

Regression in SPSS

Workshop offered by the Mississippi Center for
Supercomputing Research and the UM Office of
Information Technology

John P. Bentley
Department of Pharmacy Administration
University of Mississippi

October 8, 2009

Outline

- Overview, background, and some terms
- Simple regression in SPSS
- Correlation analysis in SPSS
- Overview of multiple regression in SPSS
- This talk is intended as an overview and obviously leaves out a number of very important concepts and issues.

Overview, background, and some terms

Introduction to regression analysis

- **Linear regression vs. other types of regression**
 - There are many varieties of “regression analysis.” When used without qualification, “regression” usually refers to *linear* regression with estimation performed using ordinary least squares (OLS) procedures.
 - Linear regression analysis is used for evaluating the relationship between one or more IVs (classically continuous, although in practice they can be discrete) and a single, continuous DV.
 - The basics of conducting linear regression with SPSS is the focus of this talk.
- **Regression analysis is a statistical tool for evaluating the relationship of one of more independent variables X_1, X_2, \dots, X_k to a single, continuous dependent variable Y .**
 - Is a very useful tool with many applications.
 - Association versus causality
 - The finding of a statistically significant association in a particular study (no matter how well done) does not establish a causal relationship.
 - Although association is necessary for causality, it is not sufficient for causality. Association does not imply causality!
 - True cause may be another (unmeasured?) variable.
 - Much work has been published on causal inference making; this lecture is not directly about this topic.

Introduction to regression analysis

- **A model describes the relationship between variables; when we use regression analysis, we are developing statistical models.**
 - In a study relating blood pressure to age, it is unlikely that persons of the same age will have exactly the same observed blood pressure.
 - This doesn't mean that we can't conclude that, on average, blood pressure increases with age or that we can't predict the expected blood pressure for a given age with an associated amount of variability.
 - Statistical models versus deterministic models.

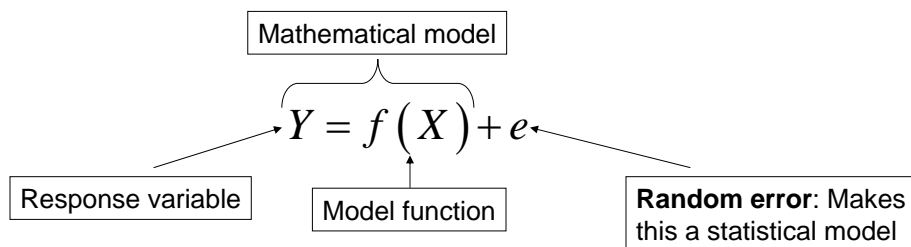
October 8, 2009

Regression in SPSS

5

Introduction to regression analysis

- **Formal representation**



- Y: dependent variable, response
- X: independent variables, explanatory variables, predictors

October 8, 2009

Regression in SPSS

6

Introduction to the GLM

- **GLM**: General Linear Model
- **General** = wide applicability of the model to problems of estimation and testing of hypotheses about parameters
- **Linear** = the regression function is a linear function of the parameters
- **Model** = provides a description between one response and at least one predictor variable (technically, we are dealing with the General Linear Univariate Linear Model (GLUM) – one response).

October 8, 2009

Regression in SPSS

7

Introduction to the GLM

- Simple linear model in scalar form:
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$
- Multiple regression in scalar form:
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$
- Multivariate: multiple y's
- Multivariable: multiple x's

October 8, 2009

Regression in SPSS

8

Simple regression in SPSS

Simple linear model

- The simplest (but by no means trivial) form of the general regression problem deals with one dependent variable Y and one independent variable X .
- Given a sample of n individuals (or other study units), we observe for each a value of X and a value of Y . We thus have n pairs of observations that can be denoted by $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, where subscripts refer to different individuals.
- These can (and should!) be plotted on a graph – the scatterplot.

TABLE 3.3 FOUR HYPOTHETICAL DATA SETS

Case No.	Data Set					
	a-c		Variable			
	X	Y	Y	Y	X	Y
1	10.0	8.04	9.14	7.46	8.0	6.58
2	8.0	6.95	8.14	6.77	8.0	5.76
3	13.0	7.58	8.74	12.74	8.0	7.71
4	9.0	8.81	8.77	7.11	8.0	8.84
5	11.0	8.33	9.26	7.81	8.0	8.47
6	14.0	9.96	8.10	8.84	8.0	7.04
7	6.0	7.24	6.13	6.08	8.0	5.25
8	4.0	4.26	3.10	5.39	19.0	12.50
9	12.0	10.84	9.13	8.15	8.0	5.56
10	7.0	4.82	7.26	6.42	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

Note: From "Graphs in Statistical Analysis," by P. J. Anascanbe, 1973, *American Statistician*, 27, pp. 17-21. Copyright 1973 by The American Statistical Association.

