UP504 (Prof. Scott Campbell) Winter 2008
Regression Example using SPSS (v. 16 for Mac)
Research Question: What influences country-level fertility rates?
(unit of analysis: country)

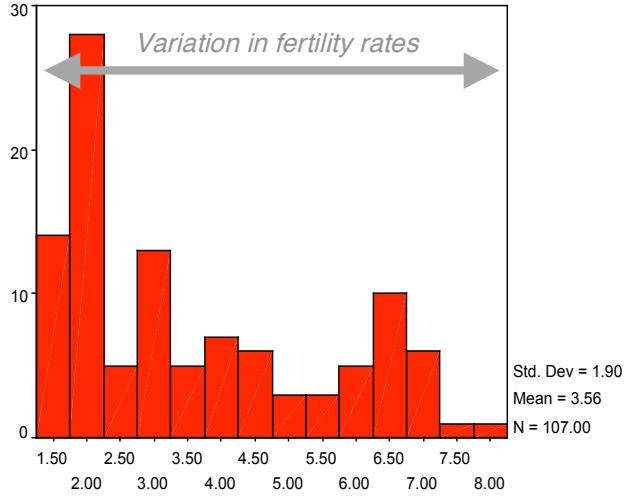
Data: World95.sav located in the SPSS directory
(n = 109 countries, but incomplete data → missing cases)

SPSS: Analyze > Descriptive Statistics > Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Number of people / sq. kilometer	109	2.3	5494.0	203.415	675.7052
People living in cities (%)	108	5	100	56.53	24.203
Average female life expectancy	109	43	82	70.16	10.572
Average male life expectancy	109	41	76	64.92	9.273
Gross domestic product / capita	109	122	23474	5859.98	6479.836
Daily calorie intake	75	1667	3825	2753.83	567.828
Fertility:average number of kids	**107**	**1.3**	**8.2**	**3.563**	**1.9025**
Males who read (%)	85	28	100	78.73	20.445
Females who read (%)	85	9	100	67.26	28.607

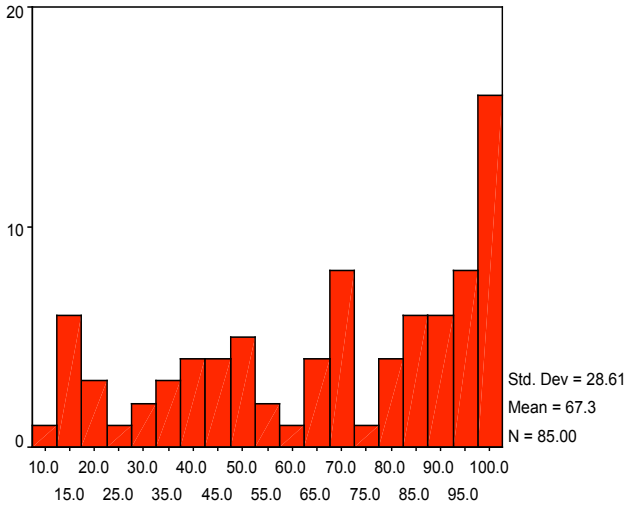
The Dependent Variable: *Fertility*. (Our goal is to explain the variation in this variable, i.e., why some countries have high fertility, and others low fertility).



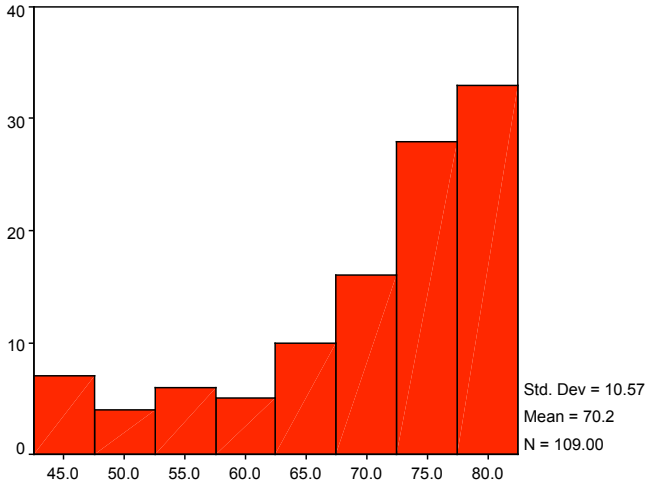
Fertility: average number of kids

Some possible independent (explanatory) variables:

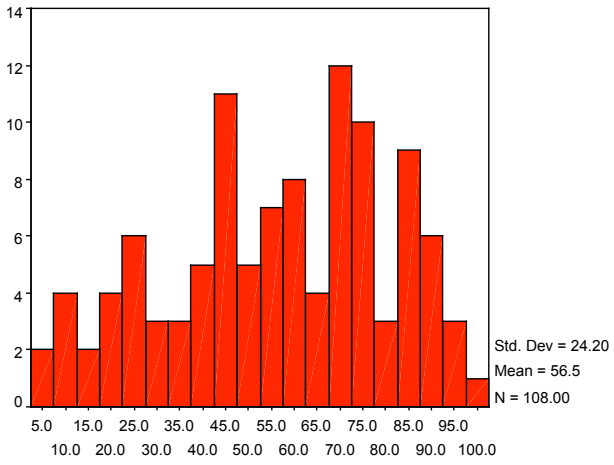
SPSS: Graphs > Histogram



Females who read (%)



Average female life expectancy

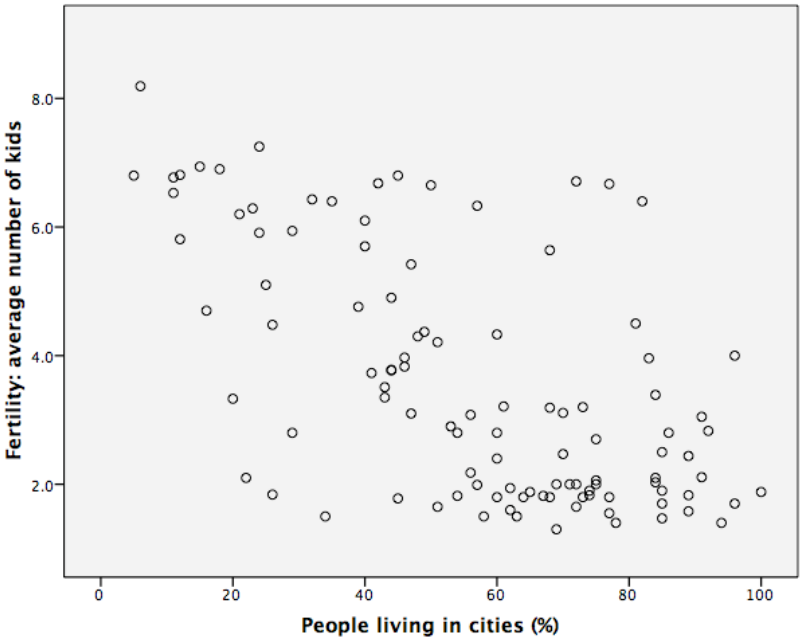


People living in cities (%)

Question: ***What influences the fertility rates (average number of kids)?***

1. One might hypothesize that fertility is related to urbanization and the "demographic transition": as the population shifts from rural to urban areas, fertility goes down. Indeed, the scatterplot shows such a relationship (below)

SPSS: Graphs > Scatter



COMMENTS:
The simple (bivariate) regression leads to the following equation:

y = a + bx
fertility rates = constant + slope (variable x)
fertility rates = 6.321 – 0.048 (percent urban)

EXAMPLE (using the U.S.)

