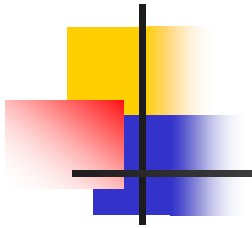
Pulling it All Together: Ordinary Least Squares Regression

Dr. Nancy Burns
Dr. Kien Le



Review of Ordinary Least Squares

Ordinary Least Squares (OLS) Regression

- *Dependent variable*, Y, what we're explaining.
- *Explanatory variable* or *independent variable*, X, what we are using to explain Y.

- When X goes up by a certain amount, on average, what happens to Y? Does it go up, go down, or not change, and by how much? And how certain are we about this effect?

The Regression Line

Source: Wonnacott and Wonnacott, 1990.

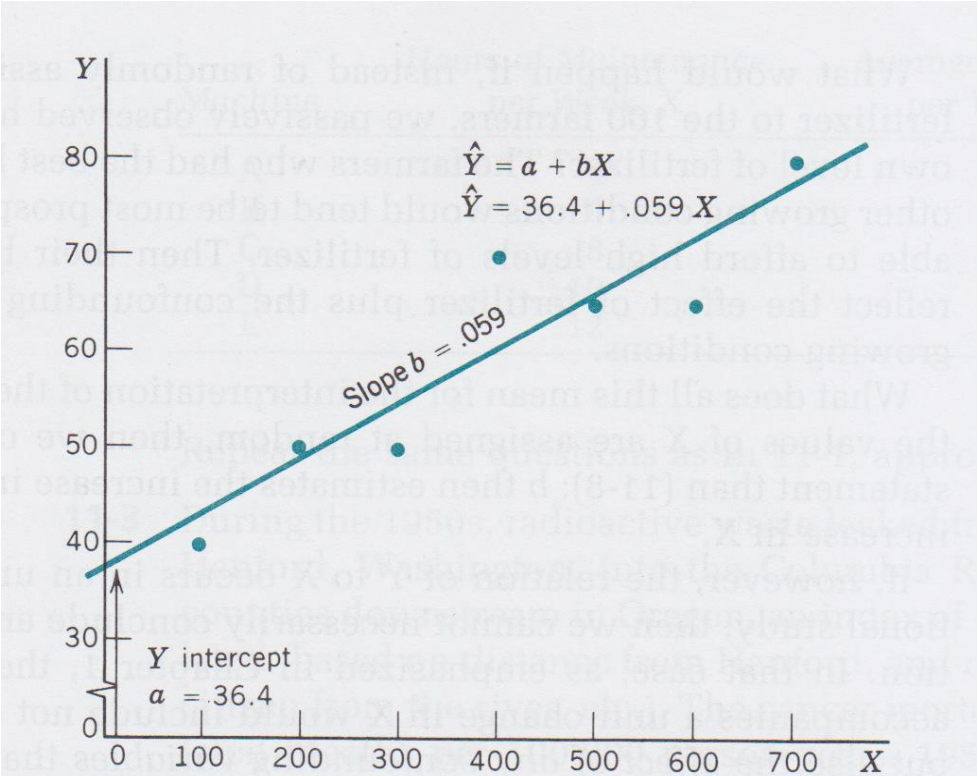


FIGURE 11-4
The least squares line fitted to the data of Example 11-1.

The Regression Line

- Source: Shively, 2005.

