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

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

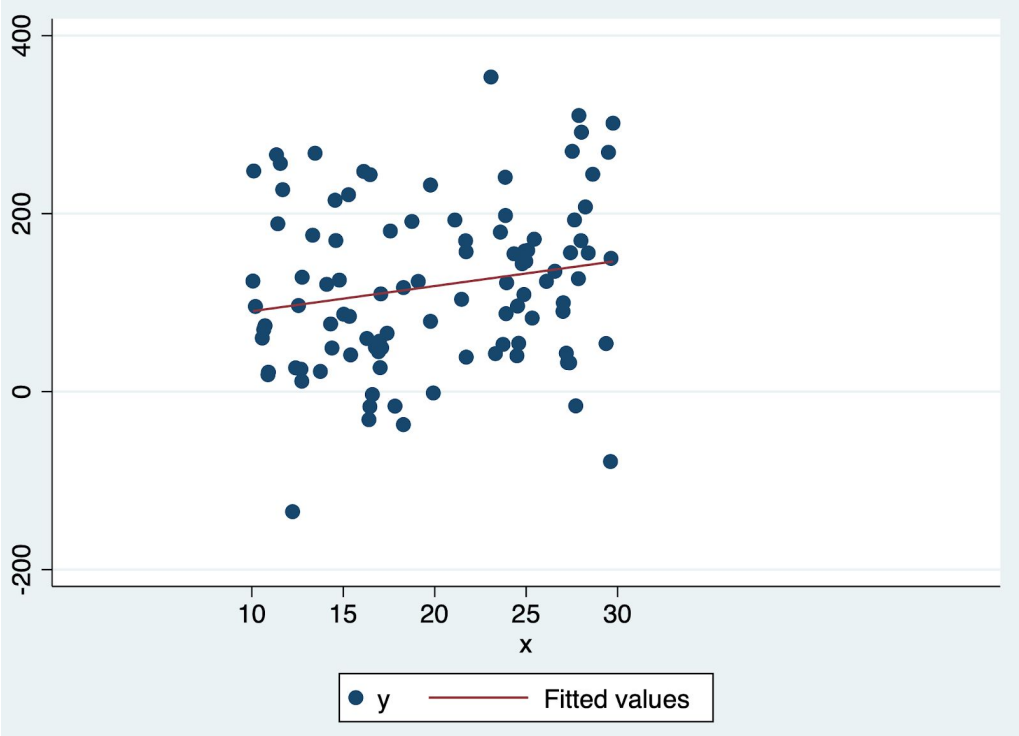
Problem 1: Bivariate linear regression.

a)

```
set obs 100

gen x = 10+20*runiform()

gen u = rnormal(0,100)

gen y = 30+5*x+u
```

b)

```
twoway (scatter y x) (lfit y x) , xscale(range(0 50))
```



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

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

Setting the xscale allows us to see that if the regression line continued, it would roughly lead to the constant of 62.09543 within the context of the problem.

c)

regress y x, robust

```
. regress y x, robust

Linear regression                               Number of obs     =        100
                                                F(1, 98)          =       2.98
                                                Prob > F          =     0.0876
                                                R-squared         =     0.0350
                                                Root MSE          =     92.115

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    2.825524   1.637496     1.73   0.088    -.4240338    6.075081
       _cons |    62.09543   33.98369     1.83   0.071     -5.34409     129.535
------------------------------------------------------------------------------
```

Each one of the three OLSE assumptions are satisfied in this case.

1) The expectation of the error term (u) is unrelated to the value of the independent variable (x). This is given in the problem.

2) The values of x (independent) and y (dependent) are randomly selected variables. This ensures that the draws of observations are independent from the other draws. This is given in the problem, which specifies random variables.

3) The independent variable x is distributed in a uniform fashion within the interval. Additionally, y is a normal variable and a linear random variable. This means that outliers are unlikely within the sample.

