This is a demonstration of data screening and transformations for a regression analysis. Our interest is in predicting current salary from education level for a sample of employees of a bank. These are real data provided by SPSS and available on Sakai as an SPSS systems file under the name BANK.SAV, or as an SPSS portable file, BANK.POR.

We should begin by examining the univariate and bivariate distributions for variables of interest.

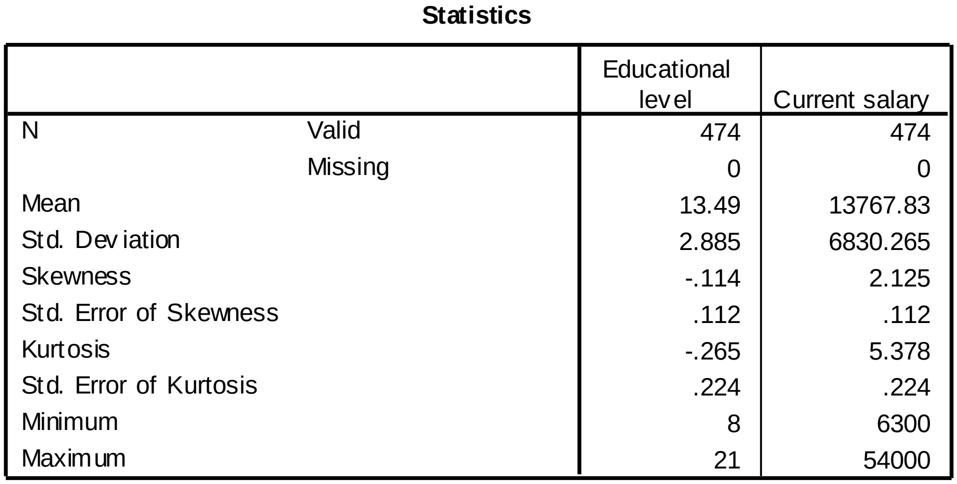
Open SPSS and the BANK.SAV data set. Click Analyze, Descriptive Statistics, Frequencies…, select Educational level (edlevel) and Current salary (salnow) as the Variable(s). Click Statistics, select Mean, Skewness, and Kurtosis, Std. Deviation, Minimum, and Maximum, and click Continue. Click Charts, select Histograms, check Show normal curve, and click Continue.

Click Format and check Suppress tables with more than n categories, and enter **20** as the Maximum number of categories, and click Continue. This will suppress the frequency table for salary where there may be well over 100 different individual salaries but will provide the frequency table for education level where there are fewer than 20 categories.

We can click Paste to save the syntax. Click Window, select the Syntax editor to see the syntax:

```
FREQUENCIES VARIABLES=edlevel salnow
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN SKEWNESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM NORMAL
  /FORMAT=LIMIT(20)
  /ORDER=ANALYSIS.
```

Run the syntax. Here is selected output.



Statistics

		Educational level	Current salary
N	Valid	474	474
	Missing	0	0
Mean		13.49	13767.83
Std. Deviation		2.885	6830.265
Skewness		-.114	2.125
Std. Error of Skewness		.112	.112
Kurtosis		-.265	5.378
Std. Error of Kurtosis		.224	.224
Minimum		8	6300
Maximum		21	54000

← !

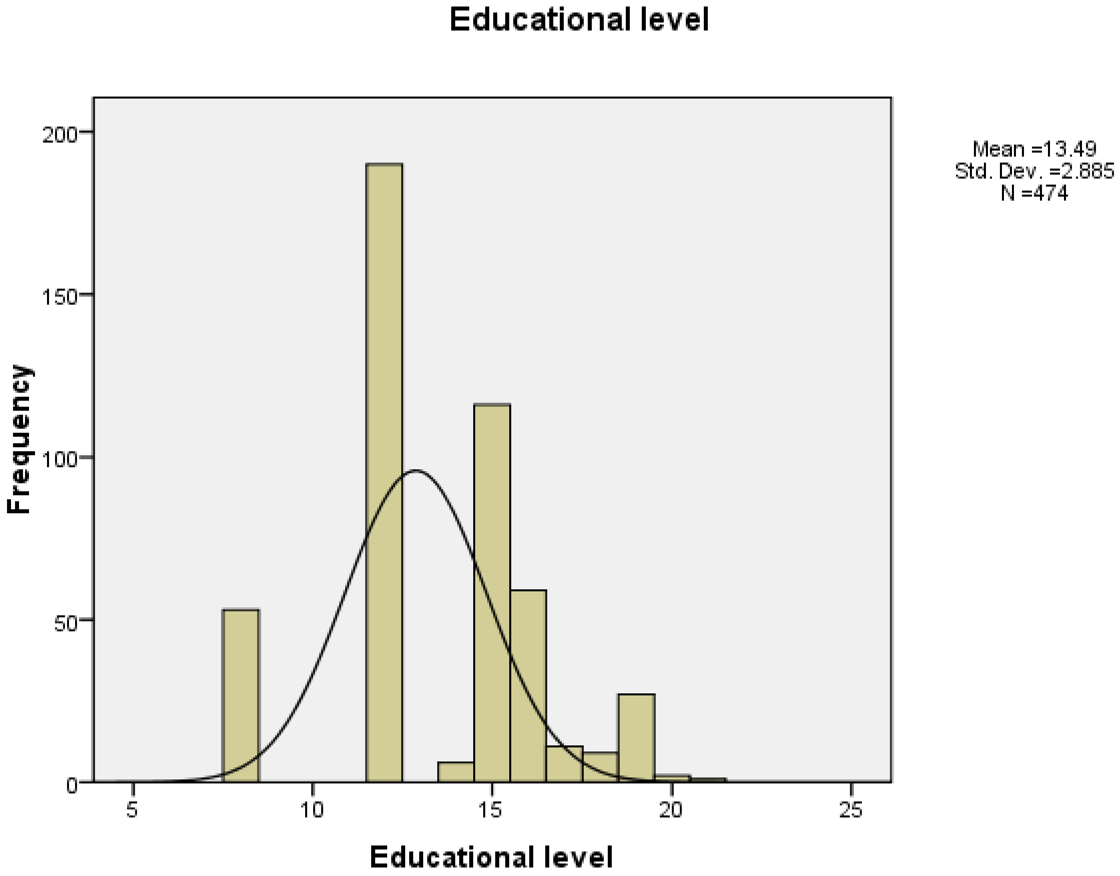
What catches your eye?
Kurtosis much greater than +1 should pique our interest. The kurtosis for Current salary is 5.378! We need to investigate this. Skew is also greater than 1 in absolute value. Outliers are the most common cause of large kurtosis, and outliers also skew a distribution if they favor one end of the distribution. Negative kurtosis indicates short tails and generally is not cause for alarm. The case with maximum salary is 54,000, which is over 5 SD greater than the mean – that is an outlier!

