**The Problem**

Bumble is concerned about infant mortality in countries around the world. He located a data set called WORLD95.SAV, provided by SPSS, that includes information on a wide variety of variables for 109 countries, collected around 1995. Two important variables for Bumble are **babymort** (deaths per 1000 live births) and gross domestic product per capita (**gdp_cap**) in $US. Bumble decided to use regression to test a model of how **babymort** can be predicted from **gdp_cap**. You can find this file on Sakai.

In SPSS, click Analyze, Regression, Linear…. Select Infant mortality as the Dependent variable and Gross domestic product / capita as the Independent. Click the Statistics tab, select Estimates, Model fit, R squared change, and Descriptives, Continue. Click OK.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .640 | .410 | .405 | 29.3834 | .410 | 74.383 | 1 | 107 | .000 |

a. Predictors: (Constant), gdp_cap Gross domestic product / capita

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 64.365 | 3.802 | | 16.927 | .000 |
| | gdp_cap Gross domestic product / capita | -.004 | .000 | -.640 | -8.625 | .000 |

a. Dependent Variable: babymort Infant mortality (deaths per 1000 live births)

As expected, wealthier countries have lower infant mortality rates. The regression equation is

Predicted Infant Mortality = 64.365 - .004 (gdp per capita in $US).

The R Square (proportion of variance in **babymort** explained by the model) is .410 and the test of statistical significance of the regression model is strongly significant, $F(1, 107) = 74.38$, $p<.001$. The correlation is -.640, and the regression weight B for **gdp_cap** yields $t(107) = 8.625$, $p < .001$. Note: t squared (df = n-1) equals F (df = 1, n-1).

Bumble concluded that there is a strong linear relationship between GDP per capita and the rate of infant mortality, and the equation is an excellent model for this relationship.

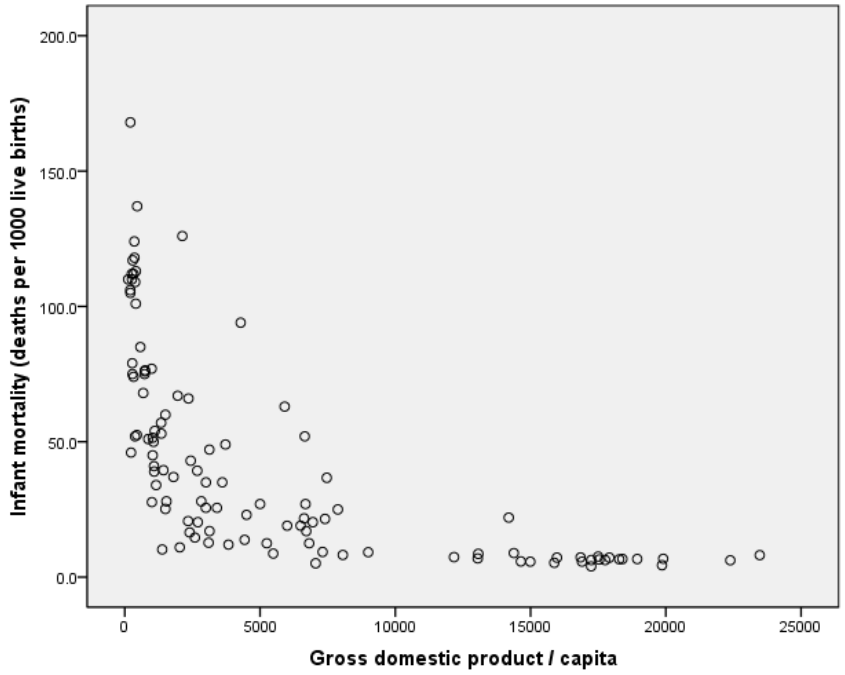
What is your advice for Bumble?

**Look at the data!**

Let's take a look at Bumble's model. We can ask for a simple scatterplot: click Graphs, Legacy Dialogs, Scatter/Dot…, click Simple Scatter, Define. Select **babymort** for the Y Axis, and **gdp_cap** for the X axis. If you Paste the syntax you will find

```
GRAPH
  /SCATTERPLOT(BIVAR)=gdp_cap WITH babymort
  /MISSING=LISTWISE.
```

Alternatively, you could just type this syntax into the syntax window and run it.