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

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

For my assessment of how well the least squares regression performs in estimating the true intercept and slope, the intercept from the Stata output: _cons = constant, which represents the intercept, which is 62.09543. This value is very far from the true intercept of 30 (not a great job for the least squares regression of estimating the true intercept). The slope estimate is 2.825524, which is reasonably close to the true value of 5, yet still a more precise estimate than the intercept. Both values are within the 95% confidence interval and have a p-value >.05 and a t-stat <1.96.

d)

The S.E.R or Standard Error of Regression is called the Root MSE or Root Mean Square in Stata. The S.E.R represents the standard deviation of the error term, u within the regression model. According to the Stata results for this regression, the Root MSE (S.E.R) = 92.115. Since the error term u was generated with a random draw from a normal distribution with a population mean 0 and population standard deviation 100, it is clear the standard deviation of 100 is reasonably close to 92.115. This means that the least squares estimation of is reasonably close to the true error term.

e)

predict uhat, residuals

summ uhat

```
. predict uhat, residuals

. summ uhat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
uhat	100	5.72e-07	91.649	-231.5412	226.1665

Mean (average) = 5.72e-07 = 0.000000572, which is extremely close to an average of zero. This confirms that the regression residuals add up to zero.

correlate uhat x

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

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

```
. correlate uhat x
(obs=100)
```

	uhat	x
uhat	**1.0000**	
x	**0.0000**	**1.0000**

Residuals are uncorrelated with regressor since the correlation of the regressor (x) is zero (0.0000).

f)

set obs 1000

gen x = 10 + 20*runiform()

gen u = rnormal(0, 100)

gen y = 30 + 5*x + u

regress y x, robust

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

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

```
. set obs 1000
number of observations (_N) was 100, now 1,000

. gen x = 10 + 20*runiform()

. gen u = rnormal(0, 100)

. gen y = 30 + 5*x + u

. regress y x, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =      72.43
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0662
                                                Root MSE          =     99.338

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    4.543496   .5338751     8.51   0.000     3.495849    5.591142
       _cons |    39.11348   10.92609     3.58   0.000     17.67273    60.55423
------------------------------------------------------------------------------
```

There is essentially a more accurate slope coefficient estimate (closer to 5) since more observations are taken in the sampled. As a result, there is a smaller standard error in the larger sample. The R^2 can increase or decrease, but in the regression that I ran, the R^2 increased from .0350 to .0662.

Problem 2: Wages and education.

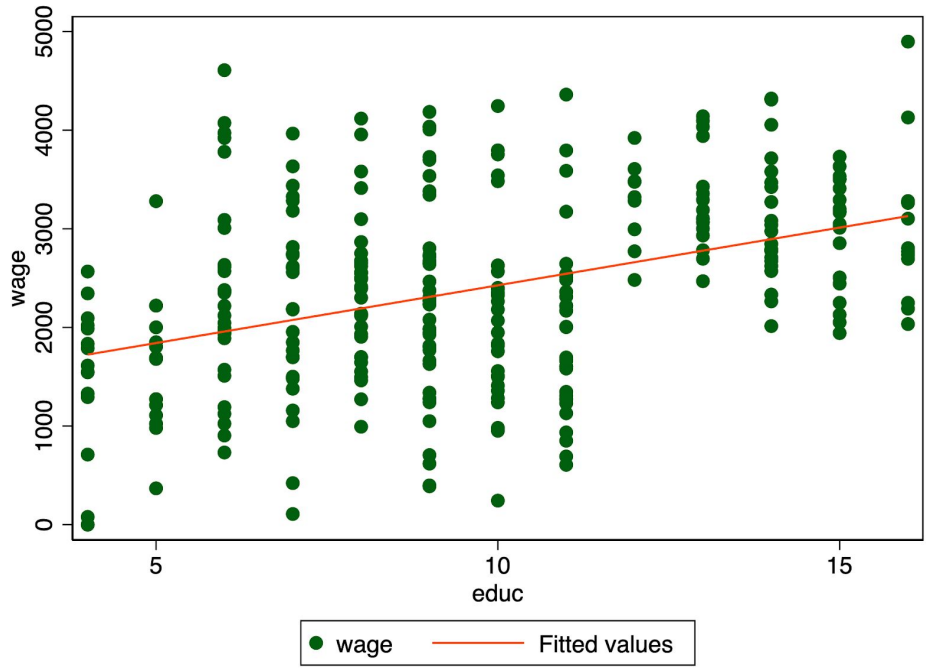
a)

graph twoway (scatter wage educ) (lfit wage educ), ytitle(wage)

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

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

It is generally expected that those with more education will have higher marginal productivity in comparison to those with less education because of the education variable. This higher productivity as a result of higher education will lead to higher wages earned. This assumed phenomenon is demonstrated on the graph which shows an upward sloping regression line for the data. The wages earned increase after high school graduation; i.e, going from years 11 to 12 and beyond after graduation from high school).



b)

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

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

regress wage educ , robust

