NORTHWESTERN UNIVERSITY

TECHNICAL NOTE

DAVID DRANOVE

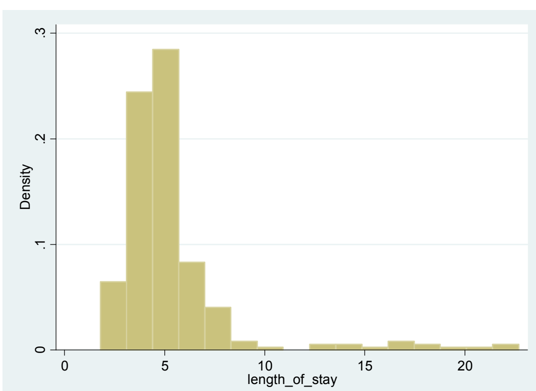
7-112-011

Practical Regression: From "Stylized Facts" to Benchmarking

This is one in a series of notes entitled "Practical Regression." These notes supplement the theoretical content of most statistics texts with practical advice on solving real world empirical problems through regression analysis.

"Stylized Facts"

Many statistical studies begin by presenting what economists call "stylized facts." Stylized facts are high-level summary statistics that raise important questions about the phenomenon under investigation. For example, a stylized fact about American hospitals is that there is considerable variation in the length of stay for inpatient admissions. Here is a histogram of length of stay for California hospitals in 2003:



Excessive lengths of stay contribute to higher costs, so hospital owners and board members are often dismayed when they learn of their facility's excessively high length of stay. Some may blame hospital administrators for the apparent inefficiencies. Administrators, on the other hand, may counter that the high length of stay reflects the types of patients served, the level of technology offered, or some other factor that the owners would not or could not change. The stylized facts about hospitals have clearly raised important questions. To answer them, we need to carefully examine the data.

Digging into the Data

Sutter Medical Center in Sacramento has an average length of stay of 8.5 days, which is 3.19 days above the state average of 5.31 days. (Let's remember this 3.19-day "gap." It comes up later in our analysis.) Does this 3.19-day gap reflect inefficiency at Sutter? Sutter's administrators might make three arguments in its defense:

- It treats patients with more severe illnesses.
- It treats older patients.
- It treats lower-income patients.

All of these factors may contribute to Sutter's above-average length of stay, but are they enough to explain the 3.19-day excess?

It turns out that we can frame this question using regression. In doing so, we have tools we can use to provide the answer. Let's begin by running the simplest possible model, a regression of stay on a constant, with no other right hand side (RHS) variables:

```
. regress length

      Source |       SS       df       MS              Number of obs =     285
-------------+------------------------------           F(  0,   284) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  2704.59507   284  9.52322206           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  2704.59507   284  9.52322206           Root MSE      =   3.086

------------------------------------------------------------------------------
length_of_~y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   5.318863   .1827972    29.10   0.000     4.959054    5.678672
------------------------------------------------------------------------------
```

Notice that the constant is just the mean length of stay. (Do you recall why this is true?) If we were to compute the residual for Sutter, it would equal 3.19.

Sutter administrators claim that its length of stay can be explained by the fact that its patients are sicker, older, and poorer than average. We can verify this with our available data, which contains an index of patient severity ("caseweight"), the percentage of patients over age 65, and the median incomes of the zip codes in which each hospital is located. (The latter is the best income measure available to us.)

Here are summary statistics for these variables:

```
. su length_of_stay caseweight pctaged medianincome
```

Variable	Obs	Mean	Std. Dev.	Min	Max
length_of_~y	285	5.318863	3.085972	1.787966	22.70664
caseweight	285	.94791	.2611971	.1327895	2.188695
pctaged	285	.1214696	.0644745	.0071883	.5474817
medianincome	285	40025.93	15685.52	15578	117369

The corresponding means for Sutter Hospital (identified as hospital number 341051) are:

```
. su length_of_stay caseweight pctaged medianincome if hospid==341051
```

Variable	Obs	Mean	Std. Dev.	Min	Max
length_of_~y	1	8.495247	.	8.495247	8.495247
caseweight	1	1.544816	.	1.544816	1.544816
pctaged	1	.1518325	.	.1518325	.1518325
medianincome	1	32020	.	32020	32020

As you can see, Sutter's patients really are sicker, older, and poorer than average. But how strong is Sutter's case? How important are caseweight, age, and income? Are the differences between Sutter and other hospitals significant enough to explain the large discrepancy in length of stay?