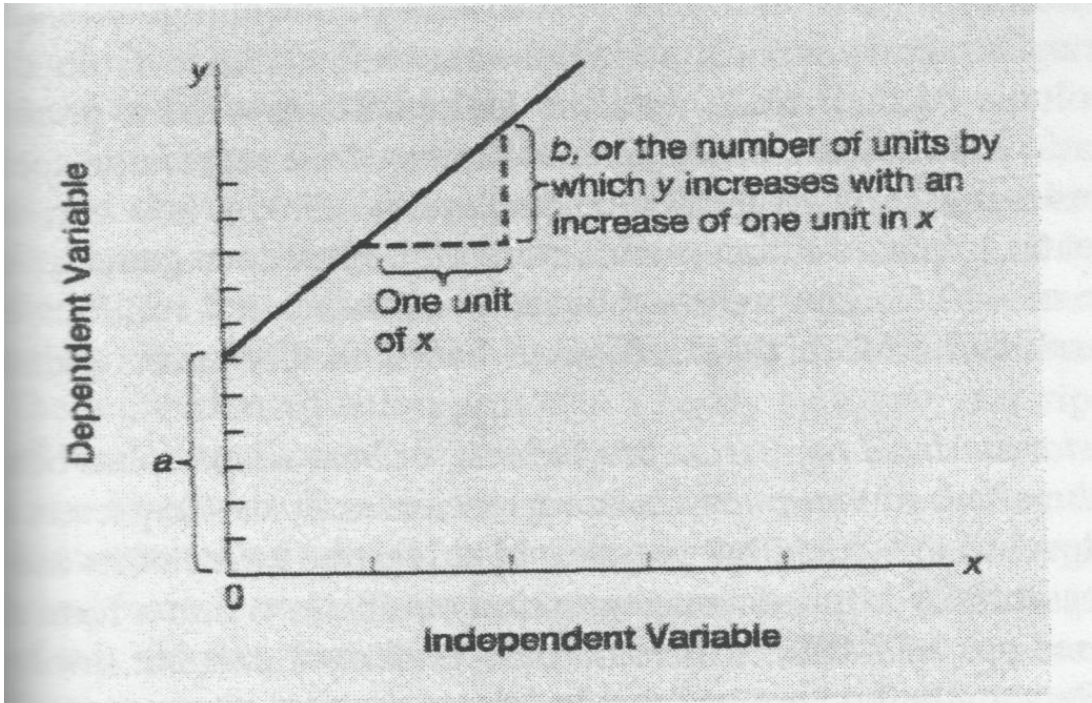


Figure 7–4 The Regression Equation

The equation of this line is $y = 6 + 3x$. The predicted value of y when x is 4, for instance, is $6 + (4 \times 3)$, or 18.

Interpreting Regression Coefficients

Imagine this is our regression model:

Predicted y = 2 + 3x

In a table, this would look like

	Coefficient (Standard Error)
Explanatory Variable	3.00* (0.10)
Intercept	2.00* (1.00)
N	1435
Adjusted R-squared	0.34

* p<.05

Drawing Inferences

Predicting Earnings, Ordinary Least Squares

Variable	Coefficient	S.E.	t
Height	1563.138	133.448	11.713
Constant	-84078.32	8901.098	-9.446

N = 1379

R-squared = .09

Source: Gelman and Nolan 2002.

Questions to ask

- On what scales are our variables measured?
- Are our coefficients statistically significant?
- Are our coefficients substantively significant?
- Are there omitted variables that will affect our estimates of the coefficients at hand?
- What information should be on this table to allow the table to speak for itself?

A Multivariate Model

Predicting Earnings in US Dollars, Ordinary Least Squares

Variable	Coefficient	S.E.	t	p-value
Height in inches	550.5448	184.57	2.983	.003
Woman	-11254.57	1448.892	-7.768	.000
Constant	-84078.32	8901.098	-9.446	.908

N = 1379
R-squared = .13
Source; Gelman and Nolan, 2002.

Group Exercise

- Working in groups, develop an interpretation of the following table.
- Use these questions as your guide:
  - On what scales are our variables measured?
  - Are our coefficients statistically significant?
  - Are our coefficients substantively significant?
  - Are there omitted variables that will affect our estimates of the coefficients at hand?
  - Describe your conclusions and your certainty about your conclusions.
  - What do you wish were on this table that isn't here?

Predicting Hours Working
Ordinary Least Squares Regression

	Women	Men
Education	4.26***	1.92***
	(.60)	(.47)
Marriage	-0.53*	1.17***
	(.25)	(.24)
Pre-school Children	-2.25***	1.54***
	(.33)	(.32)
School-aged Children	-0.14	1.65***
	(.29)	(.28)
N	1288	1177
Adjusted R-Squared	.30	.44

All explanatory variables are scaled 0 to 1.
p<.05; ** p<.01; *** p < .001. Controlling for other variables.
Standard errors in parentheses.

