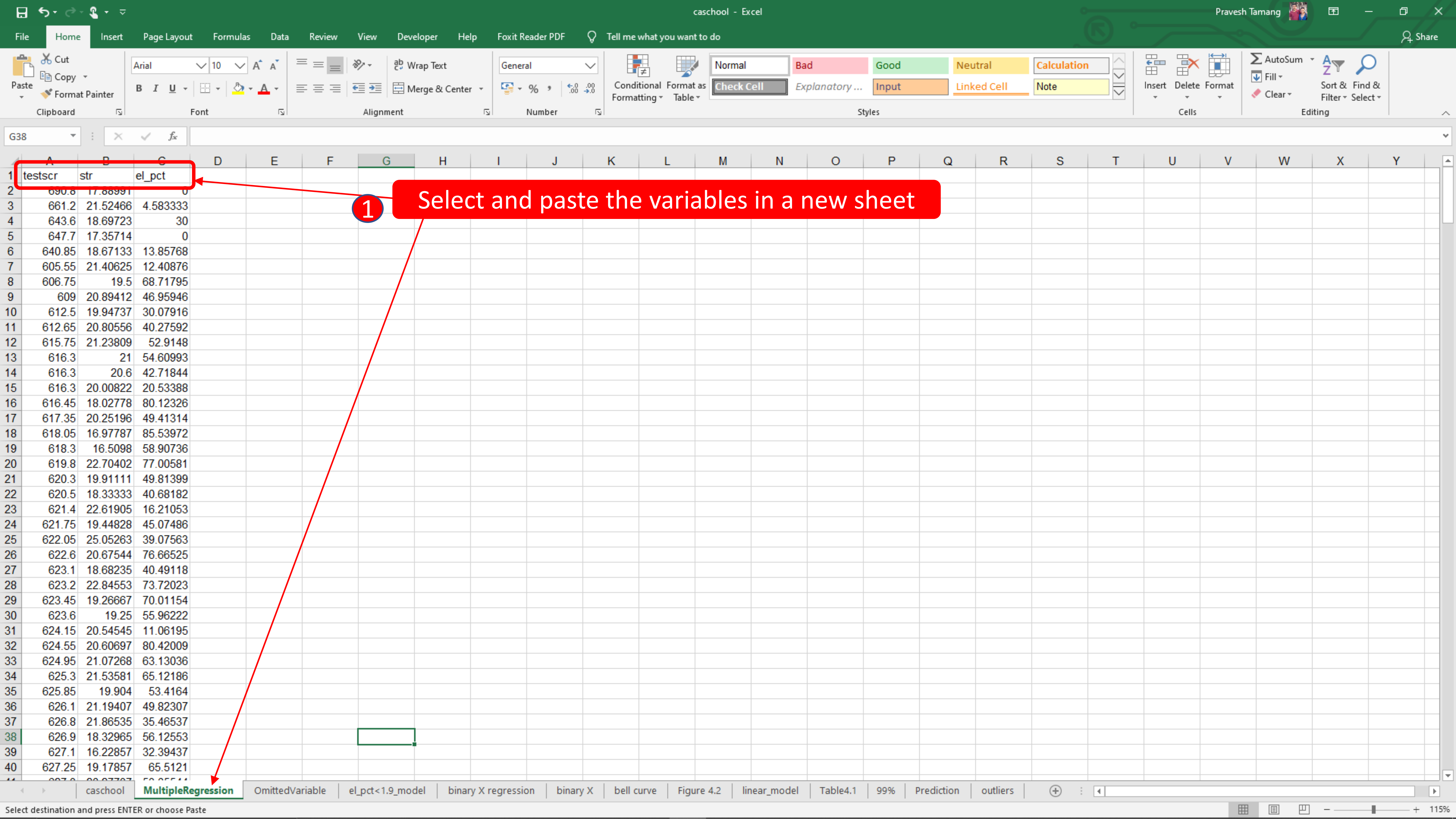
Introductory Econometrics

Using Excel

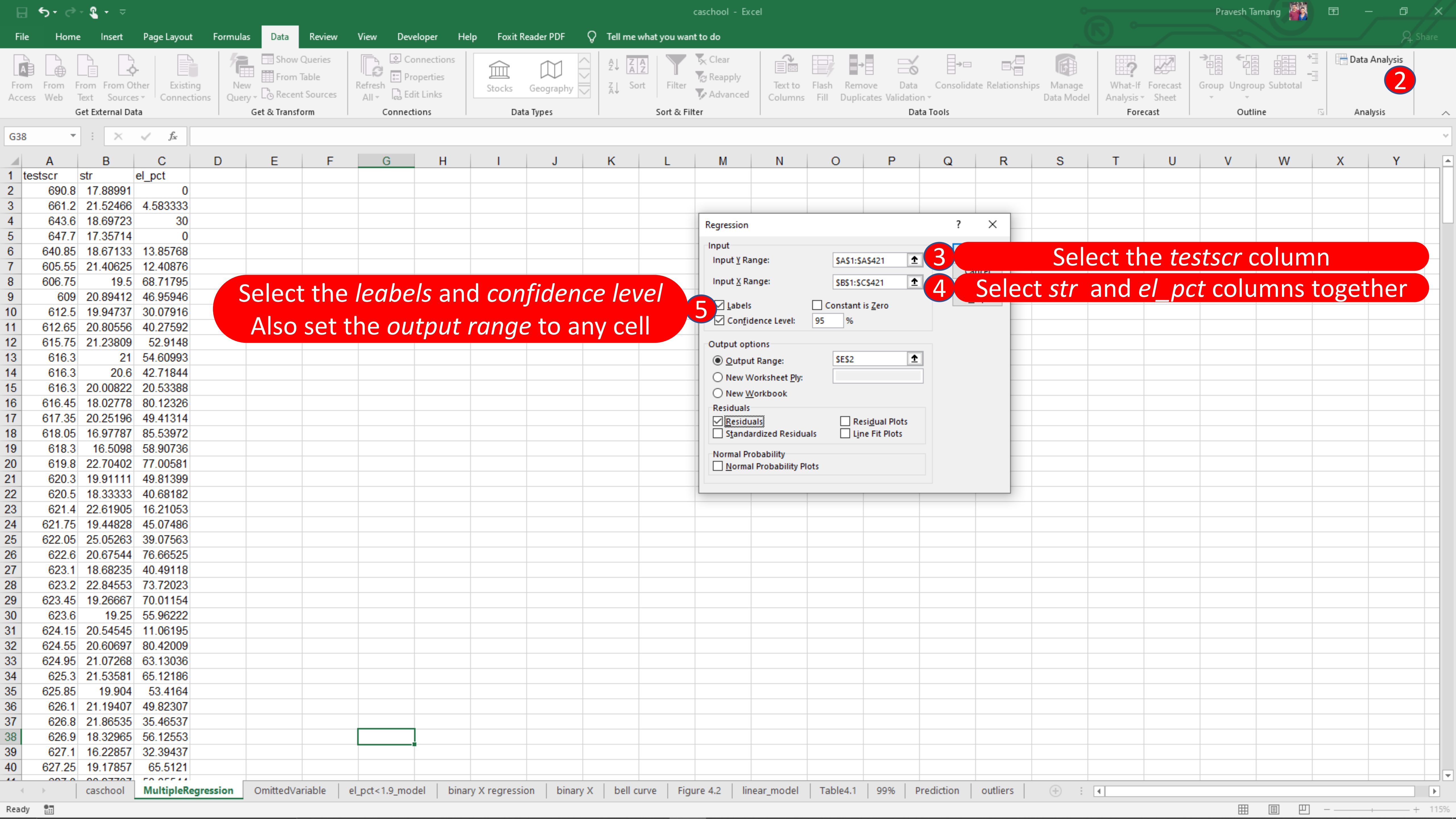
Linear Regression with Multiple Regressors

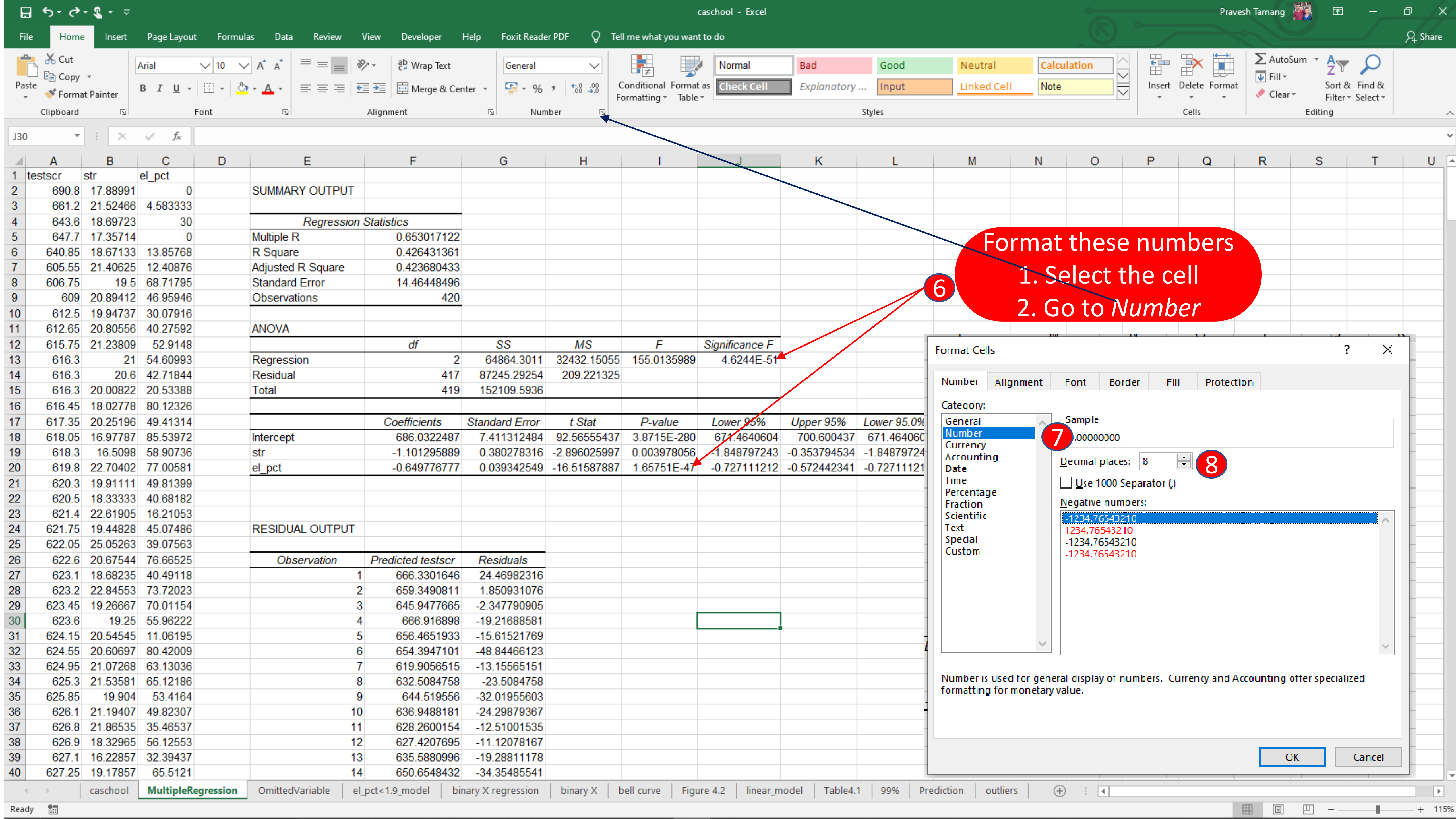
The Multiple Regression Model

- Our multiple regression model is $\widehat{TestScore} = \beta_0 + \beta_1 STR + \beta_2 PctEL$, where $PctEl$ is the percentage of students in the district who are English learners.



