Econometrics

	Course Calendar
Week	Main Content
Week 7	Extension of Simple Regression: Functional Forms I
Week 8	Extension of Simple Regression: Functional Forms II
Week 9	Extension of Simple Regression: Functional Forms III
Week 10	Multiple Regression
Week 11	Multiple Regression: Problem of Inference
Week 12	Multiple Regression: Functional Forms
Week 13	Introduction to Dummy Variables
Week 14	Introduction to Dummy Variables and Regression Methods
Week 15	Regression with Dummy Variables: Hands-on-Exercise
Week 16	Application of Regression

Econometrics

Lecture 16. Dummy Variables in Regression (Continuation) &: Application of Regression

Geetha Rani Prakasam, Ph.D

Professor,

Recap

- Several applications of dummy variables in Multiple regression in explaining interesting economic aspects with examples
- Regression with a Mixture of Quantitative and Qualitative Regressors: The ANCOVA Models
- **The Dummy Variable Alternative to the Chow Test: Structural Differences in the US Savings and Income data**
- Interaction Effects Using Dummy Variables
- **The Use of Dummy Variables in Seasonal Analysis**

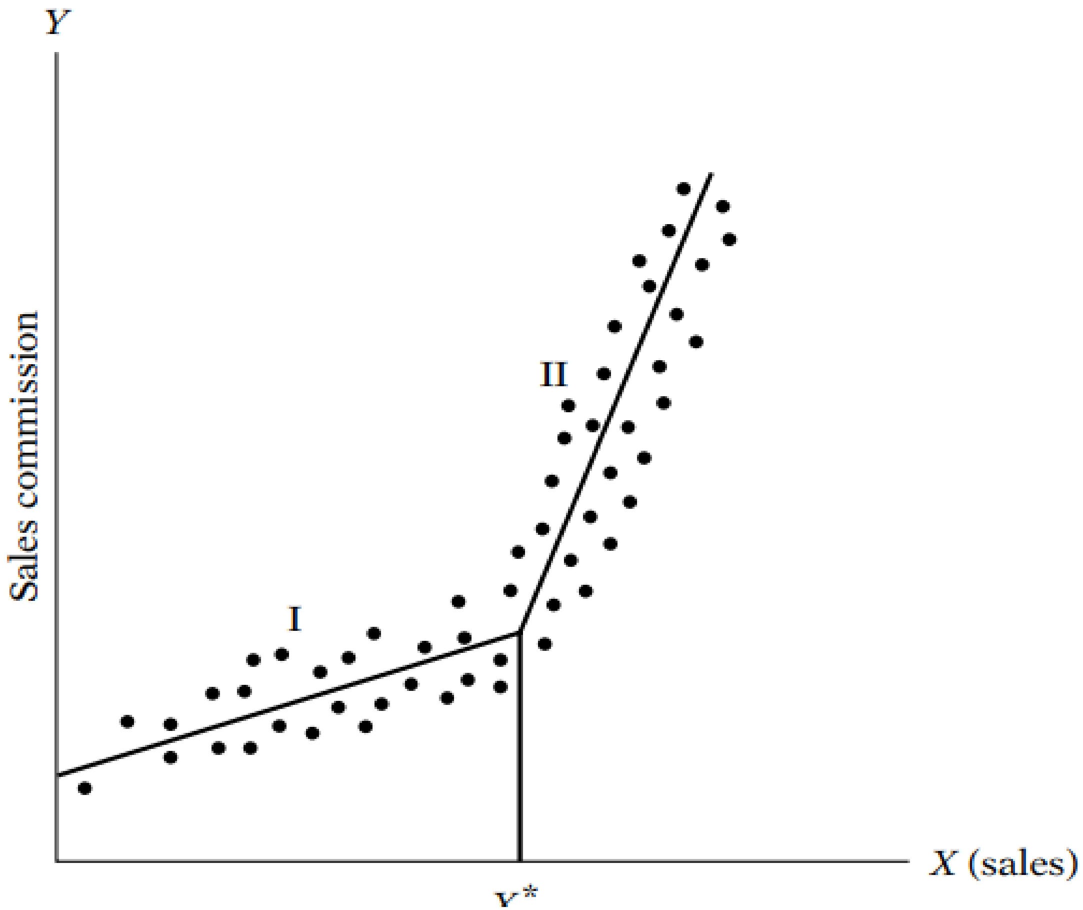
Outline

- Other DV applications
- interpretation of dummies in semilog models,
- piecewise linear regression models.
- Applications with Examples

9.8 Piecewise Linear Regression

- To illustrate yet another use of DV, consider Figure 9.5, which shows how a hypothetical company remunerates its sales representatives.
- It pays commissions based on sales in such a manner that up to a certain level, the *target, or threshold, level* $X*$*, there is* one (stochastic) commission structure and beyond that level another.
- *Note: Besides sales,* other factors affect sales commission. Assume that these other factors are represented by the stochastic disturbance term.

FIG 9.5 Hypothetical relationship between sales commission and sales volume



Source: Basic Econometrics,
Damodar Gujarti,
Page, 317

9.8 Piecewise Linear Regression

- More specifically, it is assumed that sales commission increases linearly with sales until the threshold level $X*$*, after which it continues to increase* linearly with sales but at a much steeper rate.
- Thus, we have a **piecewise linear regression** consisting of two linear pieces or segments, which are labeled I and II in Figure 9.5, and the commission function changes its slope at the threshold value.
- Given the data on commission, sales, and the value of the threshold level $X*$*, the technique of dummy variables* can be used to estimate the (differing) slopes of the two segments of the piecewise linear regression shown in Figure 9.5.