```
. regress wage educ , robust

Linear regression                               Number of obs     =        300
                                                F(1, 298)         =      70.91
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1602
                                                Root MSE          =     885.07
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	117.1024	13.9067	8.42	0.000	89.73462	144.4702
_cons	1256.172	151.039	8.32	0.000	958.9339	1553.41

Based on the three OLS assumptions $\hat{B}_0$ and $\hat{B}_1$ are both consistent and unbiased estimators of the population parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$. $\hat{B}_0$ is estimating the intercept, which is positive, significant (all confidence levels), and large based on the sample drawn of 300. $\hat{B}_0$ is the predicted average wage of someone with no education (0 years), which comes out to an average of about $1256.17 per month. This estimate does not seem completely sound since the education goes from four to sixteen years, which excludes the possibility of zero years of education. This means that the baseline prediction of the wage for someone with zero years of education is not formed based on the data used to run the regression. This means that the people in the sample probably have differences to people with 0 years of education. A person with 0 years of education is outside of the sample range to accurately predict their income based on the regression line in its present form. $\hat{B}_1$ estimates the slope. This sample has a positive, large, and significant (all confidence levels) $\hat{B}_1$ where every year of education causes wage to increase by an average of about $117.10 per month. This equates to around $1,400 in wages per year extra for each year of education. This seems to be a reasonable average increase in pay in relation to work experience, so the slope estimate appears to confirm the scatter plot above.

c)

Three OLS assumptions and examples within the problem's context and plausibility.

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

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

1) Conditional mean zero errors.
- In the context of the problem, there is no variable aside from education explaining wages and that education is correlated with it.
-Assuming this is not completely plausible within the context of this problem since other variables impact the wages that a person earns after education. Factors such as opportunities, talent, ambition, capability, ability, etc. may also be correlated with education and cause them to go farther in education and explain their wage earnings, as opposed to education being the only causal factor in higher earnings.

2) Random sampling.
- Within the context of the problem, the random sampling assumption is only satisfied if the people within the sample were selected at random from the population being sampled. For example: 300 American workers = random sample from U.S population from census data.
- The problem does not specify if random selection from the "American workers" pool is the case, so there is not a way to say that the assumption is plausibly met. All we know for sure is that 300 American workers were sampled, but not exactly how the sample was drawn from the population.

3) No large outliers likely.
- People with a very extensive (or none at all) education (in years) or extremely large/low wages are unlikely to be drawn and included in the sample.
- This is a plausible assumption within the context of this problem. People with extensive educations or extreme wages appear to be underrepresented within the population being surveyed in the context of this problem, so the assumption of no large outliers appears to be satisfied; underrepresentation means that these outliers are unlikely to be drawn. This means that people with extensive educations or large wages are not too likely to be drawn and represented in the sample.

d)

T-statistic = (117.1024 - 100)/13.9067 = 1.2298 ≈ 1.23

Given this t-statistic, we fail to reject the null hypothesis, which means that the estimate of the slope coefficient is not statistically different than 100.

