STATISTICS TOPIC NOTES

The Explanatory Power of the Regression Model

In a broad sense, the *explanatory power* of a regression model refers to the ability of the predictor variables (the X's) to provide a good explanation of the response variable (the Y). In this Topic Note, I present two metrics for quantitatively assessing the explanatory power of a regression. I also explain how these two metrics can be interpreted. The two metrics:

- **root mean square error** (**RMSE**), also called the standard deviation of the residuals and the "Standard Error of the Estimate" (by StatTools)
- **R-square**

These two metrics are printed in the output of all regression software, including Excel and StatTools.

In this Topic Note, I assume that we have a valid regression. This means that the four assumptions (L,H,I,N) are satisfied. If any of the four assumptions is not satisfied, then we do not have a valid regression model, and then RMSE and R-square may not be meaningful metrics.

Let us recall the useful statistical modeling truism from earlier Topic Notes:

Actual = Estimate + Residual

This means that the actual observed Y value equals the value that the regression equation says it "should" be, plus a leftover amount that balances the equation. The leftover amount – the Residual – is the deviation between Actual and Estimate. Consider an illustrative case of two predictor variables. Then this truism becomes:

$$\boldsymbol{Y = a + b_1 X_1 + b_2 X_2 + e}$$

For example, Y could be the rent of an apartment, X_1 could be the area of the apartment, and X_2 could be the number of bathrooms in the apartment. Then the truism says that the actual rent of an apartment (Y) equals the rent as estimated by the regression model ($a + b_1 X_1 + b_2 X_2$) plus a positive or negative leftover amount (Residual, deviation, or e) that is inserted to force the two sides of the equation to be equal.

> Example 1. In the Austin apartment dataset, the first apartment has actual rent of Y = \$519. The multiple regression estimate is 143.67 + 0.3875×*Area* + 89.93×*Bathrooms*. The first apartment has 725 square feet and 1 bathroom. The multiple regression estimates its rent to be 143.67 + 0.3875×725 + 89.93×1 = 514.51. Therefore, the residual = Y – Estimate = 519 – 514.51 = 4.49. So the truism for that apartment becomes \$519 = \$514.51 + \$4.49.

When we ask how well the regression model explains Y, we are asking how well the Estimate ($a + b_1 X_1 + b_2 X_2$) explains Y. The truism **Actual = Estimate + Residual** suggests two natural ways to answer this question – focusing on how small the Residuals are or focusing on how strongly the Estimates correlate with the Actuals, as I am about to explain.

As a specific example to illustrate ideas, Table 1 (below) shows the multiple linear regression for the Austin apartment dataset. The rents of the 60 apartments are Y. They are regressed on the area and number of bathrooms. This output also appears in the Topic Note on Multiple Linear Regression. In the output, the RMSE is indicated by the red arrow #2. StatTools calls RMSE by an alternative name, "StErr of Estimate." The R-square is indicated by the red arrow #3.

Table 1. Multiple linear regression output for 60 Austin apartments

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.8948	0.8007	0.7937	63.83
	4	3		2

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	2	932,883	466,441	114.4998	< 0.0001
Unexplained	57	232,203	4,074		
	12	11	13	14	15

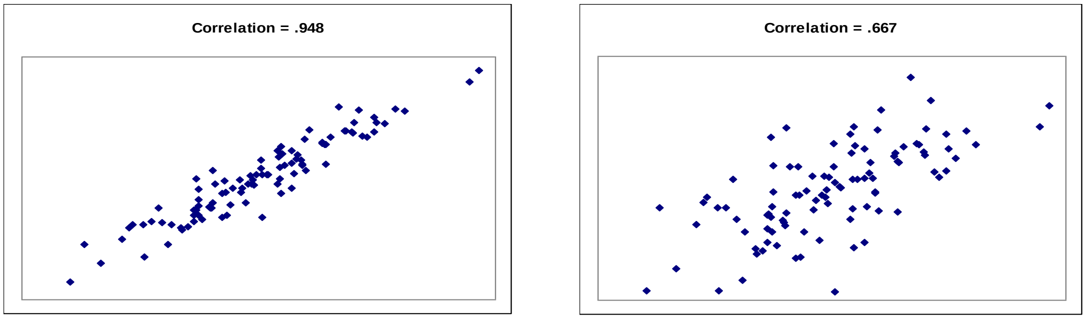
Regression Tabl	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	143.6693	29.5135	4.8679	< 0.0001	84.5696	202.7689
Area	0.3875	0.0498	7.7773	< 0.0001	0.2877	0.4872
Bathrooms	89.9290	27.7507	3.2406	0.0020	34.3592	145.4989
	1	5	6	7	8	