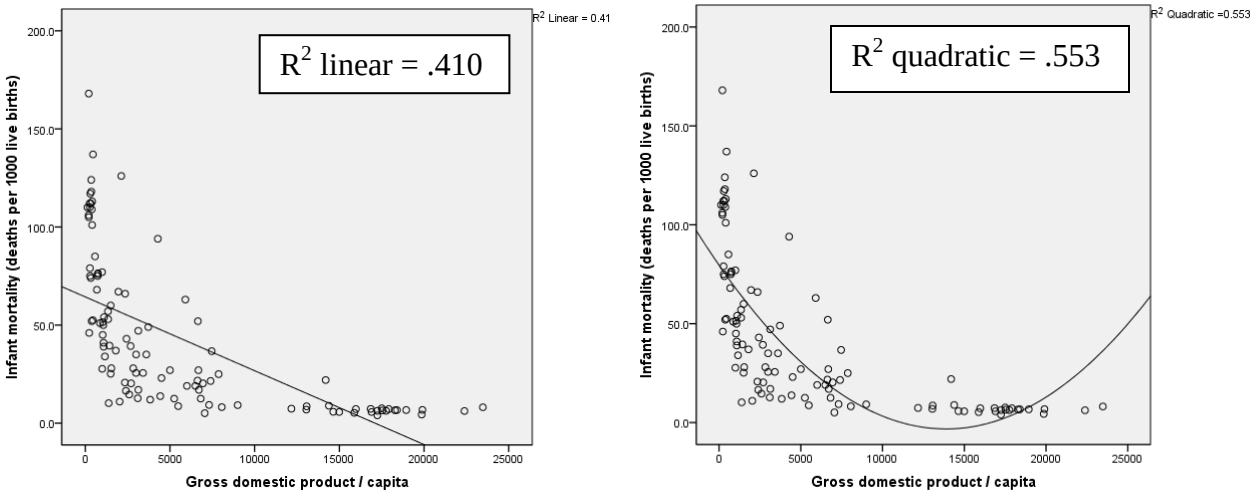



Do you think a linear model is appropriate for this bivariate relationship? The scatterplot shows quite clearly that the relationship is not linear and that the errors from a straight line will not be distributed normally with equal variance for any value of X (gpd_cap).

**Exploring a scatterplot with linear, quadratic, and loess (lowess)**

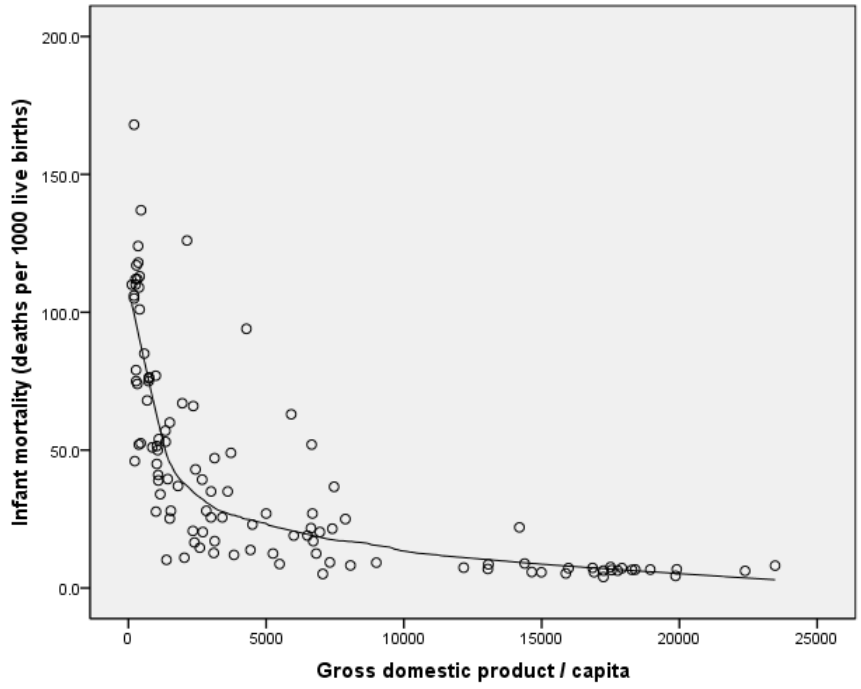
We can edit the scatterplot to reveal interesting features. If you double-click on the graph with the left mouse button, a new Chart Editor window will open. Click on Elements, Fit Line at Total to open a Properties window. Here we can select several common models. First, click Linear to check the fit of Bumble's model. This plot shows a strong departure from linearity, so the linear model fit by regression is clearly inappropriate although $R^2$ = .41. It grossly under predicts infant mortality for the very poorest countries and for the richest countries the model predicts negative infant mortality! A linear model is silly.