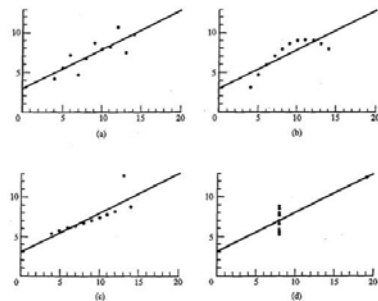


Fig. 3.3 Scatterplots for the data sets in Table 3.3. From Myers and Well (2003)

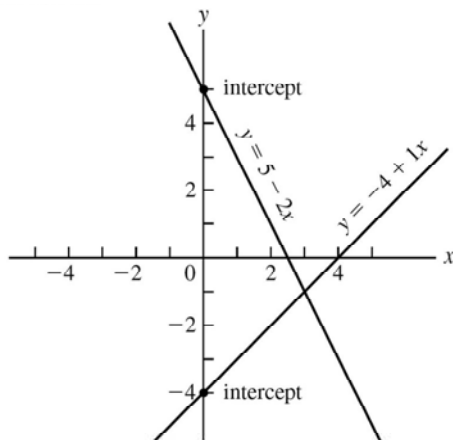
- The data points in all four panels have identical best fitting straight lines – but all tell a different story. It is always helpful to plot your data.

October 8, 2009

Regression in SPSS

11

Review of the mathematical properties of a straight Line



© 2007 Thomson Higher Education

From Kleinbaum *et al.* (2008)

- ✓ Equation: $y = \beta_0 + \beta_1 x$
- ✓ $\beta_0 =$ y-intercept (value of y when $x = 0$)
- ✓ $\beta_1 =$ slope (amount of change in y for each 1-unit increase in x – rate of change is constant given a straight line)

October 8, 2009

Regression in SPSS

12

Inference in the simple linear model

- We usually compute confidence intervals and/or test statistical hypotheses about unknown parameters.
- The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ together with estimators of their variances, can be used to form confidence intervals and test statistics based on the t distribution.

October 8, 2009

Regression in SPSS

13

Inference about the slope and intercept

- Test for zero slope – most important test of hypothesis in the simple linear model.

$$H_0 : \beta_1 = 0$$

This is a test of whether X helps to predict Y using a straight-line model.

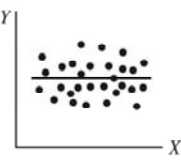
October 8, 2009

Regression in SPSS

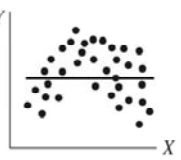
14

FTR H_0 : X provides little or no help in predicting Y .

\bar{Y} is essentially as good as $\bar{Y} + \beta_1(X - \bar{X})$ for predicting Y .



(a)



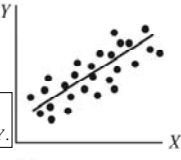
(b)

FTR H_0 : The true underlying relationship between X and Y is not linear.

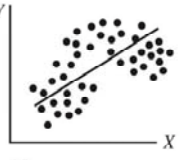
Examples when $H_0: \beta_1 = 0$ is rejected

Reject H_0 : X provides significant information for predicting Y .

The model $\bar{Y} + \beta_1(X - \bar{X})$ is far better than the naive model \bar{Y} for predicting Y .



(c)



(d)

Reject H_0 : Although there is statistical evidence of a linear component, a better model might include a curvilinear term.

© 2007 Thomson Higher Education From Kleinbaum *et al.* (2008)

Just because the null is rejected, does not mean that the straight-line model is the best model.

October 8, 2009

Regression in SPSS

15

Inference about the slope and intercept

- Test for zero intercept

$$H_0 : \beta_0 = 0$$

This is a test of whether the Y – intercept is zero.
- This test is rarely scientifically meaningful.
- However, most recommend leaving the intercept in the model even if you FTR this null.

October 8, 2009

Regression in SPSS

16

TABLE 5.1 Observations on systolic blood pressure (SBP) and age for a sample of 30 individuals

Individual (i)	SBP (Y)	Age (X)	Individual (i)	SBP (Y)	Age (X)
1	144	39	16	130	48
2	220	47	17	135	45
3	138	45	18	114	17
4	145	47	19	116	20
5	162	65	20	124	19
6	142	46	21	136	36
7	170	67	22	142	50
8	124	42	23	120	39
9	158	67	24	120	21
10	154	56	25	160	44
11	162	64	26	158	53
12	150	56	27	144	63
13	140	59	28	130	29
14	110	34	29	125	25
15	128	42	30	175	69

© 2007 Thomson Higher Education

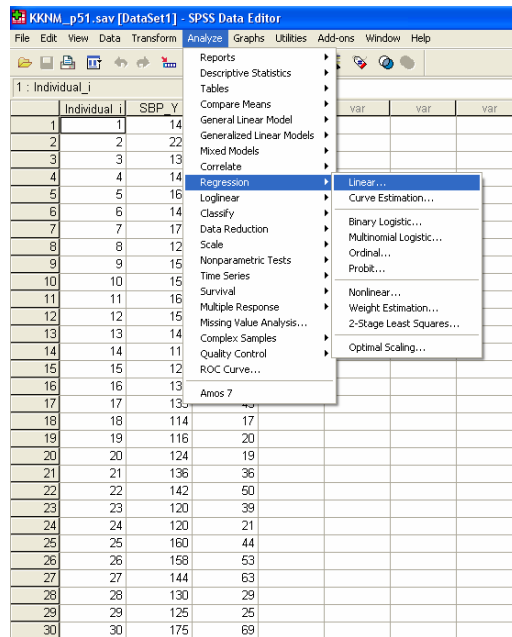
October 8, 2009

Regression in SPSS

From Kleinbaum *et al.* (2008)