RMSE

For the first metric, we focus on the residuals. The smaller the residuals, collectively, the closer the Estimates approximate the *Y*'s and the higher the explanatory power. In the extreme case, if the residuals are all zero, then each Estimate must equal its actual *Y*, and the regression model fits the *Y*'s perfectly. Conversely, the bigger the residuals, collectively, the further the Estimates are from the *Y*'s and the less well the regression model fits the *Y*'s.

This line of thinking suggests that we base a measure of explanatory power upon the collective size of the residuals. We want a collective metric instead of a metric based on one or a few residuals because we want to encourage getting the overall fit correct. If we pay attention only to a select few of the residuals, then we might distort the model to cater to those residuals and make them small, at the cost of allowing other residuals to bloat.

Figure 1. Two scatterplots with the same regression line.



Panel A.

Panel B.

Figure 1 shows two scatterplots that have the same regression equation. However, the points in Panel A collectively are clearly closer to the regression line (not shown) than are the points in Panel B. The residuals collectively are smaller in A than in B. X does a better job of explaining Y in Panel A than in Panel B.

An initial try at a metric might be the mean of the residuals. However, many of the residuals are negative. The negative residuals exactly balance the positive residuals, which makes the mean residual always equal to zero. So the mean residual would not be a good metric.

A much improved try is to take the mean of the *magnitudes* (absolute values) of the residuals. The formula for this would be $\frac{\sum_{i=1}^{n} |\,residual_i\,|}{n}$. This is a very natural metric. But for reasons that may be mysterious at the level of this course (and that you are not responsible for), statisticians prefer to square the residuals instead of taking their absolute values. This preference parallels statisticians' preference for squaring deviations from the mean in calculating the ordinary standard deviation. Then average the squared residuals. (However, divide by $n - (p + 1)$, where n is the sample size and p is the number of predictor variables – again for mysterious reasons.) Finally, take the square root. Taking the square root puts the metric in the same units as Y. The final result is the metric accurately called the **Root Mean Square Error**:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (residual_i)^2}{n-(p+1)}}$$

RMSE is computed by every statistical computer program that does regression. In StatTools, RMSE is called "Standard Error of the Estimate". (See red arrow #2 in Table 1.) In Excel, it is called simply "Standard Error". Both the StatTools and Excel names are ill-chosen, for the incorporation of the term "standard error" in their names suggests that they measure variability in a Population 3 sense. They do not. RMSE is a Population 1 metric of (individual) variability. That is why I prefer the more accurately descriptive phrase "root mean square error".

In the above formula, notice that if RMSE = 0, then all of the residuals must be zero, so the Estimates coincide with the Y's and the regression model offers a perfect explanation of Y. Moreover, the larger the RMSE, then the bigger the residuals must be, collectively, and the poorer the explanation of Y.

Interpretation of RMSE: The RMSE may be interpreted as the average error that the regression model makes when estimating the Actual values of the Y's in the data. For example, for the 60 Austin apartments, the RMSE for the regression of rent (Y) on area (X_1) and bathrooms (X_2) is $63.83. This number speaks to knowledgeable people in readily understandable terms. It says that if you use this two-predictor model to estimate individual Austin apartment rents, your estimates will be off by about ±$63.83, on average. The average – or *standard* – deviation between regression-estimated rent and actual rent is about ±$63.83. From RMSE, you can tell immediately how accurately the model estimates rent for individual apartments.