Bumble says, "Well, if the relationship isn't linear, then it must be quadratic." Let's see. Click Quadratic, Apply. Although a quadratic model fits better ($R^2$ = .55), it obviously is inappropriate, too. The quadratic model predicts that lowest infant mortality is seen for countries with gpd_cap of about $14,000. But this model predicts that infant mortality is greater for countries with gdp_cap greater than $20,000. That is silly, too, and it doesn't reflect the actual pattern in the data.




The actual relationship can be seen more clearly with a loess plot. Loess (or lowess) stands for Locally Weighted Scatterplot Smoother. Essentially, loess fits a regression line based on only limited window of X values that slides across the X axis. The size of the window can be varied (% of points to fit) and different weighting algorithms can be selected (the default is Epanechnikov).

Click Loess, Apply. We see a plot that looks roughly like the letter L. The data might be described as showing a weak negative relationship for countries with GDPpC greater than about $3000 and a very strong negative relationship for poorer countries. There are several clear outliers from this model, showing substantially greater infant mortality rates than other countries with similar GDPpC.




A loess plot is merely descriptive – it does not provide us with tests of statistical significance.

But 'merely' understates the value of this descriptive tool – loess is an excellent diagnostic tool to assess whether a modeled relationship is really linear.

We can't know that a statistically significant linear model is really silly unless we look!!

**A better approach to regression analysis**

The first step in any data analysis, including regression, should be to examine the data carefully. What do we look for? We should verify that the model provides an accurate description of the actual data and that we have satisfied the assumptions for any test of statistical significance that we apply.

A good place to begin is with univariate histograms. In SPSS, click Analyze, Descriptive Statistics, Frequencies…, select infant mortality and gross domestic product as the Variables. Click Statistics, and select Mean, Median, Minimum, Maximum, Skewness and Kurtosis, Continue. Click Charts, select Histograms, select show normal curve on histogram, Continue. Click Format…, Ascending values, Compare variables, and select Suppress tables with many categories, Maximum = 10 is fine, Continue. Click Paste to save the syntax in the syntax window, or click OK to run.