Frequency Table

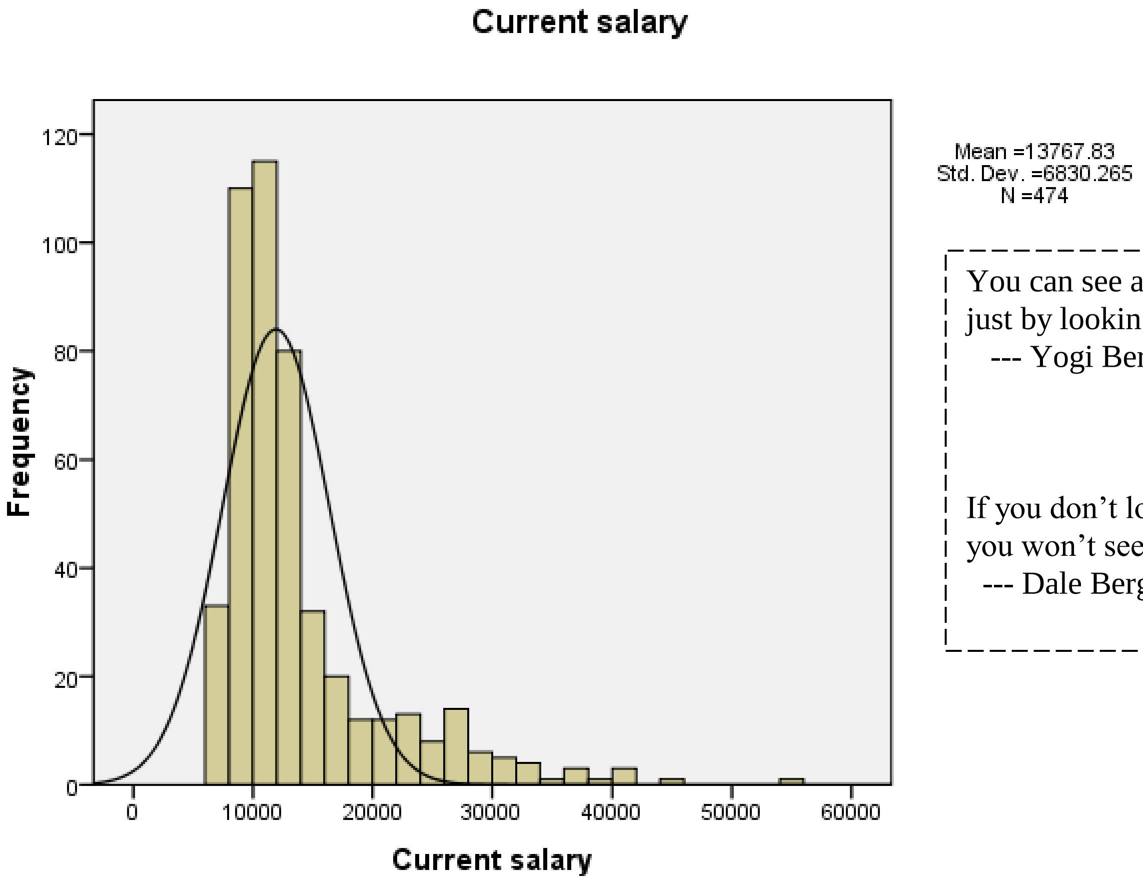
Educational level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	53	11.2	11.2	11.2
	12	190	40.1	40.1	51.3
	14	6	1.3	1.3	52.5
	15	116	24.5	24.5	77.0
	16	59	12.4	12.4	89.5
	17	11	2.3	2.3	91.8
	18	9	1.9	1.9	93.7
	19	27	5.7	5.7	99.4
	20	2	.4	.4	99.8
	21	1	.2	.2	100.0
	Total	474	100.0	100.0	

Here we have the exact frequency distribution for education level. We can see that it is not a nice continuous normal distribution because there are several spikes and gaps. We should not be surprised to see the spike at 12 because that indicates a high school graduate who has not gone on to college. The spike at 15 is more interesting. Perhaps recruiting favors people who have completed a three-year program after high school. This is something to investigate.



We can edit the graph in SPSS to change labels, intervals, colors, etc. The default labels on Education could be changed; we would not use these for presentation to other folks.



You can see a lot
just by looking.
--- Yogi Berra

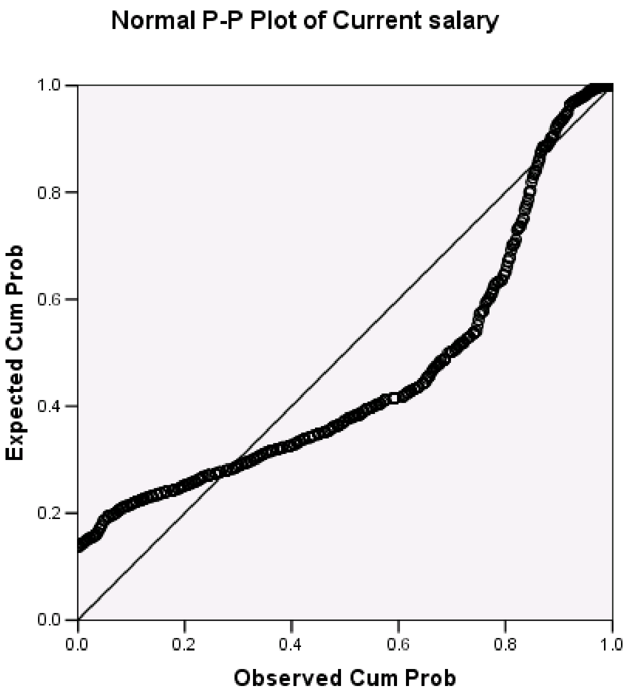
If you don’t look,
you won’t see it!
--- Dale Berger

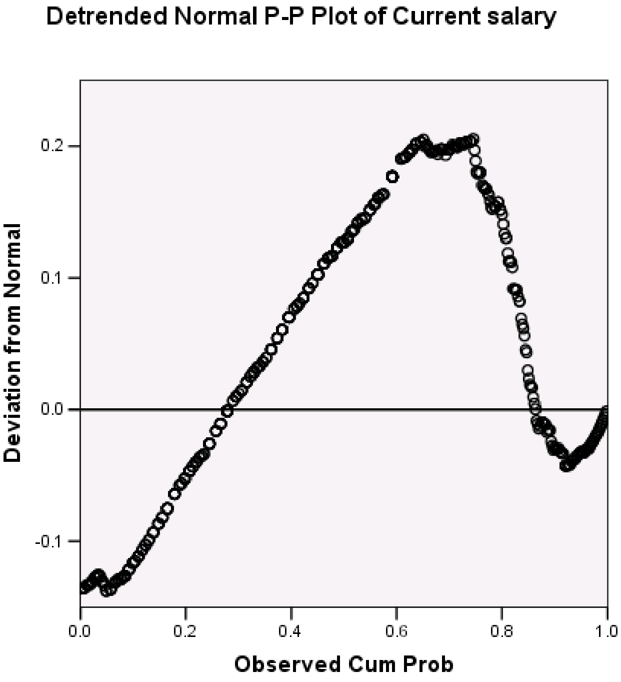
The histogram for Current salary clearly shows strong skew, with a few relatively extremely large values. These cases have a great influence on mean and variance, and potentially can also have a great influence on correlation. Statistical tests of significance assume normal distributions of errors, so these cases are likely to distort the tests substantially.

