[redacted]

Name: Ben Chu

[redacted]

ben-chu@berkeley.edu

Problem 1. Binary dependent variable.

(a) Using these 8 explanatory variables, fill out the table below by estimating a Linear Probability Model, a Probit Model and a Logit Model. For each model, enter the coefficient estimates and then their robust standard errors in parentheses. Include the sample average of each explanatory variable.

```
regress affair male age yrsmarr kids relig educ occup ratemarr , robust
probit affair male age yrsmarr kids relig educ occup ratemarr , robust
logit affair male age yrsmarr kids relig educ occup ratemarr , robust
esttab m1 m2 m3, se r2 ar2 pr2 title("Models of Affairs") mtitle("LPM" "Probit" "Logit")
```

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

Models of Affairs

	(1) LPM	(2) Probit	(3) Logit
main			
male	0.0452	0.173	0.280
	(0.0412)	(0.141)	(0.245)
age	-0.00742*	-0.0246*	-0.0443*
	(0.00316)	(0.0112)	(0.0192)
yrsmarr	0.0160**	0.0543**	0.0948**
	(0.00563)	(0.0190)	(0.0323)
kids	0.0545	0.217	0.398
	(0.0463)	(0.169)	(0.298)
relig	-0.0537***	-0.185***	-0.325***
	(0.0153)	(0.0532)	(0.0924)
educ	0.00308	0.0113	0.0211
	(0.00852)	(0.0292)	(0.0500)
occup	0.00591	0.0137	0.0309
	(0.0117)	(0.0414)	(0.0726)
ratemarr	-0.0875***	-0.272***	-0.468***
	(0.0170)	(0.0533)	(0.0915)
_cons	0.736***	0.779	1.377
	(0.163)	(0.539)	(0.938)
N	601	601	601
R-sq	0.107		
adj. R-sq	0.095		
pseudo R-sq		0.096	0.098

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

summarize male age yrsmarr kids relig educ occup ratemarr

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. summarize male age yrsmarr kids relig educ occup ratemarr
```

Variable	Obs	Mean	Std. Dev.	Min	Max
male	601	.4758735	.4998336	0	1
age	601	32.48752	9.288762	17.5	57
yrsmarr	601	8.177696	5.571303	.125	15
kids	601	.7154742	.4515641	0	1
relig	601	3.116473	1.167509	1	5
educ	601	16.16639	2.402555	9	20
occup	601	4.194676	1.819443	1	7
ratemarr	601	3.93178	1.103179	1	5

The sample average of each explanatory variable is included in the summarize table listed above. The average is listed as the mean.

N = observations

(b) What is the interpretation of the constant coefficient in the case of the LPM and the Logit Model?

The interpretation of the constant coefficient in the case of the LPM and the Logit Model is that the constant = the probability of an affair occuring when all of the explanatory variables within the regression have a value of zero. It makes sense in most cases, especially for older ages when people are generally old enough to be married. If the person is a child, they rationally do not have the ability to be in an extramarital relationship. The dependent variables in LPM, PROBIT, and LOGIT is the probability of having an extramarital affair for each person; however, the difference between all three models is the inherent relationship to the explanatory variables within the model. Within the dataset, the relig variable cannot be zero; the lower realm for this variable is 1 = anti-religious. The LOGIT has the exact same interpretation except for one feature since this model must evaluate the overall probability of an individual having an affair.

LOGIT model:

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

P(Y=1|X=0)
= 1/(1+exp(-$\beta_0$hat)
= 1/(1+exp(-1.377))
= .798508755
= .798508755*100% = 79.8508755%

PROBIT model:

P(Y=1|X=0)
= ϕ($\beta_0$hat)
= ϕ(.779)
= .782
= .782*100% = 78.2%

(c) Interpret the coefficients on "male" and "age" in the case of the LPM. Do the same for the Probit Model.

The regression is linear, so the interpretation of coefficients male and age: "male" is a binary dummy variable. This means =1 if male and =0 if not male (female). When the "male" variable is run in the regression, its coefficient is the difference of probabilities that a man has an affair vs. if a female has an affair using the binary system.

Given the LPM of 0.0452 (0.0412) for the male variable, which means that men included in the sample are 0.0452*100% = 4.52% more likely to engage in an extramarital affair in comparison to a female. The coefficient on age is -0.007 (.003). Age is not a binary variable meaning that a 1 year increase in the age of a respondent would equate to a -0.007*100% = -0.7% decrease in likelihood of an extramarital affair in the LPM model.

Interpretation of coefficients male and age for PROBIT model:

The Probit model as specified in lecture is: $PR(Y=1) = \Phi(\beta_0 + \beta_1X_1 + ... + \beta_kX_k)$

$\Phi(\beta_0 + \beta_{male} + \beta_2X_2 + ... + \beta_kX_k) - \Phi(\beta_0 + \beta_0 + \beta_2X_2 + ... + \beta_kX_k)$. β_0 = constant, β_{male} = coefficient on male binary variable, and the rest are the remaining regressors. I can take the derivative of the regressor "age" from the PROBIT Model in order to track the change within variables by using the formula $d\Phi(\beta_0 + \beta_1X_1 + ... + \beta_kX_k)/d_j = B_j\Phi(\beta_0 + \beta_1X_1 + ... + \beta_kX_k)$.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

-.025^2*100 = -.03125 by taking derivative of age

For PROBIT, the probability is .2282696 or 22.82696%.

probit affair male age yrsmarr kids relig educ occup ratemarr , robust
margins, predict(pr) atmeans

```
. probit affair male age yrsmarr kids relig educ occup ratemarr , robust

Iteration 0:   log pseudolikelihood = -337.68849
Iteration 1:   log pseudolikelihood =  -305.4326
Iteration 2:   log pseudolikelihood = -305.19796
Iteration 3:   log pseudolikelihood = -305.19796

Probit regression                               Number of obs     =        601
                                                Wald chi2(8)      =      54.97
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -305.19796               Pseudo R2         =     0.0962
```

affair	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	.1734568	.1406886	1.23	0.218	-.1022877	.4492014
age	-.0245844	.0111845	-2.20	0.028	-.0465055	-.0026632
yrsmarr	.0543435	.0189627	2.87	0.004	.0171772	.0915098
kids	.2166441	.1687474	1.28	0.199	-.1140948	.5473829
relig	-.1854684	.0531842	-3.49	0.000	-.2897074	-.0812294
educ	.0112622	.029205	0.39	0.700	-.0459786	.068503
occup	.0136686	.0414407	0.33	0.742	-.0675536	.0948908
ratemarr	-.2717912	.0533357	-5.10	0.000	-.3763273	-.1672551
_cons	.7794021	.5394882	1.44	0.149	-.2779754	1.83678

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. margins, predict(pr) atmeans

Adjusted predictions                            Number of obs     =        601
Model VCE    : Robust

Expression   : Pr(affair), predict(pr)
at           : male            =    .4758735 (mean)
               age             =    32.48752 (mean)
               yrsmarr         =    8.177696 (mean)
               kids            =    .7154742 (mean)
               relig           =    3.116473 (mean)
               educ            =    16.16639 (mean)
               occup           =    4.194676 (mean)
               ratemarr        =     3.93178 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2282696   .0182317    12.52   0.000     .1925362     .264003
------------------------------------------------------------------------------
```

.2282696 or 22.82696%*-.03125 = -.007133425*100% = -0.7133%

For the probit model coefficient on age, another year in age decreases the probability of having an affair by around -0.7133% according to the probit model.

