22401 HW3

Bin Yu

Feb 03, 2025

Question 1

(a)

Three indicator variables were created for fertilizer groups 1, 2, and 3 using the following code:

```
gen F1 = (Fertilizer == 1)
gen F2 = (Fertilizer == 2)
gen F3 = (Fertilizer == 3)
```

Group 4, corresponding to the control, is represented by observations for which all three indicators are 0.

(b)

Estimated the following model:

$$\text{Yield}_{ij} = \mu_0 + \mu_1 \text{F1} + \mu_2 \text{F2} + \mu_3 \text{F3} + \epsilon_{ij}.$$

The Stata regression output is shown in Figure 7.

```
. regress Yield F1 F2 F3
```

Source	SS	df	MS	Number of obs	=	40
				F(3, 36)	=	5.14
Model	362.6	3	120.866667	Prob > F	=	0.0046
Residual	845.8	36	23.4944444	R-squared	=	0.3001
				Adj R-squared	=	0.2417
Total	1208.4	39	30.9846154	Root MSE	=	4.8471

Yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F1	6.8	2.167692	3.14	0.003	2.403717	11.19628
F2	.1	2.167692	0.05	0.963	-4.296283	4.496283
F3	5.1	2.167692	2.35	0.024	.7037167	9.496283
_cons	29.8	1.53279	19.44	0.000	26.69136	32.90864

Figure 1: Stata regression output for the model $\text{Yield}_{ij} = \mu_0 + \mu_1 \text{F1} + \mu_2 \text{F2} + \mu_3 \text{F3} + \epsilon_{ij}$.

(c)

We first test the null hypothesis that none of the fertilizers affect the corn yield:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = 0.$$

The alternative hypothesis is that at least one of the fertilizer effects is nonzero:

$$H_1 : \text{At least one } \mu_i \neq 0 \quad (i = 1, 2, 3).$$

Under the null hypothesis, the restricted model omits the predictors $F1$, $F2$, and $F3$. The F-statistic for testing these $q = 3$ restrictions is computed as

$$F = \frac{(RSS_R - RSS_F)/q}{RSS_F/(n-k)},$$

where:

- RSS_R is the residual sum of squares from the restricted model,
- RSS_F is the residual sum of squares from the full model (with $F1$, $F2$, and $F3$),
- n is the number of observations (40 in this case), and
- k is the number of parameters estimated in the full model (including the intercept, so $k = 4$).

Using Stata's `test` command, obtained:

$$F(3, 36) = 5.14, \quad p = 0.0046.$$

Since $p < 0.05$, we reject H_0 in favor of H_1, indicating that at least one fertilizer has a statistically significant effect on yield. The output is displayed in Figure 2.

```
. test F1 F2 F3

 ( 1)  F1 = 0
 ( 2)  F2 = 0
 ( 3)  F3 = 0

       F(  3,    36) =    5.14
            Prob > F =    0.0046
```

Figure 2: F-test output for $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$ vs. H_1: at least one $\mu_i \neq 0$.

(d)

To test whether the three fertilizers have equal effects:

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

The alternative hypothesis is that at least one pair of fertilizer effects differs:

$$H_1 : \text{Not all } \mu_i \text{ are equal} \quad (i = 1, 2, 3).$$

Null hypothesis can be restated in terms of differences:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{and} \quad \mu_1 - \mu_3 = 0.$$

In matrix form, if we let $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^\top$, the restrictions can be expressed as:

$$H_0 : R\boldsymbol{\mu} = \mathbf{0}, \quad \text{with} \quad R = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

The F-statistic is calculated as:

$$F = \frac{(RSS_R - RSS_F)/q}{RSS_F/(n-k)},$$

with $q = 2$ restrictions for this test. Here, RSS_R corresponds to the residual sum of squares under the restricted model (which imposes $\mu_1 = \mu_2 = \mu_3$), and RSS_F is the residual sum of squares from the full model (with separate effects). Stata's joint F-test yielded

$$F(2, 36) = 5.16, \quad p = 0.0107.$$

Since $p < 0.05$, we reject H_0 in favor of H_1, suggesting that the effects of the three fertilizers are not equal. The corresponding output is given in Figure 3.

```
. test F1 = F2 = F3

 ( 1)  F1 - F2 = 0
 ( 2)  F1 - F3 = 0

       F(  2,    36) =    5.16
            Prob > F =    0.0107
```

Figure 3: F-test output for $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_1: not all μ_i are equal.

(e)

To test whether applying any fertilizer (groups 1, 2, or 3) has a common effect relative to the control (group 4), a new indicator variable `anyfert` is created using the following code:

```
gen anyfert = (Fertilizer < 4)
```

We then estimate the regression model:

$$\text{Yield} = \alpha + \beta\, \texttt{anyfert} + \epsilon.$$

The hypotheses for testing the overall effect of any fertilizer are:

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0.$$

Under H_0, the restricted model excludes the effect of fertilizer (i.e., it reduces to a model with only the intercept). The F-test for this single restriction (with $q = 1$) is calculated as:

$$F = \frac{(RSS_R - RSS_F)/q}{RSS_F/(n-k)},$$

where:

- RSS_R is the residual sum of squares from the restricted model,
- RSS_F is the residual sum of squares from the full model (including `anyfert`),
- n is the number of observations (40), and
- k is the number of parameters estimated in the full model (here, $k = 2$, for intercept and β).