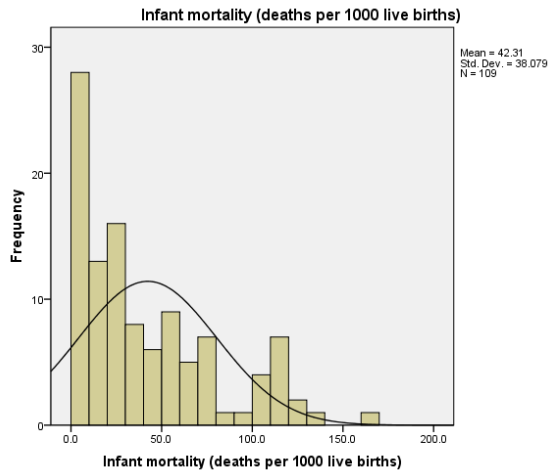
**Statistics**

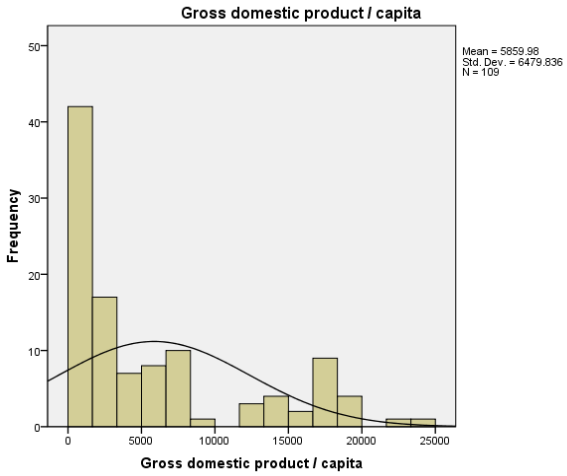
| | | babymort Infant mortality (deaths per 1000 live births) | gdp_cap Gross domestic product / capita |
|---|---|---|---|
| N | Valid | 109 | 109 |
| | Missing | 0 | 0 |
| Mean | | 42.313 | 5859.98 |
| Median | | 27.700 | 2995.00 |
| Skewness | | 1.090 | 1.146 |
| Std. Error of Skewness | | .231 | .231 |
| Kurtosis | | .365 | -.028 |
| Std. Error of Kurtosis | | .459 | .459 |
| Minimum | | 4.0 | 122 |
| Maximum | | 168.0 | 23474 |

We note that none of the 109 countries is missing data on these two variables. The means for both variables exceed the medians substantially, suggesting positive skew. The skew and kurtosis statistics help diagnose departures from a normal distribution. Both skew and kurtosis have expected values of zero for normal distributions. Here the skew statistics exceed 1.000 for both variables, suggesting possible outliers on the postive tails (the right tails). A large positive kurtosis is an indicator of outliers in either tail. Here the kurtosis looks fine. Negative kurtosis would indicate a distribution with tails shorter than a normal distribution, often seen with Likert scales. However, these statistics are inadequate to determine the best way to deal with the data. We need to examine the histograms, too.

Skew divided by its standard error is distribued approximately as Z for normal distributions. Here $Z = 1.090 / .231 = 4.72$, $p < .001$ for infant mortality, confirming the departure from normality for this variable. However, tests of statistical significance for skew and kurtosis generally are not very useful, because they depend so much on sample size. With very large samples, inconsequential departures from normal may be statistically significant, while with small samples we may miss important departures. Minimum and Maximum are useful to identify the extreme cases and spot outliers or errors.

Next we examine the histograms for the two variables of interest.

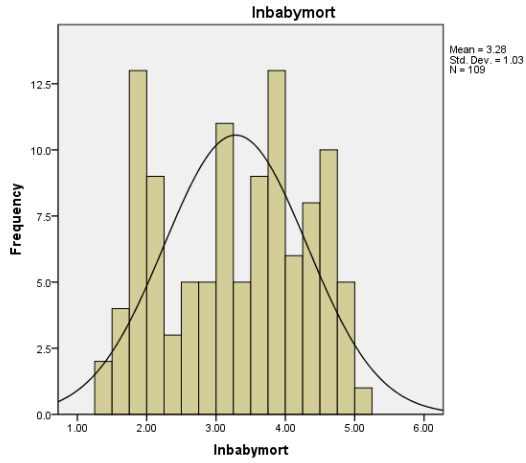


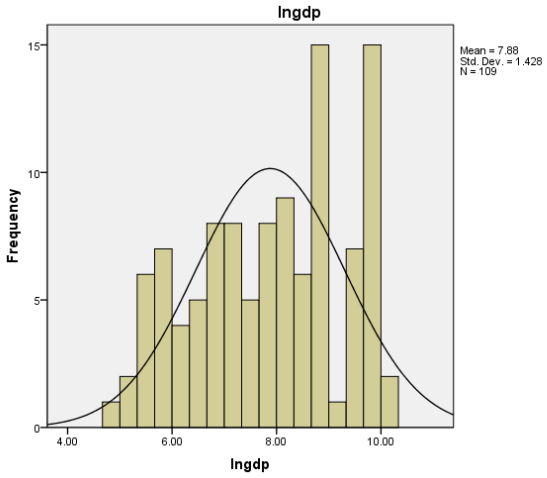




The histograms show the shapes of the distributions, which provide guidance for dealing with departures from normality. Distributions with long smooth tails on the positive end are good candidates for a log transformation, which may produce a distribution much closer to normal. Here we see a large number of cases at the very bottom of the distribution. If there are many cases tied at exactly zero, a log transformation won't solve the problem because all of those cases will remain tied at the very bottom of the transformed distribution. Let's explore.