(d) Is the coefficient on the Male dummy significant? Is this what you would expect?

regress affair male age yrsmarr kids relig educ occup ratemarr , robust

t = B0/SE
t = .0452014/.0412434 = 1.095

With a p-value of 0.274 the male dummy is not significant at any significance level, including 10%, 5%, and 1%.

I do not expect to see a significant effect for the coefficient on the Male dummy variable. Nothing in this problem provides data that men have a higher propensity to cheat despite any assumptions that people would have on this topic. Additionally, this assumption holds as long as

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

extramarital affairs are not abnormally commonly correlated among a few people within either the male or female category. That is to say, as long as unconditional means are zero. I am unconvinced of the significance of the male dummy variable until I see data which would suggest that. Intuitively, in a heterosexual situation, a man needs a corresponding partner (a woman) to cheat on his spouse, in which case, both the man and the woman are cheating (assuming the woman is also cheating). This would be an even proportion. That is to say any extramarital affair involves both a man and a woman, so either the man is cheating or the woman is cheating in a heterosexual relationship. Until data suggests otherwise, I can not just make such a broad assumption that the male dummy is significant. For this reason, I do not expect to see the coefficient on the Male dummy be significant.

(e) Why might yrsmarried be an endogenous regressor? Explain.

yrsmarried might be an endogenous regressor because factors which are not accounted for within the context of the model may strengthen a marriage and cause the chance of an extramarital affair to occur. These unobserved factors may likely be related to the years that two people remained married to each other. The longer that two people are married in years, the stronger that unobserved variables which cannot necessarily be observed, quantified, or measured may manifest and decrease the chance of an extramarital affair. There may be a connection between yrsmarried and the likelihood in percentage of affairs, which may cause yrsmarried to be an endogenous regressor, which means that it is correlated/non-zero covariance with the u (error term).

(f) Can you think of any reason why you may want to use homoskedastic standard errors when estimating your model via logit?

I cannot see why I would want to use homoskedastic standard errors when estimating my model via LOGIT unless I had an explicit reason to do so. An example where I would want to use homoscedastic standard errors is if the data presented to me clearly had homoskedastic errors or there was a strong reason presented to me which would lead me to believe that the distribution has homoskedastic errors. Since having homoskedastic errors is markedly rare in real-world situations even if given observational data, such as the variables involving marriage and extramarital affairs, then it is a safe bet to use heteroskedastic standard errors. As mentioned in class: when in doubt, choose heteroskedastic standard errors.

(g) Calculate the predicted probability that an "average" individual had an affair for each of the three models.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

To calculate the predicted probability that an "average" individual had an affair for each of the three models is as follows:

regress affair male age yrsmarr kids relig educ occup ratemarr , robust
margins, predict(xb) atmeans

```
. regress affair male age yrsmarr kids relig educ occup ratemarr , robust

Linear regression                               Number of obs     =        601
                                                F(8, 592)         =       8.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1066
                                                Root MSE          =     .41215

------------------------------------------------------------------------------
             |               Robust
      affair |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0452014   .0412434     1.10   0.274    -.0357997    .1262025
         age |  -.0074204   .0031622    -2.35   0.019     -.013631   -.0012099
     yrsmarr |   .0159811   .0056307     2.84   0.005     .0049226    .0270396
        kids |   .0544873   .0462901     1.18   0.240    -.0364255    .1454002
       relig |  -.0536979   .0153418    -3.50   0.001    -.0838289   -.0235668
        educ |   .0030784   .0085195     0.36   0.718    -.0136537    .0198106
       occup |   .0059126   .0116823     0.51   0.613    -.0170311    .0288563
    ratemarr |  -.0874553   .0170267    -5.14   0.000    -.1208955   -.0540152
       _cons |   .7361066   .1630619     4.51   0.000     .4158563    1.056357
------------------------------------------------------------------------------
```

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. margins, predict(xb) atmeans

Adjusted predictions                            Number of obs     =        601
Model VCE    : Robust

Expression   : Linear prediction, predict(xb)
at           : male            =    .4758735 (mean)
               age             =    32.48752 (mean)
               yrsmarr         =    8.177696 (mean)
               kids            =    .7154742 (mean)
               relig           =    3.116473 (mean)
               educ            =    16.16639 (mean)
               occup           =    4.194676 (mean)
               ratemarr        =     3.93178 (mean)
```

	Delta-method Margin	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	.249584	.0168121	14.85	0.000	.2165654	.2826027

probit affair male age yrsmarr kids relig educ occup ratemarr , robust
margins, predict(pr) atmeans

```
. probit affair male age yrsmarr kids relig educ occup ratemarr , robust

Iteration 0:   log pseudolikelihood = -337.68849
Iteration 1:   log pseudolikelihood =  -305.4326
Iteration 2:   log pseudolikelihood = -305.19796
Iteration 3:   log pseudolikelihood = -305.19796

Probit regression                               Number of obs     =        601
                                                Wald chi2(8)      =      54.97
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -305.19796               Pseudo R2         =     0.0962
```

affair	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	.1734568	.1406886	1.23	0.218	-.1022877	.4492014
age	-.0245844	.0111845	-2.20	0.028	-.0465055	-.0026632
yrsmarr	.0543435	.0189627	2.87	0.004	.0171772	.0915098
kids	.2166441	.1687474	1.28	0.199	-.1140948	.5473829
relig	-.1854684	.0531842	-3.49	0.000	-.2897074	-.0812294
educ	.0112622	.029205	0.39	0.700	-.0459786	.068503
occup	.0136686	.0414407	0.33	0.742	-.0675536	.0948908
ratemarr	-.2717912	.0533357	-5.10	0.000	-.3763273	-.1672551
_cons	.7794021	.5394882	1.44	0.149	-.2779754	1.83678

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. margins, predict(pr) atmeans

Adjusted predictions                            Number of obs     =        601
Model VCE    : Robust

Expression   : Pr(affair), predict(pr)
at           : male            =    .4758735 (mean)
               age             =    32.48752 (mean)
               yrsmarr         =    8.177696 (mean)
               kids            =    .7154742 (mean)
               relig           =    3.116473 (mean)
               educ            =    16.16639 (mean)
               occup           =    4.194676 (mean)
               ratemarr        =     3.93178 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2282696   .0182317    12.52   0.000     .1925362     .264003
------------------------------------------------------------------------------
```