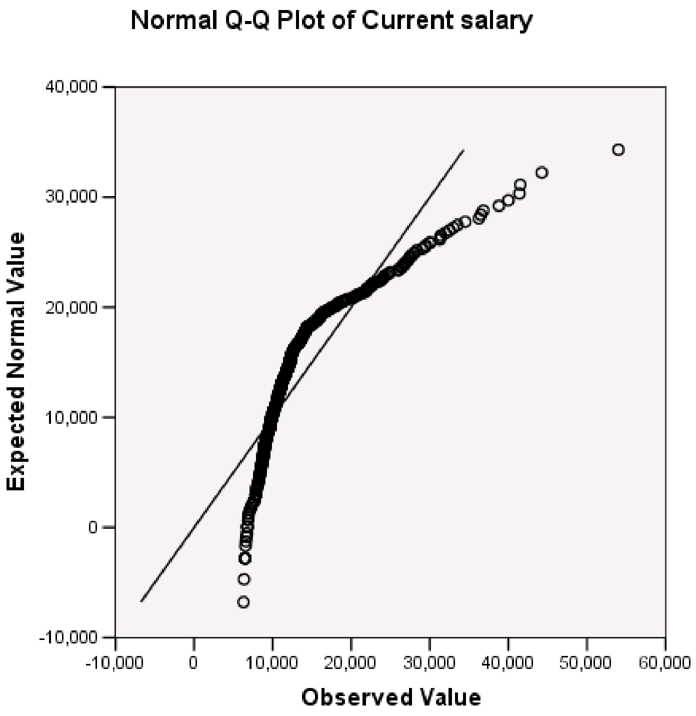
Other diagnostics to check for departures from normality are the P-P plot and Q-Q plot. You can generate a P-P plot by clicking Analyze, Descriptive Statistics, P-P Plots…and selecting **salnow** as the variable. Click OK. The P-P plot compares the expected cumulative probability assuming a normal distribution to the observed cumulative probability for each case. If the distribution is normal, the points form a line on the diagonal. Here we see that the left tail is shorter than normal, because the *Observed Cum Prob* is still zero when the expected proportion is already over .10. The middle of the distribution includes more cases than a normal distribution.

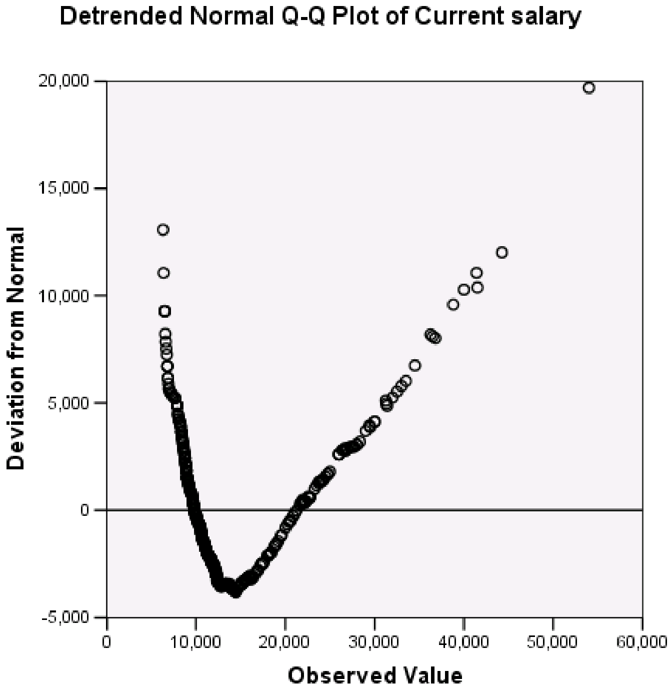
Similarly, the Q-Q plot shows the expected value vs. the observed value for each case where the expected value is calculated as the value expected for a case at the observed percentile on a normal distribution with the observed mean and SD. The Q-Q plot shows that if we had a normal distribution with the observed mean and standard deviation, the lowest expected value would be about negative $8000! The lowest observed value is positive $6300. At the upper end, the highest expected value is about $35,000 but the largest observed value is $54,000. (You can find the actual minimum and maximum in our initial summary of descriptive statistics.) Based on the mean and SD for our sample, there are fewer than expected cases at the very low values and more than expected at the very high values.

Our eyes are very good at detecting departures from a straight line, though in our example the departure from normality is strongly apparent even in the histogram.







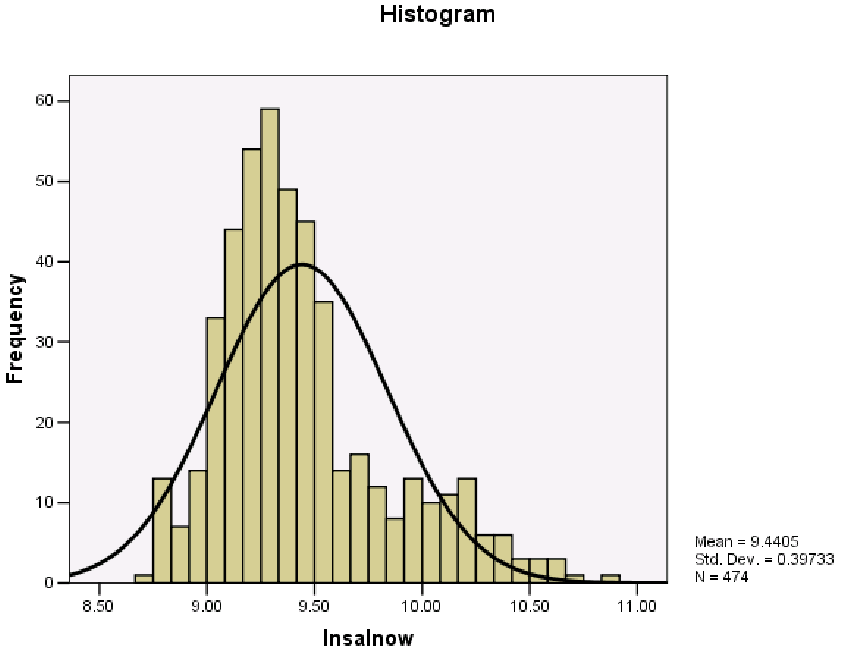


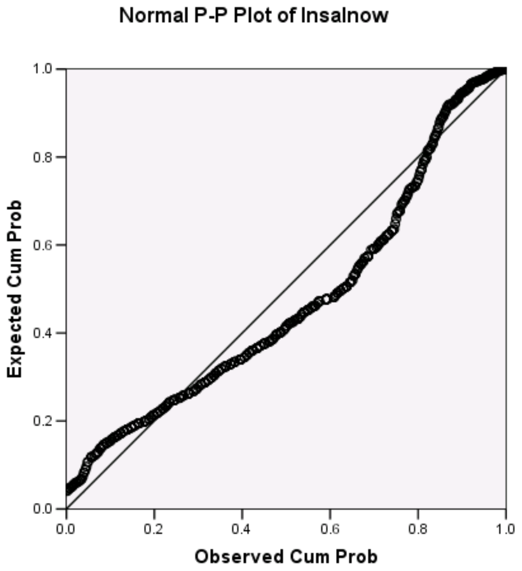
Detrended plots show difference between the observed and predicted for each case (the horizontal difference between the point and the straight line). These plots show deviations from a model and patterns in those deviations clearly, but it does take some practice to interpret them, especially because SPSS rescales these plots to fill the space – small differences become large. It is easy to see patterns in how the sample data depart from a normal distribution.

