Ordinary Least Squares Regression

March 2013

Nancy Burns (nburns@isr.umich.edu) - University of Michigan

From description to cause

Group	Sample Size	Mean Health Status	Standard Error
Hospital	7,774	3.21	.014
No Hospital	90,049	3.93	.003

Source: Angrist and Pischke, 2009.

- How would we interpret this comparison of means?

- Is this a description?

- Is a hospital stay a ***cause*** of health status?

- What are the problems with thinking about hospital stay as a cause?

Selection Bias

- When is a “treatment” not a cause?
- When is a “treatment” a cause?
- When are we in between these two?

Thought exercise about selection bias

- Working in groups, design a study to investigate the effects of primary school class size on student achievement.
- What are the things you have to do and to take into account for the results of your study to provide information about whether class size affects student achievement?
- What are the mistakes one could make?

Ordinary Least Squares (OLS) Regression

- *Dependent variable*, Y, what we're explaining.
- *Explanatory variable* or *independent variable* or *treatment*, X. This is the variable we'd like to think of as a cause, the variable 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?

What does this look like? When X goes up, what happens to Y?

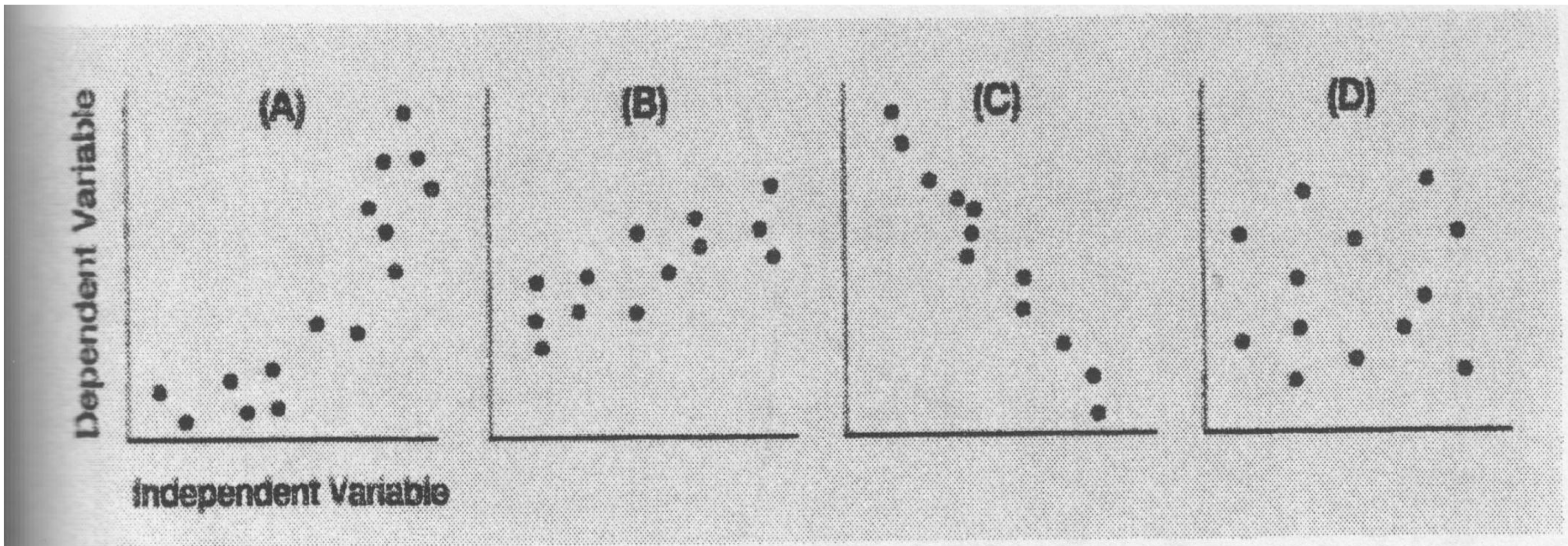


Figure 7–2 Assorted Scattergrams

Source: Shively, 2005

Scatterplots

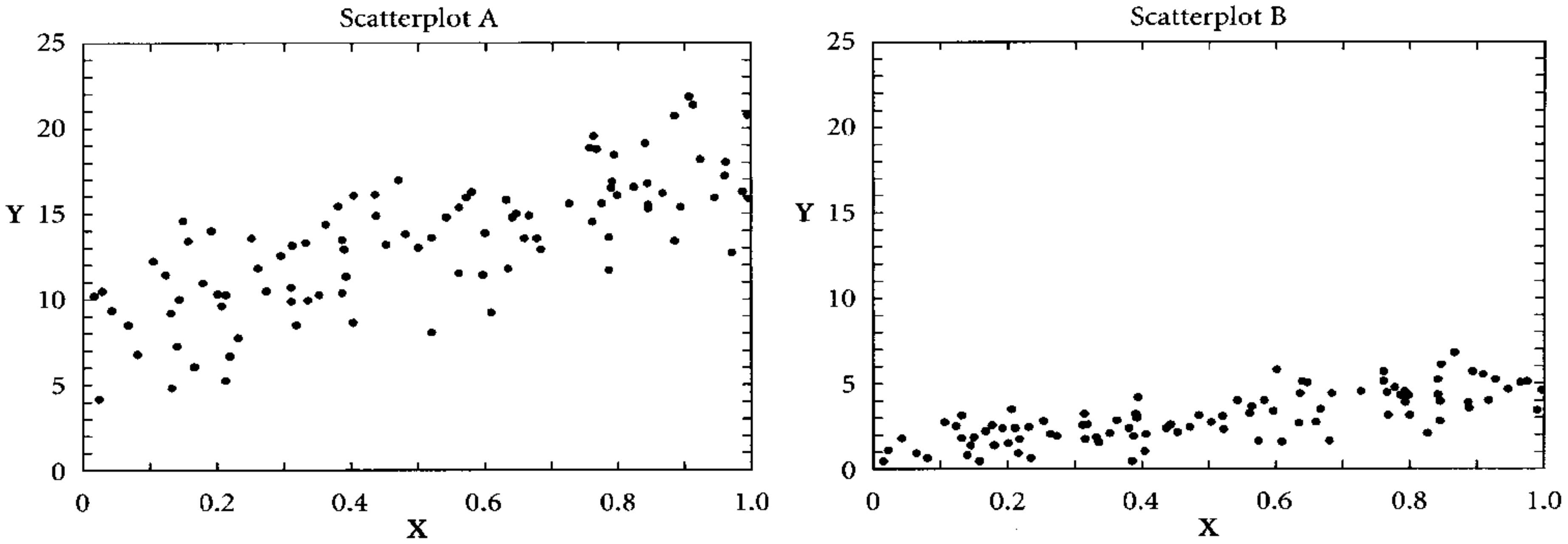


FIGURE 1.3 Two scatterplots with correlation coefficients of +0.75

Source: Berry and Sanders, 2000.

We want a way to describe this relationship.

When we use Ordinary least squares, we are describing the relationship this way:

The predicted value of $Y = a + bX$

The Regression Line

The predicted value of Y =

intercept + slope * X

Y is the dependent variable

X is the explanatory variable

The Regression Line

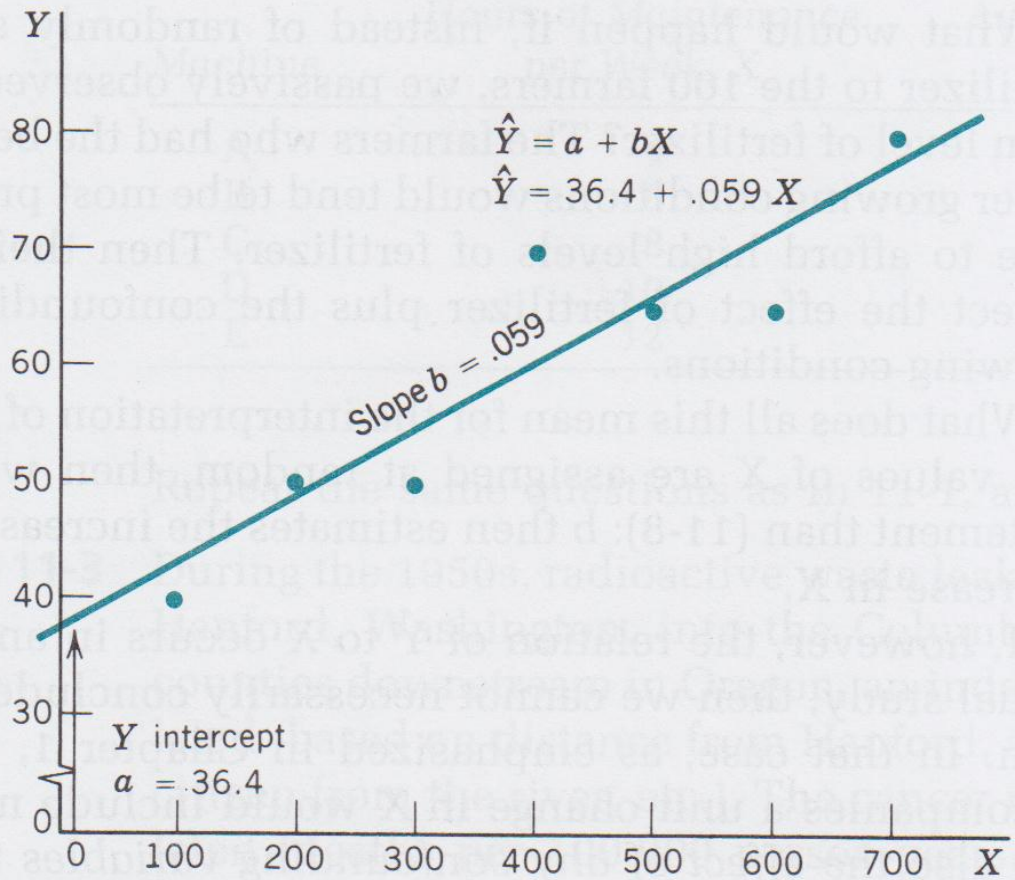


FIGURE 11-4

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

Source: Wonnacott and Wonnacott, 1990.