logit affair male age yrsmarr kids relig educ occup ratemarr , robust
margins, predict(pr) atmeans

```
.
. logit affair male age yrsmarr kids relig educ occup ratemarr , robust

Iteration 0:   log pseudolikelihood = -337.68849
Iteration 1:   log pseudolikelihood =  -305.8919
Iteration 2:   log pseudolikelihood = -304.75954
Iteration 3:   log pseudolikelihood = -304.75521
Iteration 4:   log pseudolikelihood = -304.75521

Logistic regression                             Number of obs     =        601
                                                Wald chi2(8)      =      52.95
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -304.75521               Pseudo R2         =     0.0975

------------------------------------------------------------------------------
             |               Robust
      affair |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .2802867   .2453376     1.14   0.253    -.2005662    .7611395
         age |  -.044255    .019193    -2.31   0.021    -.0818726   -.0066375
     yrsmarr |   .094773    .0323312     2.93   0.003     .031405     .158141
        kids |   .3976721   .2978786     1.34   0.182    -.1861591    .9815034
       relig |  -.3247206   .0924055    -3.51   0.000     -.505832   -.1436093
        educ |   .0210509   .0500122     0.42   0.674    -.0769713    .1190731
       occup |   .0309197   .0726332     0.43   0.670    -.1114387    .1732781
    ratemarr |  -.4684543   .0914609    -5.12   0.000    -.6477144   -.2891941
       _cons |   1.377258   .9383412     1.47   0.142    -.4618568    3.216373
------------------------------------------------------------------------------

. margins, predict(pr) atmeans
```

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. margins, predict(pr) atmeans

Adjusted predictions                            Number of obs     =        601
Model VCE    : Robust

Expression   : Pr(affair), predict(pr)
at           : male            =    .4758735 (mean)
               age             =    32.48752 (mean)
               yrsmarr         =    8.177696 (mean)
               kids            =    .7154742 (mean)
               relig           =    3.116473 (mean)
               educ            =    16.16639 (mean)
               occup           =    4.194676 (mean)
               ratemarr        =     3.93178 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2224621   .0186032    11.96   0.000     .1860006    .2589237
------------------------------------------------------------------------------
```

For LPM, the probability is .249584 or 24.9584%.
For PROBIT, the probability is .2282696 or 22.82696%.
For LOGIT, the probability is .2224621 or 22.24621%.

(h) Holding other variables constant at their sample averages, if an individual has their first child, what is the effect on the probability that the individual has an affair?

LPM: Holding other variables constant at their sample averages, if an individual has their first child, the effect on the probability that the individual has an affair is 5.45% for the LPM model. This is because the coefficient for "kids" in this model is .054, meaning that +1 kid leads to an increase in extramarital affair probability of 5.45%. Linear: 5.45%. PR(Y=1) = $\Phi(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$. Probit: 6.12%. 1/(1+e^-z) Logit: 6.15%.

(i) Compute the Percent Correctly Predicted for the Logit Model.

Based on the sample, the average observed probability of an extramarital affair is .249584 or 24.9584% according to LPM or the linear specification. According to the LOGIT model predictions, the fitted dependent variable values = the probability that an individual has an extramarital affair. Now, I will put this data with the implication predictions of the probability of someone having an extramarital affair for each individual being studied based on which of their fitted values are greater than the 50% level. Finally, when comparing these predictions with the

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

true data from the sample, it shows that I have accurately predicted the model the following percentage:

predict yl, pr
gen y2=(y1>0.5)
tab y2 affair

```
. predict yl, pr

. gen y2=(y1>0.5)
variable y2 already defined
r(110);

. tab y2 affair
```

y2	=1 if had at least one affair: 0	1	Total
0	437	127	564
1	14	23	37
Total	451	150	601

the pr stands for probit and the tab gives the breakdown of the data
601 is the number of participants in sample
437 + 23 is the number correctly predicted in the sample as being above .5
437 + 23 = 460
460/601 = .765391015*100% = 76.5391015%, which is an accurate prediction.

Alternatively, the fractional formula is: 23/601 + 437/601, which yields the exact same result since the formulas are completely equivalent.

In sum: the Percent Correctly Predicted for the Logit Model is 76.5391015%.

Problem 2. Instrumental variable estimation

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

(a) Estimate a simple linear demand equation by regressing the quantity of gas (quantgas) consumed on the price of a gallon of gas (pricegas). What is your estimate of the price coefficient from the OLS estimation?

reg quantgas pricegas, robust

```
. use "/Users/julianbertalli/Downloads/PS5data(gasoline).dta"

. reg quantgas pricegas, robust

Linear regression                               Number of obs     =        296
                                                F(1, 294)         =      13.84
                                                Prob > F          =     0.0002
                                                R-squared         =     0.0462
                                                Root MSE          =     696.03
```

quantgas	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
pricegas	7.825219	2.103573	3.72	0.000	3.685248	11.96519
_cons	6531.83	223.2809	29.25	0.000	6092.399	6971.262

The OLSE of the pricegas coefficient is 7.825219. The t-stat for pricegas is 3.72 and a p-value of 0.000, which is extremely statistically significant.

(b) Use your OLSEs to express the price elasticity of demand evaluated at the average price of gas. Does it make economic sense? [Hint: express the price elasticity when demand is linear.]

The elasticity for interpreting a Stata output is the log-log function for the demand for pricegas. Since this is a linear-linear specification for demand, the pricegas coefficient is not an in price elasticity form for interpretation. Q=A-BP. The elasticity can be written as:

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

Elasticity = PdQ/QdP
= P(-B)/(A-BP)
= -BP/(A-BP)

A = 6531.83
B = -7.825219
P = 114.8811

```
. use "/Users/julianbertalli/Downloads/PS5data(gasoline).dta"

. sum pricegas
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pricegas	**296**	**114.8811**	**19.5499**	**64.7**	**172.9**

P is the pricegas mean, which is 114.8811 from the Stata output
B = -7.825219 and not positive 7.825219 since the demand curve is sloping upwards, so B must be <0, which means the coefficient on pricegas must be negative within the context of this problem.

Elasticity = -(-7.825219)(114.8811) / (6531.83-(-7.825219)(114.8811))
= 898.9697665 / (6531.83 + 898.9697665)
= 898.9697665 / 7430.799767
= 0.120978871

Price elasticity of demand evaluated at the average price of gas = 0.120978871, which does not make economic sense. 0.120978871 does not make economic sense since it is a positive number, which means that the demand curve is sloping upwards, not downwards. The ordinary least squares estimator used to express the elasticity of demand in this problem may be bias.

(c) Now introduce per capita personal income (persincome) as a regressor in the linear demand model and re-estimate using OLS. How has your estimate of price coefficient changed?

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

reg quantgas pricegas persincome, robust

```
. reg quantgas pricegas persincome, robust

Linear regression                               Number of obs     =        296
                                                F(2, 293)         =     520.92
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7591
                                                Root MSE          =     350.44
```

quantgas	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
pricegas	-6.860633	1.360879	-5.04	0.000	-9.53897	-4.182296
persincome	.3188366	.0099482	32.05	0.000	.2992576	.3384157
_cons	6632.961	168.5702	39.35	0.000	6301.199	6964.723

After using a Stata regression to introduce per capita personal income (persincome) as a regressor in the linear demand model and re-estimate using OLS, the estimate of the price coefficient pricegas is -6.860633, whereas before persincome was introduced, the value was 7.825219. pricegas is still extremely statistically significant with a p-value of 0.000 and the demand slope remains characterized by a downward slope. Since this regressor is extremely statistically significant with the 0.000 p-value, we must continue to use this instrument as an exogenous factor for the quantgas independent variable.

