PREDICTABLE/EXPLAINABLE CRITERION VARIANCE **VERSUS**

ERROR/UNEXPLAINABLE/UNACCOUNTED-FOR CRITERION VARIANCE

Dennis Roberts
Educational Psychology

What if the context is that you have 10 institutions and, you need to predict the average salary value at each of these 10 institutions. This is the criterion value, call it Y. I make a list for you of institutions 1 to 10 ... and then you are to put down next to each one what you estimate/predict the salary value will be. But, you have no other information to go on so, it is like a crap shoot. However, I do have some descriptive stats on all 10, together, and these are:

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
salary	10	15.10	14.00	14.50	3.60	1.14

MODEL OF USING MEAN OF Y

So, I do tell you that the mean salary across the 10 institutions is 15.1, and although this is not much, it is a start. Therefore, your best estimate (given no other information) for each of the 10 institutions is 15.1. You can't do any better than that at the moment but it is somewhere in the ballpark. So, let's see how you did? How big are your errors?

Inst	salary	mean	meanerr	
1	15	15.1	-0.1	<---- estimate OVERshot by .1
2	14	15.1	-1.1	
3	12	15.1	-3.1	
4	12	15.1	-3.1	
5	17	15.1	1.9	
6	23	15.1	7.9	<---- estimate UNDERshot by 7.9
7	14	15.1	-1.1	
8	19	15.1	3.9	
9	12	15.1	-3.1	
10	13	15.1	-2.1	NOTE: Average absolute error = 2.74

While using the mean did get you close to actual salary #1, you missed by a bunch for actual salary #6.

What is the variance of the actual criterion Y salaries?

```
MTB > let k1=stdev(c3)**2
MTB > prin k1
K1    12.9889
```

The variance of the ACTUAL salaries is 12.9889

What is the variance of the errors you made when you used the mean (ie, meanerr) as your best estimate of the salaries for each of the 10 schools? We could also calculate the variance of the meanerr values.

```
MTB > let k2=stdev(c5)**2
MTB > prin k2
K2    12.9889
```

The variance of the errors is 12.9889, which is the same as the variance of the actual criterion Y values. That is, THERE IS JUST AS MUCH VARIANCE IN THE ERRORS AS THERE IS IN THE ACTUAL CRITERION SALARY VALUES. Or, another way to say this is: all of the Y criterion variance is error. Hence, none of the actual criterion variance is explainable by anything other than the mean value on the criterion. Thus, 100% of the criterion Y variance is unexplainable.

But, what if we did have some other information; ie, one or two OTHER variables that might be predictive of the salary criterion? Look below.

```
MTB > print c1-c3
```

Inst	size	prestige	salary
1	17	14	15
2	11	25	14
3	19	27	12
4	18	22	12
5	18	22	17
6	29	26	23
7	19	17	14
8	17	25	19
9	16	14	12
10	13	12	13

In this case, assume that the two other variables are size (X1) of the school and a prestige rating (X2)for the school.

```
MTB > corr c1-c3

           size   prestige
prestige  0.371
salary    0.672    0.419
```

Note that both size and prestige correlate positively with salary; ie, the larger the size of the institution or the higher the prestige rating of the institution, the higher the average salary at the institution. Can we use this information .... either using only 1 of the predictors or both of them, to estimate salary and, REDUCE the error variance? Remember, when we used the mean only, we found that the variance of the errors was the SAME (12.9889) as

the variance of the criterion salary values.

MODEL OF ONE PREDICTOR VARIABLE

Since size seems to correlate a little more highly with salary (.672)than prestige does (.419), let's start with the better predictor first. We can find a simple regression equation using size (X1) to estimate salary (Y).

```
MTB > regr c3 1 c1;
SUBC> fits c10;
SUBC> resi c11.
```

In this case, c3 is the criterion salary variable and c1 is the size predictor variable, and c10 is where I have put the predicted or estimated salary values (using size as the predictor) and c11 is where the residuals or errors are after having used the predictor.

The regression equation is: salary = 6.05 + 0.511 size

Inst	salary	preYsize	errYsize
1	15	14.7422	0.25779
2	14	11.6754	2.32459
3	12	15.7645	-3.76447
4	12	15.2533	-3.25334
5	17	15.2533	1.74666
6	23	20.8758	2.12420
7	14	15.7645	-1.76447
8	19	14.7422	4.25779
9	12	14.2311	-2.23107
10	13	12.6977	0.30233

NOTE: The average absolute error = 2.203

What is the variance of the new errors, where size as X1 is used?

```
MTB > let k3=stdev(c11)**2
MTB > prin k3
```

```
K3    7.12222
```

The actual criterion variance is 12.9889, and the error variance just using the mean as the best predictor was also 12.9889, but, when we use size as a predictor, then the error variance drops to 7.12222.