The Regression Line

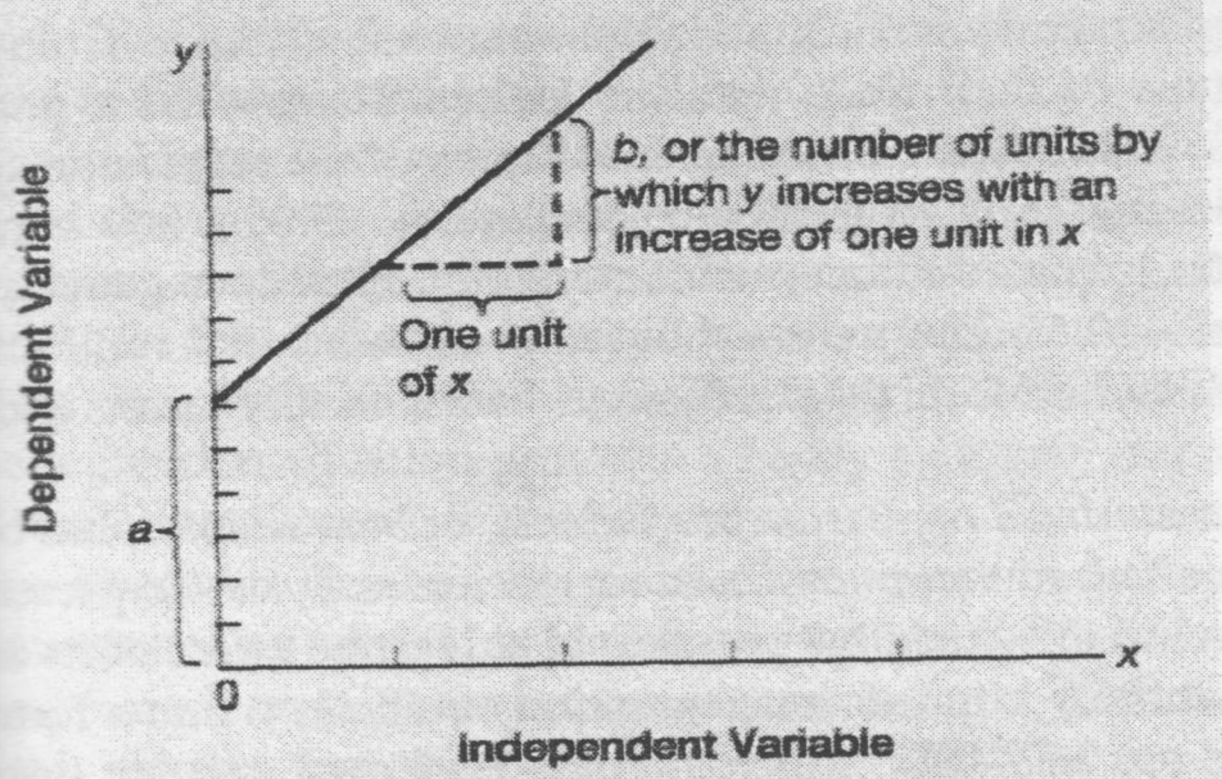
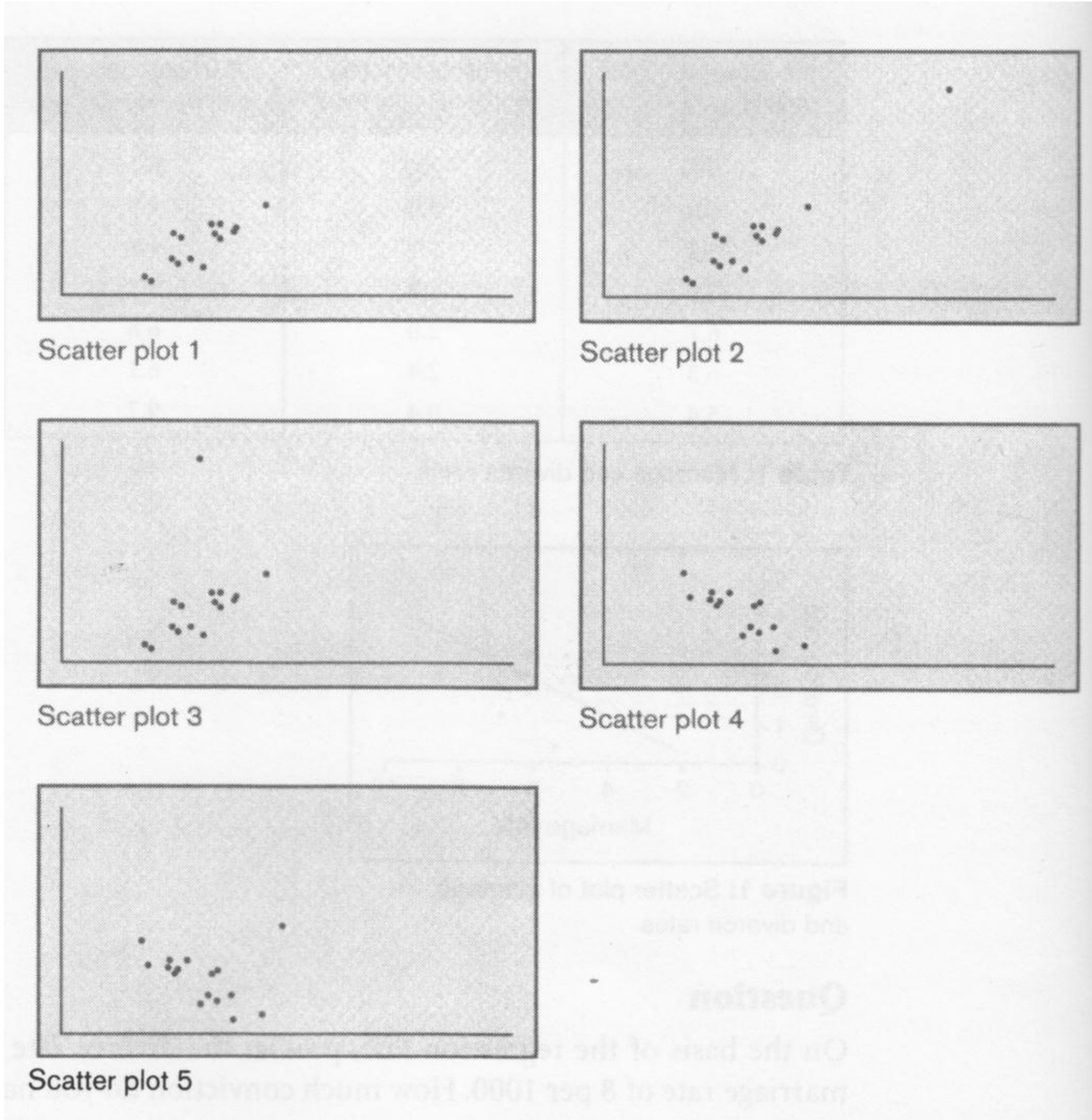


Figure 7–4 The Regression Equation

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.

Source: Shively, 2005.

Group exercise



Source: Scheaffer.

Some questions for the group exercise

- When x goes up by one unit, for which of these slides does y go down?

- If you were drawing a line to describe the points for the graphs where y goes down when x goes up, which would have the steeper slope? For which one would y go down more as x goes up?

- Three of these have exactly the same coefficient on x. Which three?

- The three have different correlations between x and y; which is higher, and which is lower?

How do we calculate a and b, the intercept (or constant) and the slope?

Minimize the sum of squared residuals.

Would use calculus and calculate partial derivatives with respect to a and b.

Residual

A residual is the difference between our observation, y, and the predicted value of y from our model.

We want the difference to be small.