caschool - Excel
1
Select and paste the variables in a new sheet
testscr str el_pct
690.8 17.88991 0
661.2 21.52466 4.583333
643.6 18.69723 30
647.7 17.35714 0
640.85 18.67133 13.85768
605.55 21.40625 12.40876
606.75 19.5 68.71795
609 20.89412 46.95946
612.5 19.94737 30.07916
612.65 20.80556 40.27592
615.75 21.23809 52.9148
616.3 21 54.60993
616.3 20.6 42.71844
616.3 20.00822 20.53388
616.45 18.02778 80.12326
617.35 20.25196 49.41314
618.05 16.97787 85.53972
618.3 16.5098 58.90736
619.8 22.70402 77.00581
620.3 19.91111 49.81399
620.5 18.33333 40.68182
621.4 22.61905 16.21053
621.75 19.44828 45.07486
622.05 25.05263 39.07563
622.6 20.67544 76.66525
623.1 18.68235 40.49118
623.2 22.84553 73.72023
623.45 19.26667 70.01154
623.6 19.25 55.96222
624.15 20.54545 11.06195
624.55 20.60697 80.42009
624.95 21.07268 63.13036
625.3 21.53581 65.12186
625.85 19.904 53.4164
626.1 21.19407 49.82307
626.8 21.86535 35.46537
626.9 18.32965 56.12553
627.1 16.22857 32.39437
627.25 19.17857 65.5121
caschool
MultipleRegression
OmittedVariable
el_pct<1.9_model
binary X regression
binary X
bell curve
Figure 4.2
linear_model
Table4.1
99%
Prediction
outliers
Select destination and press ENTER or choose Paste


caschool - Excel
Data Analysis
2
Regression
Input
Input Y Range:
$A$1:$A$421
Input X Range:
$B$1:$C$421
Labels
Constant is Zero
Confidence Level:
95
%
Output options
Output Range:
$E$2
New Worksheet Ply:
New Workbook
Residuals
Residuals
Residual Plots
Standardized Residuals
Line Fit Plots
Normal Probability
Normal Probability Plots
3
Select the testscr column
4
Select str and el_pct columns together
5
Select the leabels and confidence level
Also set the output range to any cell
testscr
str
el_pct
caschool
MultipleRegression
OmittedVariable
el_pct<1.9_model
binary X regression
binary X
bell curve
Figure 4.2
linear_model
Table4.1
99%
Prediction
outliers


Format these numbers
1. Select the cell
2. Go to *Number*
6
7
8
Format Cells
Number
Alignment
Font
Border
Fill
Protection
Category:
General
Number
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom
Sample
Decimal places: 8
Use 1000 Separator (,)
Negative numbers:
-1234.76543210
1234.76543210
-1234.76543210
-1234.76543210
Number is used for general display of numbers. Currency and Accounting offer specialized formatting for monetary value.
OK
Cancel

caschool - Excel

testscr	str	el_pct
690.8	17.88991	0
661.2	21.52466	4.583333
643.6	18.69723	30
647.7	17.35714	0
640.85	18.67133	13.85768
605.55	21.40625	12.40876
606.75	19.5	68.71795
609	20.89412	46.95946
612.5	19.94737	30.07916
612.65	20.80556	40.27592
615.75	21.23809	52.9148
616.3	21	54.60993
616.3	20.6	42.71844
616.3	20.00822	20.53388
616.45	18.02778	80.12326
617.35	20.25196	49.41314
618.05	16.97787	85.53972
618.3	16.5098	58.90736
619.8	22.70402	77.00581
620.3	19.91111	49.81399
620.5	18.33333	40.68182
621.4	22.61905	16.21053
621.75	19.44828	45.07486
622.05	25.05263	39.07563
622.6	20.67544	76.66525
623.1	18.68235	40.49118
623.2	22.84553	73.72023
623.45	19.26667	70.01154
623.6	19.25	55.96222
624.15	20.54545	11.06195
624.55	20.60697	80.42009
624.95	21.07268	63.13036
625.3	21.53581	65.12186
625.85	19.904	53.4164
626.1	21.19407	49.82307
626.8	21.86535	35.46537
626.9	18.32965	56.12553
627.1	16.22857	32.39437
627.25	19.17857	65.5121

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.653017122
R Square	0.426431361
Adjusted R Square	0.423680433
Standard Error	14.46448496
Observations	420

ANOVA

	df	SS	MS	F	Significance F
Regression	2	64864.3011	32432.15055	155.0135989	0.00000000
Residual	417	87245.29254	209.221325		
Total	419	152109.5936			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	686.0322487	7.411312484	92.56555437	3.8715E-280	671.4640604	700.600437	671.4640604	700.600437
str	-1.101295889	0.380278316	-2.896025997	0.003978056	-1.848797243	-0.353794534	-1.848797243	-0.353794534
el_pct	-0.649776777	0.039342549	-16.51587887	0.00000000	-0.727111212	-0.572442341	-0.727111212	-0.572442341

RESIDUAL OUTPUT

Observation	Predicted testscr	Residuals
1	666.3301646	24.46982316
2	659.3490811	1.850931076
3	645.9477665	-2.347790905
4	666.916898	-19.21688581
5	656.4651933	-15.61521769
6	654.3947101	-48.84466123
7	619.9056515	-13.15565151
8	632.5084758	-23.5084758
9	644.519556	-32.01955603
10	636.9488181	-24.29879367
11	628.2600154	-12.51001535
12	627.4207695	-11.12078167
13	635.5880996	-19.28811178
14	650.6548432	-34.35485541

caschool | MultipleRegression | OmittedVariable | el_pct<1.9_model | binary X regression | binary X | bell curve | Figure 4.2 | linear_model | Table4.1 | 99% | Prediction | outliers

Omitted Variable Bias

SERIES IN DATA SET *caschool*:

DIST_CODE:	DISTRICT CODE;
READ_SCR:	AVG READING SCORE;
MATH_SCR:	AVG MATH SCORE;
COUNTY :	COUNTY;
DISTRICT:	DISTRICT;
GR_SPAN:	GRADE SPAN OF DISTRICT;
ENRL_TOT :	TOTAL ENROLLMENT;
TEACHERS:	NUMBER OF TEACHERS;
COMPUTER:	NUMBER OF COMPUTERS;
TESTSCR:	AVG TEST SCORE (= (READ_SCR+MATH_SCR)/2);
COMP_STU:	COMPUTERS PER STUDENT (= COMPUTER/ENRL_TOT);
EXPN_STU:	EXPENTITURES PER STUDENT ($'S);
STR:	STUDENT TEACHER RATIO (ENRL_TOT/TEACHERS);
EL_PCT:	PERCENT OF ENGLISH LEARNERS;
MEAL_PCT:	PERCENT QUALIFYING FOR REDUCED-PRICE LUNCH;
CALW_PCT:	PERCENT QUALIFYING FOR CALWORKS;
AVGINC:	DISTRICT AVERAGE INCOME (IN $1000'S);

Example #1: Percentage of English learners

- What was our previous assessment about $testscr$?
  - We estimated the linear model $\widehat{testscr} = 698.9 - 2.27 \times str$
  - Interpretation: Class with higher str i.e. larger class size tend to have **lower** $testscr$.

- Reasons for considering el_pct as omitted variable;
  - Large immigration population in California
  - Students in the district who are still learning English.
    - Students who are still learning English might perform worse on standardized tests as compared to native English speakers.
    - The larger class sizes str might also have students still learning English. Therefore, considering the OLS estimate $\widehat{testscr} = 698.9 - 2.27 \times str$, the policy would be considering lowering the str value to improve $testscr$. But this might not solve the problem because of the English learning students in the class.

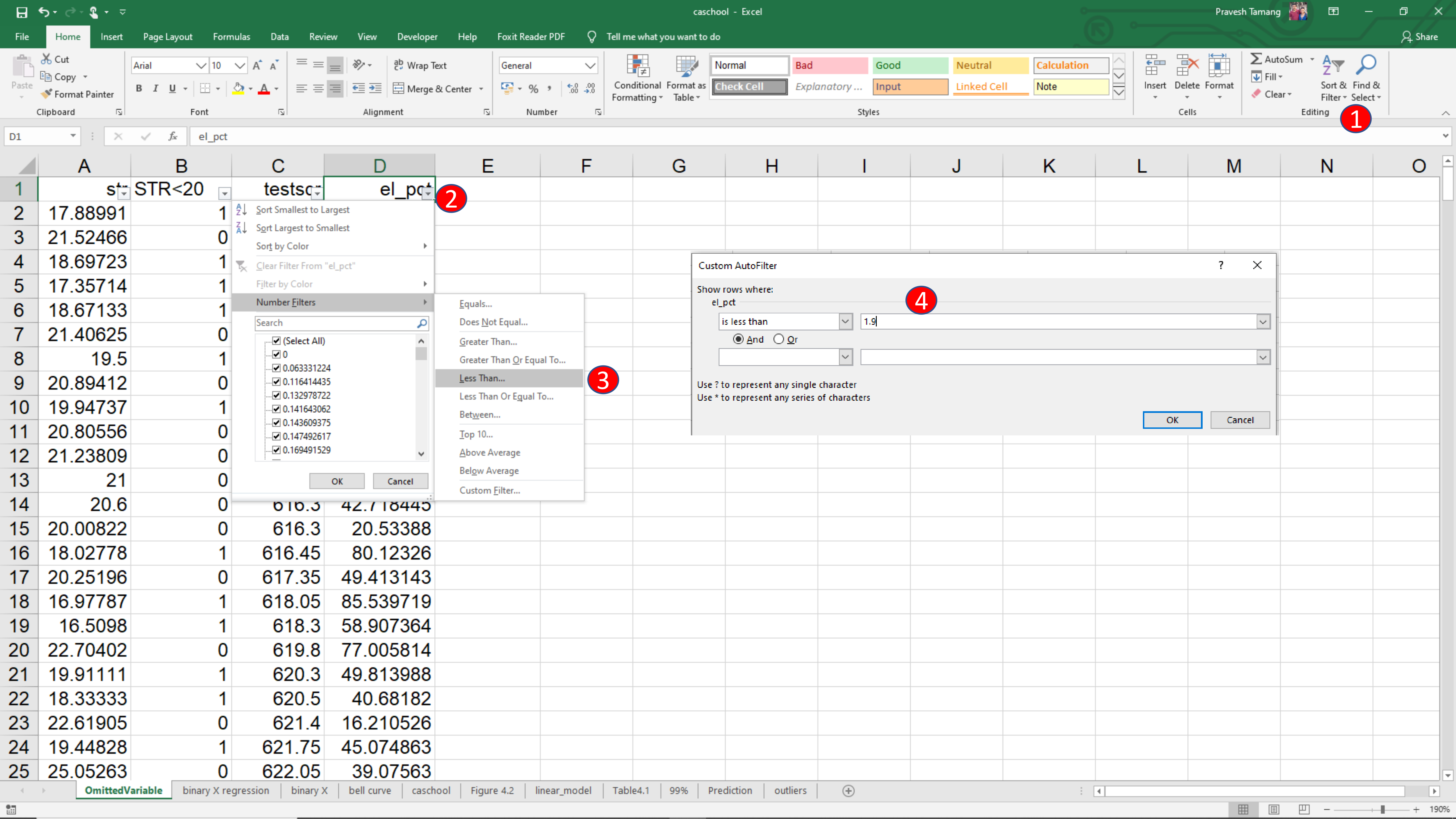
Omitted Variable Bias in Regression with a Single Regressor

Omitted variable bias is the bias in the OLS estimator that arises when the regressor, X, is correlated with an omitted variable. For omitted variable bias to occur,
two conditions must be true:
1. X is correlated with the omitted variable.
2. The omitted variable is a determinant of the dependent variable, Y