9.8 Piecewise Linear Regression

- We proceed as follows:
- $Y_i = \alpha_1 + \beta_1 X_i + \beta_2(X_i - X*)D_i + u_i$ **(9.8.1)**
- where Y_i = *sales commission;* X_i = *volume of sales generated by the sales person;* $X*$ = *threshold value of sales also known as a* ***knot (known in advance);*** $D = 1$ *if* $X_i > X* = 0$ if $X_i < X*$
- Assuming $E(u_i) = 0$, *we see at once that*
- $E(Y_i \mid D_i = 0, X_i, X*) = \alpha_1 + \beta_1 X_i$ **(9.8.2)**

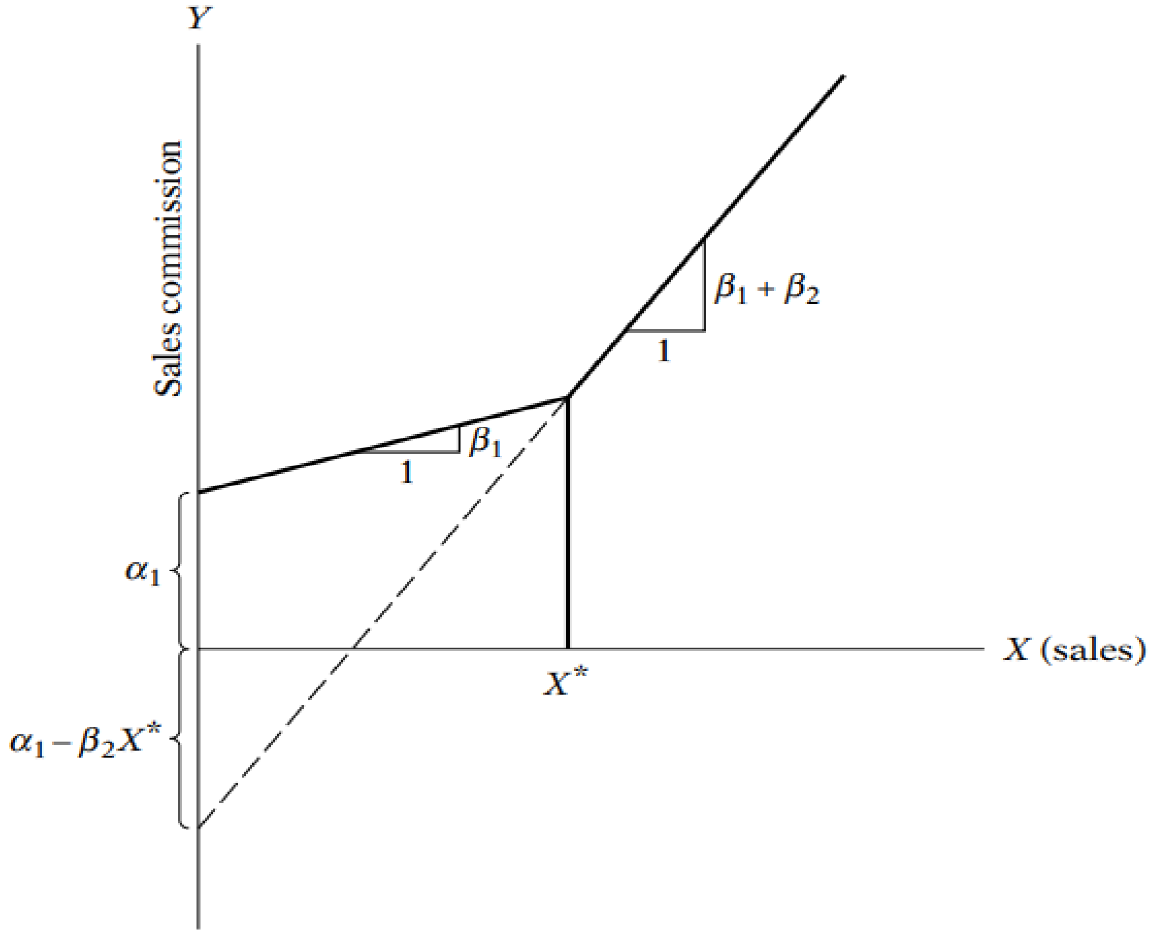
  which gives the mean sales commission up to the target level $X*$ *and*
- $E(Y_i \mid D_i = 1, X_i, X*) = \alpha_1 - \beta_2 X* + (\beta_1 + \beta_2)X_i$ **(9.8.3)**

  which gives the mean sales commission beyond the target level $X*$.

9.8 Piecewise Linear Regression

- Thus, β_1 *gives the slope of the regression line in segment I, and* $\beta_1 + \beta_2$ *gives the slope* of the regression line in segment II of the piecewise linear regression shown in Figure 9.5.
- A test of the hypothesis that there is no break in the regression at the threshold value *X* can* be conducted easily by noting the statistical significance of the estimated differential slope coefficient $\hat{\beta}_2$ *(see Figure 9.6 in the next slide).*
- Incidentally, the piecewise linear regression we have just discussed is an example of a more general class of functions known as **spline functions.**

FIG 9.6 Parameters of the piecewise linear regression



Source: Basic Econometrics, Damodar Gujarti, Page, 319

EX 9.7: Total Cost in Relation to Output

- Application of the piecewise linear regression, consider the hypothetical total cost–total output data given in Table 9.6.
- We are told that the total cost may change its slope at the output level of 5,500 units.
- Letting Y in Eq. (9.8.4) represent total cost and X total output, we obtain the following results:
- $\hat{Y}_i = -145.72 + 0.2791X_i + 0.0945(X_i - X^*_i)D_i$

  $t = \quad (-0.8245) \quad (6.0669) \quad (1.1447)$ **(9.8.4)**

  $R^2 = 0.9737 \quad X^* = 5{,}500$

EX 9.7: Total Cost in Relation to Output

- As these results show, the marginal cost of production is about 28 cents per unit and although it is about 37 cents (28 + 9) for output over 5,500 units, the difference between the two is not statistically significant because the dummy variable is not significant at 5 % level.
- For all practical purposes, one can regress total cost on total output, dropping the dummy variable.