Before, NONE of the variance was predictable, all was error. Now, using size as a predictor, only 7.12222/12.9889 = 0.548331 or, only about 55% of the criterion variance can be considered error ... which means that we have accounted for or 'explained' about 45% of the criterion variance.

THUS, USING SIZE AS A PREDICTOR, WE HAVE ABOUT 45% EXPLAINED OR PREDICTABLE VARIANCE AND ABOUT 55% ERROR OR UNEXPLAINED VARIANCE. THERE HAS BEEN AN ERROR

REDUCTION OF ABOUT 45%. (From 100% error when using only the mean value.)

Remember, that the square of the correlation coefficient is equal to the proportion of the criterion variance that is predictable or explainable ... and in this case that is: .672 squared = .45 or about 45%.

MODEL OF TWO PREDICTORS

Finally, what if we not only used size as a predictor of salary but, we also add prestige. This gives us a multiple regression situation where TWO predictors are being used (X1 and X2), rather than only 1. The issue here is: will using both size and prestige as predictors, even lower the amount of error variance more or, to put it another way, will it INCREASE the amount of predictable variance? Let's see.

```
MTB > regr c3 2 c1 c2;
SUBC> fits c12;
SUBC> resi c13.
```

In the Minitab command, c3 again is the salary criterion variable, c1 and c2 are the size (X1)and prestige (X2) predictors respectively.

The regression equation is
salary = 4.47 + 0.456 size + 0.126 prestige

If we now look at the error or residual values when we have used two predictors, we get (c12 = predicted values; c13 = error values):

```
MTB > name c12='preYsipr' c13='errYsizpr'
MTB > prin c3 c12 c13
```

Inst	salary	preYsizpr	errYsizpr
1	15	13.9777	1.02233
2	14	12.6240	1.37603
3	12	16.5209	-4.52088
4	12	15.4376	-3.43755
5	17	15.4376	1.56245
6	23	20.9527	2.04731
7	14	15.2657	-1.26569
8	19	15.3584	3.64163
9	12	13.5219	-1.52193
10	13	11.9037	1.09630

NOTE: Average absolute error = 2.149

What is the error or residual/error or unexplained variance now?

```
MTB > let k30=stdev(c13)**2
MTB > prin k30
```

K30 6.68978

It looks like adding the second independent or predictor variable of prestige (X2) has lowered the error variance even more ... though clearly not as much as compared to when we used the first predictor, size (X1).

Now we have 6.68978/12.9889 = percent of variance that is error = 51.5%. But, if error is about 51.5%, then the predictable or explainable must be about 48.5% or as a proportion, .485.

SUMMARY

When we used only the mean of the criterion as the model, ie, the simplest model we could think of, we found that all 12.9889 variance of the criterion was error (100%). But, when we used size (X1)as a predictor, we REDUCED the error variance from 12.9889 to 7.12222 or about 55% error (a 45% reduction in error from the simplest model). Finally, when we added a second predictor prestige (X2) to the model, we reduced the error a bit further to about 51.5% (about a 3.5% extra amount of error reduction compared to only using the single size predictor model). Also notice that the average of the actual size of the errors went down too ... from 2.74 (mean model) to 2.149 (2 predictor model).

Model	Crit. Var.	Error Var.	%Error	%Pred	%ErrReduction	Aver. Error
Mean Only	12.9889	12.9889	100%	0%	0%	2.740
Size	12.9889	7.12222	55%	45%	45%	2.203
Size+ Prestige	12.9889	6.68978	51.5%	48.5%	3.5% more (TOTAL=48.5%)	2.149

In the literature, when you hear terms like error variance or unexplained variance or Unaccounted for variance, what it refers to is how much of the criterion variance CANNOT be predicted or estimated USING SOME regression MODEL. Using different models (from simple to more complex) will allow us to estimate the criterion with lesser or greater accuracy (ie, more or less error variance or unexplained variance).

The reason why the model using both size and prestige did not reduce error as much as when we just used size is partly because the two predictors correlate with each other AND, the fact that size correlated more highly with salary in the first place. So, there is not as much ‘extra’ predictability to be gained (though there is some) by adding this second predictor in this case.