17

One Way to do Simple Linear Regression in SPSS



October 8, 2009

Regression in SPSS

18

One Way to do Simple Linear Regression in SPSS

Linear Regression

Dependent: SBP_Y

Independent(s): AGE_X

Method: Enter

Linear Regression: Statistics

Regression Coefficients: Estimates, Confidence intervals, Covariance matrix

Model fit: Model fit, R squared change, Descriptives, Part and partial correlations, Collinearity diagnostics

Residuals: Durbin-Watson, Carewre diagnostics

Outliers outside: 1 (Standard deviations)

All cases

To get CIs for β_0 and β_1 .

19	19	116	20
20	20	124	19
21	21	136	36
22	22	142	50
23	23	120	39
24	24	120	21
25	25	160	44
26	26	160	53
27	27	144	63
28	28	130	29
29	29	125	25
30	30	175	69
31			

October 8, 2009

Regression in SPSS

19

SPSS Output

Model Summary^a

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.658 ^a	.432	.412	17.314

a. Predictors: (Constant), AGE_X
b. Dependent Variable: SBP_Y

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6394.023	1	6394.023	21.330	.000 ^a
	Residual	8393.444	28	299.766		
	Total	14787.467	29			

a. Predictors: (Constant), AGE_X
b. Dependent Variable: SBP_Y

t tests and p values for tests on the intercept and slope

Model	Unstandardized Coefficients	Std. Error	Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
						B	Lower Bound
1	(Constant)	98.715	10.000	9.871	.000	78.230	119.200
	AGE_X	.971	.210	4.618	.000	.540	1.401

a. Dependent Variable: SBP_Y

October 8, 2009

Regression in SPSS

20

The ANalysis Of VAriance (ANOVA) summary table

- Included in the linear regression output is a table labeled ANOVA.
- It is called an ANOVA table primarily because the basic information in the table consists of several estimates of variance.
- We can use these estimates to address the inferential questions of regression analysis.
- People usually associate the table with a statistical procedure called analysis of variance.

October 8, 2009

Regression in SPSS

21

The ANalysis Of VAriance (ANOVA) summary table

- Regression and analysis of variance are closely related.
- Analysis of variance problems can be expressed in a regression framework.
- Therefore, we can use an ANOVA table to summarize the results from analysis of variance and from regression.

October 8, 2009

Regression in SPSS

22

The ANOVA summary table for the simple linear model

- There are slightly different ways to present the ANOVA summary table; we will work with the most common form.
- In general, the mean square terms are obtained by dividing the sum of squares by its degrees of freedom.
- The F statistic is obtained by dividing the regression mean square (i.e., the model mean square) by the residual mean square.

October 8, 2009

Regression in SPSS

23

SPSS Output

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6394.023	1	6394.023	21.330	.000 ^a
	Residual	8393.444	28	299.766		
	Total	14787.467	29			

a. Predictors: (Constant), AGE_X
b. Dependent Variable: SBP_Y

Annotations:

- SSY - SSE
- df = # of estimated parameters - 1 (or just the # of predictor variables)
- Regression mean square
- F statistic (F value)
- SSE
- SSY
- df = sample size (n) - number of estimated parameters (or n - # of predictor variables - 1).
- Residual mean square
- p-value

October 8, 2009

Regression in SPSS

24

The ANOVA summary table for the simple linear model

$$r^2 = \frac{SSY - SSE}{SSY}$$

- This quantity varies between 0 and 1 and represents the proportionate reduction in SSY due to using X to predict Y instead of \bar{Y} .
- r^2 indicates % of the variation in Y explained with the help of X .

Where $SSY = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the sum of the squared deviations of the observed

Y 's from the mean \bar{Y} and $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is the sum of the squared deviations of observed Y 's from the fitted regression line.

- Because SSY represents the total variation of Y before accounting for the linear effect of X , SSY is called the *total unexplained variation* or the *total sum of squares about (or corrected for) the mean*.
- Because SSE represents the amount of variation in the observed Y 's that remain after accounting for the linear effect of X , $SSY - SSE$ is called the *sum of squares due to (or explained by) regression* (SSR).

October 8, 2009

Regression in SPSS

25

The ANOVA summary table for the simple linear model

- With a little math, it turns out that $SSY - SSE = SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$, which represents the sum of the squared deviations of the predicted values from the the mean \bar{Y} .

Total unexplained variation (SSY a.k.a. SST) = Variation due to regression (SSR)
+ Unexplained residual variation (SSE)

Or symbolically:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Fundamental equation of regression analysis

October 8, 2009

Regression in SPSS

26

The ANOVA summary table for the simple linear model

- The residual mean square (SSE/df) is the estimate $S_{Y|X}^2$ provided earlier and is given by:

$$S_{Y|X}^2 = \frac{1}{n-r} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-r} SSE, \text{ where } r = \text{number}$$

of estimated parameters (in the simple linear model case $r = 2$).