Since the Estimate is the best explanation of Y that the predictors can offer, then the predictors do not explain any of the residual. (If the predictors did explain any of the residual, then the Estimate would not be the best explanation of Y that the predictors can offer.) Thus, factors other than the predictors must account for the residual. In the case of the regression of apartment rent on area and bathrooms, \$63.83 is the average portion of rent that is *un*accounted for by area and bathrooms. Some apartments have larger and others have smaller unaccounted for portions. The apartment in Example 1 above has an unaccounted for portion of only \$4.49, which is a much smaller unaccounted for portion than average. \$63.83 of the rent of the typical apartment is accounted for by factors other than area and bathrooms; all of the rest of the rent is explained by area and bathrooms.[1] \$63.83 is a measure of the average uncertainty about the value of rent that remains after taking area and bathrooms into account.

RMSE can also be interpreted as the standard deviation of the residuals. The reason is that when you compute the standard deviation of the residuals using the formula that defines standard deviation, you compute

$$\sqrt{\frac{\sum_{i=1}^{n}(residual_i - mean\ residual)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(residual_i - 0)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(residual_i)^2}{n-1}}$$

In going from the left-hand expression to the middle expression, recall that the average residual is zero. But the right-hand expression is just RMSE after you replace $n - 1$ by $n - (p + 1)$. Because of this, it is customary to *define* the standard deviation of the residuals to be RMSE, replacing the division by $n - 1$ with division by $n - (p + 1)$. So the average size deviation of the residuals from zero – i.e., their standard deviation – is just the average size of the residuals. The replacement of $n - 1$ by $n - (p + 1)$ makes little difference if n is large.

R-SQUARE

For the second metric of explanatory power, we focus on the Estimates. The higher the correlation between the Estimates and the *Y*'s, the more closely their relationship approximates a straight line. In one extreme case, if the correlation between the estimates and the *Y*'s is 1.00, then the estimates and the *Y*'s coincide on a straight line, and the regression model fits the *Y*'s perfectly.[2] In another extreme case, if the correlation between the estimates and the *Y*'s is 0, then there is no relationship between the estimates and their target *Y*'s, and the regression predictors fail to explain *Y* at all. This line of thinking suggests that we can base a metric of explanatory power upon the correlation between the estimates and the *Y*'s.

Figure 1 (above) shows two scatterplots that have the same regression equation. However, the correlation between X and Y is clearly higher in Panel A than in Panel B. X does a better job of explaining Y in Panel A than in Panel B.

An initial try at a metric might be simply the correlation between the estimates and the *Y*'s. Many software regression programs calculate this correlation. It is called **multiple R** or the **multiple correlation coefficient**. Both StatTools and Excel calculate and print out this number. (See red arrow #4 in Table 1.) This number is adequate for the intended purpose. But there is an even better metric. **R-square** is the square of the correlation between the estimates and the *Y*'s. That is, R-square is the square of multiple R. All software regression programs compute and print out R-square. (See red arrow #3 in Table 1.) Because R-square is a simple function of the multiple R,

[1] And by the intercept.

[2] Likewise if the correlation is -1.00.

R-square captures all of the relationship between the estimates and the *Y*'s that the multiple R does. But R-square has three additional advantages:

1. Since R-square is the square of multiple R, the scale of R-square runs from 0 to 1 instead of from -1 to 1. Thus, the two opposite extremes of no explanatory power (0) and total explanatory power (1) are also at opposite extremes of the measurement scale. However, this is a relatively minor advantage.
2. As you have seen in my illustrative scatterplots of correlations of various sizes,[3] as the correlation declines from 1, there is an initial, rapid spreading out of the points around the regression line. For example, a plot with a correlation of 0.90 shows a lot more spread than you would think that such a plot should have with a correlation only 10% reduced from perfection. Squaring the correlation helps to give a more accurate picture of this rapid spreading. For example, the correlation of 0.90 becomes an R-square of 0.90 × 0.90 = only 0.81.
3. Most importantly, R-square can be interpreted as the *proportion* of the *Y* that is explained by the predictors. Thus, for the illustrative regression of apartment rent on area and bathrooms, the R-square is 0.8007. (See red arrow #3 in Table 1.) We interpret this to say that 80% of rent is explained by area and bathrooms together. What allows us to make this interpretation? The full reason is mathematical, and you are not responsible for it. But I can sketch the ideas that are involved in an aside, which follows. If you are willing to accept this interpretation, you may skip the following lengthy aside without harm.