We can use syntax to create a new variables as the log of original skewed variables:

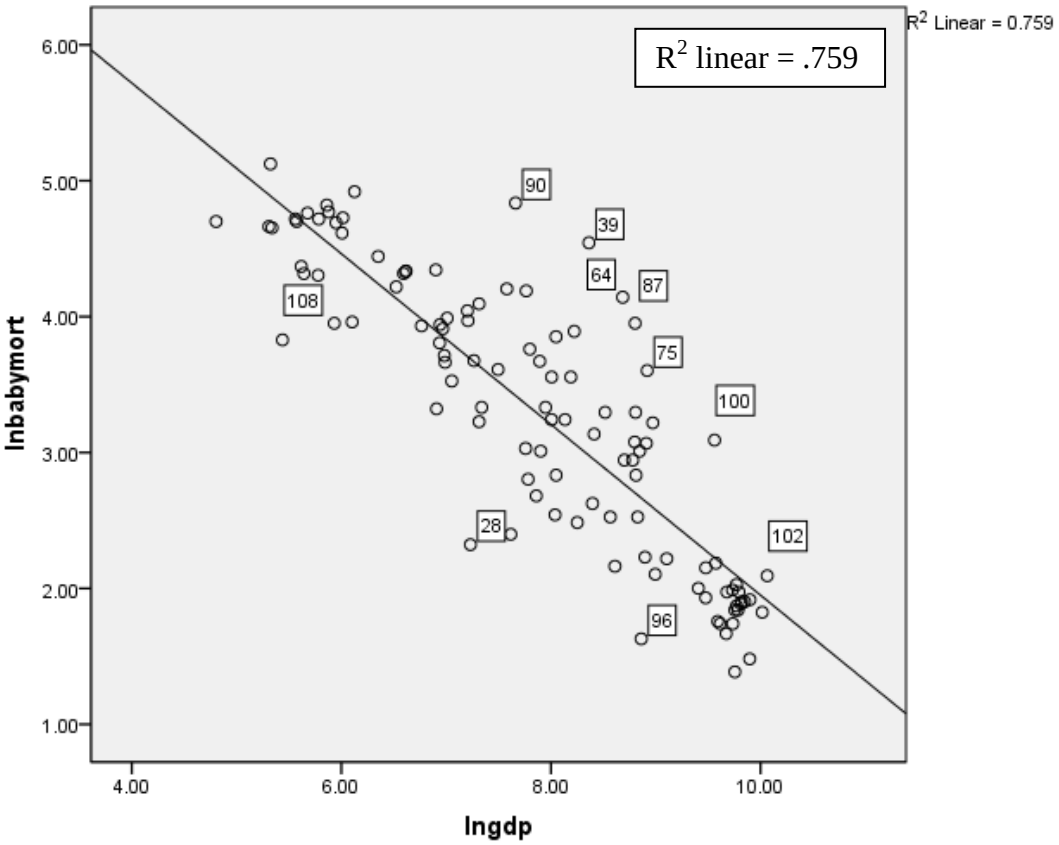
**compute lnbabymort = ln(babymort).**
**compute lngdp = ln(gdp_cap).**
**execute.**




The skew for these two variables is -.114 and -.243, respectively, and the kurtosis is -1.245 and -1.048, respectively. The histograms show distributions with no outliers, so we expect these variables should be suitable for regression modeling.

Next, we check a plot to examine how well a linear regression model fits the transformed data.