Financial data often have a positive skew and a log transformation is commonly applied to produce a measure that is better for modeling and hypothesis testing. We can create a new log transformed variable where lnsalnow = ln(salnow) by clicking Transform, Compute variable…, type **lnsalnow** into Target Variable, under Function Group select All, under Functions and Special Variables select Ln, click the curved arrow that points up, select Current salary [salnow] and click the curved arrow that points right, and click Paste to save the syntax.

```
COMPUTE lnsalnow=LN(salnow).
EXECUTE.
```

We need to run the procedure that defines the variable – you can go to the syntax window, highlight the two lines and click the triangle to run. Next we examine the shape of the new variable. Click Analyze, Descriptive Statistics, Frequencies…, select **lnsalnow** as the only variable, click Statistics, select Mean, Skewness, and Kurtosis, Std. Deviation, Minimum, and Maximum, and click Continue. Click Charts, select Histograms, check Show normal curve, and click Continue. Click Format…, select Suppress tables with a maximum of 10 categories. Run it.





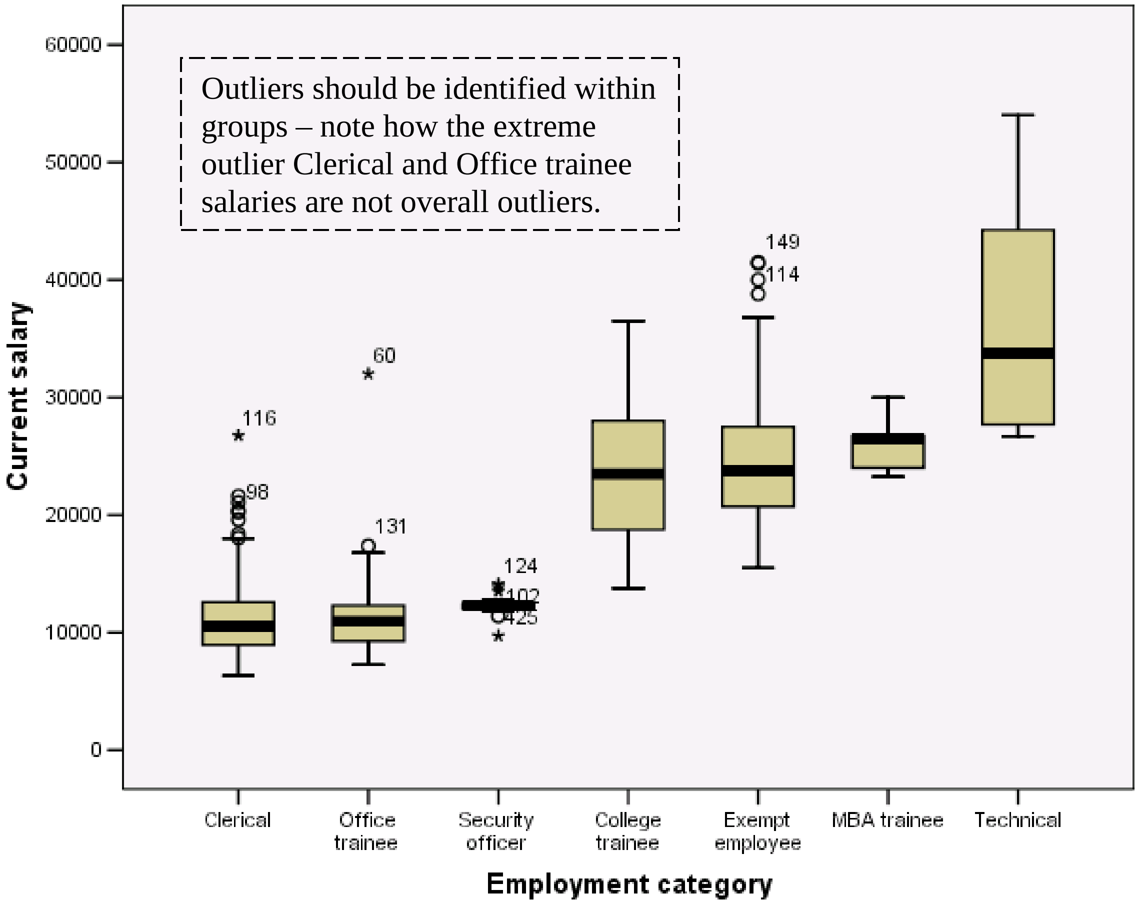
The summary statistics and the plot look much better. Skew is 1.00 and kurtosis is .68. The plot shows an interesting departure from normality in that it appears to be somewhat bimodal. This suggests that we may have more than one population in our sample.

The bank employees include clerical workers, office trainees, security officers, college trainees, exempt employees, MBA trainees, and technical staff. Boxplots provide a useful tool for taking a quick look for possible differences between these groups. Click Graphs, Chart Builder…, Boxplot…, drag the Simple Boxplot into the Chart window, Drag Employment category into the X axis, drag Current salary into the Y axis, click OK (or PASTE syntax and run from the syntax window).

The bottom and top of the box are the first and third quartile, respectively, and the heavy line in the box is the median (the 50th percentile). Some programs extend the 'whiskers' from the ends of the box all the way out to the most extreme score. SPSS does not allow a whisker to extend beyond a box more than 1.5 times the distance between Q1 and Q3 (called the Inter Quartile

Range, or IQR). Cases between 1.5 IQR and 3.0 IQR are indicated with a hollow circle (outliers), and cases beyond 3.0 IQR from the end of the box are indicated with an asterisk (extreme outliers). Some programs use other rules, so make sure you know what the rules are, and you should indicate what rules you used (e.g., SPSS21) when you report box plots. Some statisticians follow Tukey's terminology and call the quartiles "hinges."

We could do the same analysis with **lnsalnow**, but it is easier to interpret the untransformed measure of salary. In the boxplots we see positive skew within most categories, and we see that there are sizable group differences. A check on the frequency distribution shows that the largest group, by far, is Clerical with 227 cases. Some groups (MBA trainee and Technical) have only five or six cases. For further analyses here we will limit our model building to clerical staff to avoid the large confound that job category brings to salary.