The Interpretation of DV in Semilogarithmic Regressions

- In Lecture 8 we discussed the log–lin models, where the regressand is logarithmic and the regressors are linear.
- In such a model, the slope coefficients of the regressors give the semielasticity, that is, the percentage change in the regressand for a unit change in the regressor.
- This is only so if the regressor is quantitative. What happens if a regressor is a dummy variable?
- To be specific, consider the following model:
- $\ln Y_i = \beta_1 + \beta_2 D_i + u_i$ **(9.10.1)**

  where Y = hourly wage rate ($) and D = 1 for female and 0 for male.

The Interpretation of DV in Semilogarithmic Regressions

- How do we interpret such a model?
- Assuming $E(u_i) = 0$, we obtain:
- Wage function for male workers:

$$E(\ln Y_i \mid D_i = 0) = \beta_1 \qquad \textbf{(9.10.2)}$$

- Wage function for female workers:

$$E(\ln Y_i \mid D_i = 1) = \beta_1 + \beta_2 \qquad \textbf{(9.10.3)}$$

9.10. Some Technical Aspects of the DV Technique

- Therefore, the intercept β_1 gives the mean log hourly earnings and the "slope" coefficient gives the difference in the mean log hourly earnings of male and females.
- This is a rather an awkward way of stating things.
- But if we take the antilog of β_1, what we obtain is not the mean hourly wages of male workers, but their median wages.
- As you know, mean, median, and mode are the three measures of central tendency of a random variable.
- And if we take the antilog of (β_1+ β_2), we obtain the median hourly wages of female workers.

EX. 9.8: Logarithm of Hourly Wages in Relation to Gender

- To illustrate Eq. (9.10.1), we use the data that underlie Example 9.2. The regression results based on 528 observations are as follows:
- $\ln Y_i = 2.1763 - 0.2437D_i$

  $t = (72.2943)^* \ (-5.5048)^*$ **(9.10.4)**

  $R^2 = 0.0544$; where * indicates p values are practically zero.
- Taking the antilog of 2.1763, we find 8.8136 ($), which is the median hourly earnings of male workers, and taking the antilog of [(2.1763 − 0.2437) = 1.92857], we obtain 6.8796 ($), which is the median hourly earnings of female workers.
- Thus, the female workers' median hourly earnings are lower by about 21.94 percent compared to their male counterparts [(8.8136 − 6.8796)/8.8136].

EX. 9.8: Logarithm of Hourly Wages in Relation to Gender

- Interestingly, we can obtain semielasticity for a dummy regressor directly by the device suggested by Halvorsen and Palmquist.
- Take the antilog (to base e) of the estimated dummy coefficient and subtract 1 from it and multiply the difference by 100.
- Therefore, if you take the antilog of −0.2437, we will obtain 0.78366.
- Subtracting 1 from this gives −0.2163.
- After multiplying this by 100, we get −21.63 percent, suggesting that a female worker's (D = 1) median salary is lower than that of her male counterpart by about 21.63 percent, the same as we obtained previously, save the rounding errors.

What Happens If the Dependent Variable Is a DV?

- So far we have considered models in which the regressand is quantitative and the regressors are quantitative or qualitative or both.
- But there are occasions where the regressand can also be qualitative or dummy.
- Consider, for ex., the decision of a worker to participate in the labor force.
- The decision to participate is of the yes or no type, yes if the person decides to participate and no otherwise.
- Thus, the labor force participation variable is a DV.
- Of course, the decision to participate in the labor force depends on several factors, such as the starting wage rate, education, and conditions in the labor market (as measured by the unemployment rate), etc.