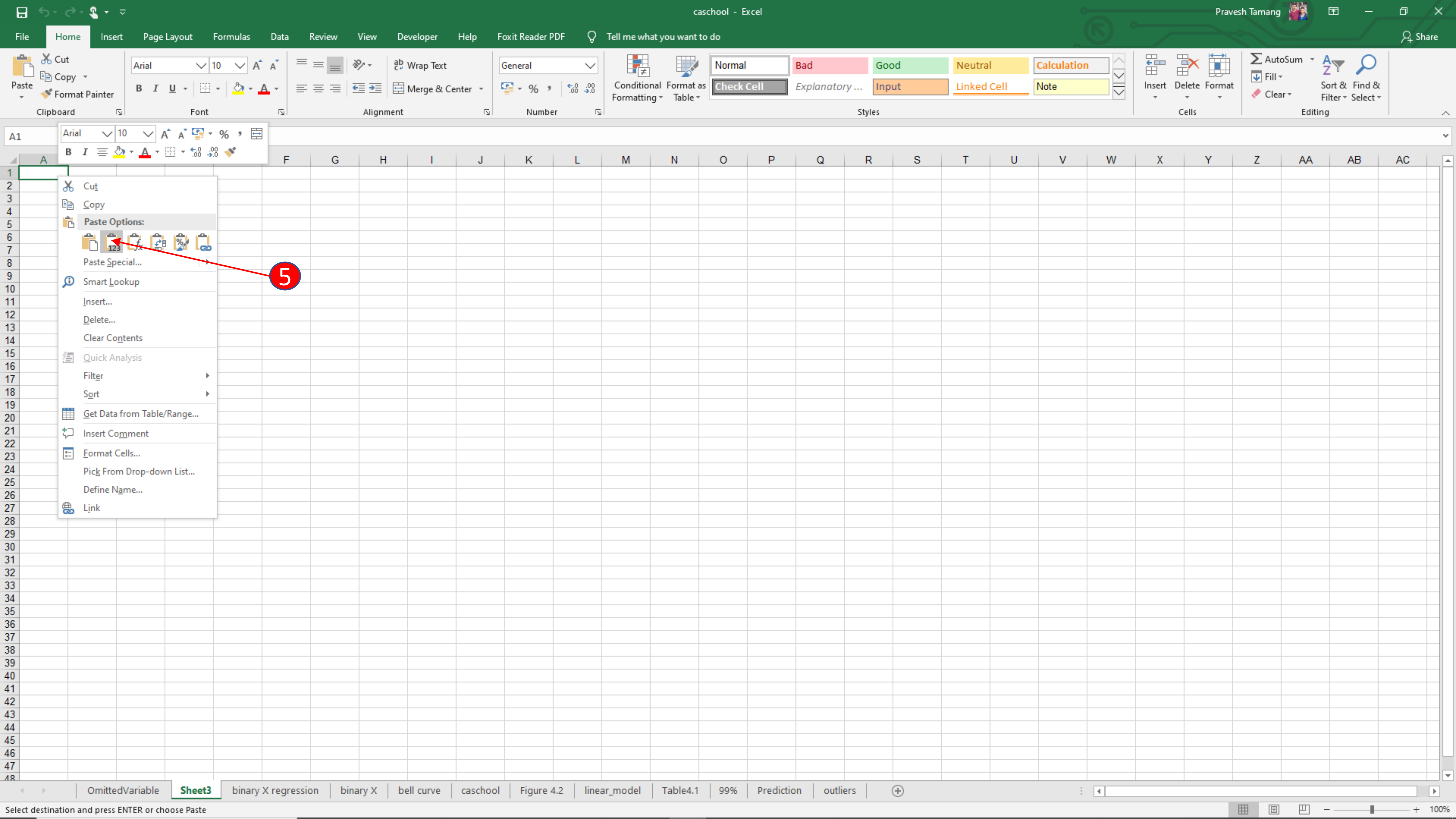
Addressing Omitted Variable Bias by Dividing the Data into Groups

- Relationship between class size $\boldsymbol{str}$ and test scores $\boldsymbol{testscr}$ within districts with comparable percentages of English learners.
  - First, the districts are broken into four categories that correspond to the quartiles of the distribution of the percentage of English learners across districts.
  - Second, within each of these four categories, districts are further broken down into two groups, depending on whether the student–teacher ratio is small $str < 20$ or large $str \geq 20$.
  -

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	n	Average Test Score	n	Difference	t-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68



caschool - Excel
Pravesh Tamang
File Home Insert Page Layout Formulas Data Review View Developer Help Foxit Reader PDF Tell me what you want to do
Share
Sort & Filter
1
D1
el_pct
st STR<20 testsc el_pct
2
Sort Smallest to Largest
Sort Largest to Smallest
Sort by Color
Clear Filter From "el_pct"
Filter by Color
Number Filters
Search
(Select All)
0
0.063331224
0.116414435
0.132978722
0.141643062
0.143609375
0.147492617
0.169491529
OK
Cancel
Equals...
Does Not Equal...
Greater Than...
Greater Than Or Equal To...
Less Than...
3
Less Than Or Equal To...
Between...
Top 10...
Above Average
Below Average
Custom Filter...
Custom AutoFilter
Show rows where:
el_pct
is less than
1.9
4
And
Or
Use ? to represent any single character
Use * to represent any series of characters
OK
Cancel
OmittedVariable binary X regression binary X bell curve caschool Figure 4.2 linear_model Table4.1 99% Prediction outliers


caschool - Excel
Pravesh Tamang
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Help
Foxit Reader PDF
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Font
Alignment
Number
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Normal
Bad
Good
Neutral
Calculation
Check Cell
Explanatory ...
Input
Linked Cell
Note
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Cut
Copy
Paste Options:
Paste Special...
Smart Lookup
Insert...
Delete...
Clear Contents
Quick Analysis
Filter
Sort
Get Data from Table/Range...
Insert Comment
Format Cells...
Pick From Drop-down List...
Define Name...
Link
5
OmittedVariable
Sheet3
binary X regression
binary X
bell curve
caschool
Figure 4.2
linear_model
Table4.1
99%
Prediction
outliers
Select destination and press ENTER or choose Paste
100%

	B	C	D
1	STR<20	testscr	el_pct
2	1	690.799988	0
3	1	647.700012	0
4	1	635.450012	0
5	1	635.599976	0
6	1	636.599976	1.79533219
7	1	637.099976	0.95923263
8	0	637.349976	0.66666669
9	1	637.949951	0
10	1	641.549988	1.38328528
11	1	643.5	0.92592591
12	1	644.199951	0.13297872
13	1	645.25	0
14	1	645.549988	0
15	1	645.75	0
16	0	646	0.75187969
17	1	646.400024	0
18	1	647.299988	0
19	1	650.650024	0
20	1	650.900024	0
21	1	650.900024	0
22	1	651.349976	0
23	1	651.849976	0
24	1	652.349976	0
25	0	652.400024	1.76470602
26	1	652.400024	0
27	1	652.5	0
28	1	653.949951	0
29	1	654.199951	1.3691932

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.026663625
R Square	0.000710949
Adjusted R Square	-0.009183002
Standard Error	14.98551621
Observations	103

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	16.13660078	16.13660078	0.071856927	0.789198743
Residual	101	22681.13531	224.5656961		
Total	102	22697.27191			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	665.3703772	2.88396394	230.7138338	2.5429E-139	659.6493685	671.0913858	659.6493685	671.0913858
STR<20	-0.899986431	3.357388836	-0.268061424	0.789198743	-7.560142753	5.760169892	-7.560142753	5.760169892

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

Repeat steps 1 to 6 for the rest of these values

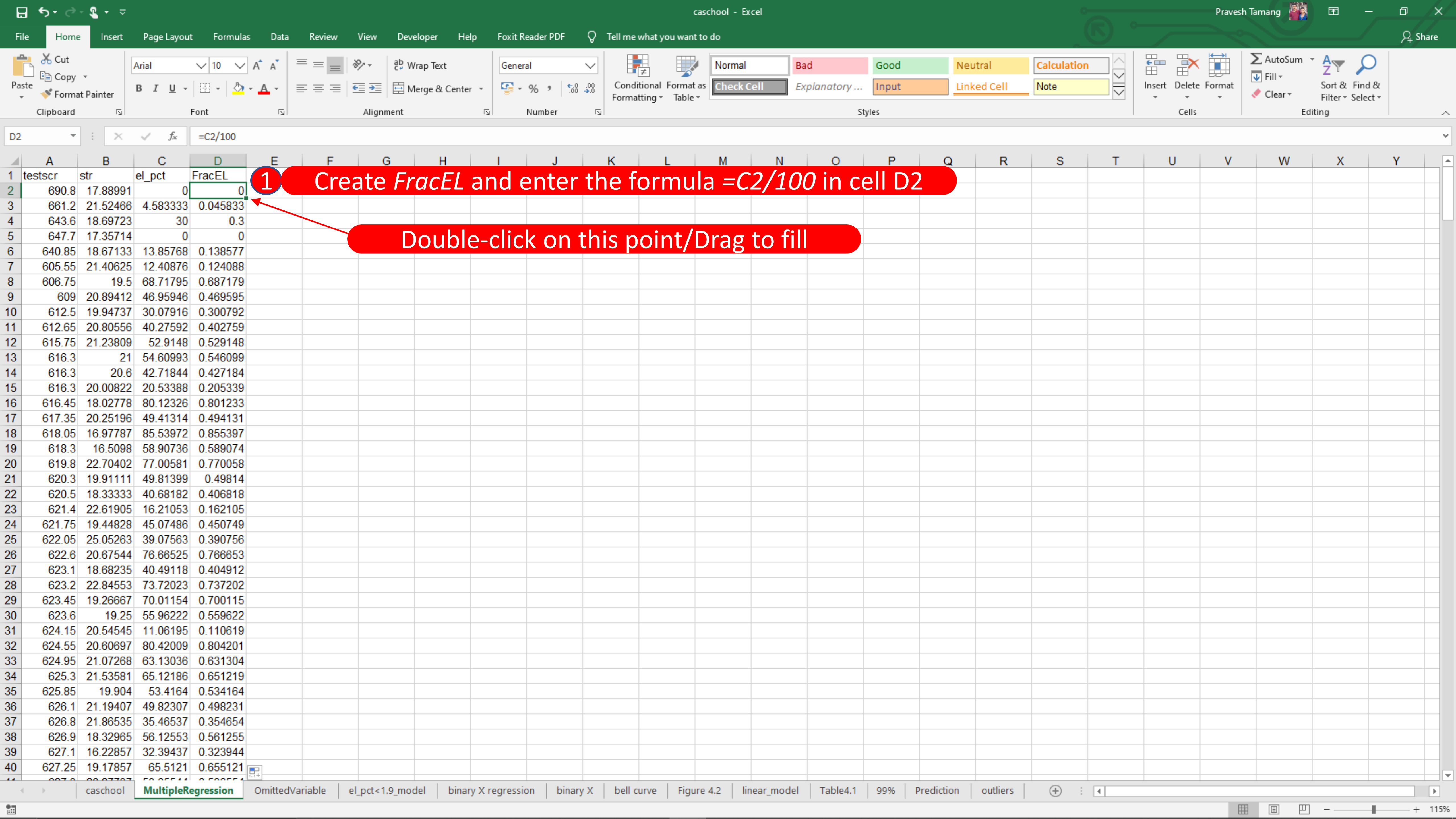
OmittedVariable | el_pct<1.9_model | binary X regression | binary X | bell curve | caschool | Figure 4.2 | linear_model | Table4.1 | 99% | Prediction | outliers

Multicollinearity

Example #1: Fraction of English Learners

Consider the following model, where we add another variable $FracEL = \frac{el\_pct}{100}$, i.e. the fraction of English learners.