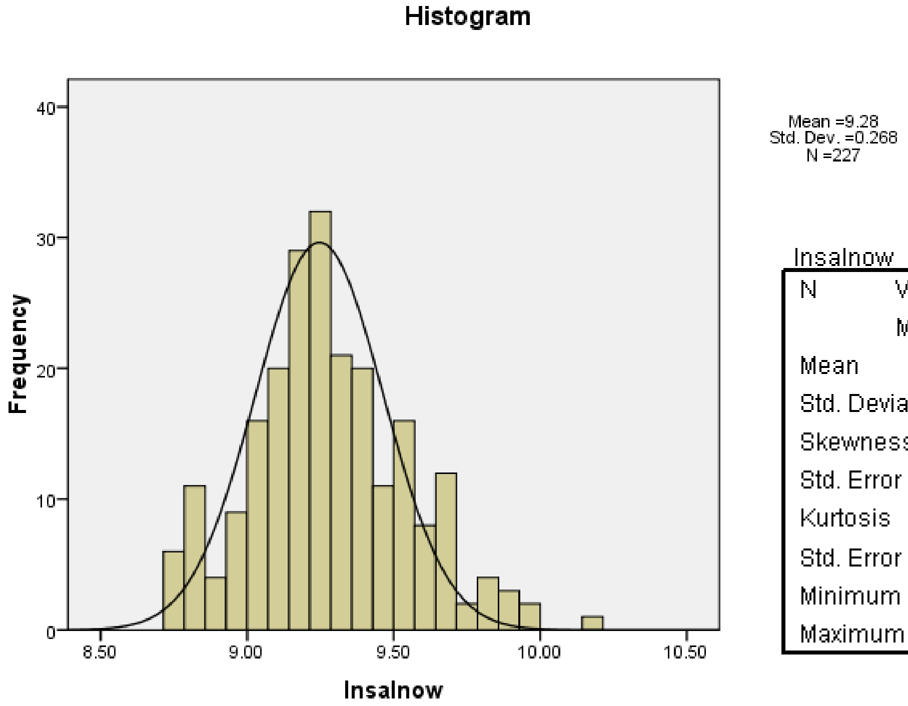
Outliers should be identified within groups – note how the extreme outlier Clerical and Office trainee salaries are not overall outliers.

Clerical staff is coded as 1 on the variable jobcat. To select only clerical staff, click Data, Select Cases…, select If condition satisfied, Click If…, select Employment category [jobcat] and click the arrow, click =, 1, Continue, Paste. Running this syntax creates a new **filter_$** variable in your data set. **filter_$** = 1 for cases where jobcat=1 and **filter_$** = 0 for all other cases. After you run this syntax, this filter will stay on for all subsequent analyses until you change the filter setting.

```
USE ALL.
COMPUTE filter_$=(jobcat = 1).
VARIABLE LABEL filter_$ 'jobcat = 1 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
```

After we run the filter syntax, let’s check the distribution of **lnsalnow**. We can rerun the P-P and Q-Q plots as well. To keep an accurate record of analyses, it is good practice to copy the appropriate syntax to the current end of the syntax file.

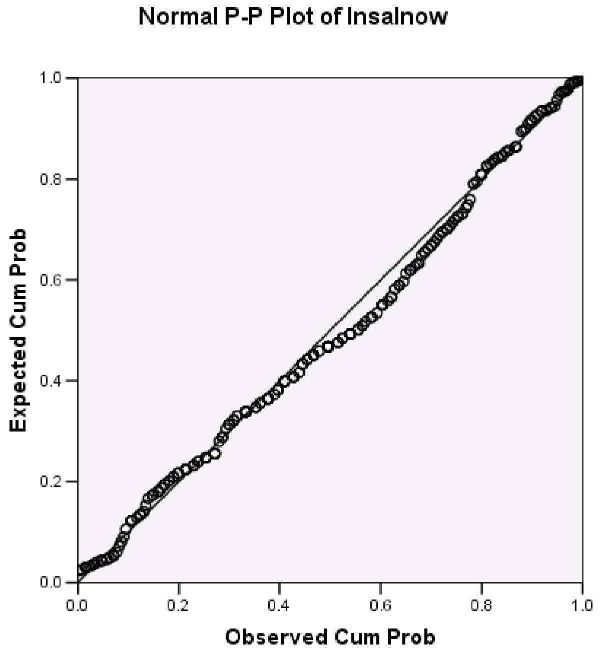
An option in SPSS includes the syntax for each procedure with the output. You can turn this on by clicking Edit, Options…, select the Viewer tab, on the bottom left check the box labeled Display commands in the log. **I strongly recommend using this option.** This will help you keep track of what commands generated specific output.

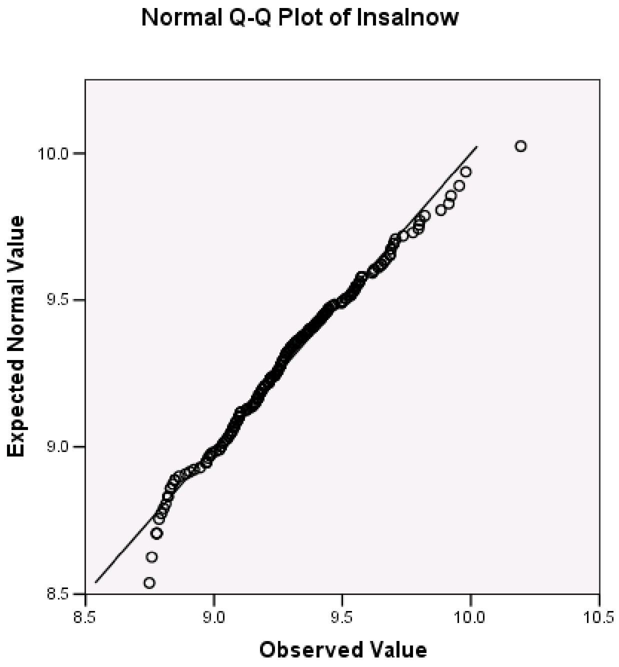


Statistics

lnsalnow

N	Valid	227
	Missing	0
Mean		9.2809
Std. Deviation		.26771
Skewness		.380
Std. Error of Skewness		.162
Kurtosis		.175
Std. Error of Kurtosis		.322
Minimum		8.75
Maximum		10.19





These distributions look much better. The histogram shows that the sample is quite close to normal, the skew and kurtosis are well under 1, and the P-P and Q-Q plots are quite linear with only a few points that are somewhat off. The lower tail is still a bit short, the upper tail a bit long, and there is a hint of a little subpopulation at the lower end, but all in all this looks pretty good.

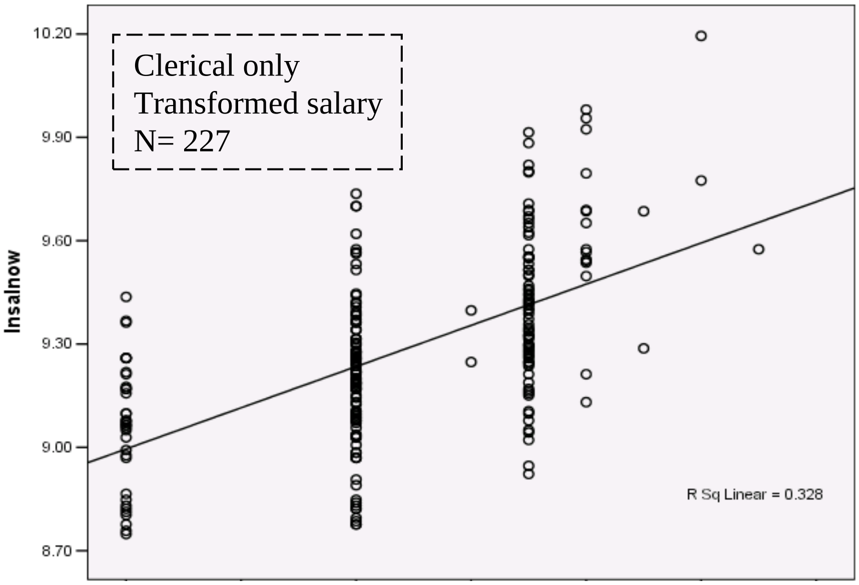
Now let’s check the bivariate relationship between **edlevel** and **lnsalnow**. Click Graphs, Scatter/Dot…, select Simple Scatter, click Define, select **lnsalnow** for the Y axis and **edlevel** for the X axis.