This linear regression model looks much more appropriate. The data are distributed close to linear, as reflected with the large multiple R squared of .759. Outliers or cases of interest can be identified with the ID code by opening the Chart Editor by double-clicking the graph, then selecting Elements, and Data Label Mode. Here some of the outlying cases have been identified.

| | | |
|---|---|---|
| 28 Cuba | 87 Saudi Arabia | 102 USA |
| 39 Gabon | 90 Somalia | 108 Vietnam |
| 64 Libya | 96 Taiwan | |
| 75 Oman | 100 United Arab Emirates | |

We can look for common themes and perhaps identify moderating variables. What accounts for unexplained variation? Do countries with higher than predicted infant mortality have greater wealth disparities among their populations than average? Does big oil production predict? Region of the world? Religion? Women's literacy? Women's literacy relative to men's literacy?

```
*Search for mediator variables.
recode region (4=1)(else=0) into Africa.
RECODE religion ('Muslim'=1) (ELSE=0) INTO Muslim.
EXECUTE.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT lnbabymort
  /METHOD=ENTER lngdp /ent lit_male /ent lit_fema /ent Africa /ent muslim
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(3).
```

**Correlations**

| | | lnbabymort | lngdp |
|---|---|---|---|
| Pearson Correlation | lnbabymort | 1.000 | -.823 |
| | lngdp | -.823 | 1.000 |
| | lit_male Males who read (%) | -.727 | .611 |
| | lit_fema Females who read (%) | -.761 | .632 |
| | Africa | .545 | -.469 |
| | Muslim | .310 | -.122 |
| Sig. (1-tailed) | lnbabymort | . | .000 |
| | lngdp | .000 | . |
| | lit_male Males who read (%) | .000 | .000 |
| | lit_fema Females who read (%) | .000 | .000 |
| | Africa | .000 | .000 |
| | Muslim | .002 | .133 |
| N | lnbabymort | 85 | 85 |
| | lngdp | 85 | 85 |
| | lit_male Males who read (%) | 85 | 85 |
| | lit_fema Females who read (%) | 85 | 85 |
| | Africa | 85 | 85 |
| | Muslim | 85 | 85 |

**Model Summary[f]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .823 | .678 | .674 | .50073 | .678 | 174.578 | 1 | 83 | .000 |
| 2 | .871 | .758 | .752 | .43656 | .080 | 27.196 | 1 | 82 | .000 |
| 3 | .880 | .775 | .767 | .42361 | .017 | 6.089 | 1 | 81 | .016 |
| 4 | .881 | .776 | .765 | .42550 | .001 | .282 | 1 | 80 | .597 |
| 5 | .885 | .784 | .770 | .42025 | .008 | 3.012 | 1 | 79 | .087 |

a. Predictors: (Constant), lngdp

b. Predictors: (Constant), lngdp, lit_male Males who read (%)

c. Predictors: (Constant), lngdp, lit_male Males who read (%), lit_fema Females who read (%)

d. Predictors: (Constant), lngdp, lit_male Males who read (%), lit_fema Females who read (%), Africa

e. Predictors: (Constant), lngdp, lit_male Males who read (%), lit_fema Females who read (%), Africa, Muslim

f. Dependent Variable: lnbabymort

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | | |
|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Zero-order | Partial | Part |
| 1 | (Constant) | 7.714 | .315 | | 24.465 | .000 | | | |
| | lngdp | -.549 | .042 | -.823 | -13.213 | .000 | -.823 | -.823 | -.823 |
| 2 | (Constant) | 7.833 | .276 | | 28.396 | .000 | | | |
| | lngdp | -.403 | .046 | -.605 | -8.813 | .000 | -.823 | -.697 | -.479 |
| | lit_male Males who read (%) | -.015 | .003 | -.358 | -5.215 | .000 | -.727 | -.499 | -.283 |
| 3 | (Constant) | 7.131 | .390 | | 18.266 | .000 | | | |
| | lngdp | -.380 | .045 | -.570 | -8.379 | .000 | -.823 | -.681 | -.442 |
| | lit_male Males who read (%) | .005 | .009 | .106 | .530 | .598 | -.727 | .059 | .028 |
| | lit_fema Females who read (%) | -.015 | .006 | -.502 | -2.467 | .016 | -.761 | -.264 | -.130 |
| 4 | (Constant) | 7.046 | .424 | | 16.630 | .000 | | | |
| | lngdp | -.378 | .046 | -.566 | -8.239 | .000 | -.823 | -.678 | -.436 |
| | lit_male Males who read (%) | .005 | .009 | .110 | .547 | .586 | -.727 | .061 | .029 |
| | lit_fema Females who read (%) | -.015 | .006 | -.485 | -2.342 | .022 | -.761 | -.253 | -.124 |
| | Africa | .079 | .148 | .037 | .531 | .597 | .545 | .059 | .028 |
| 5 | (Constant) | 6.988 | .420 | | 16.644 | .000 | | | |
| | lngdp | -.393 | .046 | -.588 | -8.519 | .000 | -.823 | -.692 | -.445 |
| | lit_male Males who read (%) | .002 | .009 | .048 | .240 | .811 | -.727 | .027 | .013 |
| | lit_fema Females who read (%) | -.010 | .007 | -.341 | -1.546 | .126 | -.761 | -.171 | -.081 |
| | Africa | .161 | .154 | .076 | 1.047 | .298 | .545 | .117 | .055 |
| | Muslim | .209 | .120 | .109 | 1.736 | .087 | .310 | .192 | .091 |