- If we assume the straight-line model is appropriate, this is an estimate of σ^2 .
- The regression mean square (SSR/df) provides an estimate of σ^2 only if X does not help to predict Y - that is only if $H_0: \beta_1 = 0$ is true.
- If $\beta_1 \neq 0$, the regression mean square will be inflated in proportion to the magnitude of β_1 and will correspondingly overestimate σ^2 .

October 8, 2009

Regression in SPSS

27

The ANOVA summary table for the simple linear model

- With a little statistical theory, we can show that the residual mean square (SSE/df) and the regression mean square (SSR/df) are statistically independent of one another.
- Thus, if $H_0: \beta_1 = 0$ is true, their ratio represents the ratio of two independent estimates of the same variance σ^2 .
- Under the normality and independence assumptions, such a ratio has the F distribution and the calculated value of F (the F statistic) can be used to test H_0 : "No significant straight-line relationship of Y on X " (i.e., $H_0: \beta_1 = 0$ or $H_0: \rho = 0$).
- This test is *equivalent* to the two-sided t test discussed earlier. Because, for ν degrees of freedom:

$$F_{1,\nu} = T_{\nu}^2 \quad \text{so} \quad F_{1,\nu,1-\alpha} = t_{\nu,1-\alpha/2}^2$$

October 8, 2009

Regression in SPSS

28

SPSS Output

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6394.023	1	6394.023	21.330	.000 ^a
	Residual	8393.444	28	299.766		
	Total	14787.467	29			

a. Predictors: (Constant), AGE_X

b. Dependent Variable: SBP_Y

$$4.618^2 = 21.33$$

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	98.715	10.000		9.871	.000	78.230	119.200
	AGE_X	.971	.210	.658	4.618	.000	.540	1.401

a. Dependent Variable: SBP_Y

NOTE: The F -test from the ANOVA summary table will test a different H_0 in *multiple regression*; as we shall see, it is a test for significant overall regression.

October 8, 2009

Regression in SPSS

29

Inferences about the regression line

- In addition to making inferences about the slope and the intercept, we may also want to perform tests and/or compute CIs concerning the regression line itself.

For a given $X = X_0$, we may want a confidence interval for $\mu_{Y|X_0}$, the mean of Y at X_0 .

We may also want to test the hypothesis $H_0: \mu_{Y|X_0} = \mu_{Y|X_0}^{(0)}$, where $\mu_{Y|X_0}^{(0)}$ is some hypothesized value of interest.

- In addition to drawing inferences about the specific points on the regression line, researchers find it useful to construct a CI for the regression line over the entire range of X -values – Confidence bands for the regression line.

October 8, 2009

Regression in SPSS

30

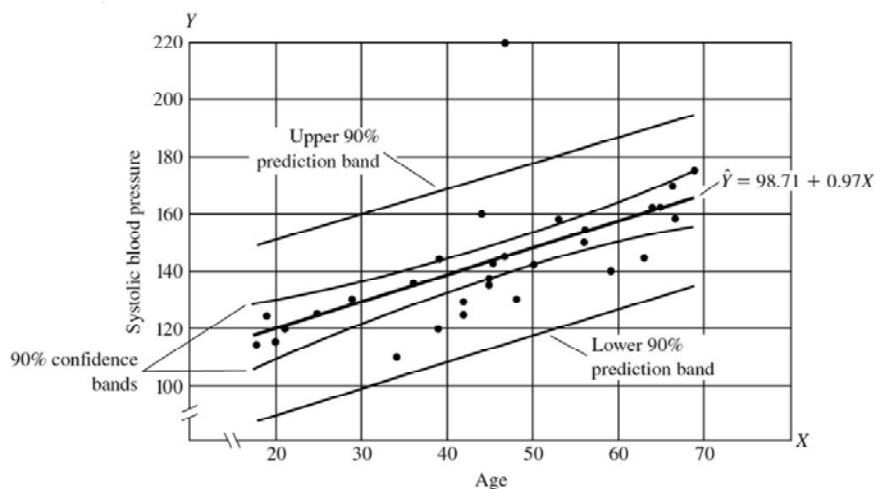
Prediction of a new value

- We have just dealt with estimating the mean $\mu_{Y|X_0}$ at $X = X_0$.
- We may also (or instead) want to estimate the response Y of a single individual; that is, predict an individual's Y given his or her $X = X_0$.
- The point estimate remains the same: $\hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$.
- However, to construct a prediction interval (technically is not a CI since Y is not a parameter), we need to include the additional variability of the Y scores around their conditional means. Thus, an estimator of an individual's response should naturally have more variability than an estimator of a group's mean response and thus prediction intervals are larger than confidence intervals given the same $1-\alpha$ coverage.
- Like confidence bands for the regression line, we can construct prediction bands – these will be wider than the confidence bands.