(d) Do you think that the above regression suffers from omitted variable bias? If so, can you determine the sign of the bias?

In part (c), the coefficient value was -6.860633, while in part (a), the value was 7.825219. This means that the coefficient estimate on pricegas went from positive to negative in almost the same magnitude. Given this large change and the change in the sign of the coefficient on pricegas, it does appear that OVB (Omitted Variable Bias) is at play in this problem. Additionally, the sign of the bias must be negative since the coefficient is clearly downwardly biased.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

(e) Give reasons why you should suspect that the gasoline price would be correlated with error term even after you introduced personal income into the regression. Evaluate the monthly sales of autos in the U.S. (carsales) serve as a good instrument for price of gas? Explain.

I would suspect that gas price would be correlated with the error term even if personal income is introduced into the regression for several reasons. First, as demonstrated with the OVB that is clear from part (d), several key factors have been omitted from the regression model which clearly have a significant impact on the price of gas, which may cause gas prices to be correlated with the error term since they are unknown. The monthly sales of autos in the U.S (carsales) is likely correlated with the supply of the car company manufacturers response. This correlation will also carry over with the price that autos sell for on the market. The data is from gas consumption from 1978 to 2002. During this time, autos undoubtedly became more efficient since cars have been continuing to push for better mileage and thus more fuel efficiency for decades now, which is directly related to the prices of gas. The impact of supply and demand on the price of gas can have an impact which is indicative of simultaneous causality, which is a potential threat to "internal validity." Simultaneous causality is when an independent variable within the regression model may also serve as a dependent variable within another regression model, which would make the variable essentially endogenous. This is a Keynesian reaction to the market functions since the supply would be in a different equation. Intuitively, if the price of gas goes up, the propensity to drive will go down and so too will the need or desire to buy new cars because of the price of gas. Additionally, the ramifications of COVID-19 are also heavily impacting the supply and demand of gas, so this may lend itself to simultaneous causality, which is another good reason that the price of gas can be correlated with the error term if the error term includes pandemics and shocks to the economy due to diseases. The monthly sales of autos in the U.S. (carsales) can be subject to bias, so it may not be a good instrument.

(f) Estimate the first stage of a two stage least squares estimation by regressing price of gasoline on the sales of cars. Perform a test that determines whether car sales is a "strong instrument."

reg pricegas carsales persincome, robust

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. reg pricegas carsales persincome, robust

Linear regression                               Number of obs     =        296
                                                F(2, 293)         =      43.63
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3080
                                                Root MSE          =     16.318
```

pricegas	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
carsales	-6.337825	.9567319	-6.62	0.000	-8.220763	-4.454888
persincome	.0022527	.0005947	3.79	0.000	.0010824	.0034231
_cons	162.2362	10.13177	16.01	0.000	142.2959	182.1765

The F-statistic F(2, 293) = 43.63 (which is much greater than 10) and has a small p-value of 0.0000, which is extremely statistically significant. In light of this, I must reject the null hypothesis that car sales is a weak instrument for pricegas. Car sales is a “strong instrument.” The F test for strength of an instrument shows the same conclusion that the reg pricegas carsales persincome, robust demonstrated: a high F-stat of F(1, 293) = 43.88 and an extremely low p-value of 0.0000. In sum, car sales is a “strong instrument.”

test carsales

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. test carsales

 ( 1)  carsales = 0

       F(  1,   293) =   43.88
            Prob > F =    0.0000
```

F-stat of 43.88 with a p-value of 0.000

(g) Can you suggest another instrument that is likely to be a better instrument than car sales?

I can suggest another instrument which is likely to be a better instrument that car sales within the context of this problem. The estimated demand function is both supply and demand within the context of the problem. For this reason, we have other options to look for a better instrument for car sales. Since supply and demand both factor into the equation now, I can use gas supply factor as a better instrument than car sales, such as the oil crisis. Right now, oil is at historic lows due to the corona virus due to an overabundance of supply and nowhere to store it. Full ships cannot dock and the oil sits useless while the stock crashes. This will surely have a huge impact on the price of oil. So in this case, the supply of oil may be a much better instrument than car sales. These supply factors are extremely likely to be correlated with the price of gas in some way. The supply factor may not predict demand since demand for oil is unpredictable since less people are driving right now due to the nature of the country-wide shut down, but people are not driving less because of any factor in the supply such as an abundance of oil, it must be something else. In sum: the supply of the oil is likely correlated with the pricegas and is likely to be a better instrument than car sales. Additionally, the supply of oil would not change the demand for driving and increase or decrease gas consumption. This exact situation is playing out in our economy today with the crude oil prices and the excess supply and driving habits.

Note: a good instrument must satisfy two components:
1. Relevance = endogenous regressor and high correlation
2. Exogeneity - low correlation with error term u

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

Checking both of these components, the supply of oil holds up since it would likely be endogenous regressor and high correlation, which means that it has (1) relevance and has a low correlation with error term u, which means that it has (2) exogeneity. Since both hold up, it is likely a good instrument.

(h) Now perform the second stage of the TSLS estimation and report any change in the size of the coefficient on gasoline price as a result of using the instrumental variable.

predict pghat, xb
reg quantgas pghat persincome, robust

```
. predict pghat, xb

. reg quantgas pghat persincome, robust

Linear regression                               Number of obs     =        296
                                                F(2, 293)         =     415.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7507
                                                Root MSE          =     356.45

------------------------------------------------------------------------------
             |               Robust
    quantgas |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       pghat |  -14.94912   3.923322    -3.81   0.000    -22.67058   -7.227652
  persincome |   .3514918    .016071    21.87   0.000     .3198626    .383121
       _cons |   7399.737   402.8345    18.37   0.000     6606.922    8192.553
------------------------------------------------------------------------------
```

In this regression, pghat represents the fitted value for the pricegas coefficient in the model from the first stage. The OLSE is biased towards zero. This bias would happen if the pricegas coefficient had a positive correlation with the error term in the regression u. This new price of -14.94912 is actually more elastic than the previous OLS estimate of -6.860633. Given the new regression, the pricegas coefficient is -14.94912 with a robust standard error of -3.81.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

(i) Is the TSLS estimate of the price coefficient statistically significant? Do you have any reason to doubt the reported values of the standard errors from the second stage? Explain.