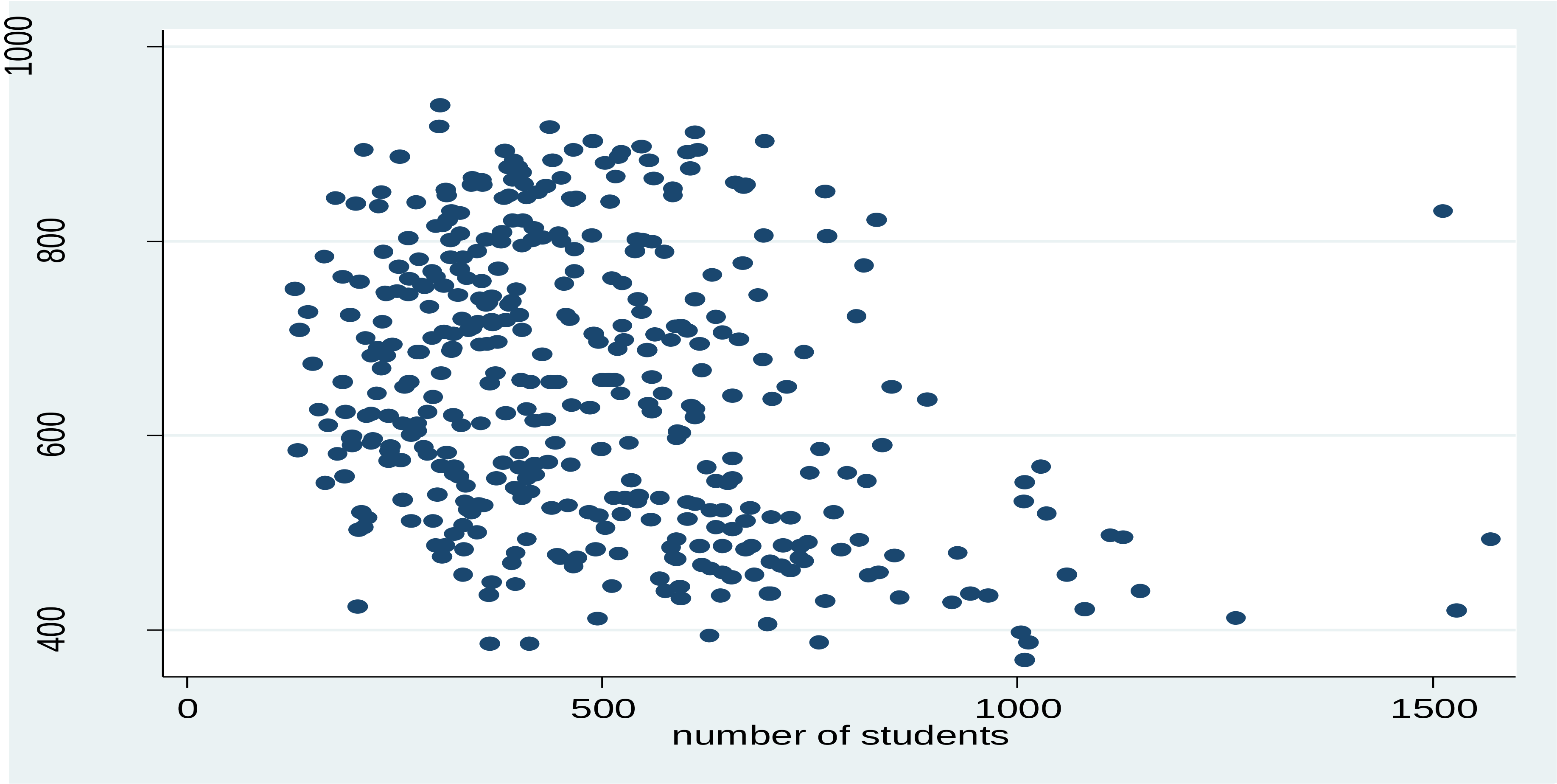
What Happens If the Dependent Variable Is a DV?

- Can we still use OLS to estimate regression models where the regressand is dummy?
- Yes, mechanically, we can do so. But there are several statistical problems that one faces in such models.
- And since there are alternatives to OLS estimation that do not face these problems, we will discuss this topic in logit and probit models.
- Models in which the regressand has more than two categories; for example, the decision to travel to work by car, bus, or train, or the decision to work part-time, full time, or not work at all.
- Such models are called polytomous dependent variable models in contrast to dichotomous dependent variable models in which the dependent variable has only two categories.

Simple and Multiple Linear Regression

- Diagnostics Examining Data: Plotting the data for Regression Analysis
- Specifying the Simple linear regression
- Estimation
- Understanding and Interpreting the Estimated Results
- Example data elemapi2.dta

twoway (scatter api00 enroll)



regress api00 enroll

```
. regress api00 enroll

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(1, 398)       =     44.83
       Model |  817326.293         1  817326.293   Prob > F        =    0.0000
    Residual |   7256345.7       398  18232.0244   R-squared       =    0.1012
-------------+----------------------------------   Adj R-squared   =    0.0990
       Total |     8073672       399  20234.7669   Root MSE        =    135.03

------------------------------------------------------------------------------
       api00 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      enroll |  -.1998674   .0298512    -6.70   0.000    -.2585532   -.1411817
       _cons |   744.2514   15.93308    46.71   0.000     712.9279    775.5749
------------------------------------------------------------------------------
```

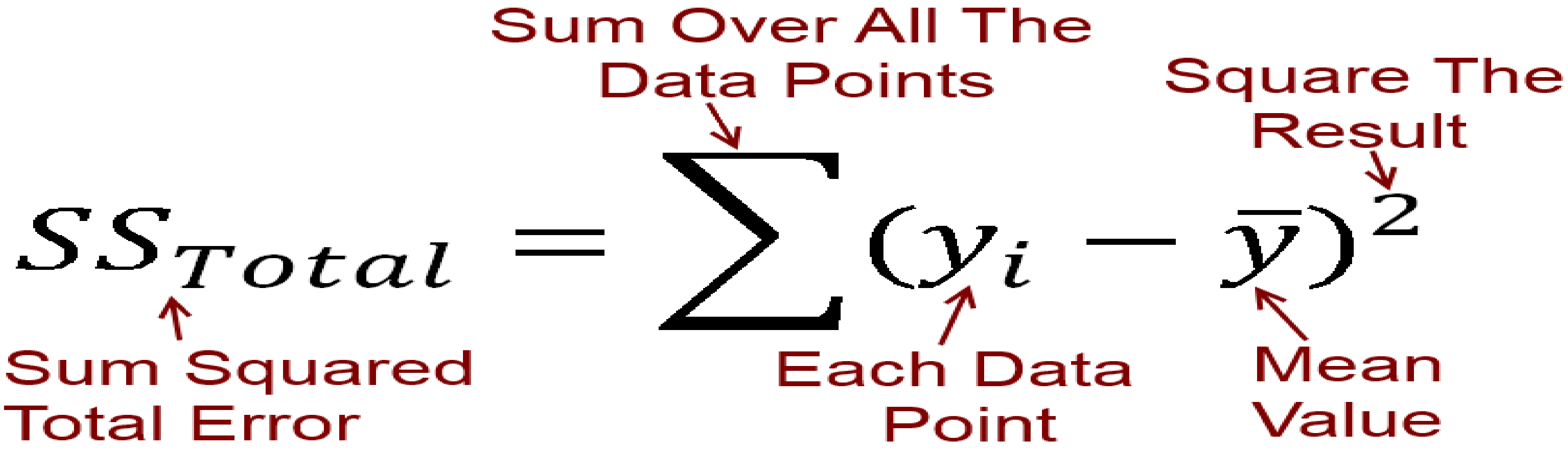
Anova Table: Source Column

Source	SS	df	MS	F
Model	817326.29	1	817326.3	44.83
Residual	7256345.70	398	18232.0	
Total	8073671.99	399	835558.3	

- Source: First Column
- This is the source of variance, Model, Residual, and Total.
- The Total variance is divided into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables.