Intercept shifts

Source: Hanushek and Jackson

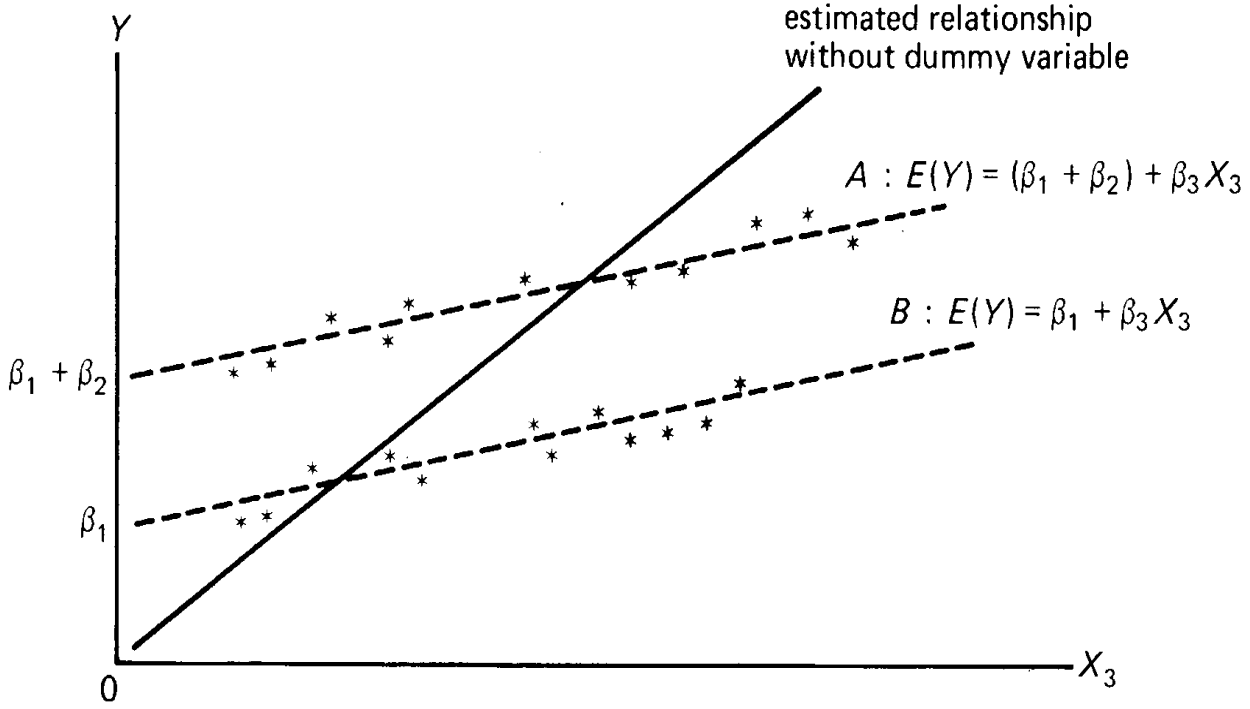


FIGURE 4.4 Estimation of misspecified bivariate relationship excluding dummy variable.

An example of an intercept shift

Personal economic conditions as a function of strata,

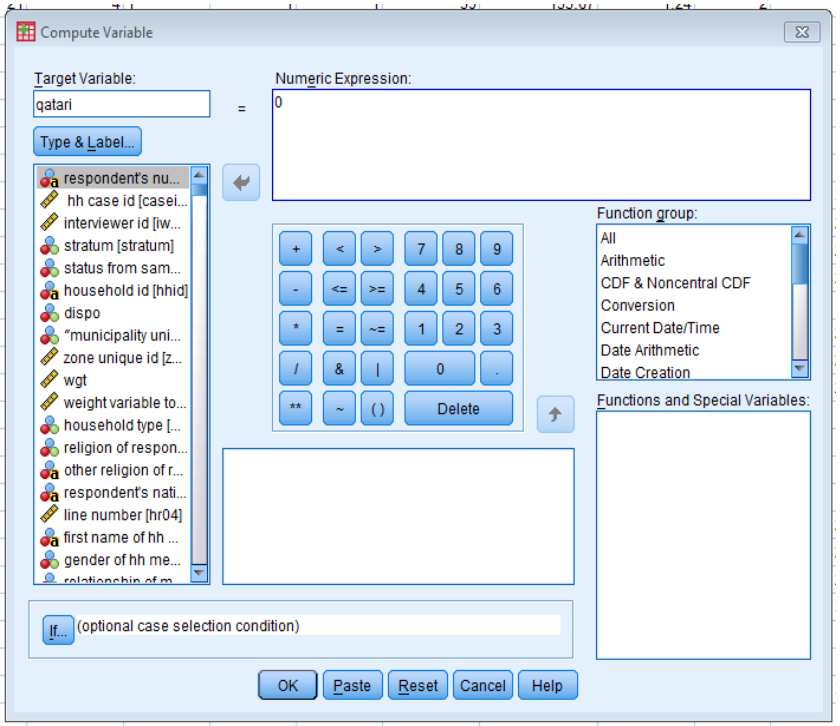
Qatari

White collar

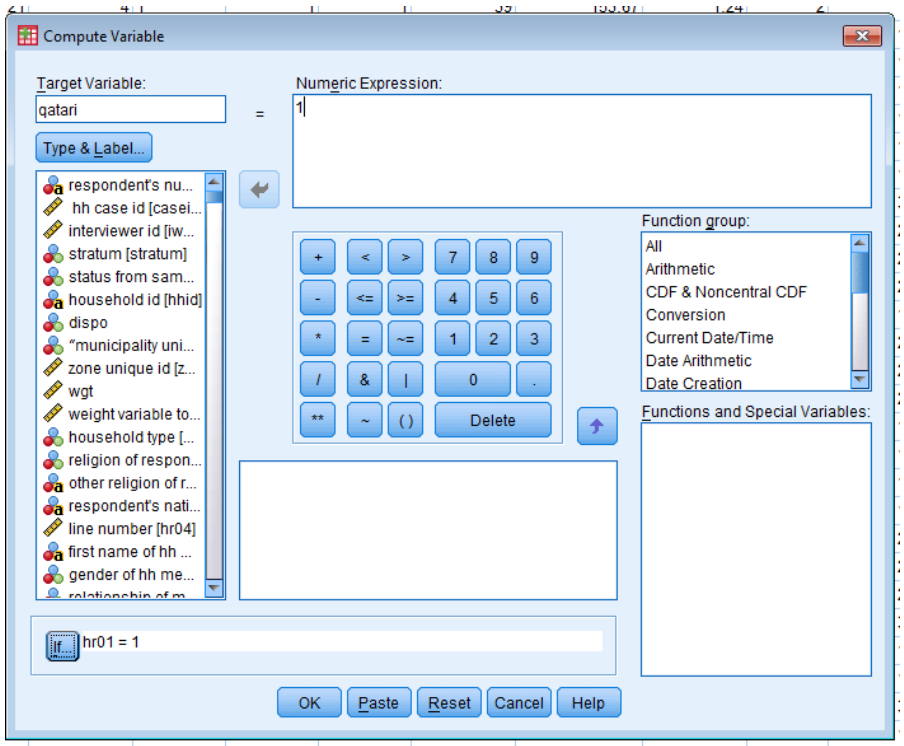
Blue collar

How would we do this?