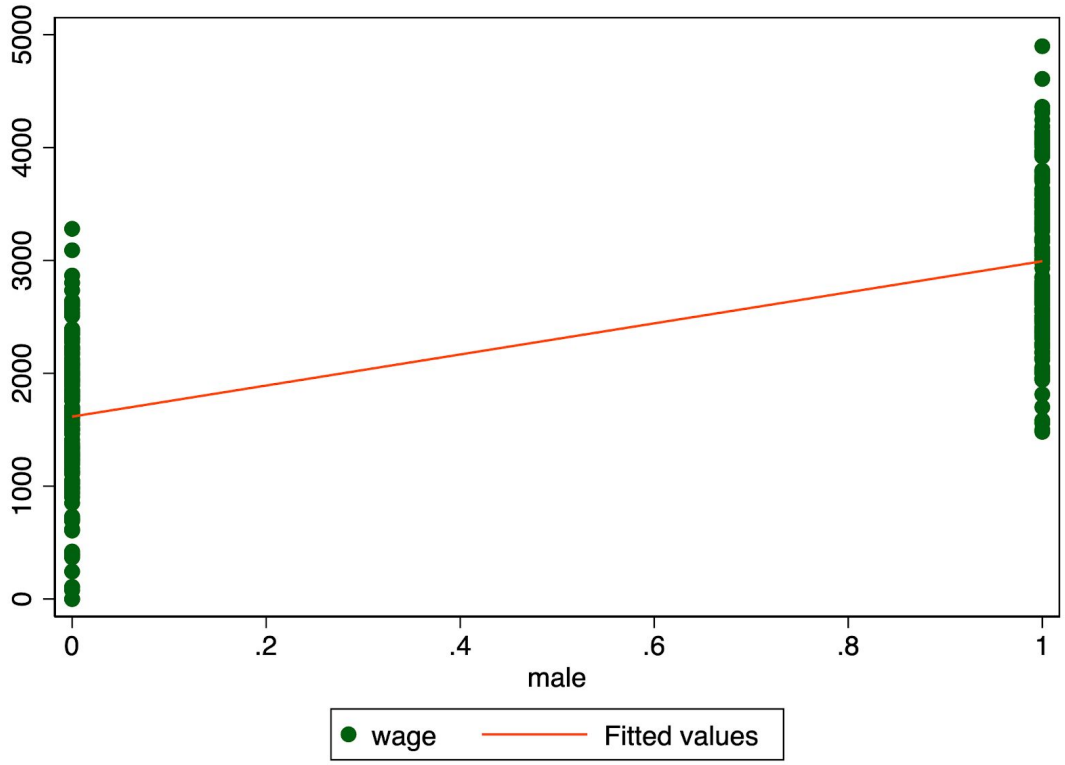
1.23<1.96, which is .05 level to reject, so we must fail to reject because 1.23 is within the range, not outside for 5% significance.

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

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

e)

twoway (scatter wage male) (lfit wage male)



The graph shows that women within the sample earn less wage (on average) than men earn. This is evident because the higher variable values for males are typically associated with higher values of wages. The covariance among the wages and men are positive within the context of the problem since the relationship appears to be linear. 0 = female, while 1 = male; the visualization clearly shows the upward trend of wages for males as opposed to females. The graph also does not present any obvious nonlinearities between the relationships for men and women's wages.

f)

regress wage male , robust

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

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

```
.
. regress wage male , robust

Linear regression                               Number of obs     =        300
                                                F(1, 298)         =     308.34
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5050
                                                Root MSE          =     679.52
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1376.893	78.41304	17.56	0.000	1222.579	1531.206
_cons	1616.227	56.84131	28.43	0.000	1504.366	1728.088

Interpretation: Given the three OLS assumptions, both $\hat{B}_0$ and $\hat{B}_1$ are both unbiased and consistent estimators of the population parameters: intercept ($\boldsymbol{\beta}_0$) and slope ($\boldsymbol{\beta}_1$). The OLS coefficient differs from the education coefficient since "male" is the dummy variable. The $\hat{B}_0$ is the intercept, which is positive, large, and significant. This intercept is the context of this problem is the predicted wage for a woman on average, which means that a woman should earn an average of about $1616.23 per month based on the given data. $\hat{B}_1$ represents the slope of the data. The $\hat{B}_1$ estimator is also positive, large, and significant. Within the context of this problem, the slope represents the difference of the average wage between women and men included in the sample. This allows the conclusion that based on the sample, men earn an average of about $1,376.89 more than women sampled.

Causality: All three OLS assumptions must be met for a relationship to be deemed "causal." On average, males have more education than women within this sample. This is evident when education is tabulated by gender. OLS assumption #1, conditional mean zero errors, is violated via the assumption that education has an influence on wages (which is very likely in the real world). The positive slope coefficient is not favorable towards either sex. The positive slope means that on average, men have a higher education than women do within the sample. Within the sample, women have a lower educational attainment (on average), which may explain the lower wages that women within the sample have on average. Another explanation may be that women within the sample may choose to pursue lower-paying jobs after their education. These real-world factors may influence the results of job selection and pay.

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

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

g)

ttest wage , by(male) unequal