Anova Table: Sums of Squares (SS) Column

- Sums of Squares (SS) for the Model & Residual add up to the Total Variance.



Anova Table: Sums of Squares (SS) Column

- SS_Total = The total variability around the mean. $\Sigma(Y - Ybar)^2$.
- SS_Residual = The sum of squared errors in prediction. $\Sigma(Y - Ypredicted)^2$.
- SS_Model = The improvement in prediction by using the predicted value of Y over just using the mean of Y. Hence, this would be the squared differences between the predicted value of Y and the mean of Y, $\Sigma(Ypredicted - Ybar)^2$.

- One more way: SS_Model = SS_Total – SS_Residual.
- SS_Total = SS_Model + SS_Residual.
- SS_Model / SS_Total is equal to .10, the value of R-Square.
- Because R-Square = proportion of the variance explained by the independent variables = SS_Model / SS_Total.

Anova Table: degrees of freedom (df) Column

- The total variance has N-1 degrees of freedom.
- In this case, there were N=400 observations, so the DF for total is 399.
- The model degrees of freedom corresponds to the number of predictors minus 1 (K-1).
- Including the intercept, there are 2 predictors, so the model has 2-1=1 degree of freedom.
- The Residual degrees of freedom is the DF total minus the DF model, 399 – 1 is 398.

Anova Table: Mean Squares(MS) Column

- Mean Squares = the Sum of Squares divided by their respective DF.
- For the Model, 817326.293 / 1 is equal to 817326.293.
- For the Residual, 7256345.7 / 398 equals 18232.0244.
- These are computed so that we can compute the F ratio;
- Dividing the Mean Square of Model by the Mean Square of Residual to test the significance of the predictor(s) in the model.

Second panel of the Estimated Regression output

Number of obs	=	400
F(1, 398)	=	44.83
Prob > F	=	0
R-squared	=	0.1012
Adj R-squared	=	0.099
Root MSE	=	135.03

F = MS(Model) / MS(Residual)

F = 817326.3 / 18232.0 = 44.83

- **R^2 = Explained Sum of Squares/Total Sum of Squares**
- **is coefficient of determination, it measures the proportion or percentage of the total variation in Y explained by the regression Model**

Second panel of the Estimated Regression output

- Adjusted R-square: As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance.
- We can add predictors to the model which would improve the ability of the predictors to explain the dependent variable, but some of this increase in R-square could be due to chance variation.
- The adjusted R-square attempts to yield a more better value to estimate the R-squared for the population.
- The value of R-square was .10, while the value of Adjusted R-square was .099.
- Adjusted R-squared is computed using the formula
- 1 – ((1-Rsq)*(N-1)/(N-k-1)).

Second panel of the Estimated Regression output

- From this formula, we can see that when the number of observations is small and the number of predictors is large, there will be a much greater difference between R-square and adjusted R-square, because the ratio (N-1)/(N-k-1) will be much greater than 1 and adjusted R-squared will be much smaller than unadjusted R-squared.
- By contrast, when the number of observations is very large compared to the number of predictors, the value of R-square and adjusted R-square will be much closer because the ratio (N-1)/(N-k-1) will approach 1.
- Root MSE is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error).

Third panel: Estimated Parameters

(1)	(2)	(3)	(4)	(5)	(6)	(7)
api00	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
enroll	-0.1998674	0.0298512	-6.7	0	-0.25855	-0.14118
_cons	744.2514	15.93308	46.71	0	712.9279	775.5749

- **First column shows the dependent variable at the top (api00) with the predictor variables below it (enroll).**
- **The last variable (_cons) represents the constant, also referred to as the Y intercept, the height of the regression line when it crosses the Y axis.**

Third panel: Estimated Parameters

- Second column - These are the values for the regression equation for predicting the dependent variable from the independent variable.
- The regression equation is presented in many different ways.
- One of the way is:- **Ypredicted (Y^)= b^0 + b^1*x1**
- The column of estimates (coefficients or parameter estimates, called coefficients) provides the values for b^0 and b^1 for this equation.
- Expressed as:- **api00Predicted = 744.25 – .20*enroll**
- This estimate tells about the relationship between the independent variable and the dependent variable.
- This estimate indicates the amount of increase in api00 that would be predicted by a 1 unit increase in the predictor (enroll).

Third panel: Estimated Parameters

- Second Column: If an independent variable is not significant, the coefficient is not significantly different from 0, which should be taken into account when interpreting the coefficient.
- Third Column: These are the standard errors associated with the coefficients.
- The standard error is used for testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error to obtain a t value (see 4^{th} column with t values and 5^{th} p values).
- The standard errors can also be used to form a confidence interval for the parameter, as shown in the last 2 columns of this table.
- Fourth column with the t value and in the fifth column p value tells about testing whether the enroll coefficients are significant.
- The coefficient (parameter estimate) is -0.20.
- So, for every unit increase in enroll, a -0.20 unit decrease in api00 is predicted.