First, we would create three variables: Qatari, white collar, and blue collar.



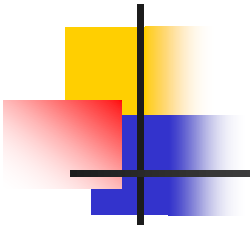
First, code the variable equal to zero.



Then, set it equal to 1 if hr01=1.

What is Going on Behind the Point-and-Click Commands?

```
COMPUTE qatari=0.
If(hr01 eq 1) qatari=1.
compute whitecollar=0.
if(hr01 eq 2) whitecollar=1.
compute bluecollar=0.
if(hr01 eq 3) bluecollar=1.
freq var=qatari whitecollar
bluecollar.
```



Why do we want to keep track of what's going on behind the point-and-click commands?

- Keeping records
- Preparing for replication
- Catching mistakes

Frequency Table:
Dummy Variables for Three Strata

qatari

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1450	67.8	67.8	67.8
	1.00	689	32.2	32.2	100.0
	Total	2139	100.0	100.0	

whitecollar

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1371	64.1	64.1	64.1
	1.00	768	35.9	35.9	100.0
	Total	2139	100.0	100.0	

bluecollar

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1457	68.1	68.1	68.1
	1.00	682	31.9	31.9	100.0
	Total	2139	100.0	100.0	

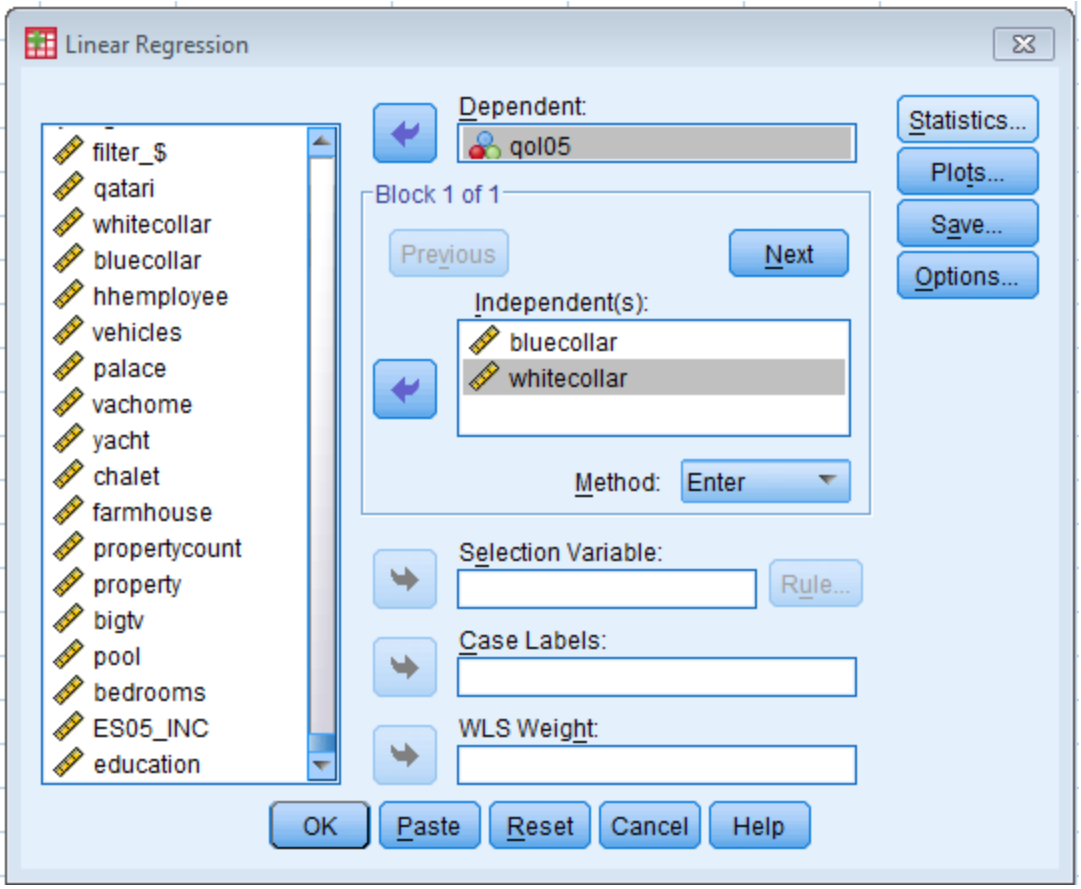
Evaluations of personal financial situation