Minimizing the Sum of Squared Residuals

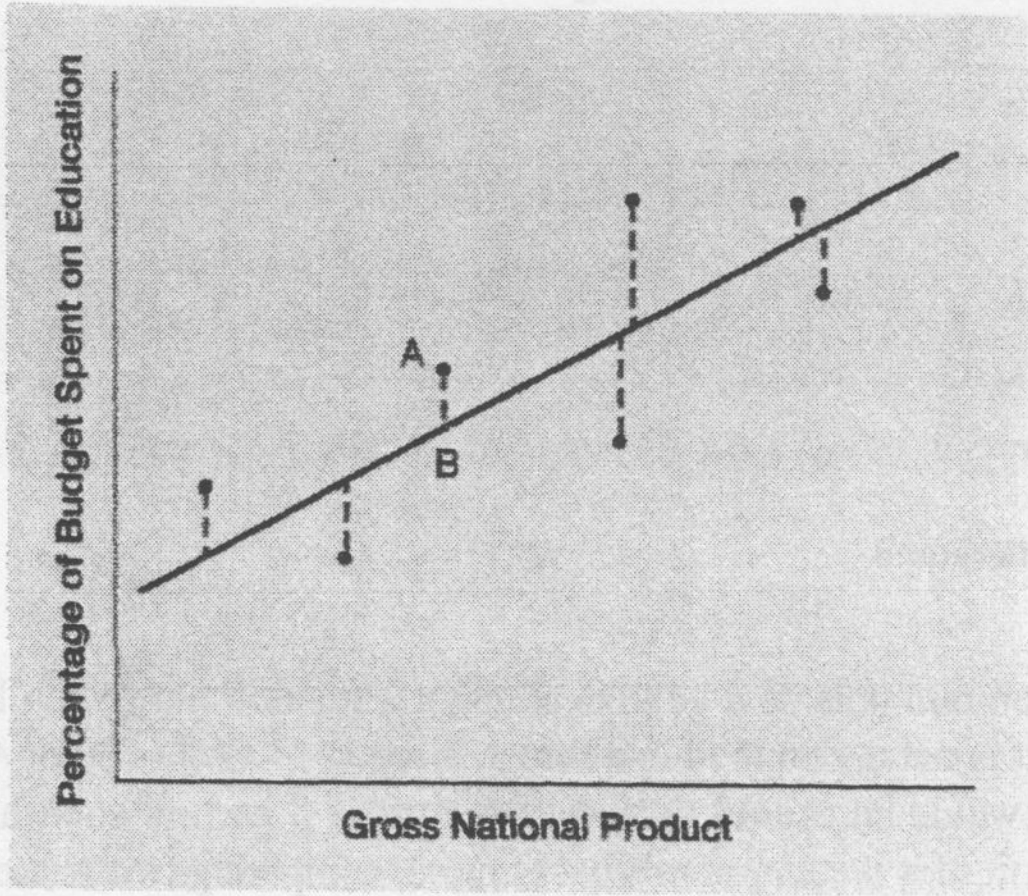


Figure 7–3 The Regression Line

Source: Shively, 2005.

Minimizing the vertical distance

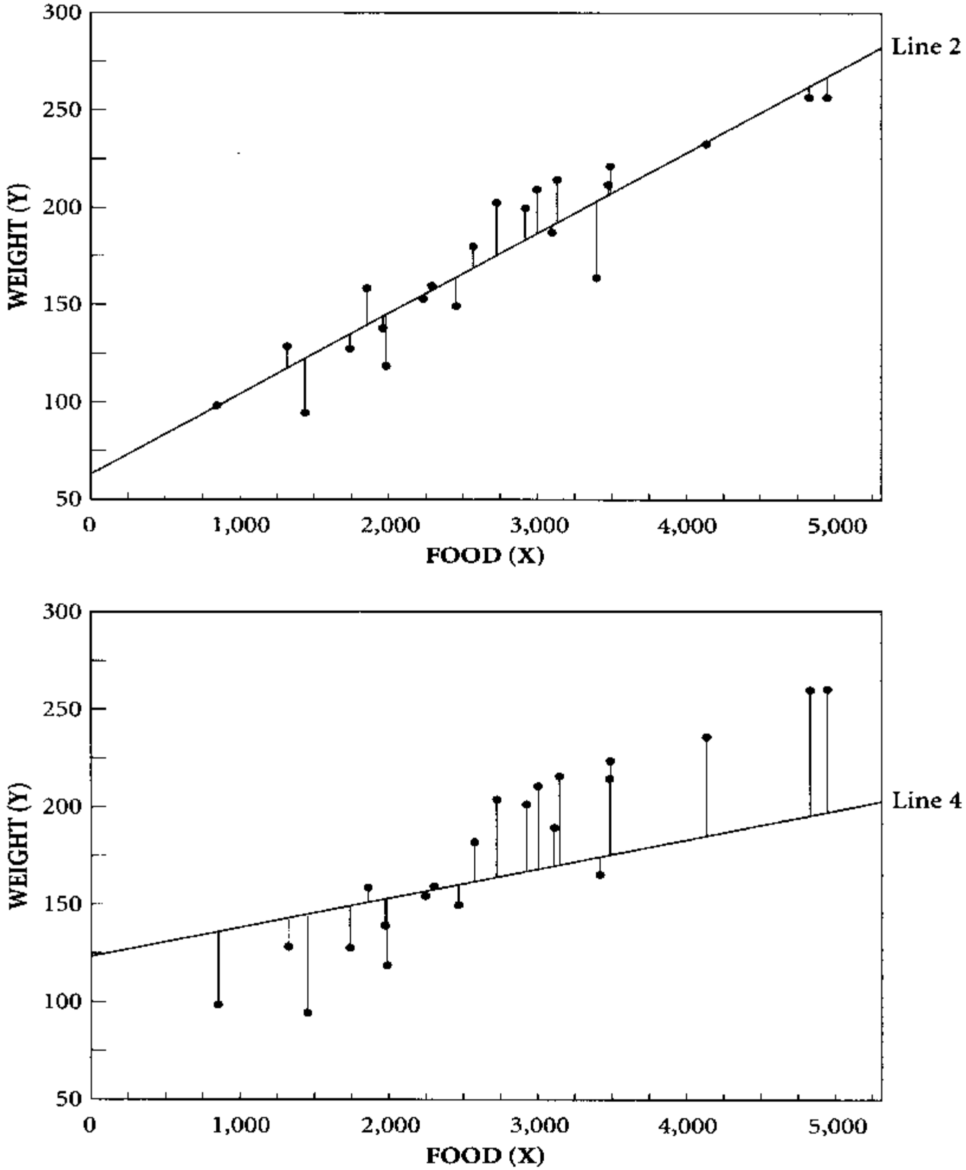


FIGURE 2.3 Vertical distances between points and two lines

Source: Berry and Sanders, 2000.

We could do the math and calculate the coefficients, but we wouldn't yet have the tools to draw inferences to data we don't have.

Without one more tool, all we have is a way to describe our data. **Without one more tool, we do not have a way to say how certain we are about that description.**

Inference

Our challenge is that we are not describing a full population.

Instead, we are drawing an inference from a sample to describe a population.

We need assumptions and tools from probability to allow us to draw these inferences.

Inference from Samples

The tools from probability and the assumptions we will make allow us to say how certain we are about the estimates we calculate with our sample.

We'll estimate the standard errors of our regression coefficients.

These standard errors are our measures of the variability of b and a. They are a function of the variability of y and x and of the sample size.

For example, when there's little variance in x, we have little certainty about b, and our estimates of the variability of b will be quite large. When our samples are small, our estimate of the variability of our coefficients will be larger.

The Value of Small Standard Errors

These standard errors describe our estimate of the sampling distribution of b and a.