By using regression, we can estimate whether and by how much these three variables predict length of stay. We can then predict what the length of stay should be for a hospital with Sutter's particular caseweight, age, and income. If Sutter's length of stay exceeds this predicted amount, then the hospital's administrators still have some explaining to do. Let's see how this analysis is done.

First, the regression:

```
. regress los caseweight pctaged medianincome
```

Source	SS	df	MS
Model	169.426191	3	56.475397
Residual	2535.16889	281	9.02195336
Total	2704.59509	284	9.52322213

Number of obs =	285
F(3, 281) =	6.26
Prob > F =	0.0004
R-squared =	0.0626
Adj R-squared =	0.0526
Root MSE =	3.0037

los	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
caseweight	1.861822	.7093999	2.62	0.009	.4654098	3.258235
pctaged	7.324587	2.887605	2.54	0.012	1.640503	13.00867
medianincome	-.0000219	.0000116	-1.89	0.059	-.0000447	8.79e-07
_cons	3.541288	.7768097	4.56	0.000	2.012183	5.070393

Now, let's compute the predicted length of stay for Sutter:

```
. predict predictedlos
(option xb assumed; fitted values)

. list predicted los if hospid==341051

     +---------------------+
     | predic~s        los |
     |---------------------|
178. | 6.828003    8.49525 |
     +---------------------+
```

The predicted stay of 6.82 days is 1.51 days more than the statewide average of 5.31 days, but well below Sutter's 8.5-day length of stay. Still, we have explained half of the 3.19-day gap with just three predictors. If we had more predictors (e.g., measures of the service mix, quality of care, additional patient demographics) we might close the gap even further.

At this point, it is useful to break down how the three predictors explain 1.51 days out of the 3.19-day gap.

- *Caseweight*: Sutter's caseweight exceeds the average by 0.5959. The coefficient on caseweight is 1.8618. Multiplied together, this implies that based only on caseweight, Sutter's length of stay should exceed the average by 1.1094 days.
- *Pctaged*: The percentage of Sutter patients who are aged exceeds the average by 0.0303. The coefficient is 7.325. Multiplied together, this implies that based only on pctaged, Sutter's length of stay should exceed the average by 0.2219 days.
- *Median income:* Sutter's median income "exceeds" the average by -8,006. The coefficient is -0.0000219. Multiplied together, this implies that based only on median income, Sutter's length of stay should exceed the average by 0.1753 days.

Adding up these numbers, we get an excess of 1.1094 + 0.2219 + 0.1753 = 1.507 days, or 1.51 if we round up. A bit of algebra reveals that the numbers have to add up in this way.

Recall that $Y_i = \sum BX_i + \varepsilon_i$. Also note that $Y^m = \sum BX^m$ (where ^{m} refers to the mean).

We can thus write $Y_i - Y^m = \sum BX_i + \varepsilon_i - \sum BX^m = \sum B(X_i - X^m) + \varepsilon_i$.

In other words, the difference between a given firm i's performance and the mean can be broken down into two components: (1) the difference between i and the average firm on the covariates, multiplied by the corresponding B's, and (2) the error term.

This suggests the following table:

Predictor	X_i	X^m	$X_i - X^m$	B	$B(X_i - X^m)$
Caseweight	1.5448	0.9479	0.5959	1.8618	1.1094
Pctaged	0.1519	0.1214	0.0303	7.325	0.2219
Median income	32,020	40,026	-8,006	-0.0000219	0.1753
Total	—	—	—	—	1.51

We call this a *deficiency table* because it breaks down the firms' "deficiency" in the raw data into components explained by regression.

Note: The sum will equal the overall shortfall in performance due to the predictors. This does not equal the overall shortfall, because there is always a residual.[1] (We will go into this further in a moment.)