a. Dependent Variable: lnbabymort

**Excluded Variables[e]**

| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics: Tolerance |
|---|---|---|---|---|---|---|
| 1 | lit_male Males who read (%) | -.358[a] | -5.215 | .000 | -.499 | .627 |
| | lit_fema Females who read (%) | -.401[a] | -5.916 | .000 | -.547 | .600 |
| | Africa | .204[a] | 3.027 | .003 | .317 | .780 |
| | Muslim | .213[a] | 3.637 | .000 | .373 | .985 |
| 2 | lit_fema Females who read (%) | -.502[b] | -2.467 | .016 | -.264 | .067 |
| | Africa | .063[b] | .888 | .377 | .098 | .596 |
| | Muslim | .125[b] | 2.172 | .033 | .235 | .852 |
| 3 | Africa | .037[c] | .531 | .597 | .059 | .581 |
| | Muslim | .089[c] | 1.484 | .142 | .164 | .763 |
| 4 | Muslim | .109[d] | 1.736 | .087 | .192 | .690 |

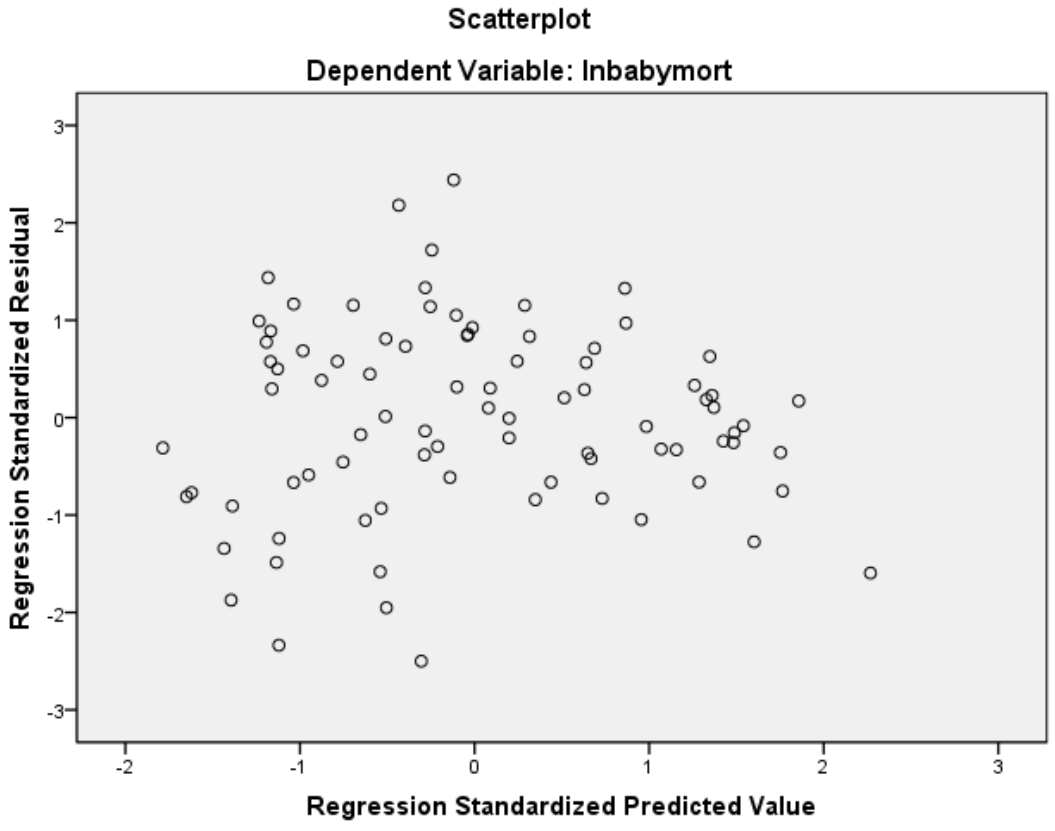
a. Predictors in the Model: (Constant), lngdp