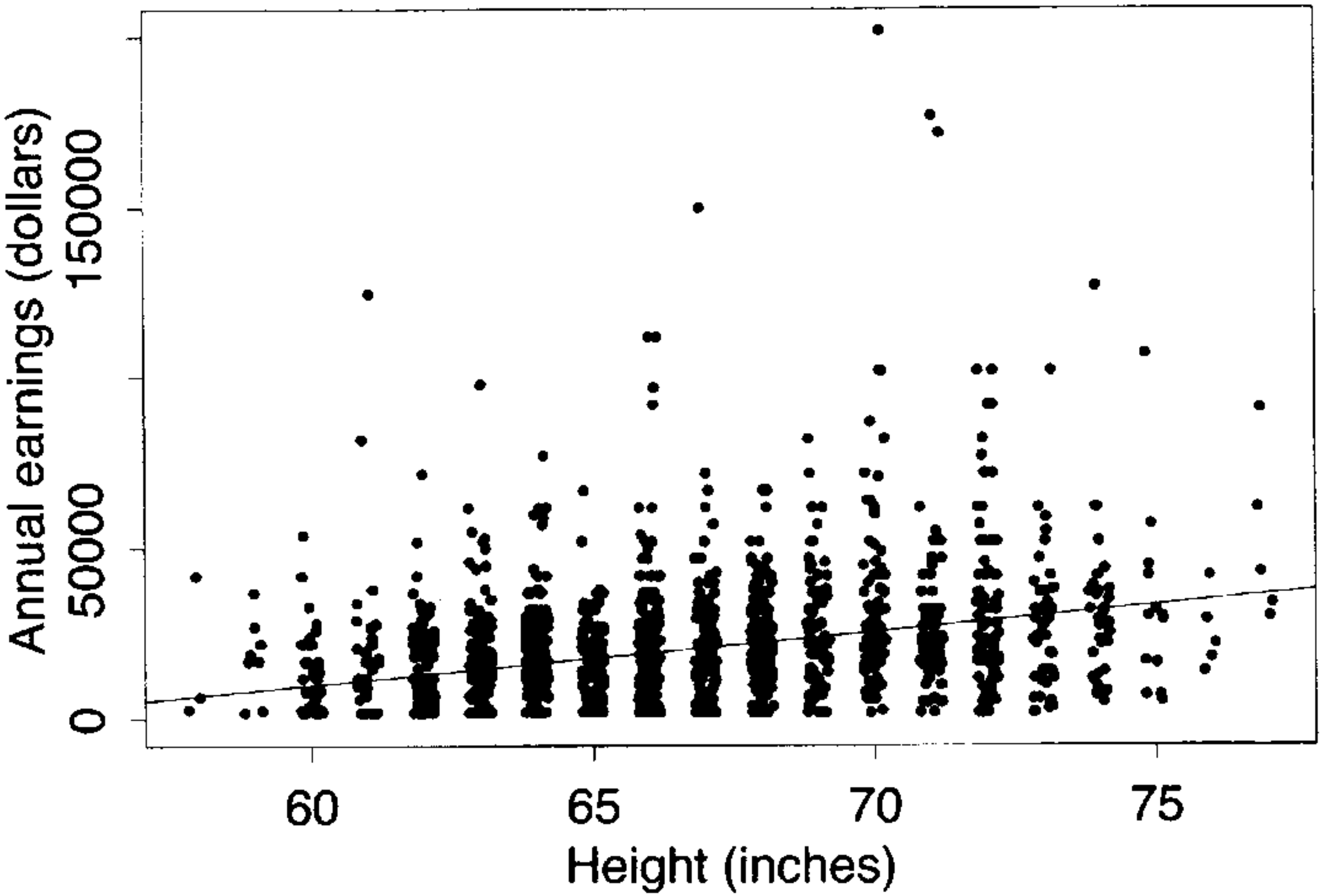
They give us the ability to describe certainty around b.

They let us say how certain we are about the estimates we’ve calculated from our sample.

When the **absolute value** of t for our coefficient is greater than or equal to the critical value of t at a particular level of confidence, we have a measure of how certain we are about the coefficient at hand.

Conventionally, we use a 95% confidence interval, or a .05 level of statistical significance. Often, we also report the level of statistical significance.

Scatterplot of height and earnings



Source: Gelman and Nolan, 2002.

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?
- Is height a treatment, a cause?

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.

Control variables

- **Bad controls** – might just as well be the dependent variable.
- **Good controls** – things we can think of as fixed by the time the value of the dependent variable came to be.
- Is gender in this model a bad control or a good control?
- What are strategies for turning a bad control into a good control?

Group Exercise
Another Multivariate Model

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

Source: Burns, Schlozman, and Verba, 2001.

* p<.05; ** p<.01; *** p < .001. Controlling for other variables.

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 – alternative explanations -- that will affect our estimates of the coefficients at hand?
  - Are the controls good controls?
  - Are these explanatory variables "treatments"?
  - Describe your conclusions and your certainty about your conclusions.
  - What do you wish were on this table that isn't here?

Predicting Free Time

	Women	**Men**
Marriage	-0.86***	-.32***
Pre-school Children	-2.29***	-.53***
School-aged Children	-0.88***	-.53***

Source: Burns, Schlozman, and Verba 2001.

Controlling for education, activity in high school, race or ethnicity, age, hours on the job, job level, and citizenship

*Coefficient significant at < .05.

**Coefficient significant at < .01.

***Coefficient significant at < .001.

Predicting Level of Education (US, from GSS data)

	1972	**2006**
Parents' Education	**.379***	**.415***
	(.028)	(.015)
Rural	**-.029***	**-.029***
	(.013)	(.013)
Age 26-35	.024	**.035***
	(.019)	(.011)
Age 36-45	.001	**.025***
	(.015)	(.011)
Age 46-55	-.006	**.028***
	(.018)	(.011)
Age 56-65	**-.040***	**.042***
	(.019)	(.012)
Age 66 and older	**-.074***	**.058***
	(.020)	(.010)
Female	-.005	-.011
	(.010)	(.007)
R-squared	.312	.293
N	597	1379

Interpreting coefficients

Ask the questions:

Compared to what?

Good control?

Treatments?

Alternative explanations?

Omitted Variables

Number of hours of TV watching per day

	B (s.e.)	t	p value
Education	-2.02 (.22)	-9.29	.0000
Age	.14 (.21)	.66	.5124
Age over 65	.75 (.19)	4.04	.0001

Adjusted R-squared .06

N 2494

Omitted variables

Number of hours of TV watching per day

	B (s.e.)	t	p value
Education	-1.55 (.21)	-7.35	.0000
Age	-.15 (.20)	-.74	.4588
Age over 65	-.06 (.19)	-.008	.7633
In the workforce	-1.57 (.11)	-14.466	.0000
Adjusted R-squared	.13		
N	2494		

Know your data

Know how it was collected.

Know what the data look like.

Know how the variables are distributed.

Know what the residuals look like.

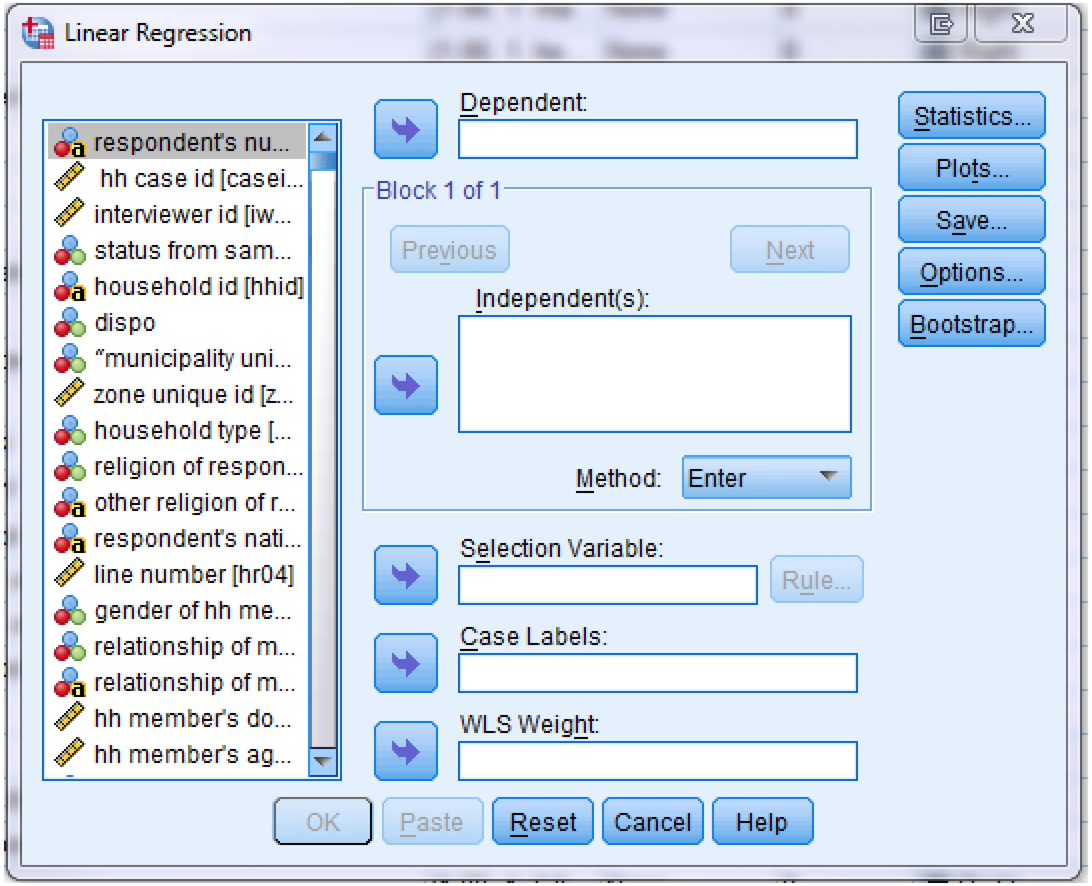
Explore the difference between observations your model predicts well and cases your model doesn’t predict well.