Third panel: Estimated Parameters

- Sixth and seventh columns provide the t value and 2 tailed p value used in testing the null hypothesis that the coefficient/parameter is 0.
- If we use a 2 tailed test, then we would compare each p value to our pre-selected value of level of significance known as alpha (α).
- Coefficients having p values less than α are significant.
- For example, if we chose alpha to be 0.05, coefficients having a p value of 0.05 or less would be statistically significant (i.e. we can reject the null hypothesis and say that the coefficient is significantly different from 0).
- If we use a 1 tailed test (i.e., we predict that the parameter will go in a particular direction), then we can divide the p value by 2 before comparing it to our preselected alpha level.
- With a 2 tailed test and alpha of 0.05, we can reject the null hypothesis that the coefficient for enroll is equal to 0.

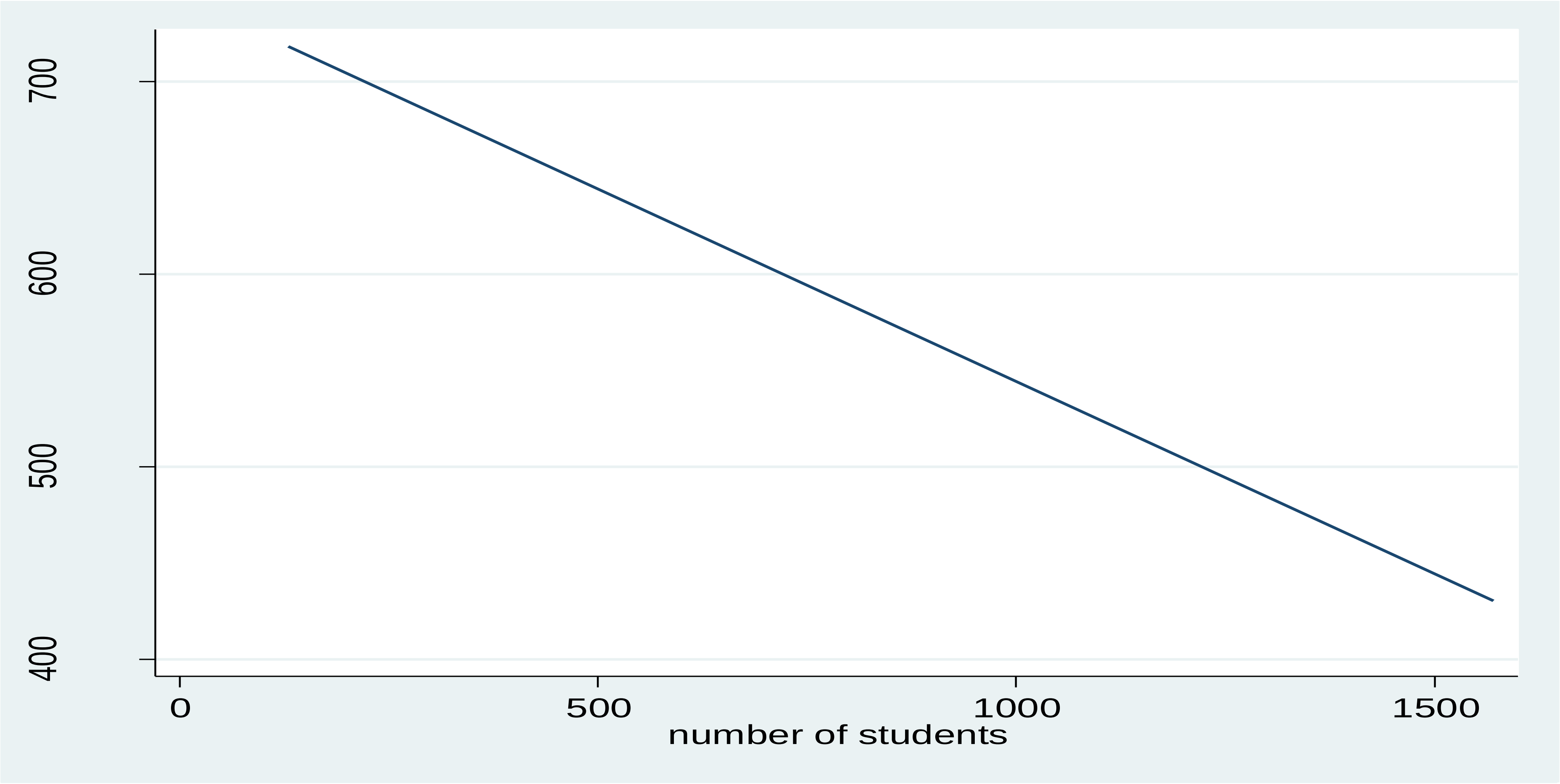
Third panel: Estimated Parameters

- The coefficient of -0.20 is significantly different from 0.
- The constant (_cons) is significantly different from 0 at the 0.05 alpha level. However, having a significant intercept is seldom interesting.
- **Column 6 and 7:-** This shows a 95% confidence interval for the coefficient.
- This is very useful as it helps us to understand how high and how low the actual population value of the parameter might be.
- Such confidence intervals help to put the estimate from the coefficient into perspective by seeing how much the value could vary.