The Stata output gived the F-statistic of

$$F(1, 38) = 4.19 \quad \text{with} \quad p = 0.0476,$$

which corresponds to a t-test for β (since for one restriction, $t^2 = F$). Since $p < 0.05$, we reject H_0 in favor of H_1, indicating that applying any fertilizer results in a statistically significant change in corn yield. The output is displayed in Figure 4.

```
. regress Yield anyfert
```

Source	SS	df	MS	Number of obs	=	40
				F(1, 38)	=	4.19
Model	120	1	120	Prob > F	=	0.0476
Residual	1088.4	38	28.6421053	R-squared	=	0.0993
				Adj R-squared	=	0.0756
Total	1208.4	39	30.9846154	Root MSE	=	5.3518

Yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
anyfert	4	1.954213	2.05	0.048	.0439032	7.956097
_cons	29.8	1.692398	17.61	0.000	26.37392	33.22608

Figure 4: Regression output for testing $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ in the model Yield $= \alpha + \beta\,$`anyfert` $+ \epsilon$.

Question 2

(a)

```
graph matrix expend_per_k income kids_per_k
correlate expend_per_k income kids_per_k urban_per_k
```

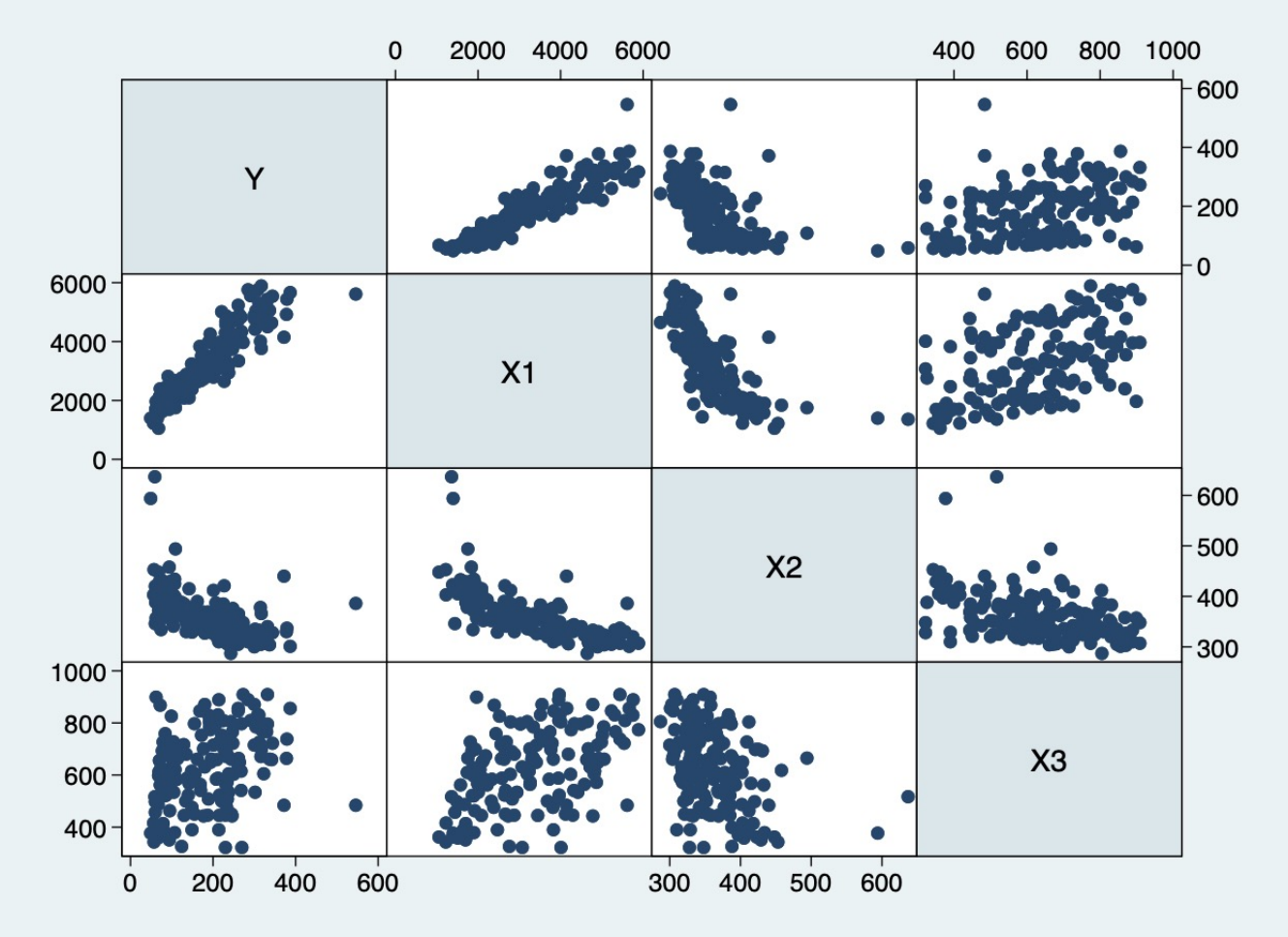

Figure 5: Scatter matrix plot for expend_per_k, income, and kids_per_k.

The correlation table indicates a strong positive correlation between income and expenditure, and a moderate correlation (negative between per capita school age children and expenditure, positive between urbanicity and expenditure).