Create a new variable *FracEL* in a new column.



testscr	str	el_pct	FracEL
690.8	17.88991	0	0
661.2	21.52466	4.583333	0.045833
643.6	18.69723	30	0.3
647.7	17.35714	0	0
640.85	18.67133	13.85768	0.138577
605.55	21.40625	12.40876	0.124088
606.75	19.5	68.71795	0.687179
609	20.89412	46.95946	0.469595
612.5	19.94737	30.07916	0.300792
612.65	20.80556	40.27592	0.402759
615.75	21.23809	52.9148	0.529148
616.3	21	54.60993	0.546099
616.3	20.6	42.71844	0.427184
616.3	20.00822	20.53388	0.205339
616.45	18.02778	80.12326	0.801233
617.35	20.25196	49.41314	0.494131
618.05	16.97787	85.53972	0.855397
618.3	16.5098	58.90736	0.589074
619.8	22.70402	77.00581	0.770058
620.3	19.91111	49.81399	0.49814
620.5	18.33333	40.68182	0.406818
621.4	22.61905	16.21053	0.162105
621.75	19.44828	45.07486	0.450749
622.05	25.05263	39.07563	0.390756
622.6	20.67544	76.66525	0.766653
623.1	18.68235	40.49118	0.404912
623.2	22.84553	73.72023	0.737202
623.45	19.26667	70.01154	0.700115
623.6	19.25	55.96222	0.559622
624.15	20.54545	11.06195	0.110619
624.55	20.60697	80.42009	0.804201
624.95	21.07268	63.13036	0.631304
625.3	21.53581	65.12186	0.651219
625.85	19.904	53.4164	0.534164
626.1	21.19407	49.82307	0.498231
626.8	21.86535	35.46537	0.354654
626.9	18.32965	56.12553	0.561255
627.1	16.22857	32.39437	0.323944
627.25	19.17857	65.5121	0.655121

testscr	str	el_pct	FracEL
690.8	17.88991	0	0
661.2	21.52466	4.583333	0.045833
643.6	18.69723	30	0.3
647.7	17.35714	0	0
640.85	18.67133	13.85768	0.138577
605.55	21.40625	12.40876	0.124088
606.75	19.5	68.71795	0.687179
609	20.89412	46.95946	0.469595
612.5	19.94737	30.07916	0.300792
612.65	20.80556	40.27592	0.402759
615.75	21.23809	52.9148	0.529148
616.3	21	54.60993	0.546099
616.3	20.6	42.71844	0.427184
616.3	20.00822	20.53388	0.205339
616.45	18.02778	80.12326	0.801233
617.35	20.25196	49.41314	0.494131
618.05	16.97787	85.53972	0.855397
618.3	16.5098	58.90736	0.589074
619.8	22.70402	77.00581	0.770058
620.3	19.91111	49.81399	0.49814
620.5	18.33333	40.68182	0.406818
621.4	22.61905	16.21053	0.162105
621.75	19.44828	45.07486	0.450749
622.05	25.05263	39.07563	0.390756
622.6	20.67544	76.66525	0.766653
623.1	18.68235	40.49118	0.404912
623.2	22.84553	73.72023	0.737202
623.45	19.26667	70.01154	0.700115
623.6	19.25	55.95222	0.559522
624.15	20.54545	11.06195	0.110619
624.55	20.60697	80.42009	0.804201
624.95	21.07268	63.13036	0.631304
625.3	21.53581	65.12186	0.651219
625.85	19.904	53.4164	0.534164
626.1	21.19407	49.82307	0.498231
626.8	21.86535	35.46537	0.354654
626.9	18.32965	56.12553	0.561255
627.1	16.22857	32.39437	0.323944
627.25	19.17857	65.5121	0.655121

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.653017122
R Square	0.426431361
Adjusted R Square	0.421282351
Standard Error	14.46448496
Observations	420

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	64864.3011	21621.4337	155.0135989	0.00000000
Residual	417	87245.29254	209.221325		
Total	420	152109.5936			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	686.0322487	7.411312484	92.56555437	3.8715E-280	671.4640604	700.600437	671.4640604	700.600437
str	-1.101295889	0.380278316	-2.896025997	0.003978056	-1.848797243	-0.353794534	-1.848797243	-0.353794534
el_pct	-0.649776777	0.039342549	-16.51587887	0.00000000	-0.727111212	-0.572442341	-0.727111212	-0.572442341
FracEL	0	0	65535	#NUM!	0	0	0	0

RESIDUAL OUTPUT

Observation	*Predicted testscr*	*Residuals*
1	666.3301646	24.46982316
2	659.3490811	1.850931076
3	645.9477665	-2.347790905
4	666.916898	-19.21688581
5	656.4651933	-15.61521769
6	654.3947101	-48.84466123
7	619.9056515	-13.15565151
8	632.5084758	-23.5084758
9	644.519556	-32.01955603
10	636.9488181	-24.29879367
11	628.2600154	-12.51001535
12	627.4207695	-11.12078167
13	635.5880996	-19.28811178

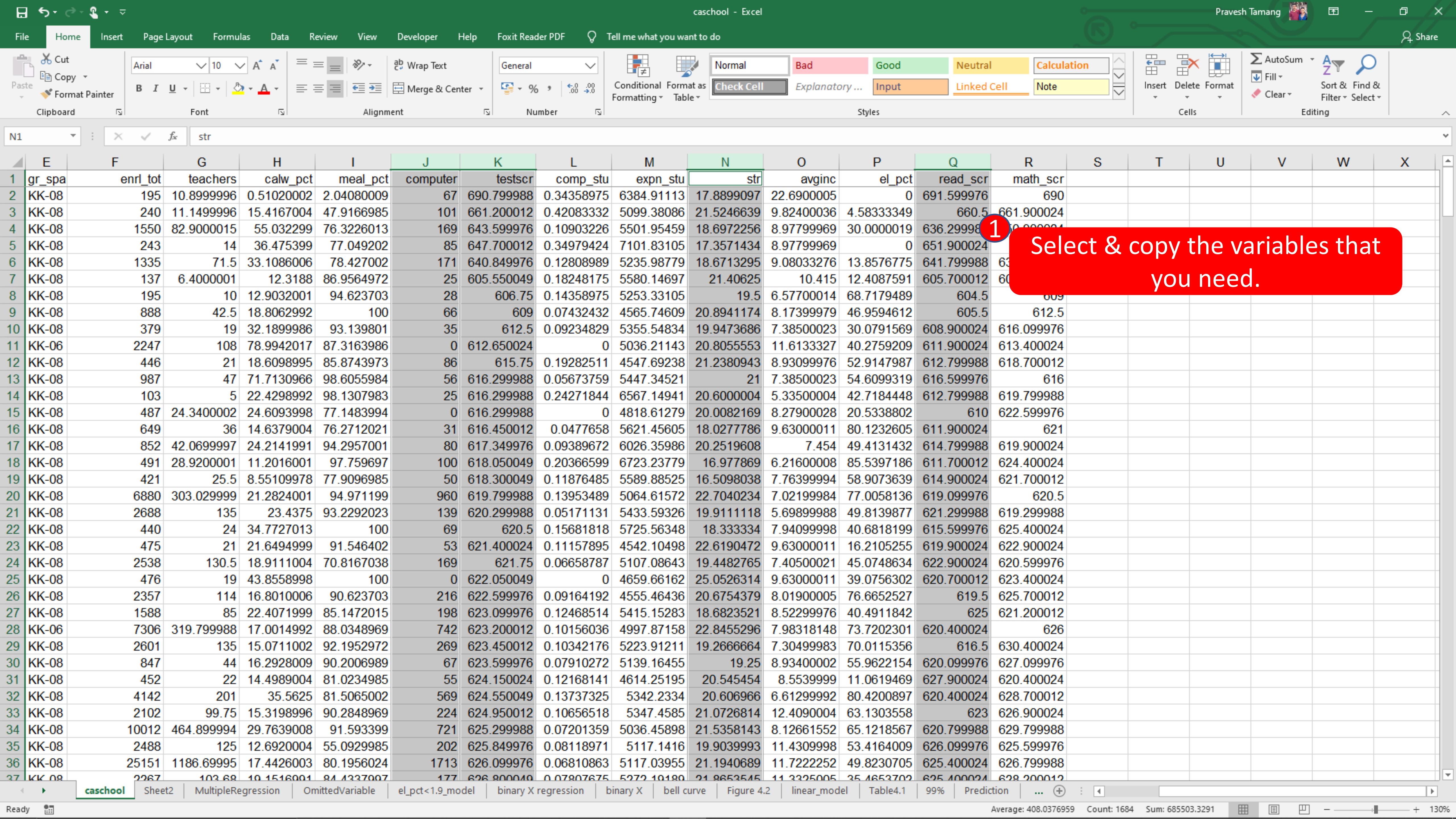
Since *FracEL=el_pct/100,* there is a relation between them (multicollinearity).

Example #2: "Not very small" classes.

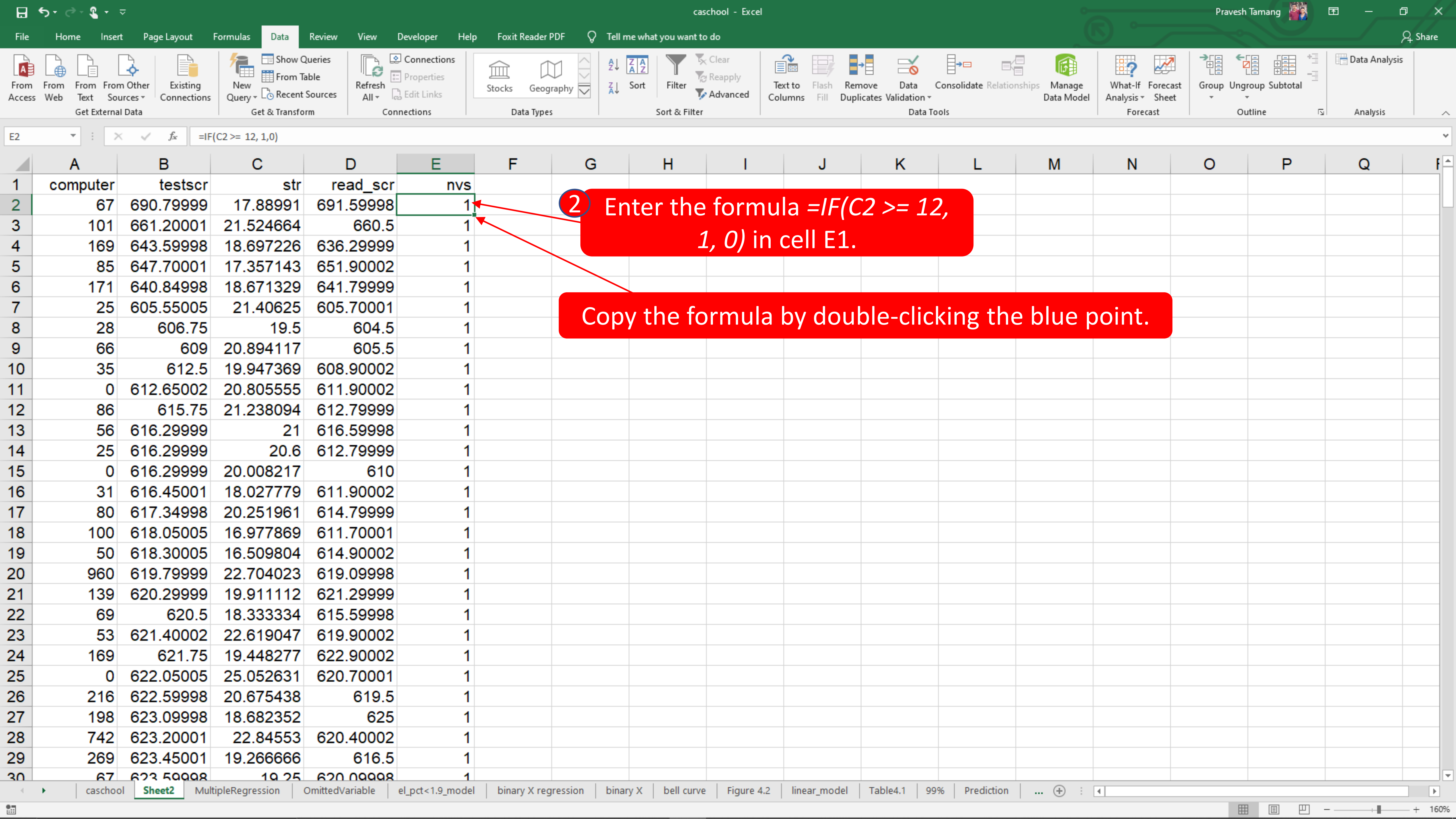
Let NVS_i be a binary variable that equals 1 if the student-teacher ratio in the i^{th} district is "not very small", specifically, NVS_i equals 1 if $STR_i \geq 12$ and equals 0 otherwise.