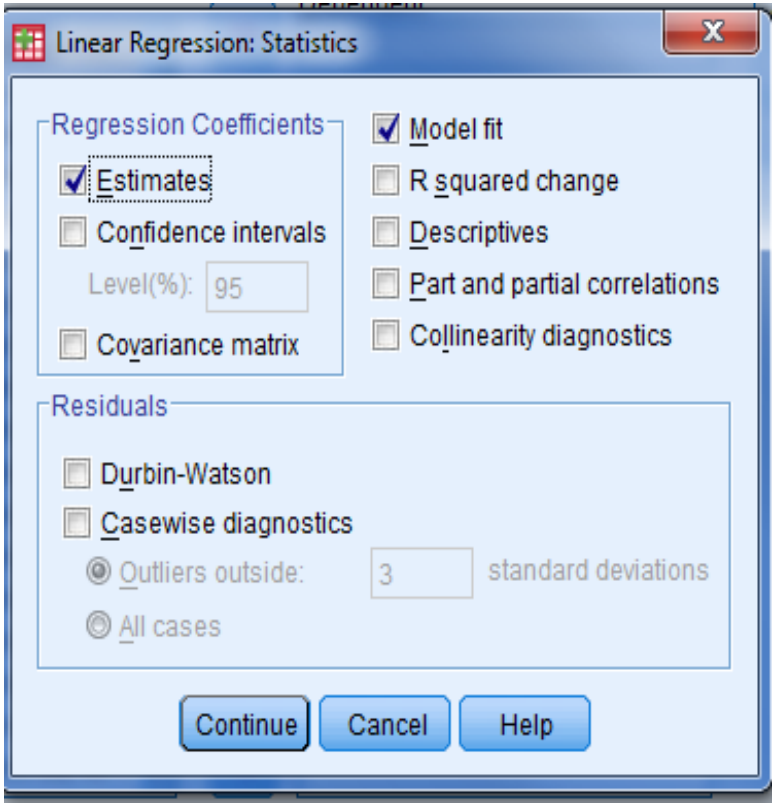
overall, how would you rate your own personal financial situation these days?

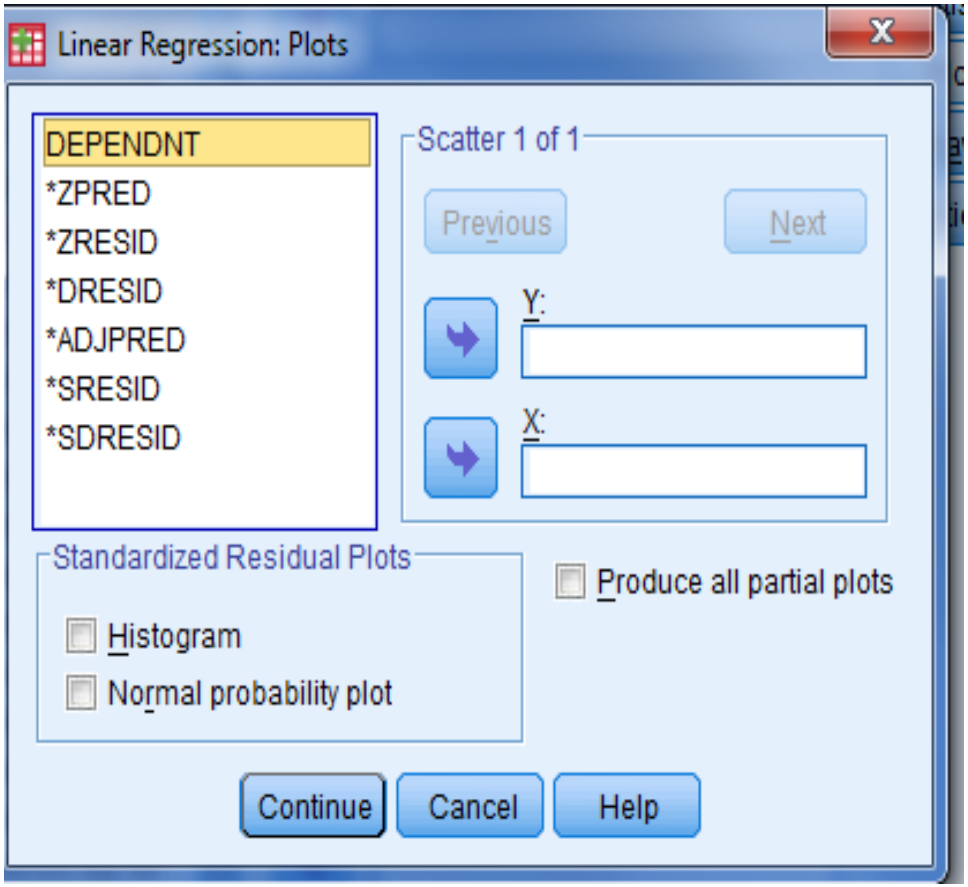
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. excellent	243	11.4	11.6	11.6
	2. good	879	41.1	41.9	53.4
	3. fair	738	34.5	35.1	88.6
	4. poor	230	10.8	11.0	99.5
	8. don't know	6	.3	.3	99.8
	9. refused	4	.2	.2	100.0
	Total	2100	98.2	100.0	
Missing	System	39	1.8		
Total		2139	100.0		