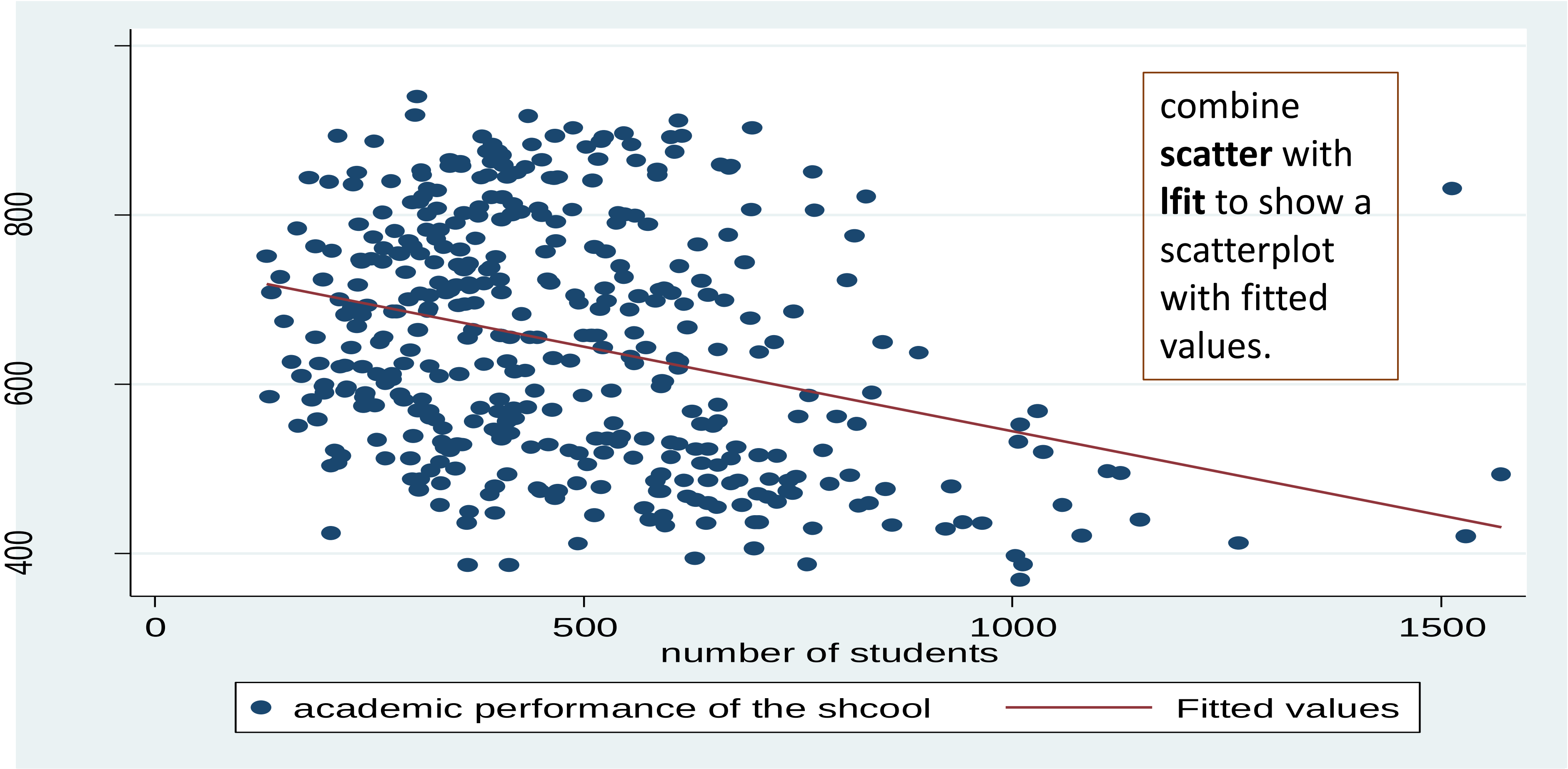
Predict the fitted values

- predict fv
- fv – fitted values
- twoway (lfit api00 enroll)

Predicted and Actual Dependent variable:
twoway (lfit api00 enroll)



twoway (scatter api00 enroll) (lfit api00 enroll)



Multiple Regression

- Multiple regression:- One outcome (dependent) variable and multiple predictors.
- We will regress the dependent variable, **api00**, on all of the predictor variables in the data set.
- Before estimating this multiple regression, a diagnostic check is required to know the correlation among these selected variables
- Correlation co-efficient matrix
- pwcorr api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll

Understanding the variables in the dataset

- des api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll

variable name	storage type	display format	value label	variable label
api00	int	%6.0g		academic performance of the shcool
ell	byte	%4.0f		english language learners
meals	byte	%4.0f		pct free meals
yr_rnd	byte	%4.0f	yr_rnd	year round school
mobility	byte	%4.0f		pct 1st year in school
acs_k3	byte	%4.0f		avg class size k-3
acs_46	byte	%4.0f		avg class size 4-6
full	byte	%8.2f		pct full credential
emer	byte	%4.0f		pct emer credential
enroll	int	%9.0g		number of students

Correlation co-efficient matrix

	api00	ell	meals	yr_rnd	mobility	acs_k3	acs_46	full	emer	enroll
api00	1									
ell	-0.7676	1								
meals	-0.9007	0.7724	1							
yr_rnd	-0.4754	0.4979	0.4185	1						
mobility	-0.2064	-0.0205	0.2166	0.0348	1					
acs_k3	0.171	-0.0557	-0.188	0.0227	0.0401	1				
acs_46	0.2329	-0.1733	-0.2131	-0.0421	0.1277	0.2708	1			
full	0.5744	-0.4848	-0.5276	-0.3977	0.0252	0.1606	0.1177	1		
emer	-0.5827	0.4722	0.533	0.4347	0.0596	-0.1103	-0.1245	-0.9057	1	
enroll	-0.3182	0.403	0.241	0.5918	0.105	0.1089	0.0283	-0.3377	0.3431	1

pwcorr api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll, star(.01)