There is also a negative correlation between income and per capita school age children, a positive correlation between income and urbanicity, a negative correlation between per capita school age children and urbanicity.

```
. correlate expend_per_k income kids_per_k urban_per_k
(obs=150)
```

	expend~k	income	kids_p~k	urban_~k
expend_per_k	1.0000			
income	0.9148	1.0000		
kids_per_k	-0.5541	-0.6963	1.0000	
urban_per_k	0.3549	0.4800	-0.3709	1.0000

Figure 6: Correlation Coefficients

(b)

We estimate the following multiple regression model:

$$Y = \beta_0 + \beta_1 \cdot \text{income} + \beta_2 \cdot \text{kids\_per\_k} + \beta_3 \cdot \text{urban\_per\_k} + \epsilon.$$

Stata Output

```
. regress expend_per_k income kids_per_k urban_per_k
```

Source	SS	df	MS	Number of obs	=	150
				F(3, 146)	=	294.66
Model	1124732.31	3	374910.77	Prob > F	=	0.0000
Residual	185764.329	146	1272.35841	R-squared	=	0.8582
				Adj R-squared	=	0.8553
Total	1310496.64	149	8795.27946	Root MSE	=	35.67

expend_per_k	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.080208	.0034482	23.26	0.000	.0733932	.0870228
kids_per_k	.2978954	.0843104	3.53	0.001	.1312689	.464522
urban_per_k	-.063398	.0220906	-2.87	0.005	-.1070566	-.0197393
_cons	-141.8916	40.59252	-3.50	0.001	-222.1165	-61.66675

Figure 7: Regression output for the model testing the overall effects of income, kids_per_k, and urban_per_k on expend_per_k.

The regression output shows that:

- The model has $F(3, 146) = 294.66$ with $p = 0.0000$, indicating that the overall regression is highly significant.
- The R^2 is 0.8582 (with an adjusted R^2 of 0.8553), meaning that approximately 85.5% of the variation in educational expenditures is explained by the predictors.

Overall Test: We test the joint null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0,$$

which implies that none of the predictors have an effect on Y. Given the highly significant overall F-statistic ($F(3, 146) = 294.66$, $p = 0.0000$), we reject H_0 and conclude that at least one of the predictors have a statistically significant effect on educational expenditures.

Individual Tests:

- **Income (X_1):** The coefficient for income is 0.080208 with a standard error of 0.0034482. The corresponding t-value is 23.26 and $p < 0.001$. This indicates that holding other variables constant, a one-unit increase in per capita school age kids is associated with an increase of 0.08 units in educational expenditures on average, which is significant.
- **per capita school age kids (X_2):** The coefficient for kids_per_k is 0.2978954 with a standard error of 0.0843104. The t-value is 3.53 with $p = 0.001$. This suggests that, holding other variables constant, a one-unit increase in per capita school age kids is associated with an increase of approximately 0.30 units in educational expenditures on average, which is statistically significant.

- **Urban per 1000 (X_3):** The coefficient for urban_per_k is -0.063398 with a standard error of 0.0220906. The t-value is -2.87 with $p = 0.005$. This implies that one unit increase in urbanicity is associated 0.063 unit decrease in educational expenditures, and the effect is significant.

(c)

We first incorporate the categorical variable `yearint` into the multivariable model using factor-variable notation so that the first level (year interval 1) serves as the baseline. The model is specified as

$$\text{expend\_per\_k} = \beta_0 + \beta_1 \cdot \text{income} + \beta_2 \cdot \text{kids\_per\_k} + \beta_3 \cdot \text{urban\_per\_k} + \delta_2 D_2 + \delta_3 D_3 + \epsilon,$$

where D_2 and D_3 are indicator (dummy) variables for year intervals 2 and 3, respectively.

Stata Output

```
. regress expend_per_k income kids_per_k urban_per_k i.yearint
```

Source	SS	df	MS	Number of obs	=	150
				F(5, 144)	=	205.23
Model	1149224.71	5	229844.942	Prob > F	=	0.0000
Residual	161271.932	144	1119.94397	R-squared	=	0.8769
				Adj R-squared	=	0.8727
Total	1310496.64	149	8795.27946	Root MSE	=	33.466

expend_per_k	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0687726	.0066163	10.39	0.000	.055695	.0818503
kids_per_k	.3633622	.080598	4.51	0.000	.2040543	.5226701
urban_per_k	-.0507201	.0242687	-2.09	0.038	-.098689	-.0027511
yearint						
2	43.08517	10.02345	4.30	0.000	23.27308	62.89727
3	43.09763	18.19469	2.37	0.019	7.134451	79.0608
_cons	-165.0158	38.40555	-4.30	0.000	-240.9273	-89.10441

Figure 8: Regression output including yearint dummies (baseline: interval 1).

The output reports the following coefficients for the year interval indicators:

- For year interval 2: $\delta_2 = 43.09$ with a standard error of 10.02, a t-statistic of 4.30, and a p-value less than 0.001.
- For year interval 3: $\delta_3 = 43.10$ with a standard error of 18.19, a t-statistic of 2.37, and a p-value of 0.019.

The coefficient $\delta_2 = 43.09$ indicates that, holding income, per capita school age children, and urbanicity constant, the average per capita educational expenditure in year interval 2 is estimated to be 43.09 units higher than in year interval 1 on average. Similarly, the coefficient $\delta_3 = 43.10$ implies that, under the same conditions, the average expenditure in year interval 3 is 43.10 units higher than that in the baseline on average. The statistical significance of these coefficients (with $p < 0.001$ for interval 2 and $p = 0.019$ for interval 3) provides strong evidence that both year intervals 2 and 3 have higher expenditures compared to year interval 1.