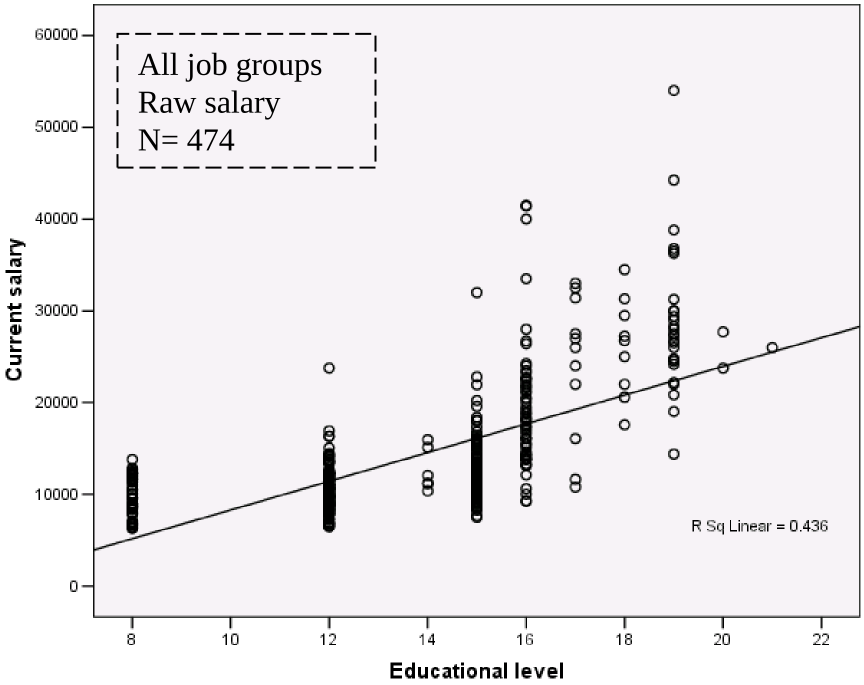
```
GRAPH
 /SCATTERPLOT(BIVAR)=edlevel WITH salnow
 /MISSING=LISTWISE .
```

Our model fits these data quite well. We have an essentially homoscedastic linear relationship. The R squared of .328 indicates that education accounts for about a third of the variance in salaries of the 227 clerical workers.

The lower graph shows the relationship between education and untransformed salary for all job groups combined (N=474).

While the overall R squared is larger in the full data set (R squared = .436 for the full sample of 474 cases), the regression model does not fit appropriately. The model systematically under predicts salary for those at the lowest education level and for those at the higher levels (curvilinearity) and the variability is much greater at the higher education levels (heteroscedasticity). Predictions and tests of statistical significance would be compromised.





Now let's generate a regression model to predict salary for clerical staff based on education level. Click Analyze, Regression, Linear…, select **lnsalnow** as the Dependent variable, select **edlevel** as the Independent variable. Click Statistics…, select Estimates, Confidence intervals, Model fit, R squared change, and Descriptives, and click Continue. Click Plots…, check Histogram, select ***ZRESID** as the Y variable, ***ZPRED** as the X variable, click Continue, click OK.

Descriptive Statistics

	Mean	Std. Deviation	N
lnsalnow	9.2809	.26771	227
Educational level	12.78	2.562	227

Check that we have the correct sample: n=227

Correlations

		lnsalnow	Educational level
Pearson Correlation	lnsalnow	1.000	.572
	Educational level	.572	1.000
Sig. (1-tailed)	lnsalnow	.	.000
	Educational level	.000	.
N	lnsalnow	227	227
	Educational level	227	227

Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.572[a]	.328	.325	.22003	.328	109.573	1	225	.000

a. Predictors: (Constant), Educational level

b. Dependent Variable: lnsalnow

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.305	1	5.305	109.573	.000[a]
	Residual	10.893	225	.048		
	Total	16.197	226			

a. Predictors: (Constant), Educational level

b. Dependent Variable: lnsalnow

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	8.517	.074		114.436	.000	8.370	8.664
	Educational level	.060	.006	.572	10.468	.000	.049	.071

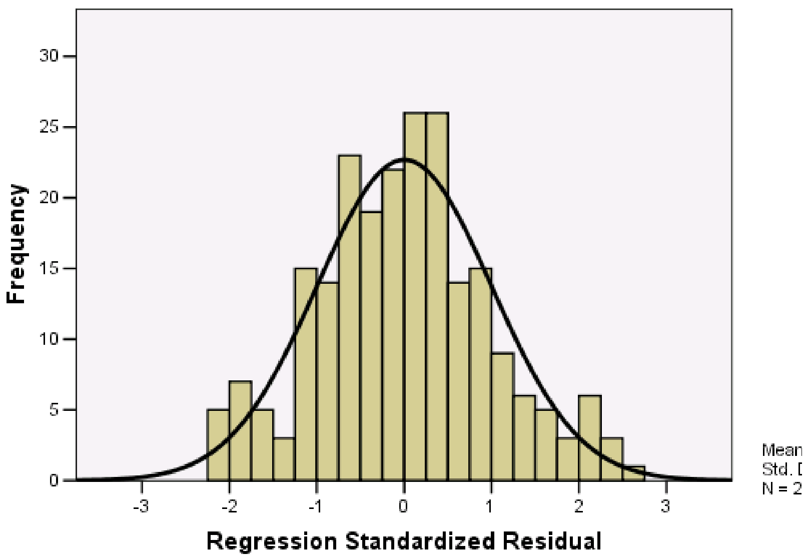
a. Dependent Variable: lnsalnow

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	8.9954	9.6531	9.2809	.15321	227
Residual	-.49130	.60094	.00000	.21954	227
Std. Predicted Value	-1.864	2.430	.000	1.000	227
Std. Residual	-2.233	2.731	.000	.998	227