The Two-Stage Least Squares or TSLS estimate of the price coefficient is statistically significant when a regression is run on the instrumental variables within the regression. To run a single-equation instrumental-variables regression, I must use ivregress. According to the Stata website: "In the language of instrumental variables, varlist1 and varlistiv are the exogenous variables, and varlist2 are the endogenous variables." In order to run a two-staged least squares test, I must follow the ivregress with 2sls notation followed by the amount of gas = quantgas, pricegas must be set equal to carsales so it shows that they do not influence one another, and finally, personal income = persincome. The final step is to call for robustness.

ivregress 2sls quantgas (pricegas = carsales) persincome, robust

```
. use "/Users/julianbertalli/Downloads/PS5data(gasoline).dta"

. ivregress 2sls quantgas (pricegas = carsales) persincome, robust

Instrumental variables (2SLS) regression          Number of obs   =        296
                                                  Wald chi2(2)    =     614.10
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.7188
                                                  Root MSE        =     376.63

------------------------------------------------------------------------------
             |               Robust
    quantgas |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pricegas |  -14.94912   3.370972    -4.43   0.000    -21.5561   -8.342134
  persincome |   .3514918   .0181088    19.41   0.000     .3159993    .3869843
       _cons |   7399.737   325.2871    22.75   0.000     6762.186    8037.289
------------------------------------------------------------------------------
Instrumented:  pricegas
Instruments:   persincome carsales
```

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

This regression allows me to identify if the Two-Stage Least Squares or TSLS estimate of the price coefficient is statistically significant. The estimate on the price coefficient is pricegas, which has a coefficient of -14.94912 and a p-value of 0.000. A p-value this close to zero is extremely statistically significant. In fact, this p-value is significant at all common levels, including 10%, 5%, and 1%, which means that the Two-Stage Least Squares or TSLS estimate of the price coefficient is absolutely statistically significant. It is worth noting that all of the other coefficients included in the TSLS regression are statistically significant as well with a p-value of 0.000.

Now, regarding the question of if I have any reason to doubt the reported values of the standard errors from the second stage: in the first stage, the SEs do not report that the regressor derives from the estimation process of the first-stage regression. Since this is the case with the SEs from the first-stage, it means that the predicted values that result from the first-stage regression values are random variables. The values are RVs because the OLSE used to get these predicted values are themselves RVs in Stata. Second stage assumes exogeneity of first stage and gives the true SEs.Since this will give me the true SEs, then I do not have any reason to doubt the reported values of the standard errors from the second stage.

(j) Suppose you were instead interested in studying how the supply of gas is influenced by its price. Would you feel comfortable regressing the quantity of gas produced on its price? Why?

I would not feel comfortable regressing the quantity of gas produced on its price. The quantity of gas is essentially the supply of gas within the context of this problem. If I did regress the quantity of gas produced on its price, I would need to account for simultaneity bias (like when the gas demand is regressed on the price for gas); however, if I regress the quantity produced (supply) on the price of gas, I am thereby failing to address the inherent problem of endogeneity within the context of this problem. For this reason, I would not feel entirely comfortable regressing the quantity of gas produced on its price.

(k) Also included in the dataset is the BLS monthly price index for consumer purchases of "transportation services" over the same sample period (transindex). Perform TSLS estimation using this price index as an instrument.

ivregress 2sls quantgas (pricegas=transindex) persincome, robust first

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. ivregress 2sls quantgas (pricegas=transindex) persincome, robust first

First-stage regressions
-----------------------

                                              Number of obs     =        296
                                              F(   2,    293)   =      99.50
                                              Prob > F          =     0.0000
                                              R-squared         =     0.3422
                                              Adj R-squared     =     0.3377
                                              Root MSE          =    15.9100

------------------------------------------------------------------------------
             |               Robust
   pricegas |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  persincome |  -.0088271   .0015476    -5.70   0.000    -.011873   -.0057813
  transindex |   1.100116   .1129407     9.74   0.000    .8778383    1.322394
       _cons |   29.34947   6.829549     4.30   0.000    15.90828    42.79066
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression      Number of obs   =        296
                                              Wald chi2(2)    =     317.11
                                              Prob > chi2     =     0.0000
                                              R-squared       =     0.4855
                                              Root MSE        =     509.47

------------------------------------------------------------------------------
             |               Robust
    quantgas |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pricegas |  -27.95674   3.434807    -8.14   0.000    -34.68884   -21.22464
  persincome |   .4040067   .0249502    16.19   0.000     .3551051    .4529083
       _cons |   8632.841   311.9091    27.68   0.000     8021.511    9244.172
------------------------------------------------------------------------------
Instrumented:  pricegas
Instruments:   persincome transindex
```

The First-stage choice in the Stata command ivregress gives the estimation results for the first stage. After running the regression in Stata, the First-stage regression shows that transindex is indeed a strong instrument since it meets the requirements of relevance and exogeneity. That is to say: transindex is a strong instrument since there is a strong correlation with the endogenous regressor and the correlation with the error term is markedly low. For the Second-stage

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

regression, pricegas is -27.95674. Additionally, this second stage is actually more elastic for the pricegas coefficient estimate.

(l) Now use both carsales and transindex as instruments for gas prices and report the results of the J-statistic test of the hypothesis that the two are valid instruments.

Using the Two-Stage Least Squares test for two instruments by themselves can produce differing results. With that being said, instrument for gas prices and use J-stat to test hypotheses of validity of both instruments: Regress pricegas on carsales, transindex, persincome in Stata to get predicted. Regress quantgas on predicted pricegas, persincome. For Two-stage least squares using carsales and transindex as instruments, I must regress the variables. The first regression will give the predicted value of the instruments. Now I must use the second-stage answers to find the error term u in the regression. Now, I must check for exogeneity by getting predicted residuals. Exogeneity is an indicator of a good instrument and means no correlation of the instruments being tested and the error terms u. To do this, I must get two instruments and an exogenous variable in the Two-Staged Least Squares test by using the original numbers from pricegas for getting predicted residuals. I must add persincome to the Two-Staged Least Squares, estimated residual, and the regression of both onto the instrument included in the problem. Additionally, I must call for a homoscedastic distribution to be estimated for the predicted values. This process will get me an F statistic. I can use this to test if an instrument equals zero or if it does not. I can find this value in Stata by running the test “test transindex carsales” which both equal zero to see if they are equal, which gives me the output of 4.75 and I can start to get the J stat. The J stat J = mF will be J is equal to the number of instruments included in a nonlinear fashion. The F makes it nonlinear, so it is not LPM. The d.o.f is the number of instruments - number of endogenous regressors included in the regression. The J must be run through a Chi-square test. There is only one endogenous regressors included in the regression, so the degrees of freedom is only 1, since the number of instruments - number of endogenous regressors included in the regression is 2 - 1. Now that I know the d.o.f is 1, I can use a Chi-squared table to find the critical value. A Chi-squared table shows that at the 95% level with a d.o.f of 1, value is 3.83. From here, I have the F stat of 4.75.

J = mF

F = 4.75

m = 2

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

J = (2)(4.75) = 9.5

With a J-statistic of 9.5, I am able to reject the hypothesis that both IVs are exogenous variables.

Since I used the J-statistic to reject the instruments, I can conclude that they are not good instruments since they do not fulfill both properties of being good instruments.

The TSLS provides the coefficient values and the test gives the F-stat of 4.75, but I still cannot know about exogeneity of either of the variables in the problem since I cannot necessarily test if one instrument is present for one endogenous regressor.

ivregress 2sls quantgas (pricegas=transindex carsales) persincome, robust