(d)

We now change the reference (baseline) level to interval 2.

Stata Output

```
. regress expend_per_k income kids_per_k urban_per_k ib2.yearint
```

Source	SS	df	MS	Number of obs	=	150
				F(5, 144)	=	205.23
Model	1149224.71	5	229844.942	Prob > F	=	0.0000
Residual	161271.932	144	1119.94397	R-squared	=	0.8769
				Adj R-squared	=	0.8727
Total	1310496.64	149	8795.27946	Root MSE	=	33.466

expend_per_k	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0687726	.0066163	10.39	0.000	.055695	.0818503
kids_per_k	.3633622	.080598	4.51	0.000	.2040543	.5226701
urban_per_k	-.0507201	.0242687	-2.09	0.038	-.098689	-.0027511
yearint						
1	-43.08517	10.02345	-4.30	0.000	-62.89727	-23.27308
3	.0124529	11.74007	0.00	0.999	-23.19268	23.21758
_cons	-121.9307	38.56866	-3.16	0.002	-198.1645	-45.69682

Figure 9: Regression output with yearint baseline set to interval 2.

In this parameterization, the model estimates the effects relative to interval 2. The output shows:

- For year interval 1: the coefficient is -43.09 (indicating that the expenditure of interval 1 on average is 43.09 units lower than interval 2, holding other predictors constant), and
- For year interval 3: the coefficient is 0.0125 (SE = 11.74, $t = 0.00$, $p = 0.999$).

To understand why coefficients and p-value of interval 3 changed, a formal test comparing the effects for intervals 2 and 3 can be performed by testing the null hypothesis that the coefficient for interval 3 is zero (i.e., that the effect in interval 3 is equal to that in the baseline, interval 2):

$$H_0 : \delta_3 = 0 \quad \text{vs.} \quad H_1 : \delta_3 \neq 0.$$

Stata Output

```
. test 2.yearint = 3.yearint

 ( 1)  2b.yearint - 3.yearint = 0

       F(  1,   144) =    0.00
            Prob > F =    0.9992
```

Figure 10: Testing output

The test result is:

$$F(1, 144) = 0.00, \quad p = 0.9992,$$

indicating no statistically significant difference between year intervals 2 and 3.

Therefore, the coefficient for year interval 3 changes dramatically when the baseline is switched from interval 1 to interval 2 because the coefficients are defined relative to the chosen baseline. In the first parameterization (baseline = 1), both intervals 2 and 3 are compared with interval 1 and yield similar positive differences (approximately 43 units) in expenditures. When interval 2 is chosen as the baseline, the difference between interval 3 and interval 2 in expenditures is approximately zero (0.0125), which is confirmed by the hypothesis test ($p \approx 1$). This reparameterization shows that intervals 2 and 3 do not differ significantly on their effects on education expenditures from each other, even though each is substantially different from interval 1 on the effects on education expenditures. .

(e)

To test if the effect of per capita school age children (X_2) on education expenditures is constant across year intervals, we create interaction terms:

$$Y = \beta_0 + \beta_1 \cdot \text{income} + \beta_2 \cdot \text{kids\_per\_k} + \beta_3 \cdot \text{urban\_per\_k} + \delta_2 D_2 + \delta_3 D_3 + \gamma_2 (D_2 \times \text{kids\_per\_k}) + \gamma_3 (D_3 \times \text{kids\_per\_k}) + \epsilon.$$

Stata Output

```
. regress expend_per_k income urban_per_k i.yearint##c.kids_per_k
```

Source	SS	df	MS	Number of obs	=	150
				F(7, 142)	=	188.80
Model	1183348.3	7	169049.757	Prob > F	=	0.0000
Residual	127148.339	142	895.410836	R-squared	=	0.9030
				Adj R-squared	=	0.8982
Total	1310496.64	149	8795.27946	Root MSE	=	29.923

expend_per_k	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0714467	.0059397	12.03	0.000	.0597051	.0831884
urban_per_k	-.0491426	.0217067	-2.26	0.025	-.0920527	-.0062325
yearint						
2	-214.2849	71.54385	-3.00	0.003	-355.7136	-72.85625
3	-403.7084	81.77559	-4.94	0.000	-565.3632	-242.0535
kids_per_k	.1500218	.0813904	1.84	0.067	-.0108717	.3109152
yearint#c.kids_per_k						
2	.6797114	.1936334	3.51	0.001	.2969349	1.062488
3	1.296788	.2355819	5.50	0.000	.8310869	1.762489
_cons	-84.67142	37.39907	-2.26	0.025	-158.6023	-10.74053

Figure 11: Regression output for the interaction model and test for $H_0 : \gamma_2 = \gamma_3 = 0$.