	api00	ell	meals	yr_rnd	mobility	acs_k3	acs_46	full	emer	enroll
api00	1									
ell	-0.7676*	1								
meals	-0.9007*	0.7724*	1							
yr_rnd	-0.4754*	0.4979*	0.4185*	1						
mobility	-0.2064*	-0.0205	0.2166*	0.0348	1					
acs_k3	0.1710*	-0.0557	-0.1880*	0.0227	0.0401	1				
acs_46	0.2329*	-0.1733*	-0.2131*	-0.0421	0.1277	0.2708*	1			
full	0.5744*	-0.4848*	-0.5276*	-0.3977*	0.0252	0.1606*	0.1177	1		
emer	-0.5827*	0.4722*	0.5330*	0.4347*	0.0596	-0.1103	-0.1245	-0.9057*	1	
enroll	-0.3182*	0.4030*	0.2410*	0.5918*	0.105	0.1089	0.0283	-0.3377*	0.3431*	1

regress api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll

Source	SS	df	MS
Model	6740702.01	9	748966.89
Residual	1240707.78	385	3222.61761
Total	7981409.79	394	20257.3852

Number of obs	=	395
F(9, 385)	=	232.41
Prob > F	=	0.0000
R-squared	=	0.8446
Adj R-squared	=	0.8409
Root MSE	=	56.768

api00	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ell	-.8600707	.2106317	-4.08	0.000	-1.274203	-.4459382
meals	-2.948216	.1703452	-17.31	0.000	-3.28314	-2.613293
yr_rnd	-19.88875	9.258442	-2.15	0.032	-38.09219	-1.685309
mobility	-1.301352	.4362053	-2.98	0.003	-2.158995	-.4437088
acs_k3	1.3187	2.252683	0.59	0.559	-3.110401	5.747801
acs_46	2.032456	.7983213	2.55	0.011	.462841	3.602071
full	.609715	.4758205	1.28	0.201	-.3258169	1.545247
emer	-.7066192	.6054086	-1.17	0.244	-1.89694	.4837019
enroll	-.012164	.0167921	-0.72	0.469	-.0451798	.0208517
_cons	758.9418	62.28601	12.18	0.000	636.4785	881.4051

Output from Multiple regression

- As with the simple regression, we look to the p-value of the F-test to see if the overall model is significant.
- With a p-value of zero to four decimal places, the model is statistically significant.
- The R-squared is 0.8446, meaning that approximately 84% of the variability of **api00** is accounted for by the variables in the model.
- In this case, the adjusted R-squared indicates that about 84% of the variability of **api00** is accounted for by the model, even after taking into account the number of predictor variables in the model.
- The coefficients for each of the variables indicates the amount of change one could expect in **api00** given a one-unit change in the value of that variable, given that all other variables in the model are held constant.

Output from Multiple regression analysis

- For example, consider the variable **ell**. We would expect a decrease of 0.86 in the **api00** score for every one unit increase in **ell**, assuming that all other variables in the model are held constant.
- The interpretation of much of the output from the multiple regression is the same as it was for the simple regression.
- The major difference is in the third panel.
- First column in this third panel shows the dependent variable at the top (**api00**) with the predictor variables below it (**ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full emer** and **enroll**).
- The last variable (**_cons**) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis.

Output from Multiple regression

- These are the values for the regression equation for predicting the dependent variable from the independent variable.
- The regression equation is presented as: **Ypredicted = b0 + b1*x1 +b2*x2 + b3*x3 . . .**
- The column of estimates (coefficients or parameter estimates, or coefficients) provides the values for b0, b1, b2, b3, b4, b5, b6, b7, b8 and b9 for this equation.
- Expressed as:- **api00Predicted =778.83-.86*ell-2.95*meals-19.89*yr_rnd-1.30*mobility+1.32*acs_k3+2.03*acs_46+.61*full-.71*emer-.01*enroll**

Output from Multiple regression

- These estimates tell about the relationship between the independent variables and the dependent variable.
- These estimates tell the amount of increase in api00 that would be predicted by a 1 unit increase in the predictor.
- For the independent variables which are not significant, the coefficients are not significantly different from 0, which should be taken into account when interpreting the coefficients.
- (See the columns with the t-value and p-value about testing whether the coefficients are significant.)
- **ell** – The coefficient (parameter estimate) is -.86. So, for every unit increase in ell, a .86 unit decrease in **api00** is predicted. Or, for every increase of one percentage point of **api00**, **ell** is predicted to be lower by .86. This is significantly different from 0.

Output from Multiple regression

- **meals** – For every unit increase in **meals**, there is a 2.95 unit decrease in the predicted **api00**.
- **yr_rnd** – For every unit increase of **yr_rnd**, the predicted value of **api00** would be 19.89 units lower.
- **mobility** – For every unit increase in **mobility**, **api00** is predicted to be 1.30 units lower.
- **acs_k3** – For every unit increase in **acs_k3**, **api00** is predicted to be 1.32 units higher.
- **acs_46** – For every unit increase in **acs_46**, **api00** is predicted to be 2.03 units higher.

Output from Multiple regression analysis

- **full** – For every unit increase in **full**, **api00** is predicted to be .61 units higher.
- **emer** – For every unit increase in **emer**, **api00** is predicted to be .71 units lower.
- **enroll** – For every unit increase in **enroll**, **api00** is predicted to be .01 units lower.
- Third to seventh columns in the third panel help us to know the statistical significance of each these multiple coefficients.
- The interpretation is the same as in simple linear regression.

References

- Basic Econometrics by Domadar Gujarati
- https://stats.oarc.ucla.edu/stata/webbooks/reg/chapter1/regressionwith-statachapter-1-simple-and-multiple-regression/