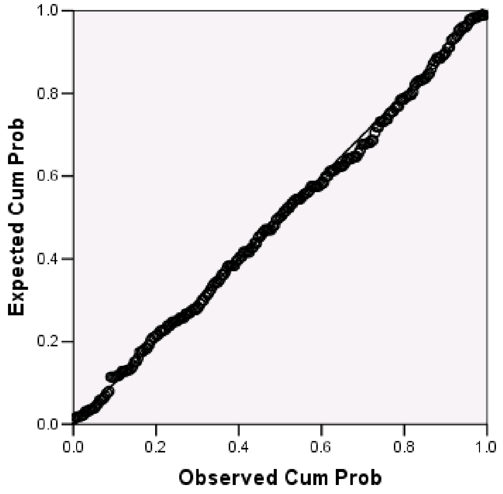
a. Dependent Variable: lnsalnow

Dependent Variable: lnsalnow

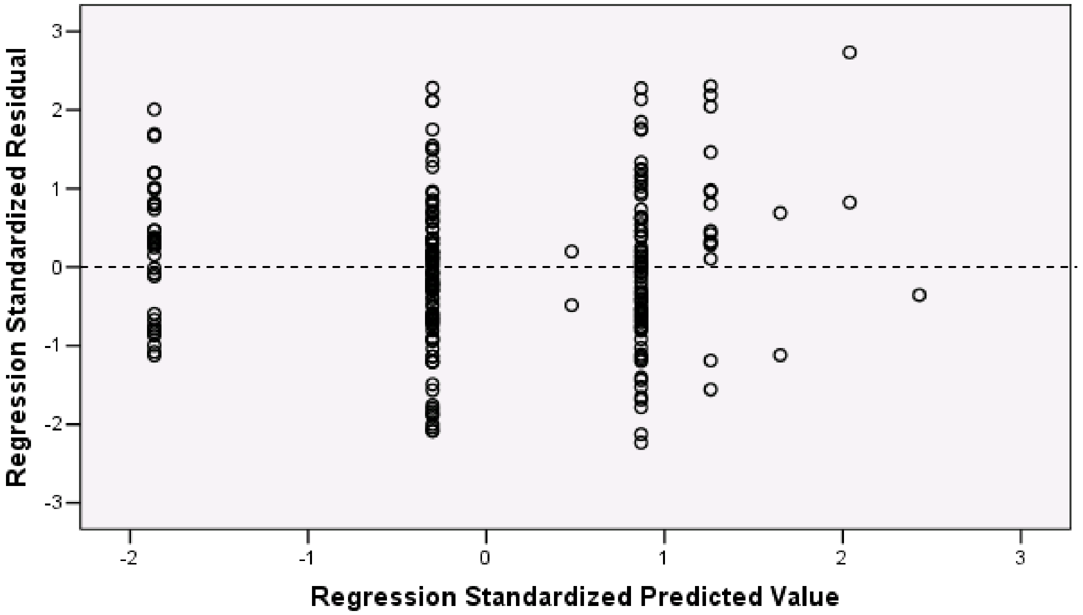


Normal P-P Plot of Regression Standardized Residual

Dependent Variable: lnsalnow



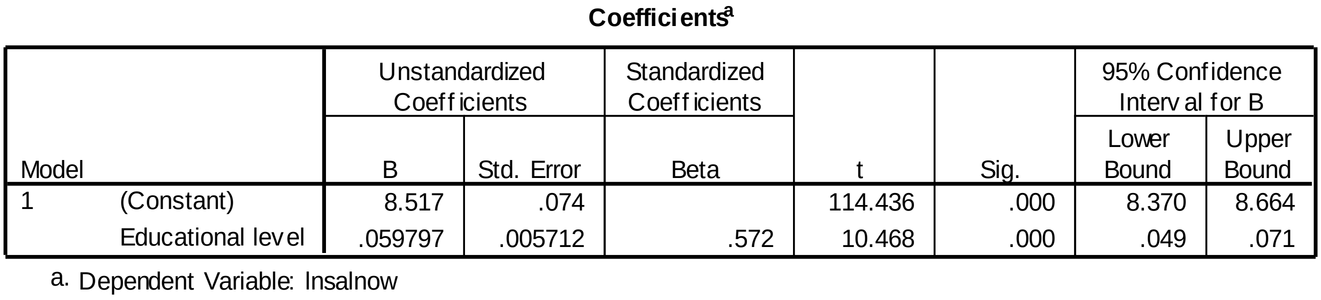
Dependent Variable: lnsalnow



I added the dashed reference line at 0. Compare deviations above and below this line.

An important assumption of regression analysis is that the residual errors are normally distributed. The residual plots look great. Now let's apply the regression model.

In the earlier Coefficients Table we found the constant = 8.517 and B for **edlevel** = .060 with standard error = 006. The standard error shows only one significant digit, which is inadequate. We need to use greater precision in our report. In the SPSS Viewer window, double-click on the coefficients table, and right-click on the cell of interest. Select Cell properties, Format Value, and change Decimals from 3 to 6. Compare the table below to the comparable table we saw earlier.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	8.517	.074		114.436	.000	8.370	8.664
	Educational level	.059797	.005712	.572	10.468	.000	.049	.071

a. Dependent Variable: lnsalnow

The regression model: **Predicted lnsalnow = 8.517 + .059797 * edlevel**.

Let's use this model to predict the salary of someone who has 10 years of education. A little arithmetic gives us the predicted lnsalnow = 8.517 + (.059797)*10 = 9.11497. That's nice but not very easy to explain to a lay audience. We need to convert from the log scale back to the familiar scale of dollars. Because lnsalnow = ln(salnow), the constant e = 2.71828 raised to the power of lnsalnow = salnow. You can do this with a calculator easily if you have an e^x button. You'll get $9090.36. You can also use Excel to do this calculation: =EXP(9.11487) gives 9090.36.

A predicted value is much more useful if we know the margin of error in the prediction. We begin by finding the appropriate formulas and values. In the text or in the formula section of the handout we find the formula for the standard error of estimate for an individual score.

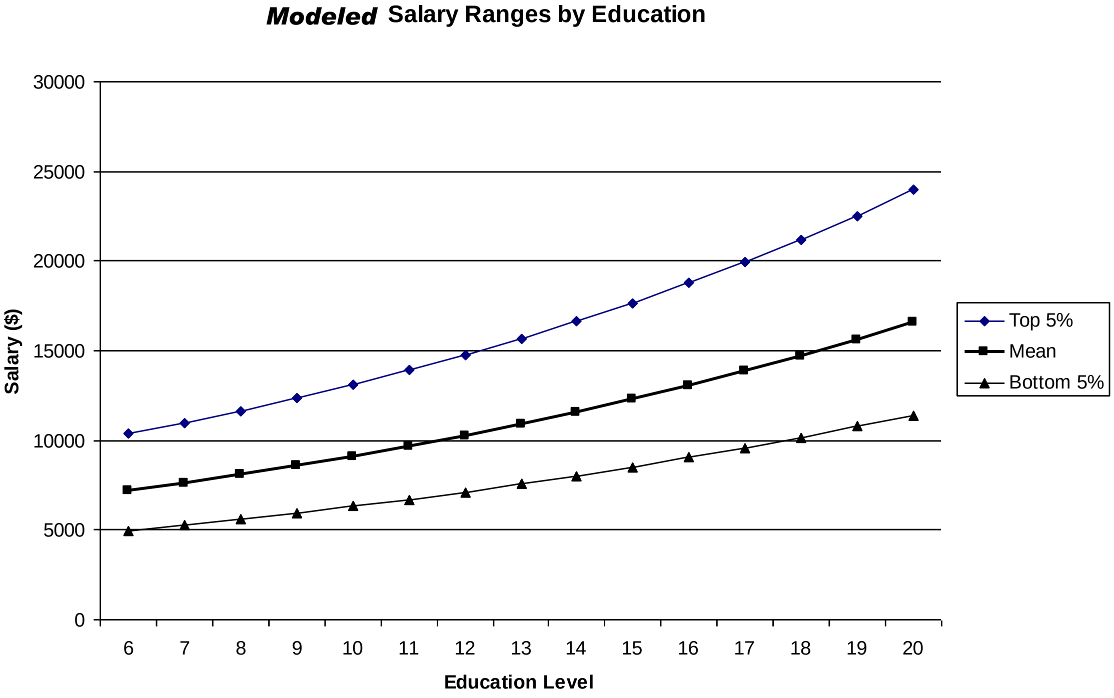
In our example, $S_{Y.X}$ =.22003 from the model summary (the standard error of estimate). Xi is the specific education level. The mean (12.78) and standard deviation (2.562) for edlevel are shown in the Descriptive Statistics table (be sure we use the table where n=227, because we are using clerical only and not the full sample).