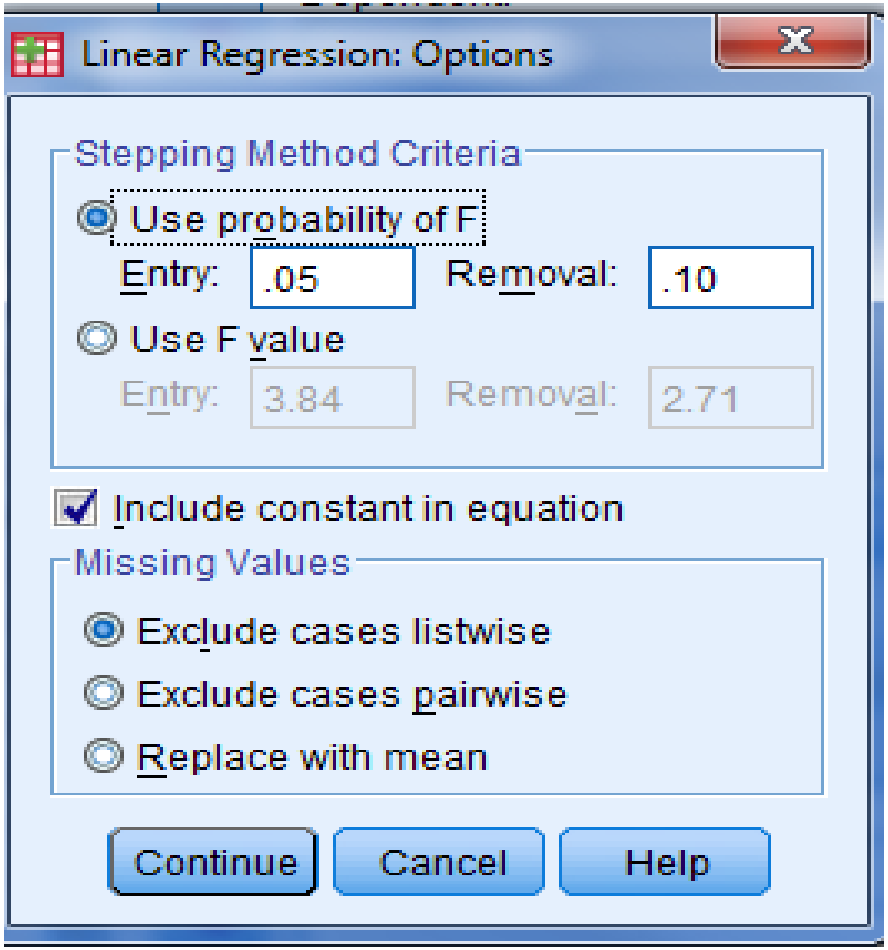
Regression

Menu option: Analyze / Regression / Linear

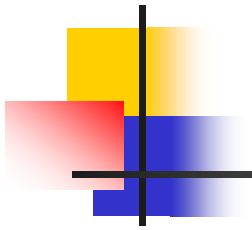





Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
X:
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help


Linear Regression: Options
Stepping Method Criteria
Use probability of F
Entry: .05
Removal: .10
Use F value
Entry: 3.84
Removal: 2.71
Include constant in equation
Missing Values
Exclude cases listwise
Exclude cases pairwise
Replace with mean
Continue
Cancel
Help

What Is Going On Behind the Point-And-Click Commands?



```
REGRESSION
  /MISSING LISTWISE
  /REGWGT=wgt_sp
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT currentfin
  /METHOD=ENTER whitecollar bluecollar.
```

SPSS Printouts from Regression Model: Two Independent Variables.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	bluecollar, whitecollar[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: subjective economic status with missings as system missing

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.426[a]	.181	.181	.76061

a. Predictors: (Constant), bluecollar, whitecollar

SPSS Printouts from Regression Model: Two Independent Variables

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	266.468	2	133.234	230.297	.000[a]
	Residual	1202.456	2078	.579		
	Total	1468.924	2080			

a. Predictors: (Constant), bluecollar, whitecollar

b. Dependent Variable: subjective economic status with missings as system missing

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.120	.030		71.270	.000
	whitecollar	.202	.041	.116	4.970	.000
	bluecollar	.851	.042	.475	20.406	.000

a. Dependent Variable: subjective economic status with missings as system missing

Predicting Subjective Economic Status
(Ordinary Least Squares)

	Coefficient
White Collar	0.20*
	(.04)
Blue Collar	0.85*
	(.04)
Constant	2.12*
	(.03)

Adjusted R-squared: .18, N=2080
* p<.05.
Subjective economic status ranges from 1 (excellent) to 4 (poor).
Standard errors in parentheses.
Source: SESRI Omnibus Survey, 2010.