Fertility = 6.321 – 0.048 (75)
= 6.321 – 2.4 = 3.92

Comparing the estimate (3.92) to the actual reported value (2.1):

Actual – estimated = residual
2.1 – 3.92 ≅ -1.8

Terminology:
y = dependent variable = endogenous variable
(a regression model has exactly one dep. variable)
x= independent variable = exogenous variable
(a regression model has 1 or more ind. variables)

a = constant = "y-intercept"
b = (unstandardized) coefficient = slope

Beta = standardized coefficient = b * (s_x / s_y)
(where s_x = std dev of x and s_y = std dev of y)
Useful to compare the relative explanatory power of different independent variables (especially when independent variables have different measurement scales).

SPSS: Analyze > Regression > Linear

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.619[a]	.383	.377	1.5031

a. Predictors: (Constant), People living in cities (%)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	145.703	1	145.703	64.493	.000[a]
	Residual	234.956	104	2.259		
	Total	380.659	105			

a. Predictors: (Constant), People living in cities (%)
b. Dependent Variable: Fertility: average number of kids

The resulting model does show a significant, negative relationship between the level of urbanization and fertility rates. But urbanization only explains about 38% of the variation in fertility (see the R-Square).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.321	.371		17.023	.000
	People living in cities (%)	-.048	.006	-.619	-8.031	.000

a. Dependent Variable: Fertility: average number of kids

The goal of Ordinary Least Squares (OLS): find the best fitting equation that links the independent variables with the dependent variable (i.e., to minimize the error of prediction). How is this error minimized? A simple approach is to minimize the **sum of squares** (i.e., "least squares") of the vertical distances between the estimate line (estimate) and the actual value of y. (This is SSE - the sum of the square of errors). There are other methods (and advantages of each): weighted least squares (WLS), 2-Stage Least Squares (2SLS), etc. So: OLS is a method that estimates an equation for the regression line by minimizing the sum of the square of differences between the actual value of each case and its predicted value.

NEXT STEP: MULTIPLE REGRESSION

Multiple regression allows us to control for other variables as well, and therefore see, simultaneously, the relative influence of several variables on fertility (the dependent variable). It also may increase the amount of variation in the dependent variable (fertility) that is explained by the model (e.g., R-Square).

Explained variation: "regression" [we want to maximize this…]

Unexplained variation: "residual" […which will in turn lower this]

Strategy One: the shotgun approach, or throw everything into the regression model and see what happens (NOT recommended). We just do it to show the mess it makes…

Why? The resulting R-Square is high (.969), and the F-score (38.348) is significant, but the independent variables are NOT statistically significant (some because they are not highly correlated with fertility, and others because of multicollinearity). As a result, you CANNOT judge the quality of a regression model simply based on the level of the R-Square. (Instead, a model as a whole should always be statistically significant (see the F Score), and each individual independent variable should be statistically significant (see the t-scores).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.894[a]	.799	.741	.9405

a. Predictors: (Constant), Tropical, cropgrow, catholic, Africa, OECD, People living in cities (%), Males who read (%), Daily calorie intake, Average male life expectancy, Gross domestic product / capita, People who read (%), Females who read (%), Average female life expectancy

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	158.634	13	12.203	13.796	.000[a]
	Residual	39.803	45	.885		
	Total	198.437	58			

a. Predictors: (Constant), Tropical, cropgrow, catholic, Africa, OECD, People living in cities (%), Males who read (%), Daily calorie intake, Average male life expectancy, Gross domestic product / capita, People who read (%), Females who read (%), Average female life expectancy