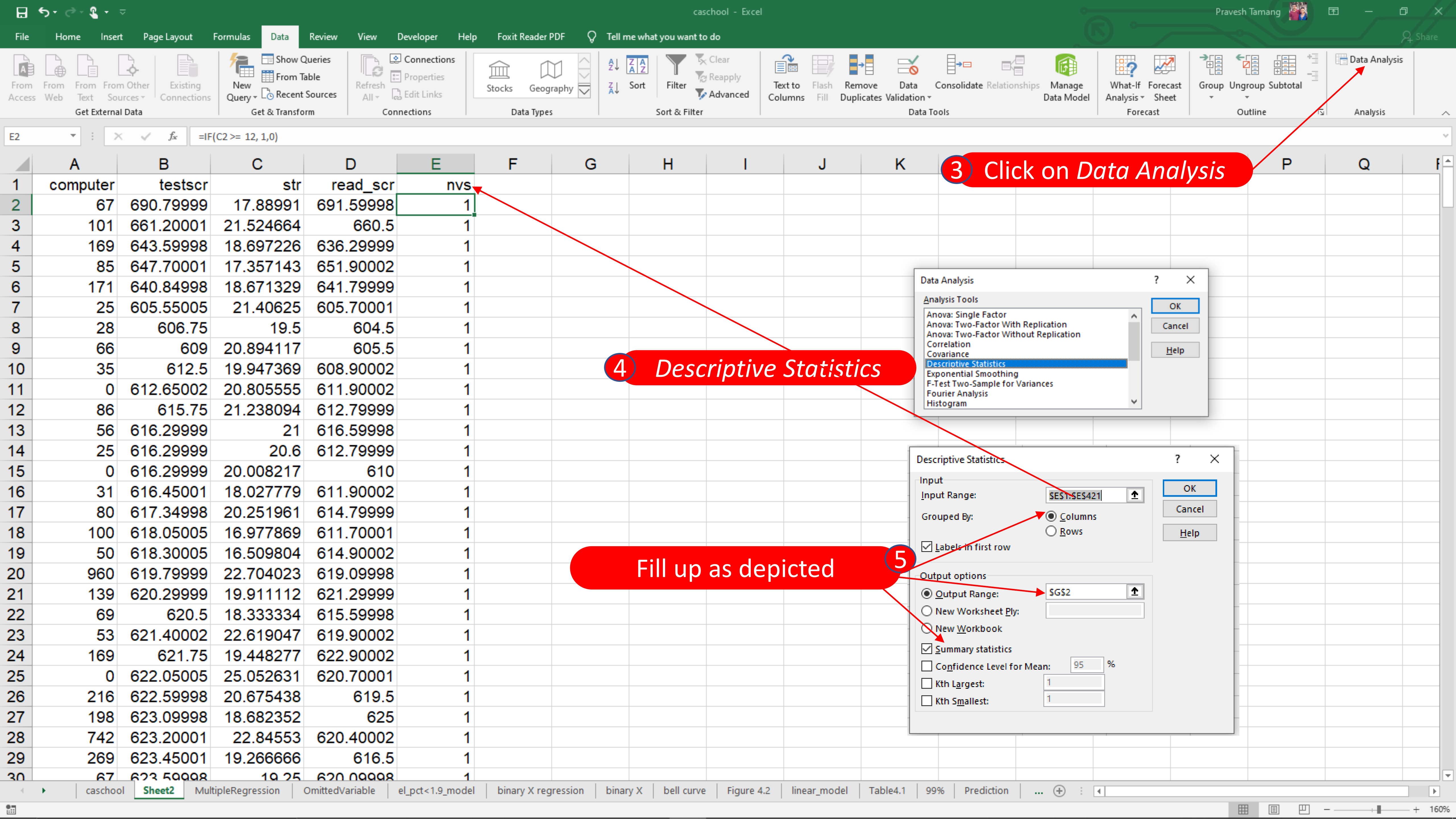
$$NVS = \begin{cases} 1, & if\ STR \geq 12 \\ 0, & otherwise \end{cases}$$

Variables used: *testscr, read_scr, computer, nvs*



	gr_spa	enrl_tot	teachers	calw_pct	meal_pct	computer	testscr	comp_stu	expn_stu	str	avginc	el_pct	read_scr	math_scr
2	KK-08	195	10.8999996	0.51020002	2.04080009	67	690.799988	0.34358975	6384.91113	17.8899097	22.6900005	0	691.599976	690
3	KK-08	240	11.1499996	15.4167004	47.9166985	101	661.200012	0.42083332	5099.38086	21.5246639	9.82400036	4.58333349	660.5	661.900024
4	KK-08	1550	82.9000015	55.032299	76.3226013	169	643.599976	0.10903226	5501.95459	18.6972256	8.97799969	30.0000019	636.29998	[illegible]
5	KK-08	243	14	36.475399	77.049202	85	647.700012	0.34979424	7101.83105	17.3571434	8.97799969	0	651.900024	[illegible]
6	KK-08	1335	71.5	33.1086006	78.427002	171	640.849976	0.12808989	5235.98779	18.6713295	9.08033276	13.8576775	641.799988	[illegible]
7	KK-08	137	6.4000001	12.3188	86.9564972	25	605.550049	0.18248175	5580.14697	21.40625	10.415	12.4087591	605.700012	[illegible]
8	KK-08	195	10	12.9032001	94.623703	28	606.75	0.14358975	5253.33105	19.5	6.57700014	68.7179489	604.5	[illegible]
9	KK-08	888	42.5	18.8062992	100	66	609	0.07432432	4565.74609	20.8941174	8.17399979	46.9594612	605.5	612.5
10	KK-08	379	19	32.1899986	93.139801	35	612.5	0.09234829	5355.54834	19.9473686	7.38500023	30.0791569	608.900024	616.099976
11	KK-06	2247	108	78.9942017	87.3163986	0	612.650024	0	5036.21143	20.8055553	11.6133327	40.2759209	611.900024	613.400024
12	KK-08	446	21	18.6098995	85.8743973	86	615.75	0.19282511	4547.69238	21.2380943	8.93099976	52.9147987	612.799988	618.700012
13	KK-08	987	47	71.7130966	98.6055984	56	616.299988	0.05673759	5447.34521	21	7.38500023	54.6099319	616.599976	616
14	KK-08	103	5	22.4298992	98.1307983	25	616.299988	0.24271844	6567.14941	20.6000004	5.33500004	42.7184448	612.799988	619.799988
15	KK-08	487	24.3400002	24.6093998	77.1483994	0	616.299988	0	4818.61279	20.0082169	8.27900028	20.5338802	610	622.599976
16	KK-08	649	36	14.6379004	76.2712021	31	616.450012	0.0477658	5621.45605	18.0277786	9.63000011	80.1232605	611.900024	621
17	KK-08	852	42.0699997	24.2141991	94.2957001	80	617.349976	0.09389672	6026.35986	20.2519608	7.454	49.4131432	614.799988	619.900024
18	KK-08	491	28.9200001	11.2016001	97.759697	100	618.050049	0.20366599	6723.23779	16.977869	6.21600008	85.5397186	611.700012	624.400024
19	KK-08	421	25.5	8.55109978	77.9096985	50	618.300049	0.11876485	5589.88525	16.5098038	7.76399994	58.9073639	614.900024	621.700012
20	KK-08	6880	303.029999	21.2824001	94.971199	960	619.799988	0.13953489	5064.61572	22.7040234	7.02199984	77.0058136	619.099976	620.5
21	KK-08	2688	135	23.4375	93.2292023	139	620.299988	0.05171131	5433.59326	19.9111118	5.69899988	49.8139877	621.299988	619.299988
22	KK-08	440	24	34.7727013	100	69	620.5	0.15681818	5725.56348	18.333334	7.94099998	40.6818199	615.599976	625.400024
23	KK-08	475	21	21.6494999	91.546402	53	621.400024	0.11157895	4542.10498	22.6190472	9.63000011	16.2105255	619.900024	622.900024
24	KK-08	2538	130.5	18.9111004	70.8167038	169	621.75	0.06658787	5107.08643	19.4482765	7.40500021	45.0748634	622.900024	620.599976
25	KK-08	476	19	43.8558998	100	0	622.050049	0	4659.66162	25.0526314	9.63000011	39.0756302	620.700012	623.400024
26	KK-08	2357	114	16.8010006	90.623703	216	622.599976	0.09164192	4555.46436	20.6754379	8.01900005	76.6652527	619.5	625.700012
27	KK-08	1588	85	22.4071999	85.1472015	198	623.099976	0.12468514	5415.15283	18.6823521	8.52299976	40.4911842	625	621.200012
28	KK-06	7306	319.799988	17.0014992	88.0348969	742	623.200012	0.10156036	4997.87158	22.8455296	7.98318148	73.7202301	620.400024	626
29	KK-08	2601	135	15.0711002	92.1952972	269	623.450012	0.10342176	5223.91211	19.2666664	7.30499983	70.0115356	616.5	630.400024
30	KK-08	847	44	16.2928009	90.2006989	67	623.599976	0.07910272	5139.16455	19.25	8.93400002	55.9622154	620.099976	627.099976
31	KK-08	452	22	14.4989004	81.0234985	55	624.150024	0.12168141	4614.25195	20.545454	8.5539999	11.0619469	627.900024	620.400024
32	KK-08	4142	201	35.5625	81.5065002	569	624.550049	0.13737325	5342.2334	20.606966	6.61299992	80.4200897	620.400024	628.700012
33	KK-08	2102	99.75	15.3198996	90.2848969	224	624.950012	0.10656518	5347.4585	21.0726814	12.4090004	63.1303558	623	626.900024
34	KK-08	10012	464.899994	29.7639008	91.593399	721	625.299988	0.07201359	5036.45898	21.5358143	8.12661552	65.1218567	620.799988	629.799988
35	KK-08	2488	125	12.6920004	55.0929985	202	625.849976	0.08118971	5117.1416	19.9039993	11.4309998	53.4164009	626.099976	625.599976
36	KK-08	25151	1186.69995	17.4426003	80.1956024	1713	626.099976	0.06810863	5117.03955	21.1940689	11.7222252	49.8230705	625.400024	626.799988


caschool - Excel
E2
=IF(C2 >= 12, 1,0)
2
Enter the formula =IF(C2 >= 12, 1, 0) in cell E1.
Copy the formula by double-clicking the blue point.