Aim for Resilience

Push your model. Are there other reasonable “specifications”, reasonable sets of ways to implement your theoretical ideas? Do your results hold up when implemented in these other ways? Know the limits of your model – which coefficients are sturdy, and which are not?

Performing a Regression in SPSS

The OLS regression option is found under the *Analyze / Regression / Linear* menu.



Exercise

The 2011 Omnibus includes several items about traffic laws and driving in Qatar. One question asks whether respondents have received a traffic ticket in the past 12 months (variable name=trafficticket. This variable was recoded from variable b2a in your codebook).

What factors might explain whether someone received a traffic ticket in the past 12 months?

One possibility is the age of the individual. What hypothesis can we generate regarding the relationship between an individual’s age and whether they received a traffic ticket?

Run a regression to test this hypothesis.

B2a. People receive Traffic penalty (ticket) from time to time for different reasons, during the past 12 months, have you received a traffic penalty within the State of Qatar?

1 YES
2 NO
3 DON'T REMEMBER
4 REFUSED

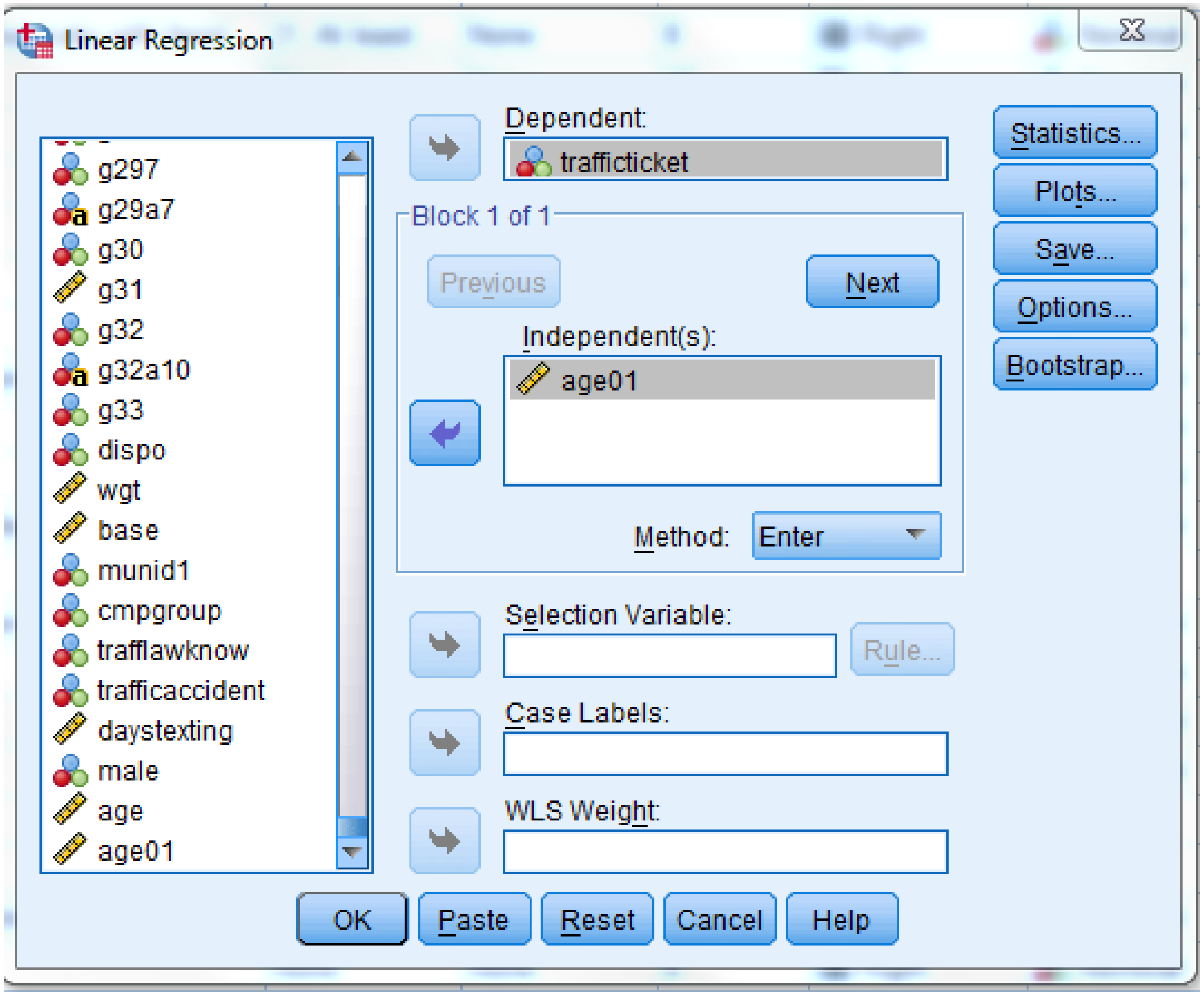
Received a traffic ticket in Qatar in past 12 months

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	662	33.1	53.6	53.6
	Yes	573	28.7	46.4	100.0
	Total	1235	61.8	100.0	
Missing	99.00	765	38.3		
Total		2000	100.0		

Note: The age variable was created by subtracting respondents' answer to question F14 in the codebook (what year were you born) from 2011 (the year the survey was conducted).

Respondent's age scaled 0 to 1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	2	.1	.1	.1
	.04	3	.2	.2	.3
	.05	26	1.3	1.3	1.6
	.07	28	1.4	1.4	3.0
	.09	32	1.6	1.6	4.6
	.11	33	1.7	1.7	6.3
	.13	29	1.5	1.5	7.7
	.14	33	1.7	1.7	9.4
	.16	28	1.4	1.4	10.8
	.18	32	1.6	1.6	12.4
	.20	40	2.0	2.0	14.5
	.21	38	1.9	1.9	16.4
	.23	44	2.2	2.2	18.6
	.25	58	2.9	2.9	21.5
	.27	56	2.8	2.8	24.4
	.29	72	3.6	3.6	28.0
	.30	54	2.7	2.7	30.7
	.32	55	2.8	2.8	33.5
	.34	68	3.4	3.4	36.9
	.36	71	3.6	3.6	40.5
	.38	82	4.1	4.1	44.7
	.39	76	3.8	3.8	48.5
	.41	67	3.4	3.4	51.9
	.43	59	2.9	3.0	54.9
	.45	64	3.2	3.2	58.1
	.46	87	4.4	4.4	62.5
	.48	62	3.1	3.1	65.6
	.50	59	2.9	3.0	68.6
	.52	43	2.2	2.2	70.8
	.54	47	2.4	2.4	73.2
	.55	47	2.4	2.4	75.5
	.57	42	2.1	2.1	77.7
	.59	46	2.3	2.3	80.0
	.61	36	1.8	1.8	81.8
	.63	37	1.8	1.9	83.7
	.64	56	2.8	2.8	86.5
	.66	33	1.7	1.7	88.2
	.68	26	1.3	1.3	89.5
	.70	29	1.5	1.5	91.0
	.71	26	1.3	1.3	92.3
	.73	28	1.4	1.4	93.7
	.75	20	1.0	1.0	94.7
	.77	5	.3	.3	94.9
	.79	16	.8	.8	95.8
	.80	18	.9	.9	96.7
	.82	14	.7	.7	97.4
	.84	4	.2	.2	97.6
	.86	7	.4	.4	97.9
	.88	8	.4	.4	98.3
	.89	4	.2	.2	98.5
	.91	5	.3	.3	98.8
	.93	3	.2	.2	98.9
	.95	4	.2	.2	99.1
	.96	2	.1	.1	99.2
	.98	5	.3	.3	99.5
	1.00	10	.5	.5	100.0
	Total	1979	99.0	100.0	
Missing	System	21	1.1		
Total		2000	100.0		


Linear Regression
g297
g29a7
g30
g31
g32
g32a10
g33
dispo
wgt
base
munid1
cmpgroup
trafflawknow
trafficaccident
daystexting
male
age
age01
Dependent:
trafficticket
Block 1 of 1
Previous
Next
Independent(s):
age01
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	age01[b]	.	Enter

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.088[a]	.008	.007	.49722

a. Predictors: (Constant), age01

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.373	1	2.373	9.598	.002[b]
	Residual	303.840	1229	.247		
	Total	306.213	1230			

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. Predictors: (Constant), age01

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.567	.036		15.791	.000
	age01	-.231	.075	-.088	-3.098	.002