```
. use "/Users/julianbertalli/Downloads/PS5data(gasoline).dta"

. ivregress 2sls quantgas (pricegas=transindex carsales) persincome, robust

Instrumental variables (2SLS) regression          Number of obs   =        296
                                                  Wald chi2(2)    =     421.75
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.6071
                                                  Root MSE        =     445.21

------------------------------------------------------------------------------
             |               Robust
    quantgas |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pricegas |  -22.58343   2.810851    -8.03   0.000    -28.0926   -17.07427
  persincome |   .3823133   .0206397    18.52   0.000     .3418603    .4227664
       _cons |   8123.46    256.4003    31.68   0.000     7620.924    8625.995
------------------------------------------------------------------------------
Instrumented:  pricegas
Instruments:   persincome transindex carsales

. predict u_tsls, residuals
```

predict u_tsls, residuals
regress u_tsls carsales transindex persincome
test transindex carsales

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. regress u_tsls carsales transindex persincome
```

Source	SS	df	MS	Number of obs	=	296
				F(3, 292)	=	3.17
Model	1849694.55	3	616564.851	Prob > F	=	0.0248
Residual	56821344.7	292	194593.646	R-squared	=	0.0315
				Adj R-squared	=	0.0216
Total	58671039.3	295	198884.879	Root MSE	=	441.13

u_tsls	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
carsales	-65.49198	24.94726	-2.63	0.009	-114.5912	-16.39273
transindex	-8.880687	3.827377	-2.32	0.021	-16.41343	-1.347944
persincome	.0854074	.044921	1.90	0.058	-.0030026	.1738174
_cons	1225.205	403.0376	3.04	0.003	431.9782	2018.432

```
. test transindex carsales

 ( 1)  transindex = 0
 ( 2)  carsales = 0

       F(  2,   292) =    4.75
            Prob > F =    0.0093
```

(m) Based on your empirical results using these data, decide what you consider the "best" estimate of the price coefficient. Explain your reasoning.

Based on my empirical results using these data, I do not consider either estimate of the price coefficient to be the "best" estimator. The Two-Staged Least Squares results suggest that neither estimate of the price coefficient from the output in question (l) can be considered the "best" in my opinion. As we learned in lecture, overidentification can be a major problem when testing results. Overidentification happens when with more than one instrument, which can still be estimated, but weak estimates can be dropped. Given the overidentification test results, at least one of the instruments used is not exogenous, which is one of the key factors of being a good instrument. Exogeneity is a low correlation with the error term. Even if one regressor is not exogenous, it is tough to know if it is exogenous in a positive or negative direction. I do not believe that either of the instruments are necessarily exogenous. With that being said, both

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

instruments can each increase due to the demand side of the equation. For example, when the COVID-19 virus gets under control and the sanctions are lifted, people will be able to leave their house more and therefore, will demand more gas in response to their increase in driving. If the demand for driving increases, so too will the demand for the gas which powers the cars. This shows that gas can be an elastic good. That is to say: if the price goes up or stays too high during a pandemic where people do not have to drive, people will not buy as much gas since they do not NEED to drive anywhere. Under normal circumstances, people must buy gas for their commute, so demand will not fall drastically with a slight change in price. This shows why a low elasticity for the carsales coefficient may not be completely valid or the "best" option. This solidifies my belief that neither of the estimates on price coefficients is an obvious "best" option. A better option would be the end of the COVID-19 epidemic so that the gas prices can be dictated by how much oil is in the economy, rather than the inability of people to drive, which indirectly influence the demand for gas in a negative way, since people are not buying less because of the price explicitly since their demand for the good is elastic and they must stay home. This which directly impacts the price of the commodity itself would be a superior or "best" instrument.

Problem 3. Experiments.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

(a) What percentage of employees volunteered to participate in the experiment?

```
. tab volunteer
```

whether volunteer for experiment (work from home)	Freq.	Percent	Cum.
0	491	49.40	49.40
1	503	50.60	100.00
Total	994	100.00	

tab volunteer

To find the percentage of employees who volunteered to participate in the experiment, I used the tabulate function in Stata. The Stata output shows that 1 means volunteered, so the corresponding percent of 50.60 means that 50.6% of employees sampled volunteered to be a part of the experiment, while 49.4% do not.

(b) Use the variables commute and tenure as a dependent variable in bivariate linear regressions where volunteer is the explanatory variable. For each regression, interpret the coefficient on volunteer and comment on its statistical significance.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

reg commute volunteer, robust

```
. reg commute volunteer, robust

Linear regression                               Number of obs     =        994
                                                F(1, 992)         =      11.46
                                                Prob > F          =     0.0007
                                                R-squared         =     0.0114
                                                Root MSE          =     56.136

------------------------------------------------------------------------------
             |               Robust
     commute |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   volunteer |    12.03184   3.554315     3.39   0.001     5.056996    19.00667
       _cons |    74.46558   2.316037    32.15   0.000     69.92069    79.01048
------------------------------------------------------------------------------
```

In this regression the coefficient on volunteer is 12.03184. This is an indication that the people who volunteered to be in the experiment have about a 12.03 minute longer commute in comparison with the people who did not volunteer to be a part of the experiment. Given the p-value of 0.001, volunteer is statistically significant at all significant leves, including 10%, 5%, and 1%.

reg tenure volunteer, robust

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. reg tenure volunteer, robust

Linear regression                               Number of obs     =        994
                                                F(1, 992)         =       7.45
                                                Prob > F          =     0.0065
                                                R-squared         =     0.0075
                                                Root MSE          =     20.909

------------------------------------------------------------------------------
             |               Robust
      tenure |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  volunteer  |  -3.623471   1.327417    -2.73   0.006    -6.228338   -1.018604
       _cons |   26.84216   .9717046    27.62   0.000     24.93533    28.74899
------------------------------------------------------------------------------
```

In this regression, the coefficient for volunteer is -3.623471. This coefficient is an indicator that the average tenure for people who volunteered to be a part of the experiment is about 3.623 years less than people who did not choose to be a part of the volunteer experiment. Given the small p-value of 0.006, the coefficient is also significant at the 1% significance level (as well as the 5% and 10% levels).

(c) Impressed by your recent econometrics training, Ctrip hires you as a consultant to analyze the results from their experiment. To begin with, you estimate a bivariate linear regression model of the productivity of workers, measured by the log of the average number of calls taken per week (call this variable ln_calls), on the variable WFHShare ("work from home" share). Interpret the regression coefficient on WFHShare in words. Is the effect statistically significant?

gen ln_calls = log(calls)

reg ln_calls WFHShare, robust

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. reg ln_calls WFHShare, robust

Linear regression                               Number of obs     =        503
                                                F(1, 501)         =     142.60
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1630
                                                Root MSE          =     .67848

------------------------------------------------------------------------------
             |               Robust
    ln_calls |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WFHShare |    .975309   .0816725    11.94   0.000     .8148463    1.135772
       _cons |    5.444198   .0624476    87.18   0.000     5.321506    5.566889
------------------------------------------------------------------------------
```

The WFHShare coefficient is .975309. The WFHShare coefficient can be interpreted as log-linear, which means that a .01 increase in the share of days worked at home leads to a .975309% increase in the average number of calls taken each week.Additionally, it is clear that the impact of WFHShare is statistically significant. This is evident due to the small 0.000 p-value, which is significant at all significant levels, including 10%, 5% and 1%. The t-statistic of 11.94 is very large and corresponds to the small p-value.