	A	B	C	D	E
1	computer	testscr	str	read_scr	nvs
2	67	690.79999	17.88991	691.59998	1
3	101	661.20001	21.524664	660.5	1
4	169	643.59998	18.697226	636.29999	1
5	85	647.70001	17.357143	651.90002	1
6	171	640.84998	18.671329	641.79999	1
7	25	605.55005	21.40625	605.70001	1
8	28	606.75	19.5	604.5	1
9	66	609	20.894117	605.5	1
10	35	612.5	19.947369	608.90002	1
11	0	612.65002	20.805555	611.90002	1
12	86	615.75	21.238094	612.79999	1
13	56	616.29999	21	616.59998	1
14	25	616.29999	20.6	612.79999	1
15	0	616.29999	20.008217	610	1
16	31	616.45001	18.027779	611.90002	1
17	80	617.34998	20.251961	614.79999	1
18	100	618.05005	16.977869	611.70001	1
19	50	618.30005	16.509804	614.90002	1
20	960	619.79999	22.704023	619.09998	1
21	139	620.29999	19.911112	621.29999	1
22	69	620.5	18.333334	615.59998	1
23	53	621.40002	22.619047	619.90002	1
24	169	621.75	19.448277	622.90002	1
25	0	622.05005	25.052631	620.70001	1
26	216	622.59998	20.675438	619.5	1
27	198	623.09998	18.682352	625	1
28	742	623.20001	22.84553	620.40002	1
29	269	623.45001	19.266666	616.5	1
30	67	623.59998	19.25	620.09998	1

computer	testscr	str	read_scr	nvs
67	690.79999	17.88991	691.59998	1
101	661.20001	21.524664	660.5	1
169	643.59998	18.697226	636.29999	1
85	647.70001	17.357143	651.90002	1
171	640.84998	18.671329	641.79999	1
25	605.55005	21.40625	605.70001	1
28	606.75	19.5	604.5	1
66	609	20.894117	605.5	1
35	612.5	19.947369	608.90002	1
0	612.65002	20.805555	611.90002	1
86	615.75	21.238094	612.79999	1
56	616.29999	21	616.59998	1
25	616.29999	20.6	612.79999	1
0	616.29999	20.008217	610	1
31	616.45001	18.027779	611.90002	1
80	617.34998	20.251961	614.79999	1
100	618.05005	16.977869	611.70001	1
50	618.30005	16.509804	614.90002	1
960	619.79999	22.704023	619.09998	1
139	620.29999	19.911112	621.29999	1
69	620.5	18.333334	615.59998	1
53	621.40002	22.619047	619.90002	1
169	621.75	19.448277	622.90002	1
0	622.05005	25.052631	620.70001	1
216	622.59998	20.675438	619.5	1
198	623.09998	18.682352	625	1
742	623.20001	22.84553	620.40002	1
269	623.45001	19.266666	616.5	1

nvs	
Mean	1
Standard Error	0
Median	1
Mode	1
Standard Deviation	0
Sample Variance	0
Kurtosis	#DIV/0!
Skewness	#DIV/0!
Range	0
Minimum	1
Maximum	1
Sum	420
Count	420

6 All the observations in this column are 1’s

Fill up as depicted

7

Regression

Input

Input Y Range: $A$1:$A$421

Input X Range: $C$1:$E$421

☑ Labels ☐ Constant is Zero

☑ Confidence Level: 95 %

Output options

◉ Output Range: $G$2

○ New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK | Cancel | Help

	str	read_scr	computer	nvs
2	17.88991	691.59998	67	1
3	21.524664	660.5	101	1
4	18.697226	636.29999	169	1
5	17.357143	651.90002	85	1
6	18.671329	641.79999	171	1
7	21.40625	605.70001	25	1
8	19.5	604.5	28	1
9	20.894117	605.5	66	1
10	19.947369	608.90002	35	1
11	20.805555	611.90002	0	1
12	21.238094	612.79999	86	1
13	21	616.59998	56	1
14	20.6	612.79999	25	1
15	20.008217	610	0	1
16	18.027779	611.90002	31	1
17	20.251961	614.79999	80	1
18	16.977869	611.70001	100	1
19	16.509804	614.90002	50	1
20	22.704023	619.09998	960	1
21	19.911112	621.29999	139	1
22	18.333334	615.59998	69	1
23	22.619047	619.90002	53	1
24	19.448277	622.90002	169	1
25	25.052631	620.70001	0	1
26	20.675438	619.5	216	1
27	18.682352	625	198	1
28	22.84553	620.40002	742	1
29	19.266666	616.5	269	1
30	19.25	620.09998	67	1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.982494662
R Square	0.965295761
Adjusted R Square	0.96504549
Standard Error	3.56223775
Observations	420

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	146830.7459	48943.58197	3857.002736	4.0198E-303
Residual	416	5278.847721	12.68953779		
Total	419	152109.5936			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	42.0620794	5.71950532	7.354146389	1.0305E-12	30.81934556	53.30481323	30.81934556	53.30481323
read_scr	0.933863698	0.008706489	107.2606497	3.1575E-305	0.9167495	0.950977895	0.9167495	0.950977895
computer	0.001454652	0.000396677	3.66709496	0.000277133	0.000674911	0.002234393	0.000674911	0.002234393
nvs	0	0	65535	#NUM!	0	0	0	0

RESIDUAL OUTPUT

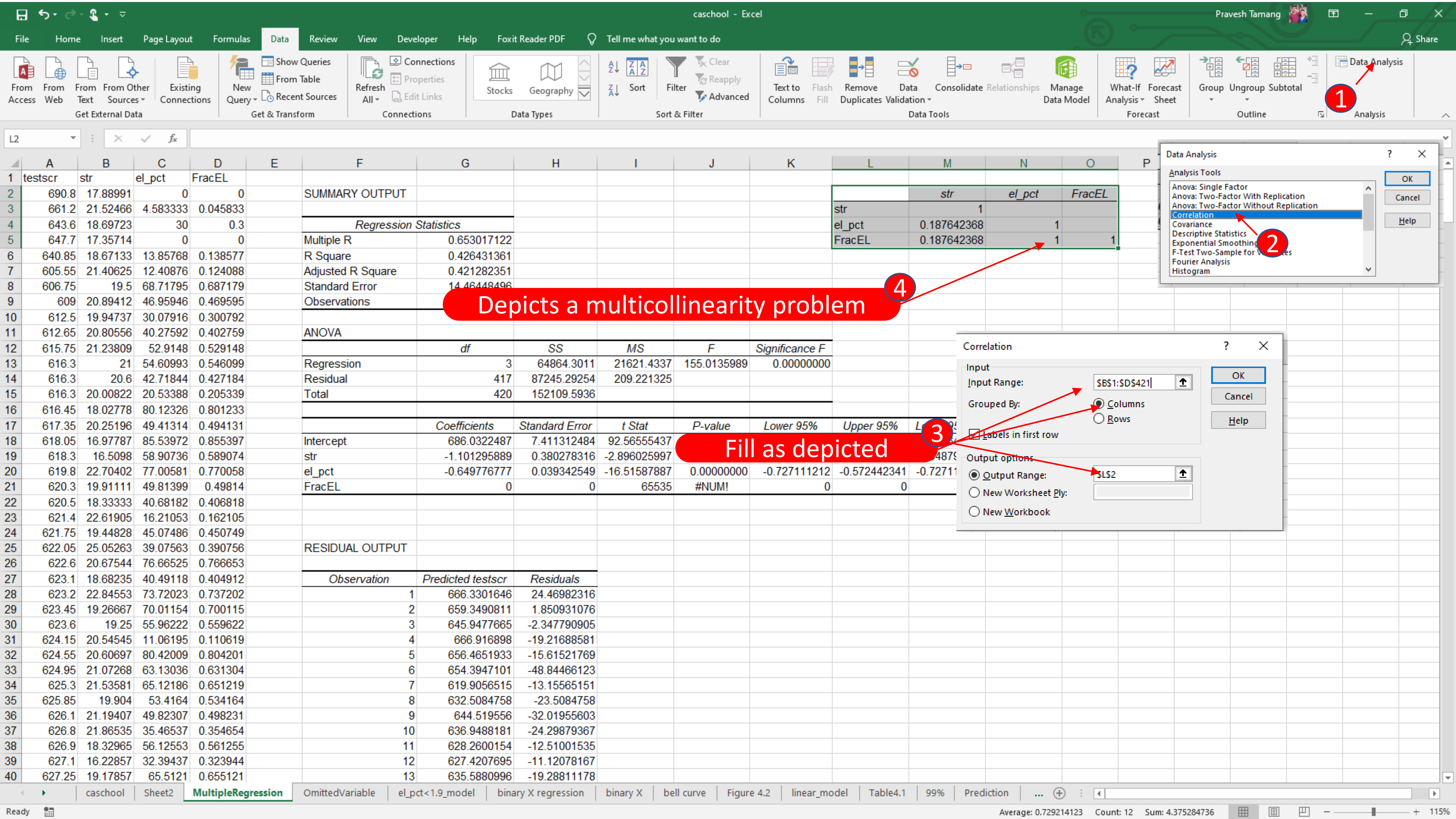
Observation	*Predicted testscr*	*Residuals*
1	688.0196516	2.780336235
2	659.0259715	2.174040678
3	636.525375	7.074600605

While it is possible to imagine a school district with fewer than 12 students per teacher, there are no such districts in our data set so we cannot analyze them in our regression.

Identifying multicollinearity

Pairwise correlation coefficients

- Calculate the sample correlation coefficient.
- As a rule of thumb, correlation coefficients around 0.8 or above may signal a multicollinearity problem.
- Use the *caschool* dataset that we used earlier.
- The independent variables are; *str, el_pct, FracEL*



	A	B	C	D
1	testscr	str	el_pct	FracEL
2	690.8	17.88991	0	0
3	661.2	21.52466	4.583333	0.045833
4	643.6	18.69723	30	0.3
5	647.7	17.35714	0	0
6	640.85	18.67133	13.85768	0.138577
7	605.55	21.40625	12.40876	0.124088
8	606.75	19.5	68.71795	0.687179
9	609	20.89412	46.95946	0.469595
10	612.5	19.94737	30.07916	0.300792
11	612.65	20.80556	40.27592	0.402759
12	615.75	21.23809	52.9148	0.529148
13	616.3	21	54.60993	0.546099
14	616.3	20.6	42.71844	0.427184
15	616.3	20.00822	20.53388	0.205339
16	616.45	18.02778	80.12326	0.801233
17	617.35	20.25196	49.41314	0.494131
18	618.05	16.97787	85.53972	0.855397
19	618.3	16.5098	58.90736	0.589074
20	619.8	22.70402	77.00581	0.770058
21	620.3	19.91111	49.81399	0.49814
22	620.5	18.33333	40.68182	0.406818
23	621.4	22.61905	16.21053	0.162105
24	621.75	19.44828	45.07486	0.450749
25	622.05	25.05263	39.07563	0.390756
26	622.6	20.67544	76.66525	0.766653
27	623.1	18.68235	40.49118	0.404912
28	623.2	22.84553	73.72023	0.737202
29	623.45	19.26667	70.01154	0.700115
30	623.6	19.25	55.96222	0.559622
31	624.15	20.54545	11.06195	0.110619
32	624.55	20.60697	80.42009	0.804201
33	624.95	21.07268	63.13036	0.631304
34	625.3	21.53581	65.12186	0.651219
35	625.85	19.904	53.4164	0.534164
36	626.1	21.19407	49.82307	0.498231
37	626.8	21.86535	35.46537	0.354654
38	626.9	18.32965	56.12553	0.561255
39	627.1	16.22857	32.39437	0.323944
40	627.25	19.17857	65.5121	0.655121

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.653017122
R Square	0.426431361
Adjusted R Square	0.421282351
Standard Error	14.46448496
Observations	

ANOVA

	df	SS	MS	F	Significance F
Regression	3	64864.3011	21621.4337	155.0135989	0.00000000
Residual	417	87245.29254	209.221325		
Total	420	152109.5936			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%
Intercept	686.0322487	7.411312484	92.56555437				
str	-1.101295889	0.380278316	-2.896025997				
el_pct	-0.649776777	0.039342549	-16.51587887	0.00000000	-0.727111212	-0.572442341	-0.72711
FracEL	0	0	65535	#NUM!	0	0	