b. Dependent Variable: Fertility: average number of kids

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	9.916	2.072		4.787	.000
	People living in cities (%)	.006	.010	.086	.640	.526
	Average female life expectancy	-.219	.104	-1.312	-2.115	.040
	Average male life expectancy	.184	.102	.986	1.808	.077
	People who read (%)	-.025	.025	-.305	-1.029	.309
	Gross domestic product / capita	1.935E-5	.000	.050	.335	.739
	Daily calorie intake	.000	.000	-.046	-.336	.739
	cropgrow	-.009	.010	-.070	-.863	.393
	Males who read (%)	.004	.030	.042	.132	.896
	Females who read (%)	-.012	.031	-.179	-.398	.692
	catholic	.285	.361	.077	.789	.434
	OECD	-1.184	.877	-.180	-1.351	.184
	Africa	.751	.463	.182	1.621	.112
	Tropical	-.592	.308	-.161	-1.925	.061

a. Dependent Variable: Fertility: average number of kids

Formula:
t = b / std. error

degrees of freedom:
for regression (explained): k
for residual (unexplained): n - k -1
total: n – 1
(where k = number of independent variables and n = number of cases)

$$R^2 = \frac{\text{RegressionSumSquares}}{\text{TotalSumSquares}}$$

$$F = \frac{\frac{\text{RegressionSumSquares}}{k}}{\frac{\text{SumSquaresError}}{n-k-1}}$$

Strategy Two:

So, let's take all those insignificant variables out of the equation and start over. Before you just throw in a bunch of variables into a regression analysis, sit down and build a hypothetical model that makes theoretical sense. For example, one might expect that fertility is related to urbanization, life expectancy, literacy (as a proxy of education levels), wealth, diet, and religion. (There are certainly other factors, but many are not in this data set.)

Then, you might look at a correlation matrix of selected variables. Focus on how each is correlated with fertility.