Aside: (Optional) Justification for interpreting R-square as the proportion of *Y* explained by the model predictors.

The main idea is to divide the part of rent explained by all predictors into two sub-parts: the part explained by the predictors included in the model and the part explained by all other predictors not included in the model. This idea is captured in the following mathematical identities:

$$\text{(i) } y = \bar{y} + (y - \bar{y}) \text{ and}$$

$$\text{(ii) } y - \bar{y} = (y - estimate) + (estimate - \bar{y})$$

If we have no predictors at all and have only the *y* values on which to base an estimate, the best we can do is to estimate each observation to be the mean $\bar{y}$, as in (i). The part of *y* that is leftover and not explained by $\bar{y}$ is the residual deviation ($y - \bar{y}$) on the right-hand side of (i). What accounts for this deviation? Why is the mean estimate wrong by this much? Since we have not yet used any predictor variables, the deviation must be because of *all possible* predictor variables. Therefore, the deviation $y - \bar{y}$ is the part of *y* that cannot be explained by the mean $\bar{y}$ and is therefore the part of *y* that is explained by *all possible* predictor variables.[4]

Let us now suppose that we have predictor variables *area* and *bathrooms*. Identity (ii) shows how to measure the improvement in explanation of rent that results from using area and bathrooms. Identity (ii) splits the part of *y* explained by all possible predictor variables (i.e.,

[3] See the Topic Note on Covariance and Correlation.

[4] And I do mean *all possible* predictors, not just those that we have collected and measured. This is based upon the reasonable idea that there is, in principle, a complete explanation for why the rent of an apartment is the value *y* instead of something else, even if the complete explanation includes not only measured variables like area, bathrooms, age, location, etc. but also unmeasured variables like the color of paint and personal psychological quirks of the landlord, for example.

$y-\bar{y}$) into the part explained by area and bathrooms and the part explained by all *other* possible predictor variables: The second term $(estimate-\bar{y})$ on the right-hand side shows how much we change our estimate of *y* when we use the model's predictor variables to estimate *y* instead of just using the mean $\bar{y}$. Therefore, $(estimate-\bar{y})$ measures the improvement in our estimate that results from using the model predictors area and bathrooms.

The first term on the right side of (ii) is $(y-estimate)$. Since *estimate* is the best explanation of *y* that can be offered by the predictor variables used in the model, then $(y-estimate)$ is the remaining part of $y-\bar{y}$ that is *un*explained by the predictor variables that are used and is therefore the part of $y-\bar{y}$ that is explained by all <u>other</u> possible predictor variables.

In summary, $(estimate-\bar{y})$ is the part of *Y* that is explained by the regression predictors; $(y-estimate)$ is the part of *Y* that is *un*explained by the regression predictors.

The above discussion is quite general. To be more specific, consider the regression of Austin apartment rent on area and bathrooms. The mean rent is \$572.27. The first apartment actually rents for \$519. Therefore, for the first apartment $y-\bar{y}$ = \$519 - \$572.27 = -\$53.27 is the part of its rent potentially explainable by predictor variables. The first apartment has 725 square feet and 1 bathroom. The multiple regression estimates its rent to be 143.67 + 0.3875×725 + 89.93×1 = 514.51. Therefore, $(y-estimate)$ = 519 – 514.51 = 4.49 remains *un*explained by area and bathrooms; and $(estimate-\bar{y})$ = 514.51 – 572.27 = -57.76 is the improvement in the estimate that results from using the predictors area and bathrooms, compared with using just the mean 572.27. The estimate using area and bathrooms is much closer to the actual rent than is the sample mean rent. So from

$$y-\bar{y}=(y-estimate)+(estimate-\bar{y})$$

we have

-\$53.27 = \$4.49 + -\$57.76

-\$57.76 is the part of the first apartment's rent that is explained by its area and bathrooms;[5] \$4.49 is the part of the first apartment's rent that is *un*explained by its area and bathrooms.[6] A similar breakdown can be made for each of the 60 apartments.

