28.43885	4.35	0.000	67.35191	180.2238
9	162.7998	29.30196	5.56	0.000	104.6511	220.9486
10	196.1103	28.80342	6.81	0.000	138.9509	253.2698
_cons	1019.204	53.6244	19.01	0.000	912.7878	1125.62

By regressing Profits on Sunday Hours, the 12 states (state 1 is the omitted baseline), and all 10 time periods (again, 1 is the omitted baseline), and evaluating the coefficient on Sunday Hours, we estimate that an increase of 1 Sunday Hour of operation will result in an average $68.42 decrease in profits.