October 8, 2009

Regression in SPSS

31



© 2007 Thomson Higher Education

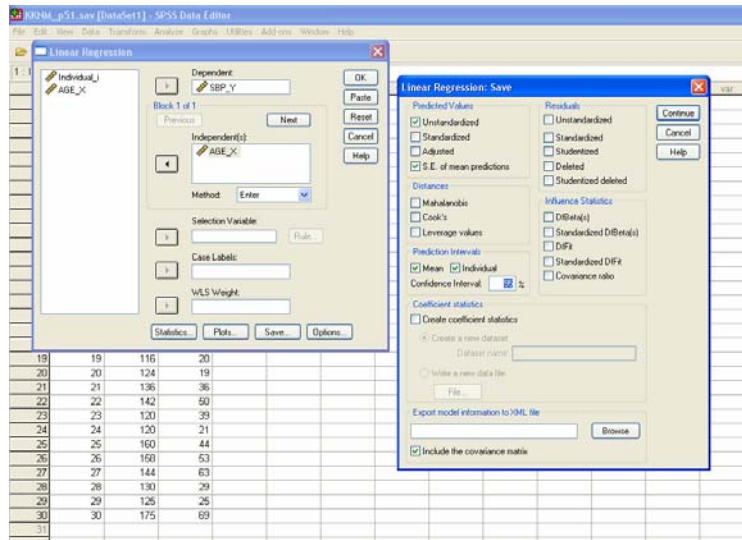
From Kleinbaum *et al.* (2008)

October 8, 2009

Regression in SPSS

32

Getting CIs and PIs in SPSS



October 8, 2009

Regression in SPSS

33

$$\hat{Y} = \bar{Y} + \hat{\beta}_1(X - \bar{X})$$

$$S_{\hat{Y}_{X_0}}$$
95% CI and PI

Individual_1	SBP_Y	AGE_X	PRE_1	SEP_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	1	144	39	136.57896	3.41388	129.58996	143.57167	100.43020
2	2	220	47	144.34562	3.18531	137.82082	150.87043	108.28481
3	3	130	45	142.40380	3.16117	135.92053	148.07924	106.35199
4	4	145	47	144.34562	3.18531	137.82082	150.87043	108.28481
5	5	162	85	161.82129	5.23771	151.09234	172.55024	124.76836
6	6	142	46	143.37475	3.16629	136.88880	149.80060	107.32097
7	7	170	67	163.76303	5.57871	152.33556	175.19050	126.50184
8	8	124	42	139.49127	3.22894	132.07709	146.10546	103.41410
9	9	168	67	163.76303	5.57871	152.33556	175.19050	126.50184
10	10	154	56	153.08346	3.90005	145.09456	161.07236	116.72921
11	11	162	64	160.05042	5.07160	150.46156	171.23920	123.08453
12	12	160	96	163.08346	3.90006	145.09456	161.07236	116.72921
13	13	140	59	155.99607	4.29993	147.10807	164.80407	119.45300
14	14	110	34	131.72431	3.93315	123.66762	139.78100	96.36510
15	15	128	42	139.49127	3.22894	132.07709	146.10546	103.41410
16	16	130	40	145.31620	3.21797	130.72470	151.90821	109.24352
17	17	135	45	142.40380	3.16117	135.92053	148.07924	106.35199
18	18	114	17	115.21951	6.70595	101.48320	128.95582	77.18670
19	19	116	20	118.13213	6.15605	105.52040	130.74305	80.49060
20	20	124	19	117.16125	6.33816	104.17813	130.14438	79.39393
21	21	136	36	133.66605	3.68044	126.09013	141.24197	97.40031
22	22	142	50	147.25824	3.32247	140.45246	154.06401	111.14563
23	23	120	39	136.57896	3.41388	129.58996	143.57167	100.43020
24	24	120	21	119.10300	5.97743	106.06079	131.34720	81.50327
25	25	160	44	141.43301	3.17001	134.93864	147.92648	105.37796
26	26	158	53	150.17085	3.56748	142.86320	157.47850	113.96020
27	27	144	63	159.07955	4.80804	149.02304	169.93525	123.01593
28	28	130	29	126.88936	4.63620	117.37314	136.36878	90.15486
29	29	125	25	122.90640	5.20251	112.16574	133.00721	85.90607
30	30	175	69	165.70477	6.92892	163.55788	177.85167	128.21669
31	.	23	121.04474	5.62499	109.52247	132.56701	83.75437	158.33511

October 8, 2009

Why this?

Regression in SPSS

34

Scatterplots, Confidence Bands, and Prediction Bands in SPSS

The screenshot shows the SPSS Data Editor window with the 'Graphs' menu open. The 'Scatterplot...' option is highlighted. The data table below shows the variables 'Individual_i', 'SBP_Y', and 'AGE'.

Individual_i	SBP_Y	AGE
1	144	39
2	220	47
3	130	45
4	145	47
5	162	65
6	142	46
7	170	67
8	124	42
9	168	67
10	154	56
11	162	64
12	160	66
13	140	59
14	110	34
15	128	42
16	130	48
17	135	45
18	114	17
19	116	20
20	124	19
21	136	36
22	142	60
23	120	39
24	120	21
25	160	44
26	158	53
27	144	63
28	130	29
29	125	25
30	175	69

October 8, 2009

Regression in SPSS

35

The screenshot shows the SPSS Data Editor window with the 'Create Scatterplot' dialog box open. The 'Assign Variables' tab is selected, and 'SBP_Y' is assigned to the Y-axis and 'AGE_X' to the X-axis. The '2-D Coordinate' option is selected. The dialog box also includes sections for 'Legend Variables' (Color, Style, Size) and 'Panel Variables'.