(d) Has the Ctrip company achieved the ideal of a randomized controlled experiment, so that we can view the estimated effects of working from home on productivity in causal terms?

Although the amount of time telecommuting was assigned at random, it may be problematic that the workers themselves had the decision of whether or not they wanted to be a part of the experiment. Furthermore, participants did in fact differ from employees who decided that they did not want to volunteer in the experiment. People who decided to volunteer in the experiment also tended to remain employed by the firm for a longer amount of time. Additionally, people who volunteered in the experiment did tend to have commute times which were longer than people who decided not to participate in the experiment.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

(e) Create a dummy variable called longcommute which is equal to one if the employee has a commute of greater than or equal to 120 (i.e., 2 hours). How would you expect that including longcommute as a second explanatory variable would alter the coefficient on WFHShare – would it increase, decrease, or stay the same? Explain.

gen longcommute = commute>119

>119 minutes since greater than or equal to 120. If it is only >12, it would not work since it could not be equal to 120.

I would expect that including longcommute as a second explanatory variable would not alter the coefficient of WFHShare. That is to say, I expect the **coefficient on** WFHShare to remain the same. I expect this outcome given the random assignment of WFHShare. Given this. The longcommute variable generated in Stata is uncorrelated with the variable. Furthermore, the variable longcommute is not actually causing OVB (Omitted Variable Bias).

(f) Management believes that commute (the travel time from home to office and back) is an important determinant of a worker's productivity. They have two hypotheses:
(i) Employees who face a longer commute time are generally less productive than workers who have shorter commute times.
(ii) The effects of WFHShare on productivity is larger for those who face a longer commute.

Estimate a regression of ln_calls, with WFHShare, longcommute, and their interaction (call it WFHShareXlongcommute) as explanatory variables. Do your results support hypothesis (i), hypothesis (ii), both hypotheses, or neither one? Explain.

gen longcommute = commute>119

Again, >119 minutes since greater than or equal to 120. If it is only >12, it would not work since it could not be equal to 120.

gen WFHShareXlongcommute = WFHShare*longcommute

This is the interaction term between WFHShare and longcommute

reg ln_calls WFHShare longcommute WFHShareXlongcommute, robust

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. reg ln_calls WFHShare longcommute WFHShareXlongcommute, robust

Linear regression                               Number of obs     =        503
                                                F(3, 499)         =     179.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1794
                                                Root MSE          =     .67314
```

ln_calls	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
WFHShare	.8640932	.1247556	6.93	0.000	.6189822	1.109204
longcommute	.0162365	.1030335	0.16	0.875	-.1861966	.2186695
WFHShareXlongcommute	.3299599	.1366225	2.42	0.016	.0615336	.5983862
_cons	5.439778	.0953329	57.06	0.000	5.252475	5.627082

This regression gives the estimate for ln_calls, WFHShare, longcommute, and WFHShareXlongcommute. The results do not actually support hypothesis (i), but they do support hypothesis (ii). This obviously also means that the results do not support both as well.

These results fail to support hypothesis (i) since the coefficient for longcommute is a positive value of .0162365. Additionally, longcommute is not even close to statistically significant with an extremely large p-value of .875, it is not significant at any level (10%, 5%, or 1%).

The results do support hypothesis (ii) since the coefficient of WFHShareXlongcommute is positive at .3299599. Additionally, tbhe value is statistically significant with a small p-value of .016, which is almost significant (but not quite) at the 1% level; however, it is well within the bounds of the 5% significance level.

(g) If the coefficient on longcommute is statistically insignificant, would this lead you to drop longcommute from the regression model in part (f)? Explain your answer.

If the coefficient on longcommute is statistically insignificant, I would not automatically drop the coefficient of longcommute from the regression model in part (f). Dropping longcommute from the model in the context of this problem would change the interpretation that I would have on the coefficients in the regression that follows. Additionally, dropping longcommute from the model

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

in the context of this problem would prohibit me from being able to use the regression to give any type of answer to hypothesis (2). That is to say, dropping longcommute from the model would make it so that I cannot answer (2) the effects of WFHShare on productivity is larger for those who face a longer commute. Furthermore, I would not drop longcommute from the regression model because dropping any variable within the context of a regression model should happen because of motives for research or theory in which longcommute no longer makes empirical sense to include in the model, just being insignificant is not enough to warrant the removal, as this can provide useful data as well.

(h) Using the regression in part (f) and without estimating any other regression, write the estimated equation for the simple regression of ln_calls on WFHShare using only data for those with a commute of fewer than 120 minutes. You must show your solution to obtain full credit.

regress ln_calls on WFHShare using only data for those with a commute of fewer than 120 minutes

Using (f) and not a new regression, so do not need to calculate gen longcommute = commute<120.

Using gen longcommute = commute<120 and reg ln_calls WFHShare longcommute WFHShareXlongcommute, robust from part (f) means that instead of using the 1s from the binary dummy which represents greater than 120, we must simply use the 0s instead. This inversion of the binary allows us to use the same regression to find <120 without changing the Stata code.

(f) ln_calls^ = _cons + WFHShare + longcommute*0 + WFHShareXlongcommute*0

(f) ln_calls^ = 5.439778 + 0.8640932*WFHShare + 0.0162365*0 + 0.3299599*0

The variables longcommute and WFHShareXlongcommute are both aggregated by 0 since the problem explicitly calls for <120, which means their binary variable would both be 0. This essentially cancels both values out within the context of the regression, which gives us:

ln_calls^ = 5.439778 + 0.8640932*WFHShare + 0.0162365*0 + 0.3299599*0

or ln_calls^ = 5.439778 + 0.8640932*WFHShare

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

To get answer, I would just plug an average value of WFHShare into the regression, which would given me ln_calls^.

5.439778 is the constant from part (f) and 0.8640932 is the regression output on WFHShare from part (f). WFHShare stays in since the question explicitly asks for a regression of ln_calls on WFHShare.

Problem 4. Natural experiments.

(a) To begin the analysis, fill out the following table (where each cell represents an average consumption of cigarettes per capita): Based on your table, what is the differences-in-differences estimate of the effect of the ban on the consumption of tobacco? Is the effect practically significant?

Copy and paste the data onto an Excel spreadsheet.

Find averages before and after the 1970 ban, which went into effect in 1971, so calculate >1970.

Column1	Column2	Column3	Column4
	Before	After	After-Before
CAN	4043.142857	3601.8	-441.3428571
USA	4280.714286	3804.05	-476.664286
USA-CAN	237.5714289	202.25	-35.32142886

Excel formulas:

CAN
=AVERAGE(B2:B8)
=AVERAGE(B9:B28)
=[@Column3]-[@Column2]

USA

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

=AVERAGE(F2:F8)
=AVERAGE(F9:F28)
=[@Column3]-[@Column2]

USA-CAN
=J4-J3
=K4-K3
=[@Column3]-[@Column2]

	A	B	C	D	E	F	G
1	YEAR	CIGSPCCAN	PRICECAN		YEAR	CIGSPCUS	PRICEUS
2	1964	3975	128		1964	4368	103
3	1965	4095	128		1965	4417	105
4	1966	4158	127		1966	4387	105
5	1967	4168	127		1967	4355	105
6	1968	3971	137		1968	4254	106
7	1969	3879	138		1969	4077	108
8	1970	4056	133		1970	4107	110
9	1971	4040	132		1971	4072	109
10	1972	4089	128		1972	4113	109
11	1973	4139	123		1973	4238	105
12	1974	4152	112		1974	4225	101
13	1975	4127	114		1975	4138	98
14	1976	4008	115		1976	4342	97
15	1977	3990	116		1977	4117	95
16	1978	3805	122		1978	3979	93
17	1979	3901	117		1979	4036	91
18	1980	3873	116		1980	3965	90
19	1981	3890	119		1981	3966	88
20	1982	3825	126		1982	3815	92
21	1983	3619	142		1983	3622	107
22	1984	3399	154		1984	3634	110
23	1985	3255	176		1985	3577	113
24	1986	2985	201		1986	3460	118
25	1987	2823	209		1987	3392	123
26	1988	2817	215		1988	3199	130
27	1989	2733	242		1989	3171	141
28	1990	2566	264		1990	3020	153

(Raw Excel data above)

Differences-in-differences estimate of the effect of the ban on the consumption of tobacco:
-35.32142886 cigarettes consumed on average per capita

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

A difference-in-difference analysis of the effect of the ban on consumption of tobacco by treating the ban in cigarette advertising as a quasi-experiment. The results on the table suggest that the United State law which banned the advertising of cigarettes on radio and television which went into effect in 1971 reduces by 35.32142886. This is also equal to -35.32142886 cigarettes consumed on average per capita. In conclusion, I believe that this is a reasonably small change from before and after. After about 20 years, the numbers are only down about 35 cigarettes per capita on average. This means a reduction of about 20/35 = .57 less cigarettes per year per capita. Since the average for Americans is between 3,804.05 and 4,280.71, then the reduction of 35.32 out of over 3,800 is extremely small. It would take 38 to have a 1% reduction after, so 35.2 means that the consumption did not even decrease by one percent. I am unsure of the goals of the lawmakers who enacted the ban, but they surely expected a decrease in consumption larger than 1% over a 20 year period.

In conclusion: I believe that any decrease in smoking is a small victory; however, I believe that the marginally small decrease in cigarette consumption is extremely small, perhaps too small for the effort (in per capita terms). Therefore, I would have to conclude that the results for the effect are NOT practically significant.

(b) Now create a dummy variable indicating whether the ban was in effect or not in the U.S., plus a dummy variable for the treatment group (i.e., the U.S.) and the control group (i.e., Canada). Regress tobacco consumption on these two dummies and on the interaction between the two. How do your results compare to your diffs-in-diffs estimator from part (a)?