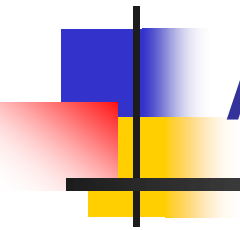
Making sense of our results

- How do we interpret the coefficients? How is this an intercept shift? Why do we exclude one category? What if we excluded a different category?
- How do we interpret the other numbers on the table? Why do we include those?
  - The n
  - The adjusted R-squared
  - The definition of the asterisk
- How can we improve this model?

Predicting Subjective Economic Status (Ordinary Least Squares)

	Old Coefficient	New Coefficient
White Collar	0.20*	-0.65*
	(.04)	(.04)
Blue Collar	0.85*	
	(.04)	
Qatari		-0.85*
		(.04)
Constant	2.12*	2.97*
	(.03)	(.03)
Adjusted R-squared:	.18	.18
N	2080	2080

* p<.05. Standard errors in parentheses.
Subjective economic status ranges from 1 (excellent) to 4 (poor).
Source: SESRI Omnibus Survey, 2010.



Adding a continuous variable:

Age

SPSS Output from Regression Model

Notes		
Output Created		23-Mar-2011 17:56:17
Comments		
Input	Data	C:\Documents and Settings burns\My Documents\Downloads\QU_UM_April_Training_Dataset_v2.sav
	Active Dataset	DataSet1
	Filter	survey=1. (FILTER)
	Weight	weight variable to be use in spss
	Split File	<none>
	N of Rows in Working Data File	2139
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax		REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT currentfinancialsituation /METHOD=ENTER whitecollar bluecollar hr09.
Resources	Processor Time	00:00:00.125
	Elapsed Time	00:00:00.125
	Memory Required	18732 bytes
	Additional Memory Required for Residual Plots	0 bytes

SPSS Printouts from Regression Model: Three Independent Variables

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	hh member's age, whitecollar, bluecollar[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: subjective economic status with missings as system missing

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.430[a]	.184	.183	.75937

a. Predictors: (Constant), hh member's age, whitecollar, bluecollar

SPSS Printouts from Regression Model: Three Independent Variables

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	270.983	3	90.328	156.646	.000[a]
	Residual	1197.941	2077	.577		
	Total	1468.924	2080			

a. Predictors: (Constant), hh member's age, whitecollar, bluecollar

b. Dependent Variable: subjective economic status with missings as system missing

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.284	.066		34.819	.000
	whitecollar	.205	.041	.117	5.034	.000
	bluecollar	.825	.043	.460	19.331	.000
	hh member's age	-.004	.002	-.058	-2.798	.005

a. Dependent Variable: subjective economic status with missings as system missing

Predicting Subjective Economic Status (Ordinary Least Squares)

	Coefficient New Model	Coefficient Old Model
White Collar	0.20*	0.20*
	(.04)	(.04)
Blue Collar	0.85*	0.85*
	(.04)	(.04)
Age	-0.004*	
	(.002)	
Constant	2.28*	2.12*
	(.07)	(.03)
Adjusted R-squared:	.18	.18

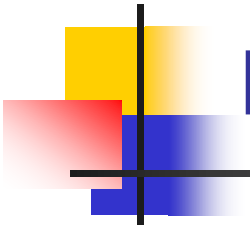
N=2080* p<.05.

Subjective economic status ranges from 1 (excellent) to 4 (poor).

Standard errors in parentheses.

Source: SESRI Omnibus Survey, 2010.

What do we learn from this new model?



- Compare the old model to the new. What's different?
- Are there hints about our next specification?
- Are there changes in presentation that would convey more information?

Predicting Subjective Economic Status (Ordinary Least Squares)

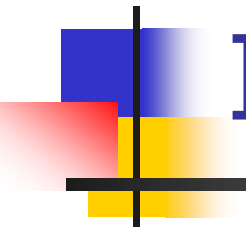
	Coefficient New Model	Coefficient Old Model
White Collar	0.20*	0.20*
	(.04)	(.04)
Blue Collar	0.85*	0.85*
	(.04)	(.04)
Age, Scaled 0 to 1	-0.32*	
	(.11)	
Constant	2.28*	2.12*
	(.07)	(.03)
Adjusted R-squared:	.18	.18

N=2080* p<.05.

Subjective economic status ranges from 1 (excellent) to 4 (poor).

Standard errors in parentheses.

Source: SESRI Omnibus Survey, 2010.