```
. ttest wage , by(male) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	133	1616.227	56.86571	655.8073	1503.741	1728.713
1	167	2993.12	53.99701	697.7952	2886.51	3099.729
combined	300	2382.697	55.66669	964.1753	2273.149	2492.245
diff		-1376.893	78.41802		-1531.233	-1222.552

```
    diff = mean(0) - mean(1)                                      t = -17.5584
Ho: diff = 0                     Satterthwaite's degrees of freedom =  289.924

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

t-stat = -17.5584

p-value = 0.0000

Assuming unequal variances gives the same test statistics for a ttest since this method is robust for heteroskedasticity standard errors for regression analysis. This essentially makes the two results (in terms of t-stat and p-value) correspond.

Problem 3: Wine prices and vintage.

a) The error term represents each of the variables aside from prices and vintage, such as the type, area of origin, brand, reputation, overall quality, etc. The first OLS assumption is conditional mean zero errors or $E[u_i|X_i] = 0$. Since the error term, u, includes variables such as the type, area of origin, brand, reputation, overall quality of wine (and other things that are important factors in

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

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

the price), the first OLS assumption may be violated. Additionally, there is likely a correlation with the vintage and price, since bottles become more valuable when they increase in age, the first OLS assumption may likely be violated. Correlation is not necessarily causation.

b) The formula suggests a linear relationship is suggested between the vintage and price. This means that we can estimate that a bottle of wine made using grapes from year 0 (same year that they were picked) gives a vintage of 0 and would cost $1.75 (price = 1.75 + 5.5(0) = $1.75). After this baseline price, the price per bottle increases by an average of $5.5 for every year increase in the vintage (vintage increase by 1 year = price increase by $5.5).

c)

$R^2 = .77$
This R^2 is fairly high (R^2 is on a scale from 1 to 0), which means that the variation of the vintage may be able to explain 77% of the variation in the prices of wine bottles. Despite this metric, correlation is not causation between the vintage and price.

d)

10 years ago: Price =1.75 + 5.5(10) = 56.75

9 years ago: 1.75 + 5.5(9) = 51.25

Difference: 56.75 - 51.25 = 5.5

e)

Marginal effect = ($d$price/$d$vintage) = $\boldsymbol{\beta}_1$

Plugging into this formula gives the estimate of the parameter of the population slope $\boldsymbol{\beta}_1$ with OLS = $\hat{B}_1 = 5.5$

This 5.5 marginal effect estimate of the population parameter is the same as the value calculated in question (d) (difference between the two values). This answer holds for linear OLS regression containing one explanatory variable; however, this does not mean that it will always hold true for all general cases in the context of this problem.

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

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

f)

5 year predicted price:
1.75+5.5*5 = 29.25

10 year predicted price:
1.75+5.5*10 = 56.75

SE = 1.02; (When 2 SD from the mean)

When vintage = 5 and CI is a 95% interval:
29.25 - (2*1.02) = 27.21
29.25 + (2*1.02) = 31.29
It is expected that 95% of the sample data is between 27.21 and 31.49

When vintage = 10 and CI is a 95% interval:
56.75 - (2*1.02) = 54.71
56.75 + (2*1.02) = 58.79
It is expected that 95% of the sample data is between 54.71 and 58.79

From these results, we are able to reject the null hypothesis that the difference in the average price for a 10-year bottle and 50year bottle is greater than 40$.

Problem 4: Family size and consumption.

a)

gen foodshare = foodpq/totexppq

regress foodshare fam_size, robust

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

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

```
. regress foodshare fam_size, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =       4.39
                                                Prob > F          =     0.0363
                                                R-squared         =     0.0053
                                                Root MSE          =     .09863

------------------------------------------------------------------------------
             |               Robust
   foodshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fam_size |   .0047069   .0022455     2.10   0.036     .0003005    .0091132
       _cons |   .1653663   .0066056    25.03   0.000     .1524038    .1783288
------------------------------------------------------------------------------
```

Increasing the size of a family by 1, then the foodshare expenses increase by .0046069. This is equivalent to an average of about .46%. Although this is not a huge percentage, the coefficient is statistically significant at the 5% significance level since the p-value is .036, which is less than the .05 significance level.

Additionally, there is a positive relationship between the foodshare and the size of a family. This may mean that bigger families may spend more on food and thereby enjoy economies of scale by allocating more of their expenses to food.

b)

1 mother + 2 children = 3 family members

.1653663+.0047069(3) = .179487
≈ .18
≈ 18%

c)

gen lfam_size = log(fam_size)

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

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

regress foodshare lfam_size, robust