Individual_i	SBP_Y	AGE_X
1	144	39
2	220	47
3	130	45
4	145	47
5	162	65
6	142	46
7	170	67
8	124	42
9	168	67
10	154	56
11	162	64
12	160	66
13	140	59
14	110	34
15	128	42
16	130	48
17	135	45
18	114	17
19	116	20
20	124	19
21	136	36
22	142	60
23	120	39
24	120	21
25	160	44
26	158	53
27	144	63
28	130	29
29	125	25
30	175	69

October 8, 2009

Regression in SPSS

36

The screenshot shows the SPSS Data Editor window with a data table and the 'Create Scatter plot' dialog box. The data table has columns for 'Individual', 'SBP Y', and 'AGE X'. The dialog box is configured as follows:

- Method: Regression
- Include constant in equation
- Prediction Lines:
 - Mean
 - Individual
 - Confidence Interval: 95.0
- Fill lines for:
 - Total
 - Subgroups

October 8, 2009 Regression in SPSS 37

Correlation analysis in SPSS

The correlation coefficient (r)

- The correlation coefficient is an often-used statistic that provides a measure of how two **random** variables are **linearly** associated in a sample and has properties closely related to those of straight-line regression.
- We are generally referring to the Pearson product-moment correlation coefficient (there are other measures of correlation).

October 8, 2009

Regression in SPSS

39

The correlation coefficient (r)

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

Sum of cross-products

Sum of squares

$$r = \frac{\text{Cov}(X, Y)}{S_X S_Y}, \text{ where } \text{Cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

and $S_X = \sqrt{\frac{SSX}{n-1}}$ and $S_Y = \sqrt{\frac{SSY}{n-1}}$

$$r = \frac{S_X}{S_Y} \hat{\beta}_1$$

October 8, 2009

Regression in SPSS

40

Hypothesis tests for r

- Test of $H_0: \rho = 0$
- This is mathematically equivalent to the test of the null hypothesis $H_0: \beta_1 = 0$ described in a previous lecture.

- Test statistic:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \text{ which has the } t \text{ distribution with } df = n-2$$

when the null hypothesis is true.

- Will give the same numerical answer as:

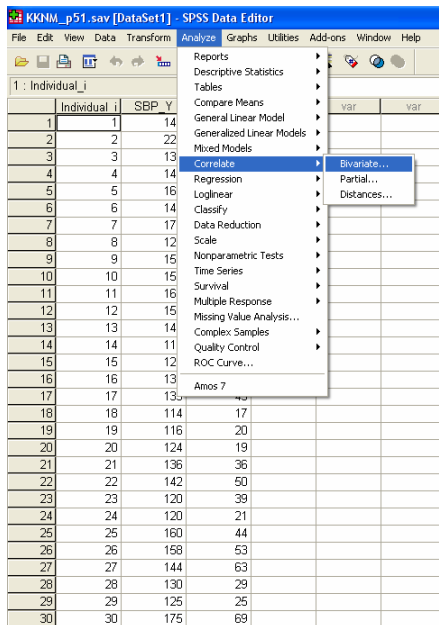
$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{S_{\hat{\beta}_1}} \text{ when } \beta_1^{(0)} = 0$$

October 8, 2009

Regression in SPSS

41

How to get a correlation coefficient in SPSS



October 8, 2009

Regression in SPSS

42

How to get a correlation coefficient in SPSS

Individual_i SBP_Y AGE_X

Individual_i	SBP_Y	AGE_X
1	144	39
2	220	47
3	138	45
4	145	47
5	162	65
6	142	46
7	170	67
8	124	42
9	158	67
10	154	56
11	162	64
12	150	56
13	140	59
14	110	34
15	128	42
16	130	48
17	135	45
18	114	17

Bivariate Correlations

Variables: Individual_i

Variables: SBP_Y, AGE_X

Correlation Coefficients: Pearson, Kendall's tau-b, Spearman

Test of Significance: Two-tailed, One-tailed

Flag significant correlations

October 8, 2009

Regression in SPSS

43

Output from correlation analysis

Correlations

		SBP_Y	AGE_X
SBP_Y	Pearson Correlation	1	.658**
	Sig. (2-tailed)		.000
	N	30	30
AGE_X	Pearson Correlation	.658**	1
	Sig. (2-tailed)	.000	
	N	30	30

** . Correlation is significant at the 0.01 level

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.6576\sqrt{30-2}}{\sqrt{1-0.6576^2}} = 4.618$$

Output from simple linear regression – see last lecture

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.715	10.000		9.871	.000
	AGE_X	.971	.210	.658	4.618	.000

a. Dependent Variable: SBP_Y

October 8, 2009

Regression in SPSS

44

Overview of multiple regression in SPSS

Multiple regression analysis

- Multiple regression analysis can be looked upon as an extension of straight-line regression analysis (the simple linear model - which involves only one independent variable) to the situation in which more than one independent variable must be considered.
- We can no longer think in terms of a line in two-dimensional space; we now have to think of a hypersurface in multidimensional space – this is more difficult to visualize, so we will rely on the multiple regression equation – the GLM.