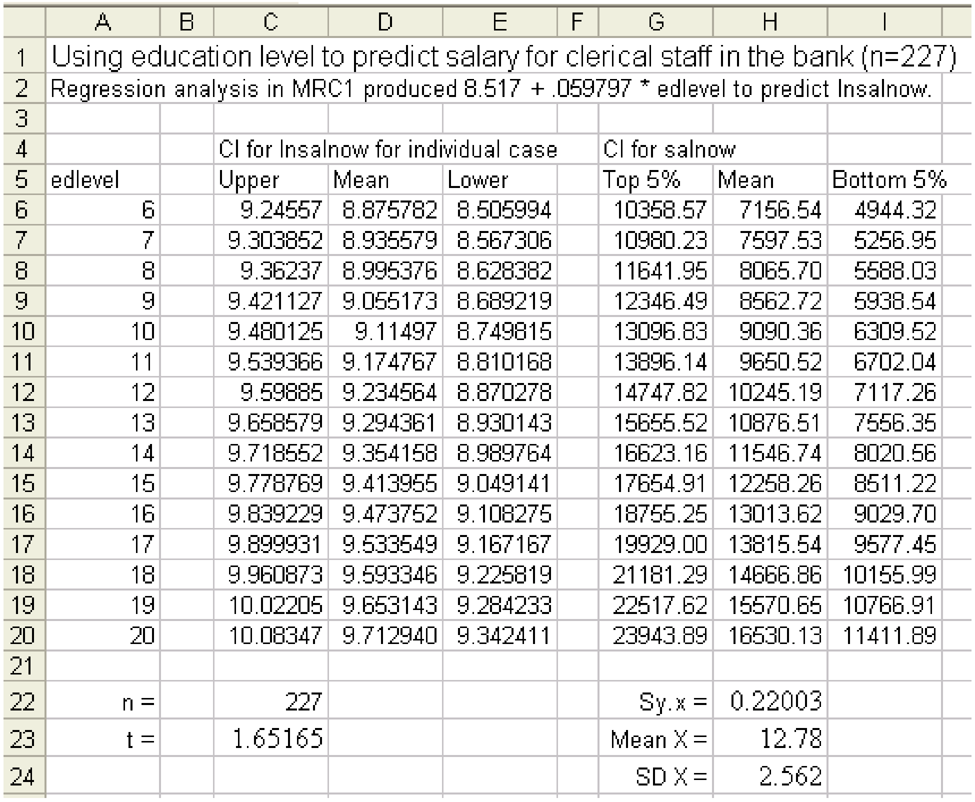
$$S'_{Y.X} = S_{Y.X}\sqrt{1+\frac{1}{N}+\frac{(Xi-\overline{X})^2}{(N-1)S_X^2}} = .22003\sqrt{1+\frac{1}{227}+\frac{(10-12.78)^2}{(227-1)(2.562)^2}} = .22003\sqrt{1.0096} = .221085$$

To construct a confidence interval, we find the upper and lower limits around the predicted value by adding or subtracting $(t_{\alpha/2})(S'_{Y.X})$. For a 95% CI with N = 227 (df = 225) we can use StatWISE to find $t_{\alpha/2}$ =1.97057. For someone with edlevel = 10, the predicted lnsalnow = 9.11497 plus or minus (1.97057)(.221085) = .43566. These limits are 8.67931 and 9.55063. Thus we can say that the probability is 95% that the interval 8.67931 to 9.55063 captures the lnsalnow for a clerical worker at the bank who has 10 years of formal education.

When we translate these limits on lnsalnow to limits on salnow, we get $5879.99 and $14,053.55. We should round these values off to whole dollars, $5880 to $14054. Note that the range is greater above the predicted value than below, reflecting the skew in the original scale.

A useful tool for a manager who would like to use these data would be a table or graph showing percentile intervals for predicted values of salnow for individuals with various education levels. Hand calculations are tedious and subject to error. If we need to do many such calculations, it is much better to use a computer than to do them by hand. Excel works very well for applications like this. Below is a chart that I edited to remove the background grey, place the limits to 90% CI, changed colors to black so they would reproduce better in black and white, and ordered the series with the top first so it would appear on the top in the labels as well.





	A	B	C	D	E	F	G	H	I
1	Using education level to predict salary for clerical staff in the bank (n=227)								
2	Regression analysis in MRC1 produced 8.517 + .059797 * edlevel to predict lnsalnow.								
3									
4			CI for lnsalnow for individual case				CI for salnow		
5	edlevel		Upper	Mean	Lower		Top 5%	Mean	Bottom 5%
6	6		9.24557	8.875782	8.505994		10358.57	7156.54	4944.32
7	7		9.303852	8.935579	8.567306		10980.23	7597.53	5256.95
8	8		9.36237	8.995376	8.628382		11641.95	8065.70	5588.03
9	9		9.421127	9.055173	8.689219		12346.49	8562.72	5938.54
10	10		9.480125	9.11497	8.749815		13096.83	9090.36	6309.52
11	11		9.539366	9.174767	8.810168		13896.14	9650.52	6702.04
12	12		9.59885	9.234564	8.870278		14747.82	10245.19	7117.26
13	13		9.658579	9.294361	8.930143		15655.52	10876.51	7556.35
14	14		9.718552	9.354158	8.989764		16623.16	11546.74	8020.56
15	15		9.778769	9.413955	9.049141		17654.91	12258.26	8511.22
16	16		9.839229	9.473752	9.108275		18755.25	13013.62	9029.70
17	17		9.899931	9.533549	9.167167		19929.00	13815.54	9577.45
18	18		9.960873	9.593346	9.225819		21181.29	14666.86	10155.99
19	19		10.02205	9.653143	9.284233		22517.62	15570.65	10766.91
20	20		10.08347	9.712940	9.342411		23943.89	16530.13	11411.89
21									
22	n =		227				Sy.x =	0.22003	
23	t =		1.65165				Mean X =	12.78	
24							SD X =	2.562	