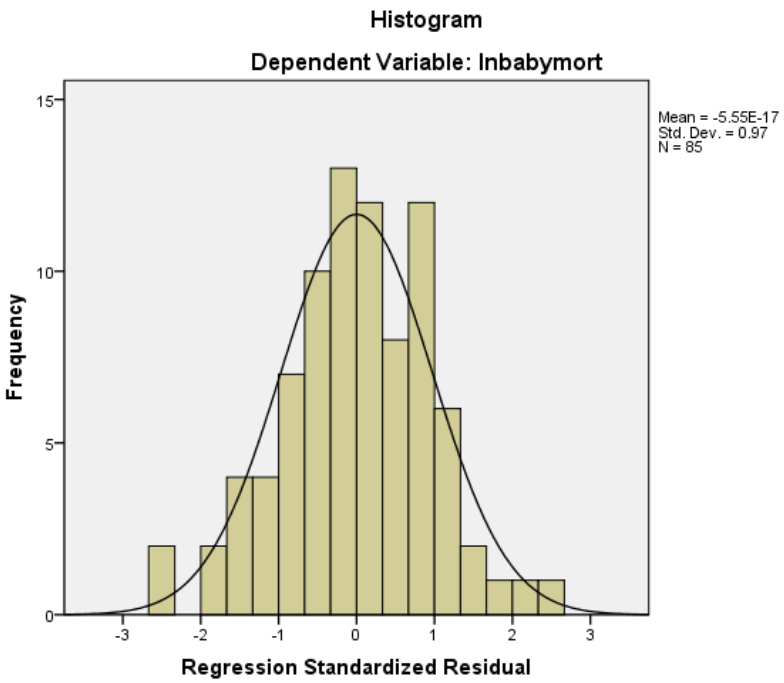
b. Predictors in the Model: (Constant), lngdp, lit_male Males who read (%)

c. Predictors in the Model: (Constant), lngdp, lit_male Males who read (%), lit_fema Females who read (%)

d. Predictors in the Model: (Constant), lngdp, lit_male Males who read (%), lit_fema Females who read (%), Africa

e. Dependent Variable: lnbabymort






Questions:

1. Is the relationship between infant mortality and gross domestic product linear?
2. What proportion of the variance in infant mortality can be predicted by gross domestic product using a linear model?
3. Is the relationship between the log of infant mortality and the log of gross domestic product linear?
4. What proportion of variance in the log of infant mortality can be predicted with the log of gross domestic product?
5. How is female literacy related to the log of infant mortality? Hint: check the Excluded variables table before and after male literacy is added to the model.
6. How is being an African country related to the log of infant mortality?
7. How is being a predominantly Muslim country related to the log of infant mortality?
8. How many countries were lost when we used the final model?

Questions not addressed in this output:
How would you characterize the countries that were missing in the final analysis? Hint: An easy diagnostic is to go to the SPSS .SAV file, click the Data View tab at the bottom, right click on a variable of interest, and click the option to sort. The entire cases are kept entact when they are sorted on any one variable. We see that many European countries did not report literacy rates. Can you find the data elsewhere, or make reasonable assumptions regarding the missing data? However you decide to deal with missing data, you must describe what you did clearly.