Applying the Principles: Benchmarking

"Benchmarking" refers to a comparison of performance across entities (i.e., individuals, organizations, firms). Often, the goal is to identify "best practices" to implement.

One approach to benchmarking is to identify good performers and to list what they do well; this list is then considered best practices. This is not a good idea, however, for two reasons:

- Bad firms may undertake the same activities as good ones.
- It is unclear which activities are associated with higher profits.

Regression can help you identify which factors are most strongly associated with performance. In strategy parlance, these predictors can be considered *key success factors* (KSFs). The residual can be interpreted as unmeasured managerial competence, as we discuss below. As with all regression, causality is a concern. When regression is used for benchmarking, causality becomes an even bigger issue, as discussed below.

There are four key steps to performing a benchmarking analysis:

1. Identify your performance measure and the key factors that contribute most to performance.
2. Select an appropriate sample of entities to include in your analysis.
3. Estimate your regression model.
4. Create a deficiency table to assess the performance of the firm in question.

Step 1: Determinants of Performance

The first thing you need to do is decide how you will measure performance. Whatever measure you choose, it will be the dependent variable in your regression model. As such, it is best to choose a measure that has some action your model can explain.

---

[1] If you use a log left hand side (LHS) variable, you should first complete the table without transforming anything to make sure the total equals your deficiency score minus the residual. Next, calculate $[\exp(B(X_i - X^m)) - 1] * 100\%$ for each row. These will *not* sum to $[\exp(\sum(B(X_i - X^m)) - 1)] * 100\%$ (because exponentiating a sum of values is not the same as summing up the individual exponentiated values). However, this calculation will give you a sense of each factor's contribution to the overall deficiency score.

Some examples include:

- *Retail*: Operating margins per store; employee turnover per store
- *Manufacturing*: Number of hours to manufacture a given part; share of parts that fail specification requirements
- *Services*: Fraction of customers who return for service; fraction of customers whose problems are resolved in a given amount of time

Having identified your performance measure, you can now start making a list of the key factors that affect performance (i.e., the KSFs). Try to divide these into factors you believe are not under management's control or that the owners would not want to change (so-called *exogenous* factors), and those that are under management's control. Exogenous factors might include market characteristics or the political environment. In the hospital example, all three predictors are exogenous.

Factors under managerial control might include staffing levels, product selection, or organizational structure; these are the kinds of measures that consulting firms routinely study. If the coefficient on any of these variables is positive, then it is positively *associated* with performance. Be careful when interpreting these coefficients, however. Because they are under management control, they are no longer "excuses" for poor performance the way that caseweight, age, and income "explained" the long length of stay at Sutton Hospital.

Both types of variables can help you understand a firm's performance. For example, in the case of retail stores, you may view the location of a given store as an exogenous determinant of performance. (Of course, in the long run, such a decision is clearly under management's control, but in the short run at least it is fixed.) Suppose further that you discover that mall locations have much higher operating margins. This information may help you to decide that future stores should be located in malls, or that non-mall-based stores should be closed if they aren't exceeding the opportunity cost of your resources.

The logic for a factor under managerial control is similar. Suppose you discover that performance tends to be lower when managers have college degrees. This suggests you should conduct a more thorough exploration of hiring practices. Perhaps, on average, college-educated managers are not interested in staying in the long term, so they are less motivated. Perhaps a college education is correlated with some omitted factor that is driving the negative relationship between college education and performance (e.g., the college-educated managers aren't "in touch" with the high-school students buying the clothing sold in the stores). In any case, it is something to think further about. We will discuss predictor variables again in Step 3.

Step 2: Identifying Your Analysis Sample

There are two key concerns when deciding what entities to include in your analysis sample: (1) ensuring all entities operate similarly, and (2) ensuring there is enough variation in the KSFs of interest. These two objectives have a natural tension between them: you want to keep the sample of entities fairly similar, but you need enough data (and enough action in the KSFs within your sample) to be able to estimate the effects of all your KSFs.