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

Other tools we have in Regression Analysis

Intercept shifts

Source: Hanushek and Jackson

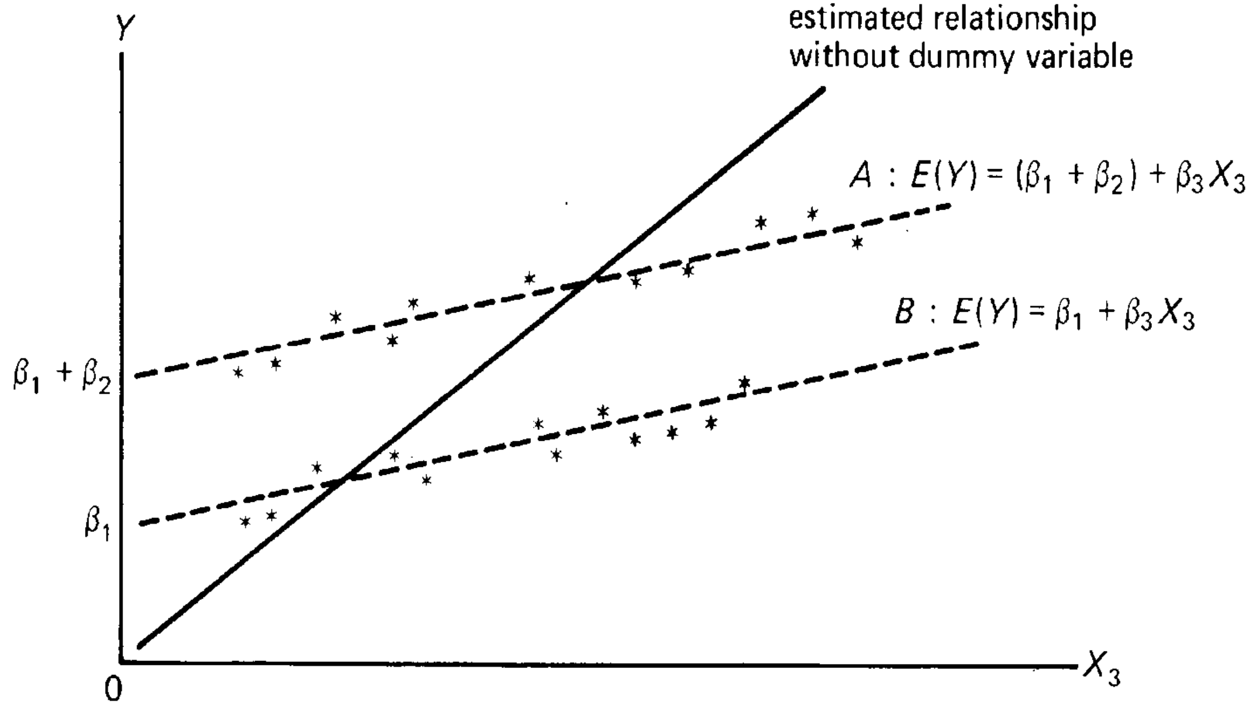


FIGURE 4.4 Estimation of misspecified bivariate relationship excluding dummy variable.

An example of an intercept shift

Is gender a factor in receiving a traffic ticket in the past 12 months in Qatar? Are men more likely to receive a traffic ticket than women?

How would we do this?

Our independent variable, male, has two values:

male

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	980	49.0	49.0	49.0
	1.00	1020	51.0	51.0	100.0
	Total	2000	100.0	100.0	

1 = Respondent is male

0 = Respondent is female

Our dependent variable, trafficticket:

Received a traffic ticket in Qatar in past 12 months

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	662	33.1	53.6	53.6
	Yes	573	28.7	46.4	100.0
	Total	1235	61.8	100.0	
Missing	99.00	765	38.3		
Total		2000	100.0		

0 = No
1= Yes

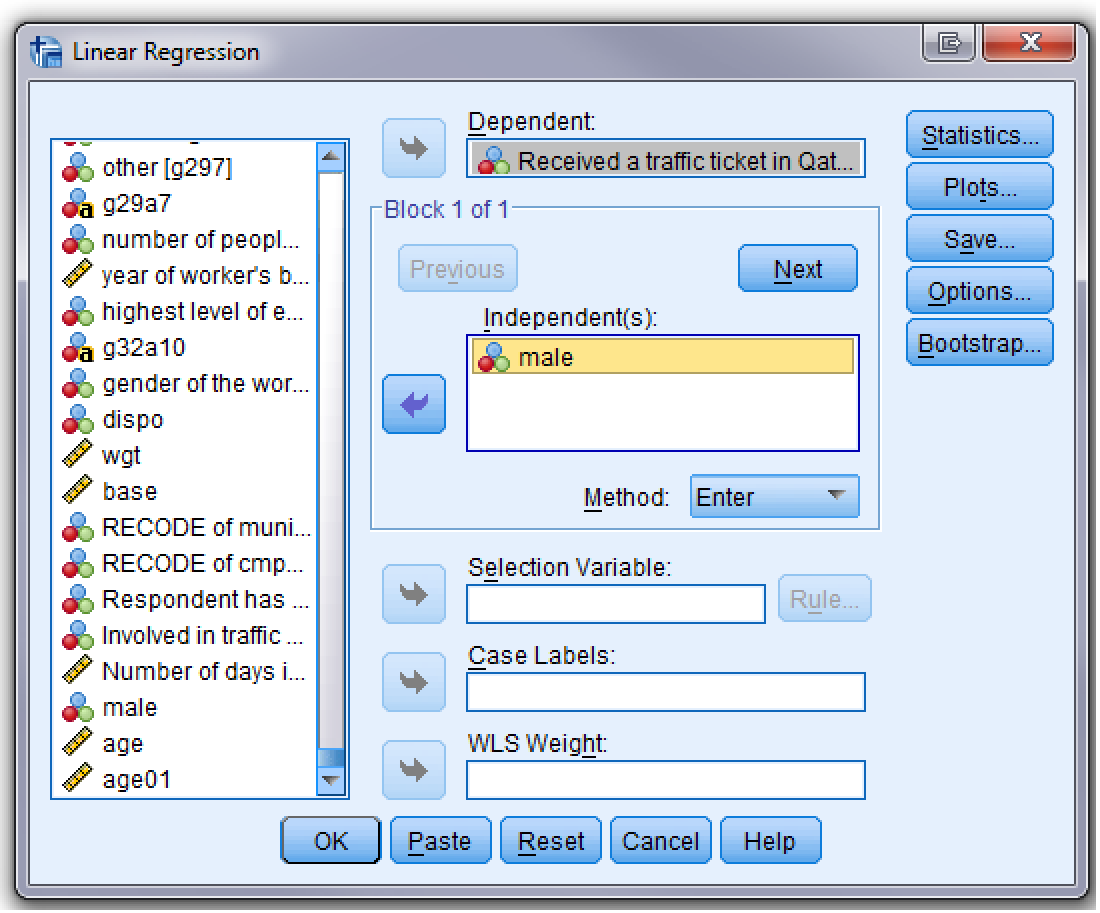
Comparing receiving a traffic ticket in the past 12 months by gender.

Received a traffic ticket in Qatar in past 12 months * male Crosstabulation

			male		Total
			.00	1.00	
Received a traffic ticket in Qatar in past 12 months	No	Count	199	463	662
		% within Received a traffic ticket in Qatar in past 12 months	30.1%	69.9%	100.0%
		% within male	63.4%	50.3%	53.6%
	Yes	Count	115	458	573
		% within Received a traffic ticket in Qatar in past 12 months	20.1%	79.9%	100.0%
		% within male	36.6%	49.7%	46.4%
Total		Count	314	921	1235
		% within Received a traffic ticket in Qatar in past 12 months	25.4%	74.6%	100.0%
		% within male	100.0%	100.0%	100.0%