SPSS: Analyze > Correlate > Bivariate

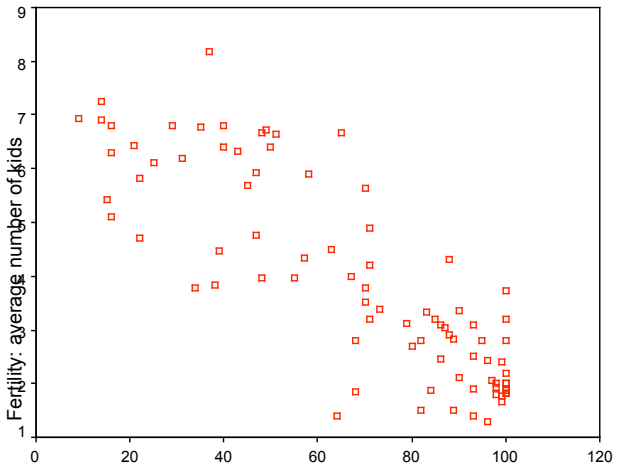
Correlations

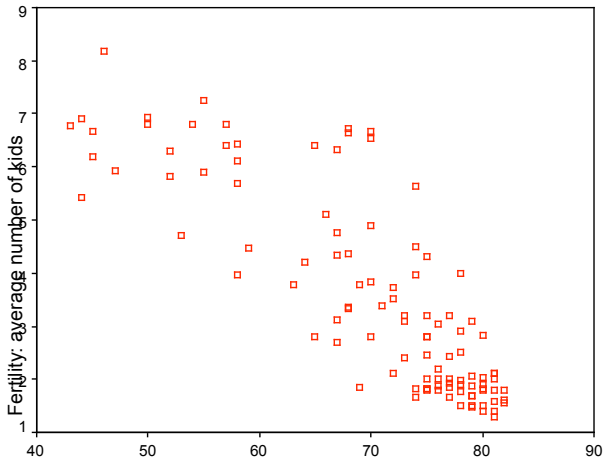
		Number of people / sq. kilometer	People living in cities (%)	Average female life expectancy	Average male life expectancy	Gross domestic product / capita	Daily calorie intake	Fertility: average number of kids	Males who read (%)	Females who read (%)
Number of people / sq. kilometer	Pearson Correlation	1.000	.223*	.128	.151	.201*	.067	-.162	.085	.029
	Sig. (2-tailed)		.020	.186	.117	.036	.570	.095	.437	.795
	N	109.000	108	109	109	109	75	107	85	85
People living in cities (%)	Pearson Correlation	.223*	1.000	.743**	.730**	.605**	.692**	-.619**	.587**	.612**
	Sig. (2-tailed)	.020		.000	.000	.000	.000	.000	.000	.000
	N	108	108.000	108	108	108	74	106	85	85
Average female life expectancy	Pearson Correlation	.128	.743**	1.000	.982**	.642**	.775**	-.838**	.777**	.819**
	Sig. (2-tailed)	.186	.000		.000	.000	.000	.000	.000	.000
	N	109	108	109.000	109	109	75	107	85	85
Average male life expectancy	Pearson Correlation	.151	.730**	.982**	1.000	.639**	.765**	-.783**	.717**	.745**
	Sig. (2-tailed)	.117	.000	.000		.000	.000	.000	.000	.000
	N	109	108	109	109.000	109	75	107	85	85
Gross domestic product / capita	Pearson Correlation	.201*	.605**	.642**	.639**	1.000	.751**	-.583**	.417**	.429**
	Sig. (2-tailed)	.036	.000	.000	.000		.000	.000	.000	.000
	N	109	108	109	109	109.000	75	107	85	85
Daily calorie intake	Pearson Correlation	.067	.692**	.775**	.765**	.751**	1.000	-.696**	.576**	.548**
	Sig. (2-tailed)	.570	.000	.000	.000	.000		.000	.000	.000
	N	75	74	75	75	75	75.000	75	59	59
Fertility: average number of kids	Pearson Correlation	-.162	-.619**	-.838**	-.783**	-.583**	-.696**	1.000	-.796**	-.839**
	Sig. (2-tailed)	.095	.000	.000	.000	.000	.000		.000	.000
	N	107	106	107	107	107	75	107.000	85	85
Males who read (%)	Pearson Correlation	.085	.587**	.777**	.717**	.417**	.576**	-.796**	1.000	.964**
	Sig. (2-tailed)	.437	.000	.000	.000	.000	.000	.000		.000
	N	85	85	85	85	85	59	85	85.000	85
Females who read (%)	Pearson Correlation	.029	.612**	.819**	.745**	.429**	.548**	-.839**	.964**	1.000
	Sig. (2-tailed)	.795	.000	.000	.000	.000	.000	.000	.000	
	N	85	85	85	85	85	59	85	85	85.000

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

The strongest correlations with fertility seem to be female literacy rates and female life expectancies. We can run scatterplots to see this visually (and also check to see if any relationships are nonlinear, thus requiring nonlinear transformations, e.g., taking the log of the variable).





SPSS: Analyze > Regression > Linear (using the optional method "forward" to enter the variables one at a time. The default is "Enter", which enters all simultaneously. The benefit of "forward" here is simply to see the incremental (intermediate) change in the model from one to two independent variables.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Females who read (%)	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	Average female life expectancy	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: Fertility: average number of kids

Our R-Square is now up to .755 (that is, over 75% of the variation of fertility explained by the two variables). Note that only 85 cases (out of the 109 countries in the data set) used in the model (since there are missing values). The F-score is significant. Importantly, the t scores for each independent variable significant (at the .000 level). Be sure to be able to write the regression equation from the slope and intercept values (see the coefficients table) and therefore know how to predict fertility based on the independent variables.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.839[a]	.704	.700	1.0316
2	.869[b]	.755	.749	.9446

a. Predictors: (Constant), Females who read (%)

b. Predictors: (Constant), Females who read (%), Average female life expectancy

Note on statistical significance:
You want a high "F" score (stat. significance of the entire model) and high (absolute values of) t scores (stat. significance of each independent variable).