To test the effect of interaction terms, the null hypothesis is:

$$H_0 : \gamma_2 = \gamma_3 = 0 \quad \text{versus} \quad H_1 : \text{At least one } \gamma_i \neq 0.$$

If H_0 holds, the relationship between `kids_per_k` and Y is the same regardless of the value of `yearint`; that is, the slope for `kids_per_k` remains β_2 across all groups. If we reject H_0, it implies that the impact of `kids_per_k` on educational expenditure differs in at least one of the non-baseline year intervals.

```
. testparm i.yearint#c.kids_per_k

 ( 1)  2.yearint#c.kids_per_k = 0
 ( 2)  3.yearint#c.kids_per_k = 0

       F(  2,   142) =   19.05
            Prob > F =    0.0000
```

Figure 12: F-test

The F-test for the interaction yields $F(2, 142) = 19.05$ with $p = 0.0000$. Since the p-value is small, we reject H_0 and conclude that at least one of the interaction coefficients is significantly different from zero. This indicates that the effect of per capita school-age children on educational expenditures is not constant across the year intervals. In

other words, the relationship between `kids_per_k` and Y varies among different year intervals, indicating that the impact of this predictor is moderated by time.

(f)

Separate regressions are run for each year interval to obtain the effect of kids_per_k within each subgroup:

Stata Output

```
. by yearint, sort: regress expend_per_k income kids_per_k urban_per_k

-------------------------------------------------------------------------------
-> yearint = 1

      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(3, 46)        =     13.71
       Model |  10133.6139         3   3377.8713   Prob > F        =    0.0000
    Residual |  11332.3061        46  246.354481   R-squared       =    0.4721
-------------+----------------------------------   Adj R-squared   =    0.4376
       Total |   21465.92        49      438.08   Root MSE        =    15.696

------------------------------------------------------------------------------
expend_per_k |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0449327   .0076673     5.86   0.000     .0294992    .0603662
  kids_per_k |   .0662226   .0488342     1.36   0.182    -.0320756    .1645208
 urban_per_k |  -.0289536   .0192932    -1.50   0.140    -.0677888    .0098816
       _cons |  -11.40463   28.97616    -0.39   0.696    -69.73062    46.92137
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> yearint = 2

      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(3, 46)        =     32.96
       Model |  71922.5923         3  23974.1974   Prob > F        =    0.0000
    Residual |  33460.2877        46  727.397558   R-squared       =    0.6825
-------------+----------------------------------   Adj R-squared   =    0.6618
       Total |  105382.88        49  2150.67102   Root MSE        =     26.97

------------------------------------------------------------------------------
expend_per_k |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0808861   .0094739     8.54   0.000     .0618161     .099956
  kids_per_k |   .8184112   .1616285     5.06   0.000       .49307    1.143752
 urban_per_k |  -.1037663   .0350621    -2.96   0.005    -.1743426     -.03319
       _cons |  -289.1793   66.16956    -4.37   0.000    -422.3717   -155.9868
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> yearint = 3

      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(3, 46)        =     22.19
       Model |  109020.418         3  36340.1394   Prob > F        =    0.0000
    Residual |  75347.5819        46  1637.99091   R-squared       =    0.5913
-------------+----------------------------------   Adj R-squared   =    0.5647
       Total |     184368        49  3762.61224   Root MSE        =    40.472

------------------------------------------------------------------------------
expend_per_k |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0723853   .0116024     6.24   0.000     .0490308    .0957398
  kids_per_k |   1.552054   .3146716     4.93   0.000     .9186534    2.185456
 urban_per_k |   -.004269   .0513929    -0.08   0.934    -.1077175    .0991794
       _cons |   -556.568   123.1953    -4.52   0.000    -804.5472   -308.5889
------------------------------------------------------------------------------
```


Figure 13: Stratified regression outputs by year interval.

The output from these regressions is summarized below.

For Year Interval 1 (`yearint` = 1):

- `kids_per_k` coefficient: 0.0662
- Standard error: 0.0488
- t-value: 1.36
- p-value: 0.182
- 95% Confidence Interval: $[-0.0321,\ 0.1645]$

Interpretation: In year interval 1, the coefficient for `kids_per_k` is small (approximately 0.0662) and not statistically significant ($p = 0.182$). This suggests that, for year interval 1, there is little evidence that a one unit increase in the per-capita school-age children are associated with 0.0662 unit increase in educational expenditures on average, holding income and urbanicity constant.

For Year Interval 2 (`yearint` = 2):

- `kids_per_k` coefficient: 0.8184
- Standard error: 0.1616
- t-value: 5.06
- p-value: 0.000
- 95% Confidence Interval: $[0.4931,\ 1.1438]$

Interpretation: In year interval 2, the coefficient for `kids_per_k` is approximately 0.8184 and highly significant ($p < 0.001$). This indicates that for year interval 2, a one unit increase in the per-capita school-age children is associated with an increase of about 0.82 units in educational expenditures on average, holding the other variables constant.

For Year Interval 3 (`yearint` = 3):

- `kids_per_k` coefficient: 1.5521
- Standard error: 0.3147
- t-value: 4.93
- p-value: 0.000
- 95% Confidence Interval: $[0.9187,\ 2.1855]$

Interpretation: In year interval 3, the coefficient for `kids_per_k` is approximately 1.5521, and it is also statistically significant ($p < 0.001$). This suggests that in year interval 3, an additional one unit increase in per capita school-age children is associated with an increase of about 1.55 units in educational expenditures on average, holding other variables constant.

Summary The stratified analyses indicate that the impact of per capita school-age children on educational expenditures varies significantly in different year interval. In the earliest period (year interval 1), the effect is small and insignificant. However, in later intervals (2 and 3), the effect becomes larger and significant year by year.