The general linear model

- Simple linear model in scalar form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

- Multiple regression in scalar form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the regression

coefficients that need to be estimated and

X_1, X_2, \dots, X_p may all be separate basic variables

or some may be functions of a few basic variables.

The general linear model

Interpretation of β_p :

- In the simple linear model β_1 was the slope - or the amount change in Y for each 1 - unit change in X .
- In multiple regression, β_p , the regression coefficient, refers the amount of change in Y for each 1-unit change in X_p , holding all of the other variables in the equation constant.

Overview of hypothesis testing in multiple regression

- There are three basic types of tests in multiple regression:
 1. **Overall test:** Taken collectively, does the *entire set* of independent variables (or equivalently, the fitted model itself) contribute significantly to the prediction of Y ?
 2. **Tests for addition of a single variable:** Does the addition of *one* particular independent variable of interest add significantly to the prediction of Y achieved by other independent variables already present in the model?
 3. **Tests for addition of a group of variables:** Does the addition of some *group* of independent variables of interest add significantly to the prediction of Y obtained through other independent variables already present in the model?
- We will address these questions by performing statistical tests of hypotheses. The null hypotheses for the tests can be stated in terms of the unknown parameters (the regression coefficients) in the model. The form of these null hypotheses depends on the question being asked.
- There are alternative, but equivalent, ways to state these null hypotheses in terms of population correlation coefficients.

October 8, 2009

Regression in SPSS

49

Overview of hypothesis testing in multiple regression

- Questions 2 and 3 can also be subdivided into questions concerning the order of entry of the variables of interest (variables-added-in-order tests vs. variables-added-last tests).
- Occasionally, questions concerning the intercept might arise.
- It is also possible that questions more complex than those described above may arise.
 - These questions might concern linear combinations of the parameters or test null hypotheses where the set of hypothesized parameter values is something other than zero.
 - It is certainly possible to do such tests, but we will not cover these during this talk.
- Statistical tests of our questions can be expressed as F tests; that is, the test statistic will have an F distribution when the stated null hypothesis is true. In some cases, the test may be equivalently expressed as a t test.
- Let's cover questions 1 and 2 and leave question 3 for a later discussion.

October 8, 2009

Regression in SPSS

50

Overview of hypothesis testing in multiple regression

- All of these tests can also be interpreted as a comparison of two models.
- One of these models will be referred to as the *full or complete model* and the other will be called the *reduced model* (i.e., the model to which the complete model reduces under the null hypothesis).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \longleftarrow \text{Full Model}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \longleftarrow \text{Reduced Model}$$

- Under $H_0 : \beta_2 = 0$, the larger (full model) reduces to the smaller (reduced) model. Thus, a test of $H_0 : \beta_2 = 0$ is essentially equivalent to determining which of these two models is more appropriate.
- The set of IVs in the reduced model is a subset of the IVs in the full model. The smaller of the two models is said to be *nested within* the larger model.

October 8, 2009

Regression in SPSS

51

Test for significant overall regression

- Consider this model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$
- $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ or equivalently
- H_0 : All p independent variables considered together do not explain a significant amount of the variation in Y .
- The full model is reduced to a model that only contains the intercept term, β_0 .

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{(\text{SSY} - \text{SSE})/p}{\text{SSE}/(n - p - 1)}$$

$$\text{where } \text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \text{ and } \text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- The computed value of F can then be compared with the critical value $F_{p, n-p-1, 1-\alpha}$ from an F table or you can look at the p -value from your output.

October 8, 2009

Regression in SPSS

52

Test for significant overall regression

- If you reject the null, you can conclude that, based on the observed data, the set of independent variables significantly helps to predict Y .
- It means that one or more individual regression coefficients *may be* significantly different from 0 (it is possible to have a significant overall regression when none of the individual predictors are significant in the given model).
- The conclusion does not mean that *all* IVs are needed for significant prediction of Y – we need further testing.
- If you fail to reject the null, then none of the individual regression coefficients will be significantly different from zero.

October 8, 2009

Regression in SPSS

53

ANOVA table for multiple regression

Source	SS	df	MS	F
Regression (Model)	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$q - 1$	$SSR / (q - 1)$	$MSR / MSE \sim F_{(q-1, n-q)}$
Error	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - q$	$SSE / (n - q)$	
Total	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$		

$SSY - SSE = SSR$
 SSE
 SSY (or SST)

$q = \#$ estimated parameters (which is $p+1$ in the terminology we used earlier in this lecture)

Under null

October 8, 2009

Regression in SPSS

54

Tests for addition of a single variable

- A partial F test can be used to assess whether the addition of any specific independent variable, given others already in the model, significantly contributes to the prediction of Y .
- Suppose we want to assess whether adding a variable X^* significantly improves the prediction of Y , given that variables X_1, X_2, \dots, X_p are already in the model.
- H_0 : X^* does not significantly add to the prediction of Y , given that X_1, X_2, \dots, X_p are already in the model
- Or equivalently: H_0 : $\beta^* = 0$ in the model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \beta^* x^* + \varepsilon_i$
- Full model contains X_1, X_2, \dots, X_p and X^* as IVs.
- Reduced model contains X_1, X_2, \dots, X_p , but not X^* as IVs.
- How much additional information does X^* provide about Y not already provided by X_1, X_2, \dots, X_p ?