Interaction terms:

Age by Strata

Slope Shifts

- Source: Hanushek and Jackson

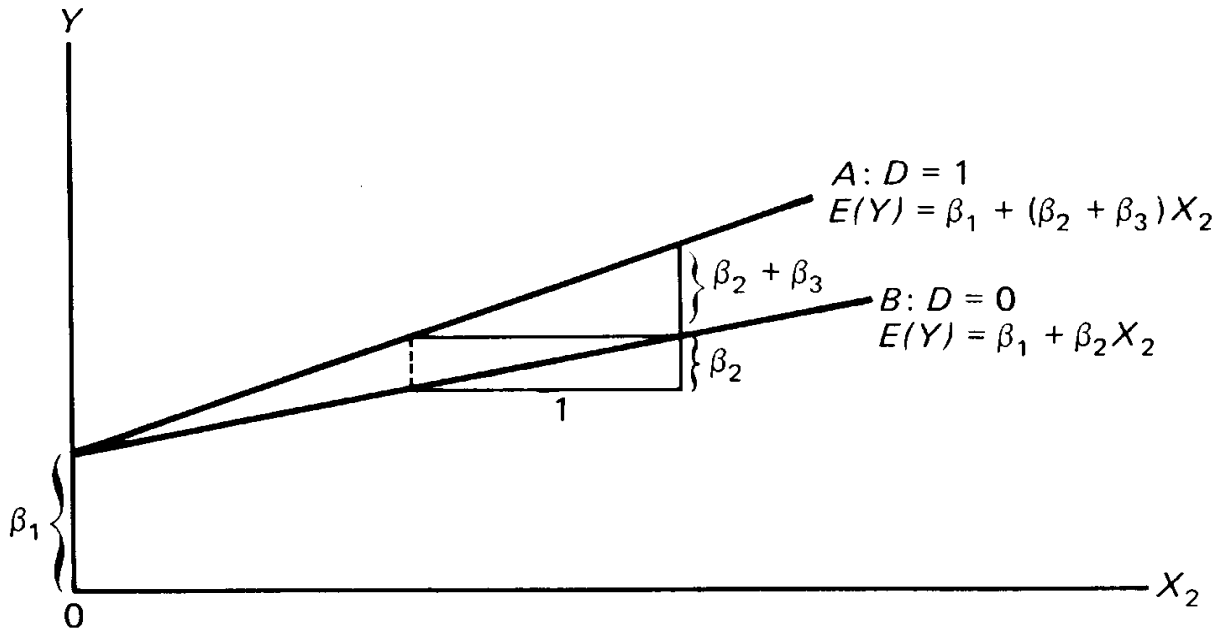


FIGURE 4.5 Bivariate relationship with slope dummy variable.

SPSS Printouts from Regression Model: Five Independent Variables, with Interaction Terms

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	agewhitecollar, hh member's age, agebluecollar, whitecollar, bluecollar[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: subjective economic status with missings as system missing

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.434[a]	.189	.187	.75781

a. Predictors: (Constant), agewhitecollar, hh member's age, agebluecollar, whitecollar, bluecollar

SPSS Printouts from Regression Model: Five Independent Variables, with Interaction Terms

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	277.031	5	55.406	96.480	.000[a]
	Residual	1191.893	2075	.574		
	Total	1468.924	2080			

a. Predictors: (Constant), agewhitecollar, hh member's age, agebluecollar, whitecollar, bluecollar

b. Dependent Variable: subjective economic status with missings as system missing

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.117	.089		23.716	.000
	whitecollar	.426	.138	.243	3.087	.002
	bluecollar	1.284	.149	.716	8.598	.000
	hh member's age	8.519E-5	.002	.001	.039	.969
	agebluecollar	-.013	.004	-.252	-3.178	.002
	agewhitecollar	-.006	.003	-.135	-1.686	.092

a. Dependent Variable: subjective economic status with missings as system missing

Predicting Subjective Economic Status (Ordinary Least Squares)

	Coefficient New Model	Coefficient Old Model	Coefficient Really Old Model
White Collar	0.43*	0.20*	0.20*
	(.14)	(.04)	(.04)
Blue Collar	1.28*	0.85*	0.85*
	(.15)	(.04)	(.04)
Age	.000	-0.004*	
	(.002)	(.002)	
Age*White Collar	-.006*		
	(.003)		
Age*Blue Collar	-.013*		
	(.004)		
Constant	2.12*	2.28*	2.12*
	(.09)	(.07)	(.03)
Adjusted R-squared:	.19	.18	.18

N=2080; * p<.05. Standard errors in parentheses.
Subjective economic status ranges from 1 (excellent) to 4 (poor).
Source: SESRI Omnibus Survey, 2010.

How do we interpret the new model, in light of the *old* and the *really old* model?

How do we interpret the interaction terms?

What does a slope shift mean?

How do we compare the three model specifications?

Do our data have enough information to carry the more elaborate specification? What are the hints?

For further reading

Wonnacott and Wonnacott. 1990. Introductory Statistics for Business and Economics, 4th edition. John Wiley and Sons.

For those comfortable with more mathematics:

William H. Greene. 2008. Econometric Analysis, 6th edition. Prentice-Hall.

Thank you!