(g)

Based on the stratified regressions reported in part (f), effects of urbanicity also differ by year interval:

Urban_per_k (Urbanicity):

- **Year Interval 1:** The coefficient is about -0.029 and not statistically significant ($p \approx 0.14$).
- **Year Interval 2:** The coefficient is more negative, approximately -0.104 and statistically significant ($p = 0.005$).
- **Year Interval 3:** The coefficient is close to zero (-0.0043) and not statistically significant ($p = 0.934$).

This pattern indicates that the effect of urbanicity on educational expenditures is not consistent across time. It appears to have a negative effect in interval 2, while its effect is minimal in intervals 1 and 3. Such a change suggests that the influence of urbanicity on educational expenditures may have altered over the time periods.

(h)

From the interaction model in Part (e), the slopes for kids_per_k can be written as follows:

$$Y = \beta_0 + \beta_1 \cdot \text{income} + \beta_2 \cdot \text{kids\_per\_k} + \beta_3 \cdot \text{urban\_per\_k} + \delta_2 D_2 + \delta_3 D_3 + \gamma_2 (D_2 \times \text{kids\_per\_k}) + \gamma_3 (D_3 \times \text{kids\_per\_k}) + \epsilon.$$

- For year interval 1 (baseline): Slope $= \beta_2$.
- For year interval 2: Slope $= \beta_2 + \gamma_2$.
- For year interval 3: Slope $= \beta_2 + \gamma_3$.

Estimated Values from the Interaction Model:

From the Stata output in part (e), we have the following estimates:

- β_2 (coefficient for `kids_per_k`) is estimated as approximately 0.1500.
- γ_2 (interaction coefficient for year interval 2) is estimated as approximately 0.6797.
- γ_3 (interaction coefficient for year interval 3) is estimated as approximately 1.2968.

Thus, the slopes are:

$$\begin{aligned} &\text{Interval 1: } 0.1500, \\ &\text{Interval 2: } 0.1500 + 0.6797 = 0.8297, \\ &\text{Interval 3: } 0.1500 + 1.2968 = 1.4468. \end{aligned}$$

Comparison with Stratified Models:

The stratified regressions reported in part (f) provided the following estimates for the coefficient on `kids_per_k`:

- **Year Interval 1:** 0.0662 (not statistically significant),
- **Year Interval 2:** 0.8184 (highly significant),
- **Year Interval 3:** 1.5521 (highly significant).

Comparing the slopes:

- **Interval 1:** The interaction model yields 0.1500 versus 0.0662 from the stratified regression.
- **Interval 2:** The interaction model gives 0.8297 compared to 0.8184 from the stratified regression. These values are very close.
- **Interval 3:** The interaction model estimates 1.4468 while the stratified model reports 1.5521. The estimates are also very close.

Overall Interpretation:

The slopes for `kids_per_k` derived from the interaction model are expressed in terms of the estimated parameters as:

$$\begin{aligned}\text{Slope in Interval 1} &= \beta_2 \approx 0.1500,\\ \text{Slope in Interval 2} &= \beta_2 + \gamma_2 \approx 0.8297,\\ \text{Slope in Interval 3} &= \beta_2 + \gamma_3 \approx 1.4468.\end{aligned}$$

These estimates are mainly consistent with the slopes obtained from the stratified regressions:

$$\begin{aligned}&\text{Interval 1 (Stratified)} \approx 0.0662,\\ &\text{Interval 2 (Stratified)} \approx 0.8184,\\ &\text{Interval 3 (Stratified)} \approx 1.5521.\end{aligned}$$

However, there are several reasons why these estimates might not be exactly the same:

1. **Joint Estimation vs. Separate Estimation:** In the joint model, the parameters are estimated simultaneously using the full sample. In contrast, the stratified regressions use only the subgroup data, which may result in different estimates due to smaller sample sizes and less precision.
2. **Control of Other Covariates:** In the joint model, the effects of other covariates (e.g., income and urban_per_k) are estimated using the full sample. However, when we are fitting the stratified model, we are assuming the effect are different for covariates among different groups, therefore, we are using samples from each group to estimate all the coefficients. Therefore, the estimation of the `kids_per_k` slope differently across groups than when each subgroup is analyzed separately.

This comparison affirms the finding that the impact of `kids_per_k` on educational expenditures is not constant across time, but rather increases in later year intervals.

Alternatively, if we fit the model in this way:

$$\begin{aligned}\text{expend\_per\_k} = \beta_0 &+ \beta_1\,\text{income} + \beta_2\,\text{kids\_per\_k} + \beta_3\,\text{urban\_per\_k}\\ &+ \delta_2\,D_2 + \delta_3\,D_3\\ &+ \gamma_2\left(D_2 \times \text{kids\_per\_k}\right) + \gamma_3\left(D_3 \times \text{kids\_per\_k}\right)\\ &+ \phi_2\left(D_2 \times \text{income}\right) + \phi_3\left(D_3 \times \text{income}\right)\\ &+ \theta_2\left(D_2 \times \text{urban\_per\_k}\right) + \theta_3\left(D_3 \times \text{urban\_per\_k}\right)\\ &+ \epsilon.\end{aligned}$$

Stata Output

```
. regress expend_per_k i.yearint##c.kids_per_k i.yearint##c.income i.yearint##c.urban_per_k
```