Regression

Menu option: Analyze / Regression / Linear



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

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT trafficticket
  /METHOD=ENTER male.
```

SPSS Printouts from Regression Model

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	male[b]	.	Enter

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.114[a]	.013	.012	.49583

a. Predictors: (Constant), male

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.021	1	4.021	16.357	.000[b]
	Residual	303.125	1233	.246		
	Total	307.147	1234			

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. Predictors: (Constant), male

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.366	.028		13.089	.000
	male	.131	.032	.114	4.044	.000

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

SPSS Printouts from Regression Model

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	male[b]	.	Enter

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.114[a]	.013	.012	.49583

a. Predictors: (Constant), male

SPSS Printouts from Regression Model

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.021	1	4.021	16.357	.000[b]
	Residual	303.125	1233	.246		
	Total	307.147	1234			

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. Predictors: (Constant), male

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.366	.028		13.089	.000
	male	.131	.032	.114	4.044	.000

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

Predicting Receiving a Traffic Ticket
(Ordinary Least Squares)

	Coefficient
Male	.131* (.032)
Constant	.366* (.028)

Adjusted R-squared: .012, N=1234
* p<.05.
Receiving a Traffic Ticket ranges from 0 (No) to 1 (Yes).
Standard errors in parentheses.
Source: SESRI Omnibus Survey, 2011.

Making sense of our results

- How do we interpret the coefficient? How is this an intercept shift?

- 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?

Adding an interaction

Slope Shifts

- Source: Hanushek and Jackson

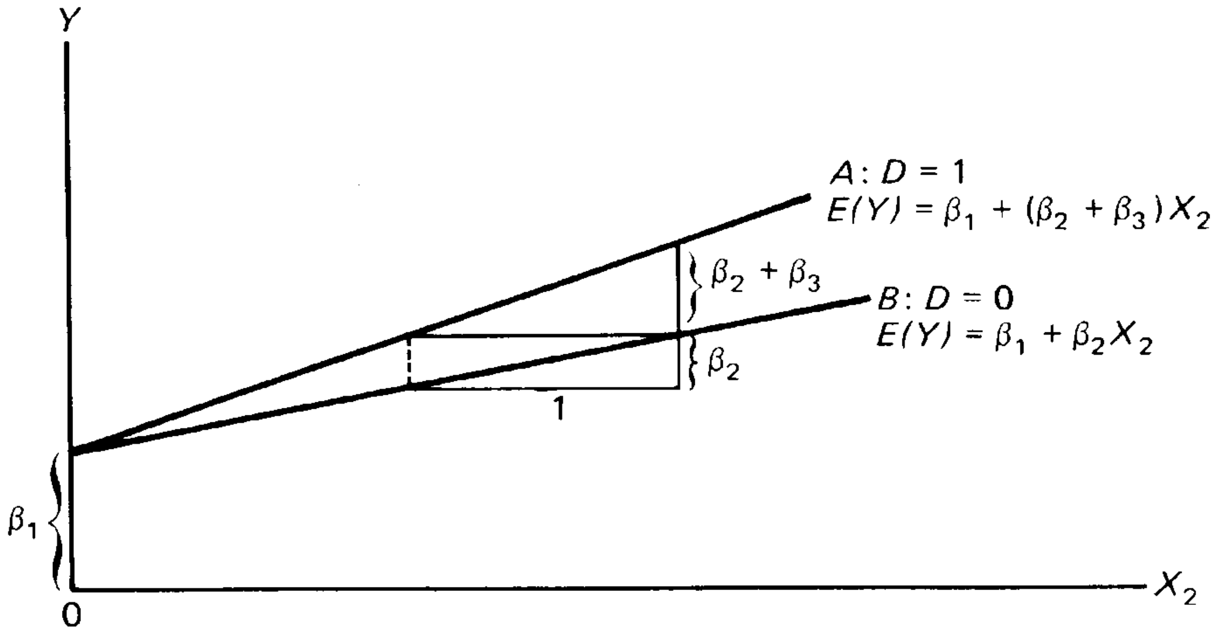
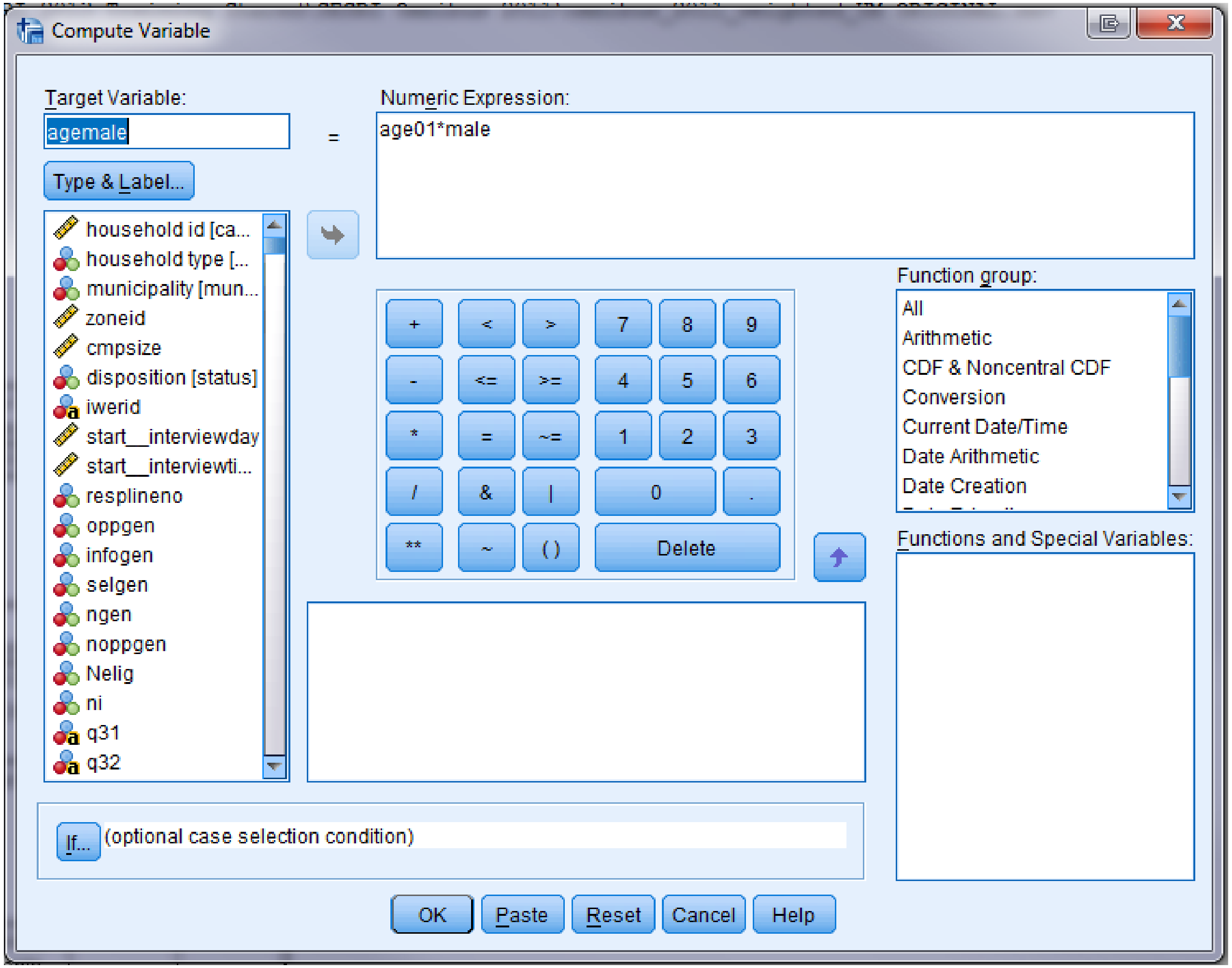


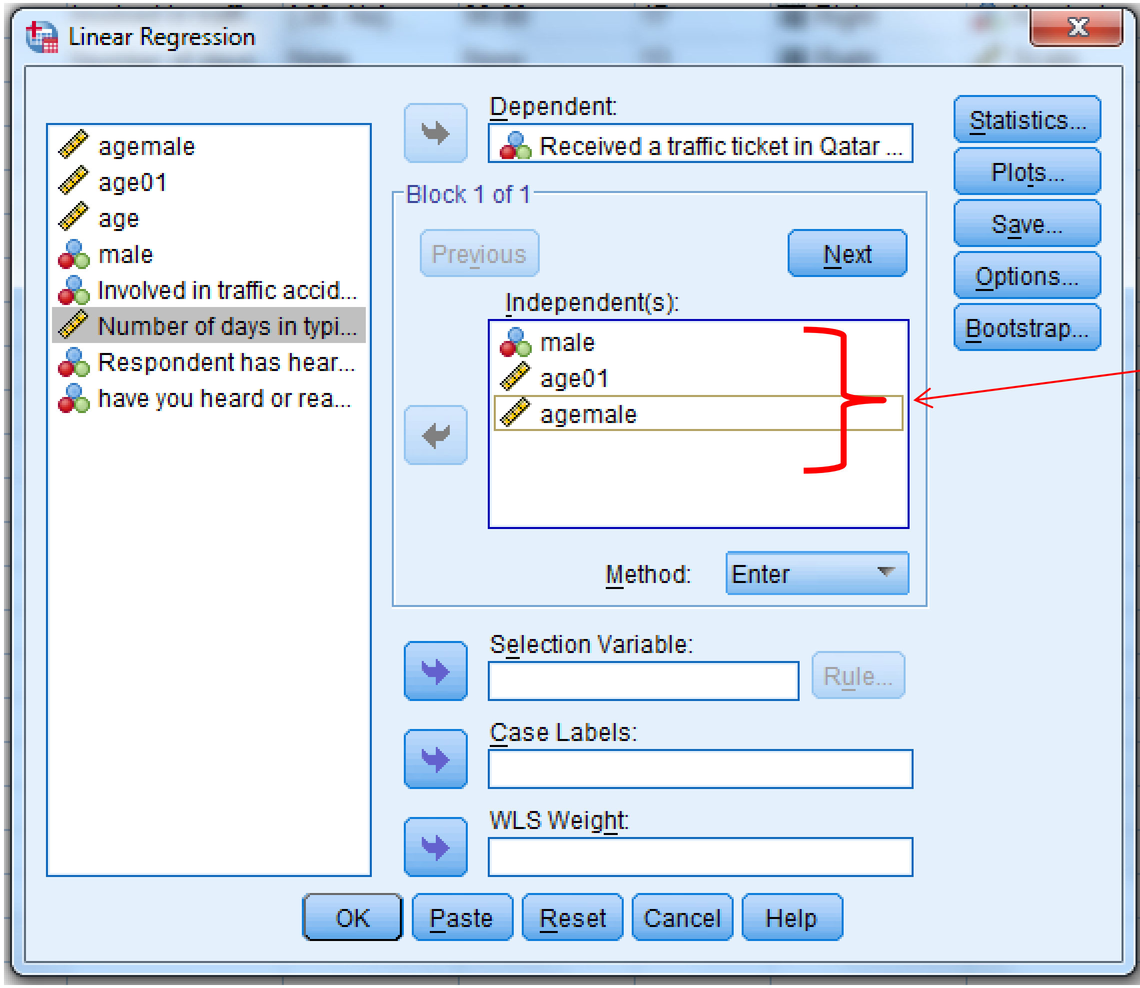
FIGURE 4.5 Bivariate relationship with slope dummy variable.

Interacting Age & Gender

We have created a variable in your dataset called agemale. To construct it, we multiplied the age variable by the male variable.

agemale = age01 * male.





Models with interaction terms should include the interaction term AND the original variables that were used to generate the interaction term.

In the above SPSS dialogue, we are specifying a regression model where receiving a traffic ticket is our dependent variable, and gender, age, and the interaction term, agemale, are our independent variables.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.172[a]	.030	.027	.49209

a. Predictors: (Constant), agemale, age01, male

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.097	3	3.032	12.522	.000[b]
	Residual	297.116	1227	.242		
	Total	306.213	1230			

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. Predictors: (Constant), agemale, age01, male