```
gen ban = 0
replace ban = 1 if YEAR > 1970
gen treat = 0
replace treat = 1 if COUNTRY == 2
gen bantreat = ban*treat
reg cigspc ban treat bantreat, robust
```

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. reg cigspc ban treat bantreat, robust

Linear regression                               Number of obs     =         54
                                                F(3, 50)          =      13.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2434
                                                Root MSE          =     416.59

------------------------------------------------------------------------------
             |               Robust
      cigspc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ban |  -441.3429   128.3388    -3.44   0.001    -699.119   -183.5668
       treat |   237.5714   63.65247     3.73   0.000    109.7217    365.4212
    bantreat |  -35.32143    164.267    -0.22   0.831   -365.2614    294.6186
       _cons |   4043.143   38.83534   104.11   0.000     3965.14    4121.146
------------------------------------------------------------------------------
```

The interaction coefficient is equal to the estimator for difference-in-difference from part (a). In the new regression, the t statistic is -.22 for bantreat. Given the p-value of .831, I cannot reject the hypothesis that the effect of the ban has no impact on the consumption of cigarettes.

(c) Finally, recognizing that price does also affect consumption, you introduce the price variable into the regression in (b). Report your results and compare to those from (b).

reg cigspc ban treat bantreat price, robust

Price does also affect consumption, so I introduce the price variable into the regression in (b). Given the regression results, the cigarette pack price increase comes with a reduction in the consumption of cigarettes. This is very expected given the laws of supply and demand, which tend to fall in line with the equilibrium price. This price increase is also statistically significant given the p-value of 0.000, which is significant at any significance level (10%, 5%, and 1%). The interaction variable of the coefficient and the difference-in-difference estimator is statistically significant at the 10%, 5%, and 1% level given the .003 p-value, which is below all significance levels, including the 1%. The interaction variable of the coefficient is also a lot larger than in the previous problem.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
.
. reg cigspc ban treat bantreat price, robust

Linear regression                               Number of obs     =         54
                                                F(4, 49)          =      72.98
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8543
                                                Root MSE          =     184.65
```

cigspc	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ban	-191.9745	42.02207	-4.57	0.000	-276.421	-107.528
treat	-60.8905	54.98356	-1.11	0.274	-171.3841	49.60311
bantreat	-259.1679	83.12174	-3.12	0.003	-426.2073	-92.12849
price	-11.87064	.9264966	-12.81	0.000	-13.73251	-10.00878
_cons	5599.893	124.2921	45.05	0.000	5350.119	5849.667

(d) Why would you expect that the price of a pack of cigarettes might be correlated with the error term? Note that some economists have argued that the advertising ban reduced competition among cigarette makers by eliminating one dimension on which they compete for customers, which in turn led to higher prices.

The price of a pack of cigarettes may very well be correlated with the error term. As demonstrated in the previous problem, the economic laws of supply and demand have a significant impact on the price of a pack of cigarettes, meaning that cigarette price is dependent on the supply and demand laws. Some factors may influence both supply and demand. Alternatives such as e-cigarettes and their corresponding prices may also impact the error term and cause a correlation to be present. The model up to this point did not yet control for the equation for demand such as the cost of healthcare. The overall cost of healthcare would make sense in this context because they may be a factor in smoking. For example, if a person knows that smoking will cost them a lot of money due to the costs of frequent doctor visits or other healthcare-related costs, they may not be as likely to start or continue smoking. The risk of losing money to healthcare costs may make people think twice about starting to smoke, continuing to

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

smoke, and the amount that people smoke and simultaneously impact the bottom line of the cigarette companies.