Source	SS	df	MS	Number of obs	=	150
				F(11, 138)	=	124.30
Model	1190356.46	11	108214.224	Prob > F	=	0.0000
Residual	120140.176	138	870.580983	R-squared	=	0.9083
				Adj R-squared	=	0.9010
Total	1310496.64	149	8795.27946	Root MSE	=	29.506

expend_per_k	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yearint						
2	-277.7746	90.59449	-3.07	0.003	-456.9074	-98.64183
3	-545.1634	105.0409	-5.19	0.000	-752.8612	-337.4656
kids_per_k	.0662226	.0918013	0.72	0.472	-.1152964	.2477416
yearint#c.kids_per_k						
2	.7521886	.1992323	3.78	0.000	.3582458	1.146131
3	1.485832	.2470931	6.01	0.000	.9972539	1.97441
income	.0449327	.0144134	3.12	0.002	.016433	.0734324
yearint#c.income						
2	.0359534	.017753	2.03	0.045	.0008503	.0710565
3	.0274526	.0167121	1.64	0.103	-.0055923	.0604975
urban_per_k	-.0289536	.0362684	-0.80	0.426	-.1006672	.0427601
yearint#c.urban_per_k						
2	-.0748127	.0527895	-1.42	0.159	-.1791937	.0295683
3	.0246845	.0521459	0.47	0.637	-.0784236	.1277927
_cons	-11.40463	54.47099	-0.21	0.834	-119.1103	96.30105

Figure 14: Alternative model

Based on our Stata output, we have:

- β_2 (coefficient for `kids_per_k`) = 0.0662226,
- γ_2 (interaction coefficient for year interval 2) = 0.7521886,
- γ_3 (interaction coefficient for year interval 3) = 1.485832.

Thus, the estimated slopes from the interaction model are:

$$\begin{aligned}
&\text{Year Interval 1: } 0.0662226,\\
&\text{Year Interval 2: } 0.0662226 + 0.7521886 = 0.8184112,\\
&\text{Year Interval 3: } 0.0662226 + 1.485832 = 1.5520546.
\end{aligned}$$

These values are exactly the same as those obtained from the stratified regressions:

- Year Interval 1: 0.0662226,
- Year Interval 2: 0.8184112,
- Year Interval 3: 1.5520546.

Why They Are the Same:
In the joint model, the group-specific slope for `kids_per_k` is computed as the sum of its overall (main) effect, β_2, and the corresponding interaction term (i.e. γ_2 for interval 2 and γ_3 for interval 3). This partitioning mirrors what is estimated in stratified regressions when all covariates are allowed to vary by group.

In order to achieve complete equivalence between the joint model and separate (stratified) regressions, one must include interactions between the group indicator (here, `yearint`) and *all* predictors in the model. This is because the stratified regressions estimate the effects of every predictor separately within each subgroup. If we only interact the variable of interest (here, `kids_per_k`) but assume that the effects of other covariates remain constant across groups, then the joint model relies on different assumptions compared to the stratified approach.

In the alternative case, we have included interactions for `kids_per_k`, `income`, and `urban_per_k` with `yearint`. As a result, the joint model replicates the subgroup-specific estimates that would be obtained by running separate regressions.

(i)

Use the following code:

```
* Estimate the interaction model
regress expend_per_k income urban_per_k i.yearint##c.kids_per_k

* Calculate means for income and urban_per_k
summarize income, meanonly
local mean_income = r(mean)
summarize urban_per_k, meanonly
local mean_urban = r(mean)

summarize kids_per_k, meanonly
local min_kids = r(min)
local max_kids = r(max)

margins i.yearint, at(income = ‘mean_income’ urban_per_k = ‘mean_urban’ kids_per_k = (‘min_kids’(5)‘max_ki

marginsplot, xdimension(kids_per_k) recast(line) noci ///
    plotopts(msymbol(O)) ///
    title("Predicted Expenditure vs. Number of School Age Kids by Year Interval") ///
    xtitle("Number of School Age Kids") ytitle("Predicted Expenditure per 1000") ///
    legend(title("Year Interval"))
```