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.280	.079		3.537	.000
	male	.380	.089	.331	4.271	.000
	age01	.207	.180	.079	1.148	.251
	agemale	-.563	.197	-.295	-2.853	.004

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.280	.079		3.537	.000
	male	.380	.089	.331	4.271	.000
	age01	.207	.180	.079	1.148	.251
	agemale	-.563	.197	-.295	-2.853	.004

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

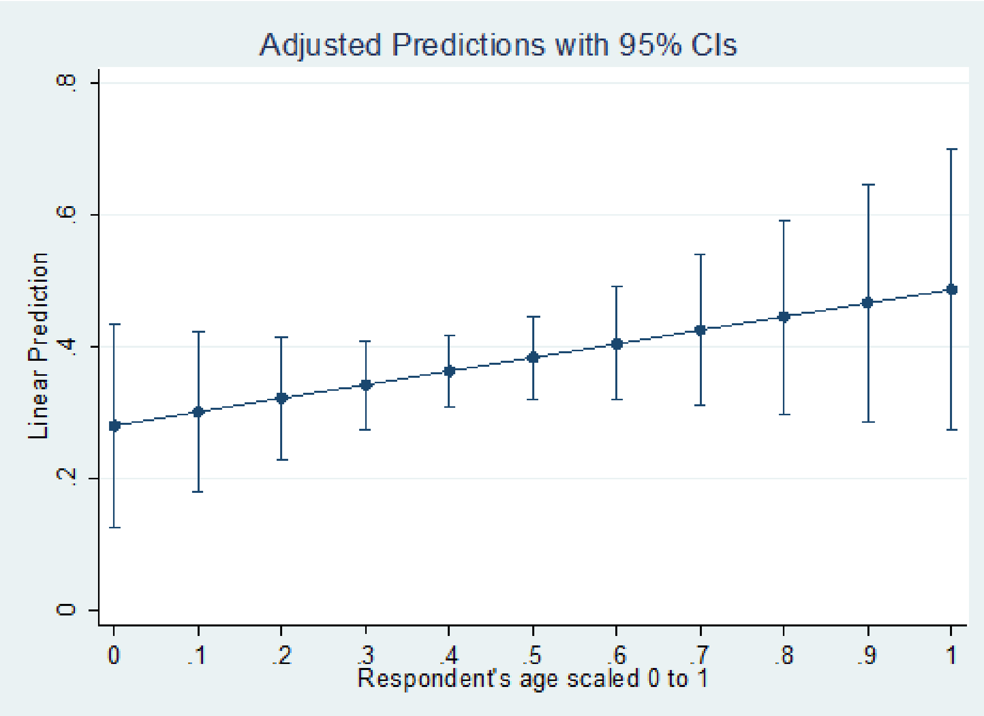
Interpreting Interaction Terms

How do we interpret the interaction terms?

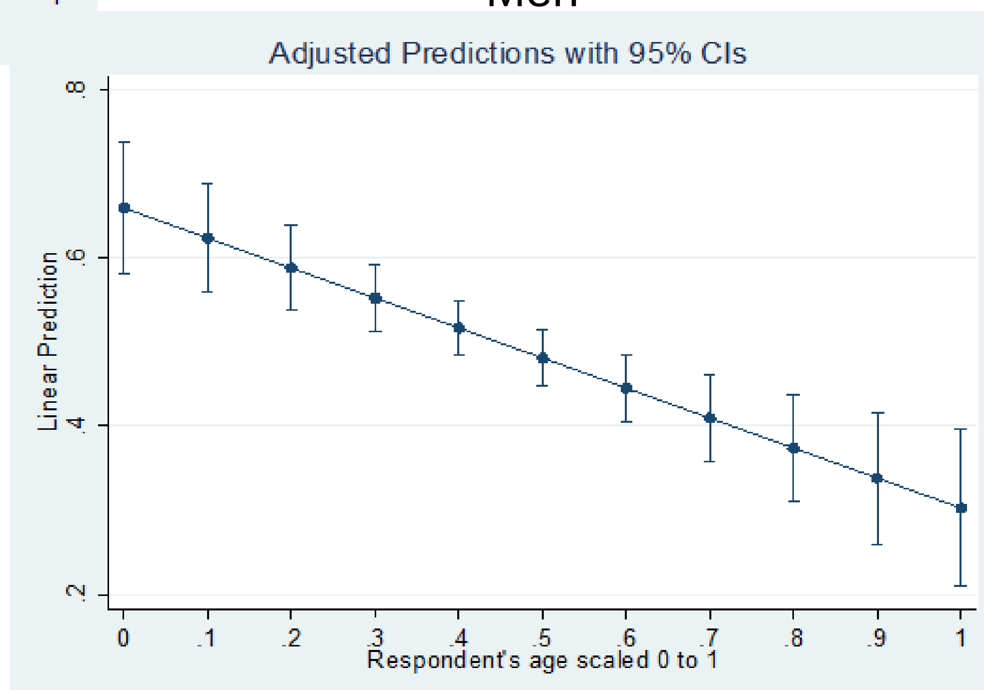
What does a slope shift mean?

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

Women


Adjusted Predictions with 95% CIs
Linear Prediction
0
.2
.4
.6
.8
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1
Respondent's age scaled 0 to 1

Men


Adjusted Predictions with 95% CIs
Linear Prediction
.2
.4
.6
.8
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1
Respondent's age scaled 0 to 1

Receiving a Traffic Ticket in Qatar in the Past 12 months as a Function of Age and Gender

	(1)	(2)	(3)	(4)
Age	-.231*		-.261*	.207
	(.036)		(.074)	(.180)
Male		.131*	.143*	.380*
		(.032)	(.032)	(.089)
Age*Male				-.563*
				(.197)
Constant	.567*	.366*	.473*	.280*
	(.036)	(.028)	(.041)	(.079)
Adjusted R-Squared	.007	.012	.022	.027
N	1230	1234	1230	1230

* p<.05

The dependent variable, receiving a traffic ticket, is coded as follows: 1=received ticket; 0=did not receive ticket.

Source: SESRI 2011 Omnibus

Class Exercise

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.090[a]	.008	.007	.49713

a. Predictors: (Constant), Number of days in typical week Respondent sends text message

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.479	1	2.479	10.033	.002[b]
	Residual	303.733	1229	.247		
	Total	306.213	1230			

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. Predictors: (Constant), Number of days in typical week Respondent sends text message

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.409	.023		18.059	.000
	Number of days in typical week Respondent sends text message	.111	.035	.090	3.167	.002

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.186[a]	.034	.031	.49112

a. Predictors: (Constant), male, Number of days in typical week Respondent sends text message, age01, agemale

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.530	4	2.633	10.914	.000[b]
	Residual	294.748	1222	.241		
	Total	305.278	1226			

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

b. Predictors: (Constant), male, Number of days in typical week Respondent sends text message, age01, agemale

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.233	.081		2.860	.004
	Number of days in typical week Respondent sends text message	.086	.035	.070	2.425	.015
	agemale	-.531	.198	-.279	-2.687	.007
	age01	.214	.180	.081	1.190	.234
	male	.367	.089	.321	4.129	.000

a. Dependent Variable: Received a traffic ticket in Qatar in past 12 months

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.