To get a collective metric for all sampled apartments, we start by squaring the identity because some parts of it are negative for some apartments:

$$(y-\bar{y})^2=[(y-estimate)+(estimate-\bar{y})]^2$$

Then we add up the squared identity for all *n*=60 apartments:

$$\sum_{i=1}^{n}(y-\bar{y})^2=\sum_{i=1}^{n}[(y-estimate)+(estimate-\bar{y})]^2$$

Then we expand the square on the right-hand side:

$$\sum_{i=1}^{n}(y-\bar{y})^2=\sum_{i=1}^{n}[(y-estimate)+(estimate-\bar{y})]^2 =$$

$$\sum_{i=1}^{n}(y-estimate)^2+\sum_{i=1}^{n}(estimate-\bar{y})^2+2\sum_{i=1}^{n}(y-estimate)(estimate-\bar{y})$$

[5] This says that we should lower the rent estimate for apartment 1 by \$57.76 below the average apartment to account for apartment 1 having smaller area and bathroom count than average.

[6] This says that all potential predictors other than area and bathrooms collectively add a small \$4.49 to the rent of apartment 1.

By mathematical computation that may seem like magic, the cross-product term $2\sum_{i=1}^{n}(y-estimate)(estimate-\bar{y})$ can be shown to equal 0. So we have the collective form of the identity:

$$\sum_{i=1}^{n}(y-\bar{y})^2 = \sum_{i=1}^{n}(y-estimate)^2 + \sum_{i=1}^{n}(estimate-\bar{y})^2$$

which we can interpret as

Total variability of *Y* potentially explainable by the model *X*'s	=	Variability *un*explained by the model *X*'s	+	Variability explained by the model *X*'s

The collective parts assume the character of their individual components:

(part potentially explainable) = (part *un*explained) + (part explained).

The regression output calculates and prints these collective explained and unexplained parts in the "ANOVA Table".[7] For example, in the regression of rent on area and bathrooms, the unexplained sum of squares = 232,203; the explained sum of squares = 932,883. (See red arrow #11 in Table 1.) So the total = 232,203 + 932,883 = 1,165,086. Thus, the collective form of the identity is:

$$\sum_{i=1}^{60}(y-\bar{y})^2 = \sum_{i=1}^{60}(y-estimate)^2 + \sum_{i=1}^{60}(estimate-\bar{y})^2$$

$$1{,}165{,}086 = 232{,}203 + 932{,}883$$

Total variability of rent potentially explainable by area and bathrooms	=	Variability *un*explained by area and bathrooms	+	Variability explained by area and bathrooms

Now, finally, to the punchline: Out of the 1,165,086 units of variability potentially explainable, 932,883 are, in fact, explained by area and bathrooms. So the proportion of *potentially* explainable variability units that are *in fact* explained by the regression model is 932,883 / 1,165,086 = 0.8007, which is R-square.

In general,

$$\text{R-square} = \frac{\sum_{i=1}^{n}(estimate-\bar{y})^2}{\sum_{i=1}^{n}(y-\bar{y})^2}$$

This is the justification for the interpretation of R-square as the proportion of *Y* explained by the model *X*'s.

One final additional quibble: You may have noticed that the formula for R-square (immediately above) has the numerator of the variance of *Y* $\left(\sum_{i=1}^{n}(y-\bar{y})^2\right)$ in the denominator on the right-hand side. So interpreting R-square as the "proportion of *Y* explained by the model *X*'s" is not quite accurate. More precisely, we should interpret R-square as the "proportion *of the variability* of *Y* explained by the model *X*'s". But this is actually what we want: The more *Y* wiggles when *X*

[7] ANOVA = ANalysis Of VAriance

wiggles, the more *X* affects *Y*. If *Y* does not wiggle when *X* wiggles, then *X* does not affect *Y*. So *X* explains *Y* by making *Y* wiggle. No wiggle, no explanatory power.
End aside.

SUMMARY

The major take-aways in this Topic Note:
There are two major metrics to assess the explanatory power of a regression model:

- RMSE can be interpreted as the average error that the model makes in estimating Y from the predictors.
- R-square can be interpreted as the proportion of Y (variability) that is explained by the predictors.

Both metrics are printed out by all regression software, including Excel and StatTools.