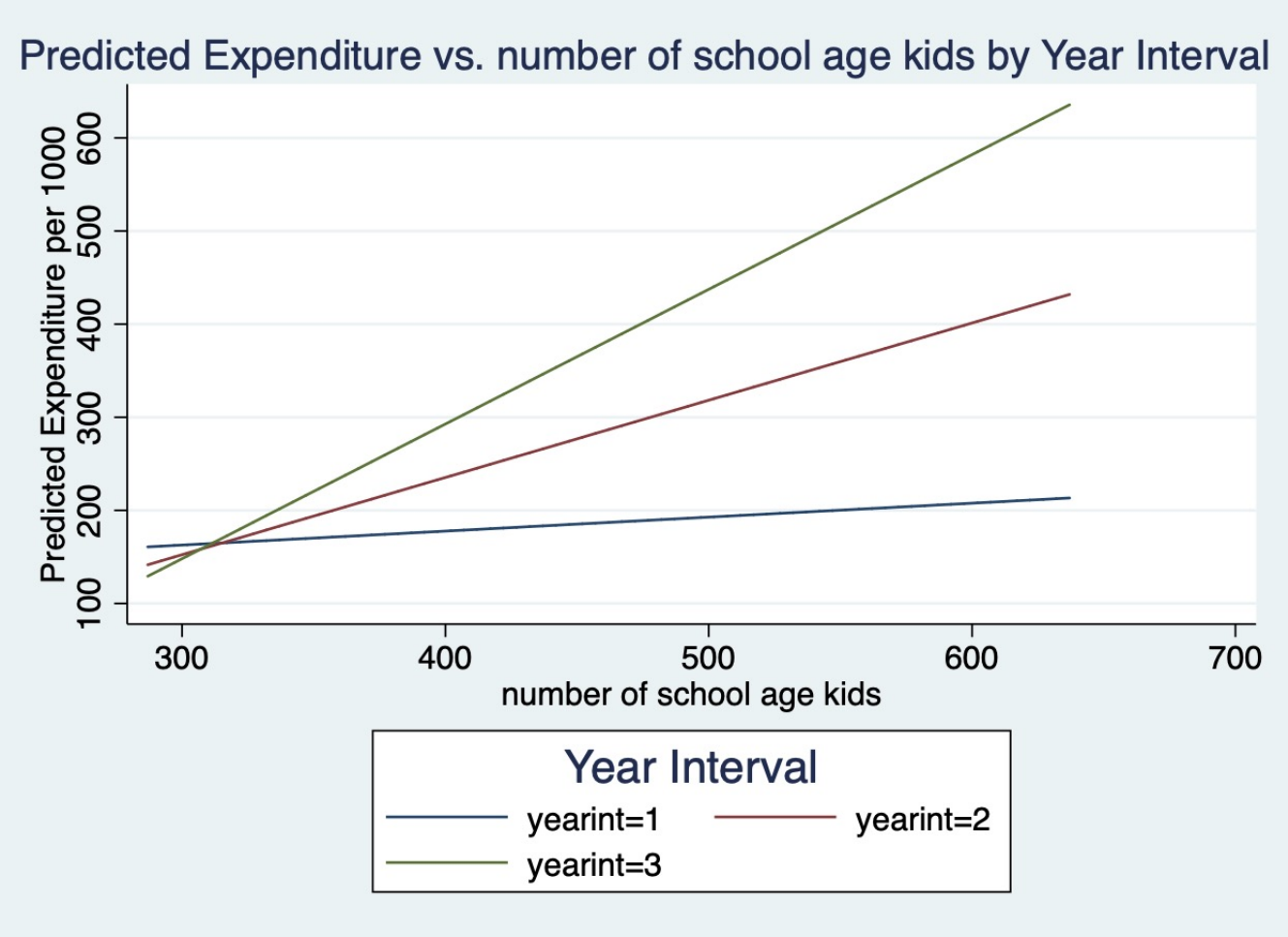

Figure 15: Predicted educational expenditures ($\hat{Y}$) for each year interval with covariates set at their means.

The graph shows that:

- In all year intervals, the predicted educational expenditures increase as the number of per capita school-age kids increases.
- The slope of the predicted line for later year intervals (e.g., intervals 2 and 3) is steeper compared to the baseline interval (year interval 1), indicating a stronger relationship between the number of school-age kids and expenditures in year interval 2 and 3.
- This pattern suggests that over time, the impact of per capita school age kids on educational expenditures has become stronger.

Question 3

(a)

For the model

$$E[Y \mid \text{Trt}, \text{EGFR}] = \beta_0 + \beta_1 \,\text{Trt} + \beta_2 \,\text{EGFR} + \beta_3 \,(\text{Trt} \times \text{EGFR}),$$

consider the two subgroups:

- When EGFR $= 0$ (EGFR-negative),

$$E[Y \mid \text{Trt}, \text{EGFR} = 0] = \beta_0 + \beta_1 \, \text{Trt}.$$

  Thus, the treatment effect (difference between Trt $= 1$ and Trt $= 0$) is

$$(\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

- When EGFR $= 1$ (EGFR-positive),

$$E[Y \mid \text{Trt}, \text{EGFR} = 1] = \beta_0 + \beta_1 \, \text{Trt} + \beta_2 + \beta_3 \, \text{Trt}.$$

  The treatment effect is

$$[\beta_0 + \beta_1 + \beta_2 + \beta_3] - [\beta_0 + \beta_2] = \beta_1 + \beta_3.$$

(b)

If the interaction term is not needed (i.e. $\beta_3 = 0$), then the model simplifies to

$$E[Y \mid \text{Trt}, \text{EGFR}] = \beta_0 + \beta_1 \, \text{Trt} + \beta_2 \, \text{EGFR}.$$

In this case, the treatment effect for both EGFR-negative and EGFR-positive patients is

$$(\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_2) = \beta_1.$$

(c)

For the model with the interaction term and when EGFR $= 1$:

$$E[Y \mid \text{Trt}, \text{EGFR}] = \beta_0 + \beta_1 \, \text{Trt} + \beta_2 \, \text{EGFR} + \beta_3 \, (\text{Trt} \times \text{EGFR}),$$

- Trt $= 0$:

$$E[Y \mid \text{Trt} = 0, \text{EGFR} = 1] = \beta_0 + \beta_2.$$

- Trt $= 1$:

$$E[Y \mid \text{Trt} = 1, \text{EGFR} = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3.$$

(d)

For the model without the interaction term (i.e., $\beta_3 = 0$) and when EGFR $= 1$:

$$E[Y \mid \text{Trt}, \text{EGFR}] = \beta_0 + \beta_1 \, \text{Trt} + \beta_2 \, \text{EGFR}.$$

- Trt $= 0$:

$$E[Y \mid \text{Trt} = 0, \text{EGFR} = 1] = \beta_0 + \beta_2.$$

- Trt $= 1$:

$$E[Y \mid \text{Trt} = 1, \text{EGFR} = 1] = \beta_0 + \beta_1 + \beta_2.$$