You want to ensure entities operate similarly because you want to be comparing apples with apples. For example, comparing the performance of mom-and-pop stores with the performance of "big box" stores such as Walmart and Kmart is unlikely to help you figure out the best practices for mom-and-pop stores. It is possible to "dummy out" certain differences (e.g., size of store within a large sample of mom-and-pop stores of various sizes), but you still have to believe the basic operation is similar. You can include slope dummies to allow certain coefficients to vary by store type (e.g., a different coefficient on number of SKUs per square foot for large and medium stores relative to small stores), but if you have too many interaction terms then you're not actually comparing apples with apples.

To put it slightly differently, by using regression you are assuming that the underlying model of performance for the comparison sample is the same as the model for the entity (or entities) of interest. *Thus, you believe that the β's are the same for all the observations in your sample.*

There is an easy two-step test to check whether the β's are the same for all observations:

1. Start with a sample of entities that you are quite certain belong in the comparison group. Run a regression $P = B_0 + \underline{B_x X}$. Call this "Model 1." The coefficients $B_{\underline{x}}$ in Model 1 tell you how the various RHS variables contribute to the performance of the entities in your sample.

   Now suppose you are considering adding a group of entities to the sample. The main reason to add them is to gain precision in your estimates. However, if the model explaining the performance of the added observations is different from Model 1, then adding these observations will "ruin" your regression; your coefficients will no longer apply to the entities you are most interested in studying. Thus, you need to see if the effects of the RHS variables are different for the new sample versus the old sample. To do this, you will need to put in slope dummies.

2. Let *New* be a dummy variable indicating whether the observations are in the new sample. Interact *New* with all of your X variables. (If this is too cumbersome, interact with the "most important" X variables—you make the call.)

   Regress $P = B_0 + \underline{B_x X} + B_1 New + \underline{B_{2x}}(\underline{X} \cdot New) + \ldots$

   Perform a partial F-test on the slope dummies $\underline{B_{2x}}$. If they are not statistically significant, then you can add the new observations to the sample. If some but not all are significant, then you can use your judgment (but be sure to add the significant slope dummies to your model).

Here are some important considerations as you build your benchmarking model:

- The regression sample should include a wide range of entities, not just the most successful ones.
- Predictors should include a range of firm activities, resources, and capabilities as well as market characteristics that affect performance.
- There should be variation in the presence and absence of these predictors so that the regression can identify those that most contribute to performance.

Step 3: Estimate Your Regression Model

This should be a piece of cake for you. It is a good idea to run a regression that only uses exogenous predictors before adding the endogenous predictors; this will aid your interpretation.

Step 4: Create and Interpret a Deficiency Table

Creating a deficiency table should also be straightforward. If your regression includes predictors that can be controlled by management, use the results to make recommendations for improving performance. (Reminder: make sure you have a well-identified model.)

Revisiting the Residual

The residual in benchmarking regressions takes on a heightened meaning. The residual contains the unmeasured factors that cause a firm to be deficient, including omitted predictors, as well as pure random chance. There is almost always one predictor that is impossible to measure—managerial competence.

It is tempting to attribute the entire residual to management skill, but this is dangerous. Bear in mind that in any regression, the residual is likely to be highly correlated with the dependent variable. This means that firms whose dependent variable is above average will usually have a positive residual, and firms whose dependent variable is below average will usually have a negative residual. Because the residual includes purely random factors, when you examine the same firm in the next period, some of that residual is likely to "regress to the mean" so that outlier performers may look more like average performers.[2]

Even if you allow for mean regression, you will almost never "explain away" a firm's extremely good or extremely bad performance. Thus, you will almost always chalk up some of the extreme performance to unobservable management skill. It is better to simply think of management skill as one of several possible unmeasured explanations. If "poor management" is a candidate explanation, challenge management to identify other candidates, then try to measure them and improve your predictive model.

[2] Suppose the difference between actual and predicted performance of a firm in period t is e_t. Regression to the mean implies that the difference between actual and predicted performance in period t + 1 will be λe_t, where $0 < \lambda < 1$. Thus, outliers in one period are smaller outliers in the next.