```
. regress foodshare lfam_size, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =       2.24
                                                Prob > F          =     0.1348
                                                R-squared         =     0.0027
                                                Root MSE          =     .09876

------------------------------------------------------------------------------
             |               Robust
   foodshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lfam_size |   .0085666   .0057233     1.50   0.135    -.0026645    .0197977
       _cons |     .17079   .0055826    30.59   0.000     .1598351    .1817449
------------------------------------------------------------------------------
```

The log of the family size coefficient is not statistically significant at the 5% significance level. This may mean that the relationship between family size and the share of food may be better demonstrated or shown by a linear model, rather than a log model.

A one unit increase in the logarithm of family size increases the food share expenditures (expected) by .9%.

These regression results differ from the results in part a), where a one unit increase in the family size increases food expenditures by about .46%.

d)

The R^2 is small for both regressions at .0027 and .0053. Many factors impact a household's expenses on food. This includes, but is not exclusive to size of the household. Other factors may also be at play that have a huge influence. Additionally, this data is also cross-sectional. A time-series set would be more smooth between transitions. Given these factors, there is reason to doubt that there is a relationship between family size and foodshare.

e)

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

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

gen exppc = totexppq/fam_size

regress foodshare fam_size if exppc<3000, robust

```
. regress foodshare fam_size if exppc<3000, robust

Linear regression                               Number of obs     =        532
                                                F(1, 530)         =       0.81
                                                Prob > F          =     0.3689
                                                R-squared         =     0.0018
                                                Root MSE          =     .10492

------------------------------------------------------------------------------
             |               Robust
   foodshare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fam_size |  -.0026552   .0029527    -0.90   0.369    -.0084557    .0031452
       _cons |   .2176878   .0102786    21.18   0.000      .197496    .2378797
------------------------------------------------------------------------------
```

Restricting the data in the sample to the households with less money (<$3,000), the fam_size coefficient is not statistically significant and the coefficient itself is negative at -.0026552. This contradicts the theory that is tested. This means that, according to the sample, the size of the family does not impact the share of income that is spent on food for poorer families (<$3,000) in the study. This does change my answer to the previous question.

f)

regress exppc fam_size, robust

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

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

```
. regress exppc fam_size, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =      94.96
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0495
                                                Root MSE          =     5004.8
```

exppc	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
fam_size	-749.0452	76.86834	-9.74	0.000	-899.8873	-598.2031
_cons	6129.586	304.9385	20.10	0.000	5531.192	6727.98

The expected per capita expenditures go down by about $750 when the size of a family increases by one person, which means that regressing the food share on fam_size allows us to illustrate the impact of per capita food expenditures on the food share. The fam_size coefficient is significant at the 1% level. Poorer households appear to make up larger families and spend a larger proportion of their income on foods, which may suggest that demand for food products is higher for people with lower incomes.

To help check the validity of the results, we can show that poorer households appear to make up larger families and spend a larger proportion of their income on foods when regressing the foodshare on its expenditure per capita.

regress foodshare exppc, robust

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

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

```
. regress foodshare exppc, robust

Linear regression                               Number of obs     =      1,000
                                                F(1, 998)         =      71.53
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1369
                                                Root MSE          =     .09187
```

foodshare	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
exppc	-7.13e-06	8.43e-07	-8.46	0.000	-8.78e-06	-5.48e-06
_cons	.2074863	.004812	43.12	0.000	.1980435	.2169292

This demonstrates the negative relationship of exppc and food share by the -7.13e-06 slope.

It is important to estimate the impact of increasing the size of the family on the share of food between houses with an equal number of exppc (ex. testing peer households).

The previous regressions from my former results only demonstrate a completely valid test if it is possible to estimate the impact of the size of a family on the share of food as exppc is held at a constant value and only allowing other variables to change. This would help to increase the validity of the results by standardizing the test across peer households or groups, essentially holding that variable constant.