October 8, 2009

Regression in SPSS

55

The t test alternative

- An equivalent way to perform the partial F test for the variable added last is to use a t test. We will demonstrate this as SPSS provides this by default.

$$H_0: \beta^* = 0 \text{ in the model: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \beta^* x^* + \varepsilon_i$$

$$T = \frac{\hat{\beta}^*}{S_{\hat{\beta}^*}}, \text{ where } \hat{\beta}^* = \text{the corresponding estimated coefficient and } S_{\hat{\beta}^*} \text{ is}$$

the estimate of the standard error of $\hat{\beta}^*$.

October 8, 2009

Regression in SPSS

56

An example

TABLE 8.1 WGT, HGT, and AGE of a random sample of 12 nutritionally deficient children

Child	1	2	3	4	5	6	7	8	9	10	11	12
WGT (Y)	64	71	53	67	55	58	77	57	56	51	76	68
HGT (X_1)	57	59	49	62	51	50	55	48	42	42	61	57
AGE (X_2)	8	10	6	11	8	7	10	9	10	6	12	9

© 2007 Thomson Higher Education

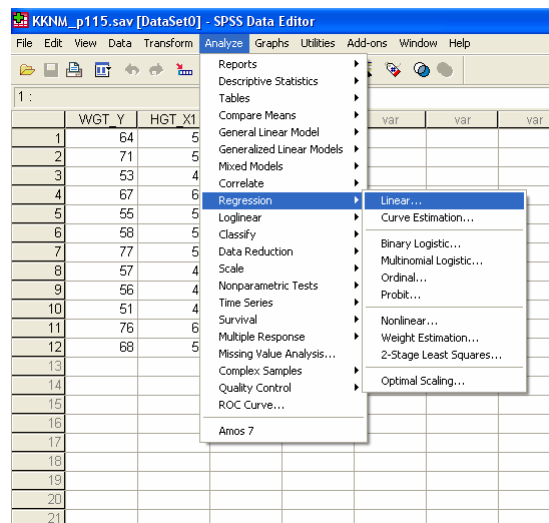
From Kleinbaum *et al.* (2008)

October 8, 2009

Regression in SPSS

57

One way to do multiple regression in SPSS



October 8, 2009

Regression in SPSS

58

SPSS Data Editor window showing a Linear Regression dialog box. The dependent variable is WGT_Y and independent variables are HGT_X1 and AGE_X2. The method is set to Enter.

	WGT_Y	HGT_X1	AGE_X2
1	64	57	8
2	71	59	10
3	53	49	6
4	67	62	11
5	55	51	8
6	58	50	7
7	77	55	10
8	57	48	9
9	56	42	10
10	51	42	6
11	76	61	12
12	68	57	9
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			

October 8, 2009

Regression in SPSS

59

SPSS output:
ANOVA Table

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	692.823	2	346.411	15.953	.001 ^a
	Residual	195.427	9	21.714		
	Total	888.250	11			

a. Predictors: (Constant), AGE_X2, HGT_X1
b. Dependent Variable: WGT_Y

Annotations:
 - **SSY - SSE**: Points to the Regression Sum of Squares (692.823).
 - **SSE**: Points to the Residual Sum of Squares (195.427).
 - **SSY**: Points to the Total Sum of Squares (888.250).
 - **ANOVA^a**: Points to the entire ANOVA table.
 - **df = # of estimated parameters - 1 (or just the # of predictor variables)**: Points to the df for Regression (2).
 - **df = sample size (n) - number of estimated parameters (or n - # of predictor variables - 1)**: Points to the df for Residual (9).
 - **Regression mean square**: Points to the Mean Square for Regression (346.411).
 - **Residual mean square**: Points to the Mean Square for Residual (21.714).
 - **F statistic (F value)**: Points to the F value (15.953).
 - **p-value**: Points to the Sig. value (.001^a).

- The critical value for $\alpha = 0.01$ is $F_{2,9,0.99} = 8.02$. Since $15.95 > 8.02$, we reject the null at the $\alpha = 0.01$ level.
- Note that this is the same as saying: since the p -value (0.001) is < 0.01 , we reject the null at the $\alpha = 0.01$ level.

October 8, 2009

Regression in SPSS

60

SPSS output: Parameter estimates and some significance tests

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	6.553	10.945		.599	.564
	HGT_X1	.722	.261	.548	2.768	.022
	AGE_X2	2.050	.937	.433	2.187	.056

a. Dependent Variable: WGT_Y

What do the results tell us?

October 8, 2009

Regression in SPSS

61

Partial *F* tests

- **One word of caution:**
 - It is possible for an independent variable to be highly correlated with a DV but have a non-significant regression coefficient in a multiple regression with other variables included in the model (i.e., a non-significant variable-added-last partial *F* test or the *t* test). Why?

October 8, 2009

Regression in SPSS

62