Remember: a **high** F → low p-value → high stat. significance
A **low** F → high p-value → low stat. significance.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	209.925	1	209.925	197.271	.000[a]
	Residual	88.324	83	1.064		
	Total	298.249	84			
2	Regression	225.081	2	112.541	126.126	.000[b]
	Residual	73.167	82	.892		
	Total	298.249	84			

a. Predictors: (Constant), Females who read (%)

b. Predictors: (Constant), Females who read (%), Average female life expectancy

c. Dependent Variable: Fertility: average number of kids

These are the p-values, i.e., the probability of getting this pattern in the sample data were there no relationship in the population as a whole. If less than 0.05 (5%), then we assume that the pattern is too unlikely to have occurred due simply to

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.654	.287		26.641	.000
	Females who read (%)	-.055	.004	-.839	-14.045	.000
2	(Constant)	10.908	.832		13.109	.000
	Females who read (%)	-.034	.006	-.518	-5.434	.000
	Average female life expectancy	-.069	.017	-.393	-4.121	.000

a. Dependent Variable: Fertility: average number of kids

Standardized Residuals (click on "statistics" within the regression dialogue window)

You can either list ALL cases (useful but LONG), or just the big outliers (here I defined by the residual – i.e., the gap between the predicted value and the actual value – being at least 2+ std. deviations). I also added "case labels" (country name) so that I knew which countries were outliers. Examining outliers can be useful to build a better model: look for systematic patterns. Here I see that the model overestimated fertility in China, and the Central African Republic, and underestimated fertility in several mid-eastern countries (Iraq, Jordan, Saudi Arabia, Syria).

Casewise Diagnostics[a]

Case Number	country	Std. Residual	Fertility: average number of kids	Predicted Value	Residual
22	Cent. Afri.R	-2.055	5.4	7.361	-1.9414
24	China	-2.107	1.8	3.830	-1.9902
53	Iraq	2.290	6.7	4.547	2.1631
58	Jordan	2.353	5.6	3.417	2.2229
87	Saudi Arabia	2.358	6.7	4.443	2.2270
95	Syria	2.299	6.6	4.479	2.1713

a. Dependent Variable: Fertility: average number of kids

NOTE on "Stepwise regression" (one method of entering variables into the model)

Stepwise enters variables, one step at a time, into the model as long as their partial correlation is significant at the .05 level. It can be useful as an initial exploratory tool, it quickly allows one to see which variables contain unique explanatory power, and how partial correlations change when other variables are entered into the model. However, there are numerous potential downsides (including problematic R-square and statistical significance). Many scholars discourage the use of stepwise, and emphasize the importance of building models based on reasoned theories rather than computerized data hunting for patterns (i.e., stepwise). Advice for the assignment: stick to the "enter" method.

A few final thoughts on regression:

See also: http://www-personal.umich.edu/~sdcamp/up504/module+regression.html

Why might an R-Square be less than 1.00?

underdetermined model (need more variables)
nonlinear relationships
measurement error
sampling error
not fully predictable/explainable even with all data available; there is a certain amount of unexplainable chaos/static/randomness in the universe.
the unit of analysis is too aggregated (e.g., if you are predicting mean housing values for a city -- you might get better results with predicting individual housing prices, or neighborhood housing prices).

Is an R-Square < 1.00 Good or bad?

This is both a statistical and a philosophical question. It is quite rare, especially in the social sciences, to get an R-square that is really high (e.g., 98%). The goal is NOT to get the highest R-square per se. Instead, the goal is to develop a model that is both statistically and theoretically sound, creating the best fit with existing data.

What is needed to run a regression

One interval scale dependent variable
At least one interval scale independent variable (plus perhaps some dummy – dichotomous – variables coded 0,1)
enough cases to be statistically significant
an understanding of the requirements of regression so that you don't violate some basic statistical rules.

Regression assumptions include:

linear relationship
error terms have a constant variance
no or only a few outliers
error terms normally distributed
error terms independent
little multicollinearity (independent variables that are highly correlated)