RESIDUAL OUTPUT

Observation	Predicted testscr	Residuals
1	666.3301646	24.46982316
2	659.3490811	1.850931076
3	645.9477665	-2.347790905
4	666.916898	-19.21688581
5	656.4651933	-15.61521769
6	654.3947101	-48.84466123
7	619.9056515	-13.15565151
8	632.5084758	-23.5084758
9	644.519556	-32.01955603
10	636.9488181	-24.29879367
11	628.2600154	-12.51001535
12	627.4207695	-11.12078167
13	635.5880996	-19.28811178

	str	el_pct	FracEL
str	1		
el_pct	0.187642368	1	
FracEL	0.187642368	1	1

Identifying multicollinearity

Auxiliary regression and the variance inflation factor (VIF)

Steps:

1. Our model is $testscr = \beta_0 + \beta_1 str + \beta_2 el\_pct + \beta_3 FracEL + u$
2. Estimate the auxiliary regressions

$$str = \alpha_0 + \alpha_1 el\_pct + \beta_3 FracEL + u_1 \ldots (1)$$

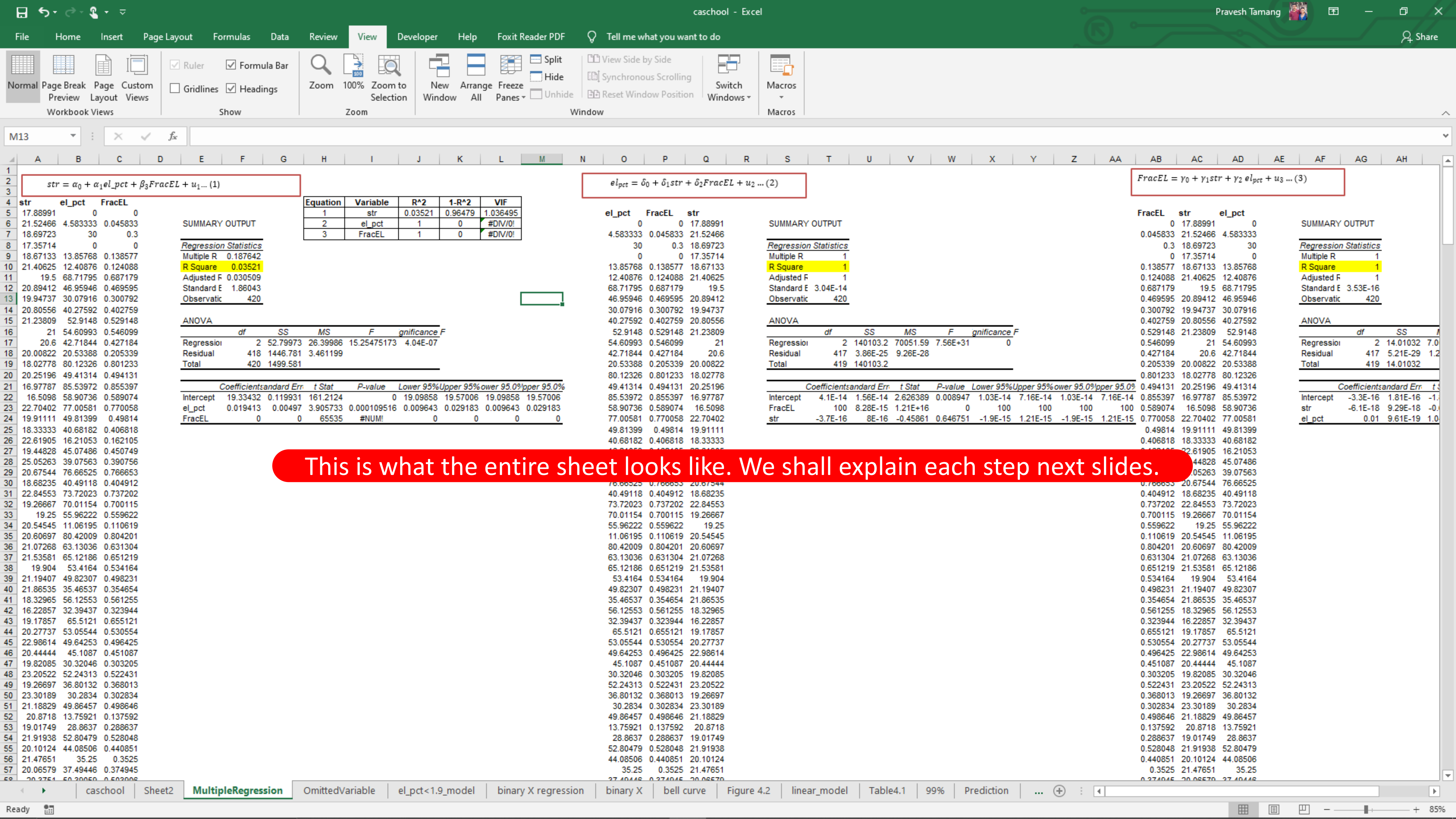
$$el_{pct} = \delta_0 + \delta_1 str + \delta_2 FracEL + u_2 \ldots (2)$$

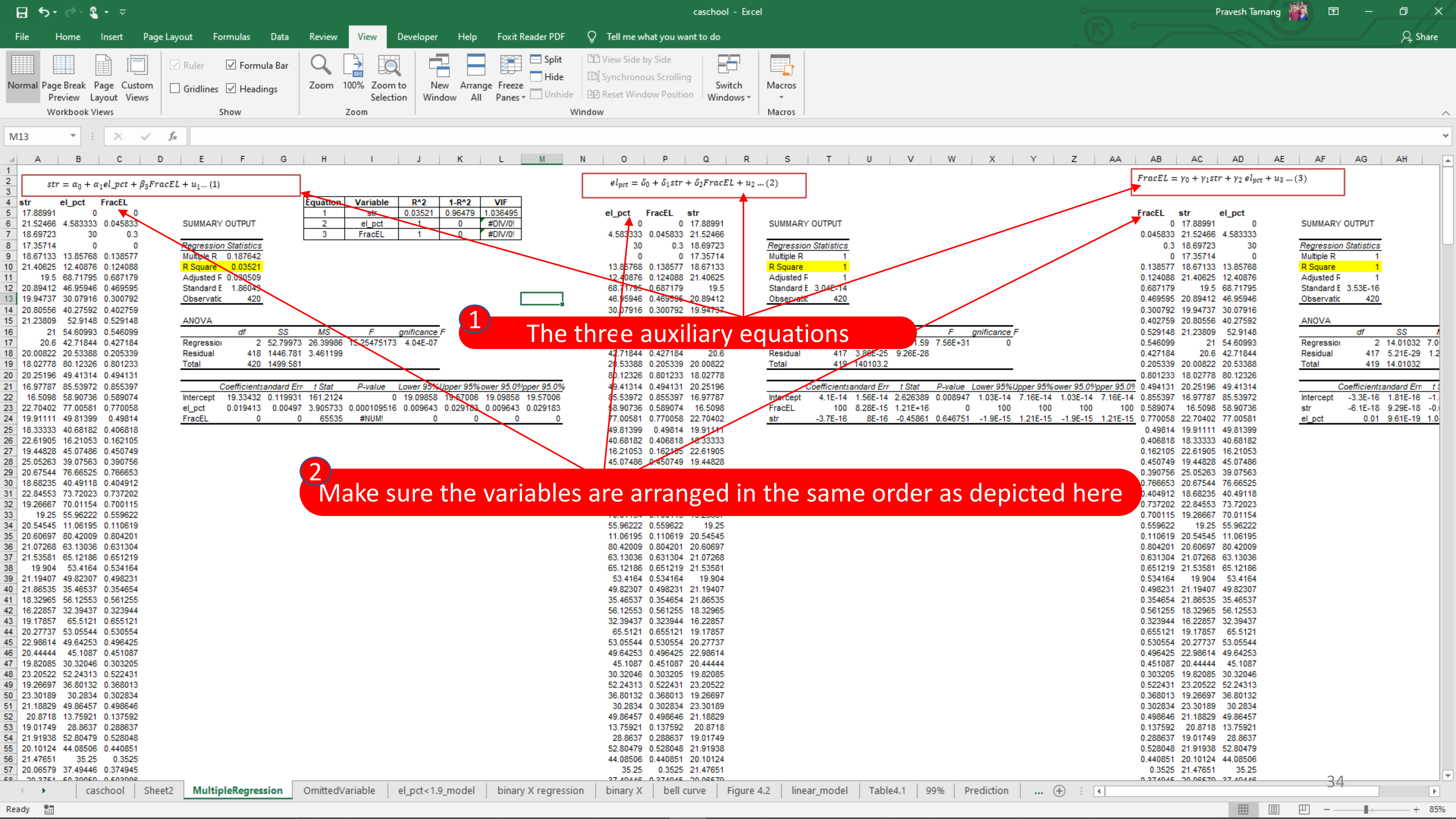
$$FracEL = \gamma_0 + \gamma_1 str + \gamma_2\, el_{pct} + u_3 \ldots (3)$$

   to obtain R_1^2, R_2^2, R_3^2.

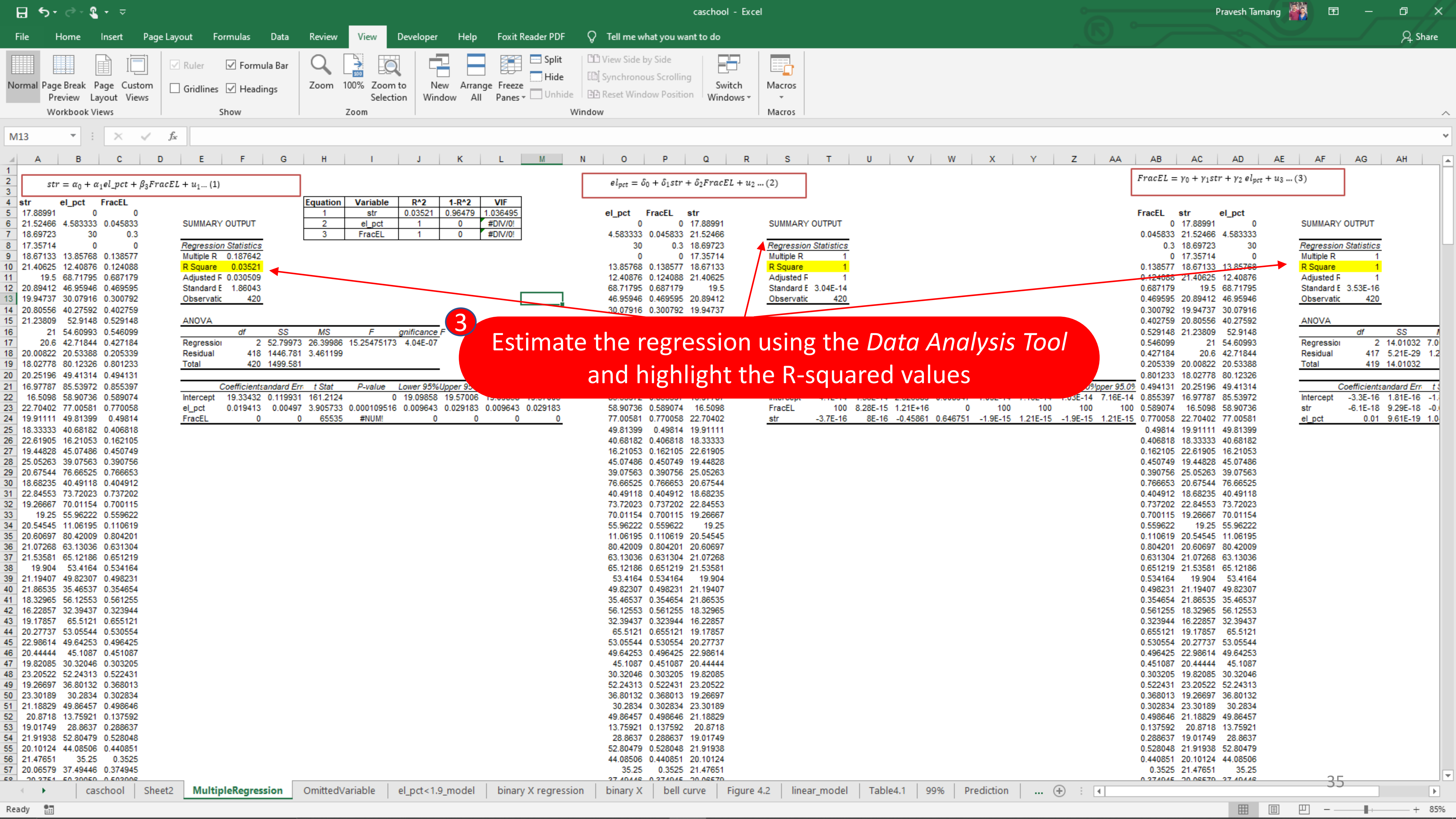
3. Obtain the VIF for each independent variable using the formula

$$VIF_k = \frac{1}{1-R_k^2}$$


caschool - Excel
$str = \alpha_0 + \alpha_1 el\_pct + \beta_3 FracEL + u_1 \dots (1)$
$el_{pct} = \delta_0 + \delta_1 str + \delta_2 FracEL + u_2 \dots (2)$
$FracEL = \gamma_0 + \gamma_1 str + \gamma_2\, el_{pct} + u_3 \dots (3)$
Equation	Variable	R^2	1-R^2	VIF
1	str	0.03521	0.96479	1.036495
2	el_pct	1	0	#DIV/0!
3	FracEL	1	0	#DIV/0!
SUMMARY OUTPUT
Regression Statistics
ANOVA
caschool
Sheet2
MultipleRegression
OmittedVariable
el_pct<1.9_model
binary X regression
binary X
bell curve
Figure 4.2
linear_model
Table4.1
99%
Prediction
This is what the entire sheet looks like. We shall explain each step next slides.



str = α0 + α1el_pct + β3FracEL + u1 ... (1)
elpct = δ0 + δ1str + δ2FracEL + u2 ... (2)
FracEL = γ0 + γ1str + γ2 elpct + u3 ... (3)
Equation Variable R^2 1-R^2 VIF
1 str 0.03521 0.96479 1.036495
2 el_pct 1 0 #DIV/0!
3 FracEL 1 0 #DIV/0!
SUMMARY OUTPUT
Regression Statistics
R Square
ANOVA
1 The three auxiliary equations
2 Make sure the variables are arranged in the same order as depicted here


caschool - Excel
str = α0 + α1el_pct + β3FracEL + u1 ... (1)
elpct = δ0 + δ1str + δ2FracEL + u2 ... (2)
FracEL = γ0 + γ1str + γ2 elpct + u3 ... (3)
Equation | Variable | R^2 | 1-R^2 | VIF
1 | str | 0.03521 | 0.96479 | 1.036495
2 | el_pct | 1 | 0 | #DIV/0!
3 | FracEL | 1 | 0 | #DIV/0!
SUMMARY OUTPUT
Regression Statistics
Multiple R 0.187642
R Square 0.03521
Adjusted R 0.030509
Standard E 1.86043
Observatio 420
ANOVA
Regression 2 52.79973 26.39986 15.25475173 4.04E-07
Residual 418 1446.781 3.461199
Total 420 1499.581
Intercept 19.33432 0.119931 161.2124 0 19.09858 19.57006
el_pct 0.019413 0.00497 3.905733 0.000109516 0.009643 0.029183 0.009643 0.029183
FracEL 0 0 65535 #NUM! 0 0 0 0
3
Estimate the regression using the Data Analysis Tool and highlight the R-squared values
MultipleRegression

$$str = \alpha_0 + \alpha_1 el\_pct + \beta_3 FracEL + u_1 \ldots (1)$$

4 Create a table as depicted here.

$$el_{pct} = \delta_0 + \delta_1 str + \delta_2 FracEL + u_2 \ldots (2)$$

Equation	Variable	R^2	1-R^2	VIF
1	str	0.03521	0.96479	1.03649
2	el_pct	1	0	#DIV/0!
3	FracEL	1	0	#DIV/0!

5 Fill in the values for R-square

6 In cell K5 use the formula =1-J5

7 In cell L5 use the formula =1/K5

str	el_pct	FracEL
17.8899	0	0
21.5247	4.58333	0.04583
18.6972	30	0.3
17.3571	0	0
18.6713	13.8577	0.13858
21.4063	12.4088	0.12409
19.5	68.7179	0.68718
20.8941	46.9595	0.46959
19.9474	30.0792	0.30079
20.8056	40.2759	0.40276
21.2381	52.9148	0.52915
21	54.6099	0.5461
20.6	42.7184	0.42718
20.0082	20.5339	0.20534
18.0278	80.1233	0.80123
20.252	49.4131	0.49413
16.9779	85.5397	0.8554
16.5098	58.9074	0.58907
22.704	77.0058	0.77006
19.9111	49.814	0.49814
18.3333	40.6818	0.40682
22.619	16.2105	0.16211
19.4483	45.0749	0.45075
25.0526	39.0756	0.39076
20.6754	76.6653	0.76665
18.6824	40.4912	0.40491
22.8455	73.7202	0.7372
19.2667	70.0115	0.70012

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.18764
R Square	0.03521
Adjusted R	0.03051
Standard E	1.86043
Observatic	420

ANOVA

	df	SS	MS	F	Significance F
Regressio	2	52.7997	26.3999	15.2547517	4E-07
Residual	418	1446.78	3.4612		
Total	420	1499.58			

	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	19.3343	0.11993	161.212	0	19.0986	19.5701	19.0986	19.5701
el_pct	0.01941	0.00497	3.90573	0.00010952	0.00964	0.02918	0.00964	0.02918
FracEL	0	0	65535	#NUM!	0	0	0	0

el_pct	FracEL	str
0	0	17.8899
4.58333	0.04583	21.5247
30	0.3	18.6972
0	0	17.3571
13.8577	0.13858	18.6713
12.4088	0.12409	21.4063
68.7179	0.68718	19.5
46.9595	0.46959	20.8941
30.0792	0.30079	19.9474
40.2759	0.40276	20.8056
52.9148	0.52915	21.2381
54.6099	0.5461	21
42.7184	0.42718	20.6
20.5339	0.20534	20.0082
80.1233	0.80123	18.0278
49.4131	0.49413	20.252
85.5397	0.8554	16.9779
58.9074	0.58907	16.5098
77.0058	0.77006	22.704
49.814	0.49814	19.9111
40.6818	0.40682	18.3333
16.2105	0.16211	22.619
45.0749	0.45075	19.4483
39.0756	0.39076	25.0526
76.6653	0.76665	20.6754
40.4912	0.40491	18.6824
73.7202	0.7372	22.8455

SUMMARY OUTP

Regression Statis	
Multiple R	
R Square	
Adjusted R	
Standard E	3E
Observatic	4

ANOVA

	df
Regressio	
Residual	4
Total	4

	Coeffici
Intercept	4.1E
FracEL	1
str	-3.7E

caschool - Excel

U10 | 13.8576774597167

$$str = \alpha_0 + \alpha_1 el\_pct + \beta_3 FracEL + u_1 \ldots (1)$$

$$el_{pct} = \delta_0 + \delta_1 str +$$

str	el_pct	FracEL
17.88991	0	0
21.52466	4.583333	0.045833
18.69723	30	0.3
17.35714	0	0
18.67133	13.85768	0.138577
21.40625	12.40876	0.124088
19.5	68.71795	0.687179
20.89412	46.95946	0.469595
19.94737	30.07916	0.300792
20.80556	40.27592	0.402759
21.23809	52.9148	0.529148
21	54.60993	0.546099
20.6	42.71844	0.427184
20.00822	20.53388	0.205339
18.02778	80.12326	0.801233
20.25196	49.41314	0.494131
16.97787	85.53972	0.855397
16.5098	58.90736	0.589074
22.70402	77.00581	0.770058
19.91111	49.81399	0.49814
18.33333	40.68182	0.406818
22.61905	16.21053	0.162105
19.44828	45.07486	0.450749
25.05263	39.07563	0.390756
20.67544	76.66525	0.766653
18.68235	40.49118	0.404912
22.84553	73.72023	0.737202
19.26667	70.01154	0.700115
19.25	55.96222	0.559622
20.54545	11.06195	0.110619
20.60697	80.42009	0.804201
21.07268	63.13036	0.631304

Equation	Variable	R^2	1-R^2	VIF	Decision
1	str	0.03521	0.96479	1.036495	VIF < 5, there is little or no evidence of multicollinearity of *str* with other variables
2	el_pct	1	0	#DIV/0!	VIF > 10, there is evidence of high multicollinearity of str with other variables
3	FracEL	1	0	#DIV/0!	VIF > 10, there is evidence of high multicollinearity of str with other variables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.187642
R Square	0.03521
Adjusted R	0.030509
Standard E	1.86043
Observatioı	420

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	2	52.79973	26.39986	15.25475173	4.04E-07
Residual	418	1446.781	3.461199		
Total	420	1499.581			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	19.33432	0.119931	161.2124	0	19.09858	19.57006	19.09858	19.57006
el_pct	0.019413	0.00497	3.905733	0.000109516	0.009643	0.029183	0.009643	0.029183
FracEL	0	0	65535	#NUM!	0	0	0	0

el_pct	FracEL
0	0
4.583333	0.045833
30	0.3
0	0
13.85768	0.138577
12.40876	0.124088
68.71795	0.687179
46.95946	0.469595
30.07916	0.300792
40.27592	0.402759
52.9148	0.529148
54.60993	0.546099
42.71844	0.427184
20.53388	0.205339
80.12326	0.801233
49.41314	0.494131
85.53972	0.855397
58.90736	0.589074
77.00581	0.770058
49.81399	0.49814
40.68182	0.406818
16.21053	0.162105
45.07486	0.450749
39.07563	0.390756
76.66525	0.766653
40.49118	0.404912
73.72023	0.737202
70.01154	0.700115
55.96222	0.559622
11.06195	0.110619
80.42009	0.804201

caschool | Sheet2 | MultipleRegression | OmittedVariable | el_pct<1.9_model | binary X regression | binary X | bell curve | Figure 4.2 | linear_model | Table4.1 | 99% | Prediction

Exercises

E6.1 Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions.

(a). Regress Birthweight on Smoker. What is the estimated effect of smoking on birth weight?

(b). Regress Birthweight on Smoker, Alcohol, and Nprevist.

(i). Using the two conditions in Key Concept 6.1, explain why the exclusion of Alcohol and Nprevist could lead to omitted variable bias in the regression estimated in (a).

(ii). Is the estimated effect of smoking on birth weight substantially different from the regression that excludes Alcohol and Nprevist? Does the regression in (a) seem to suffer from omitted variable bias?

(iii). Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.

E6.2 Using the data set **Growth** described in Empirical Exercise E4.1, but excluding the data for Malta, carry out the following exercises.

(a). Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series Growth, TradeShare, YearsSchool, Oil, Rev_Coups, Assassinations, and RGDP60. Include the appropriate units for all entries.

(b). Run a regression of Growth on TradeShare, YearsSchool, Rev_Coups, Assassinations, and RGDP60. What is the value of the coefficient on Rev_Coups? Interpret the value of this coefficient. Is it large or small in a real-world sense?

(c). Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

(d). Repeat (c) but now assume that the country's value for TradeShare is one standard deviation above the mean.

(e). Why is Oil omitted from the regression? What would happen if it were included?