

Interpreting regression models with reciprocals, proportions, percentages, other ratios and log-transformed predictors, and using standardized predictors, illustrated using the SENIC data set

Table 1

Name	Definition	Mean (SD)	Min	Max
???	$100 \times (1/\text{avg length of stay in days})$	10.7 (1.8)	5.1	14.9
age	avg age of patients, yr	53.2 (4.5)	35.8	65.9
risk	pct of patients who acquire infection	4.4 (1.3)	1.3	7.8
xray	$100 \times (\text{no. x-rays})/(\text{no. asymptomatic patients})$	81.6 (19.4)	39.6	133.5
logcen	natural log(avg daily census)	4.9 (0.8)	3.0	6.7
occupancy	$(\text{avg daily census})/(\text{no. beds})$	0.73 (0.11)	0.36	0.98
npratio	$(\text{avg no. nurses})/(\text{avg daily census})$	0.95 (0.32)	0.21	2.8

The data are from the 1975-1976 study period. How do we interpret the variable $100 \times (1/\text{avg length of stay in days})$? Now, average length of stay = $(\text{no. patient-days over study period})/(\text{no. patients discharged over study period})$. So the reciprocal is $(\text{no. patients discharged over study period})/(\text{no. patient-days over study period})$ = average proportion of patients in the hospital on a given day who are discharged. Or we can think of this as the probability that a given patient will be discharged on any given day. Multiply by 100 and we get this as a percent. Now, it doesn't always happen that when we take the reciprocal of a variable, we end up with a variable that had a meaningful interpretation. However, the reciprocal of a variable involving length can often be interpreted as a rate. In this case, it is a proportion, which we can think of as closely related to a rate.

Suppose we fit the following regression model using this as the dependent variable:

Variable	Parameter	Standard	T for H0:	
	Estimate	Error	Parameter=0	Prob > T
INTERCEP	22.349	2.032	11.00	<.0001
age	-0.082	0.027	-3.01	0.0033
risk	-0.400	0.119	-3.36	0.0011
xray	-0.022	0.007	-3.05	0.0029
logcen	-0.615	0.196	-3.14	0.0022
occupancy	-2.209	1.333	-1.66	0.1004
npratio	0.937	0.433	2.16	0.0328

How do we interpret the regression coefficients? Let's consider the easier variables first.

- The intercept is the estimated mean value of the dependent variable when the values of all of the predictors are zero. This is meaningless here, since that would be outside the scope of the data.
- For age, one unit corresponds to one year. A one-year increase in average age of patients is associated with a mean decrease in daily patients discharged of about 0.08%, controlling for

the other variables. That is a rather small amount. It might be better to enter age in decades (10-year increments). A 10-year increase in average age is associated with a mean decrease of about 0.8%, or close to 1%.

- Risk is in units of percent; one unit corresponds to 1%. A 1% increase in the number of patients acquiring infection is associated with a mean decrease in daily patients discharged of 0.4%. A 10% increase in number of patients acquiring infection is associated with a mean decrease of 4%.
- The variable \log_{cen} is the natural log of the average daily census (average number of patients in the hospital). When we use the natural log of a predictor in the model, we get a nice interpretation in terms of the effect of a 1% increase in that predictor. A 1% increase means multiply by 1.01: $\hat{y} = b_0 + b_1 \log(1.01c) = b_0 + b_1 \log(c) + b_1 \log(1.01) \approx b_0 + b_1 \log(c) + b_1 0.01 = b_0 + b_1 \log(c) + b_1/100$, so a 1% increase in the predictor is associated with an additive increase in the mean of y of $b_1/100$. (Note that it is easy to forget the factor of 0.01!! And that the interpretation is different than it is for a $\log(y)$ - $\log(x)$ model.) In this model, a 1% increase in the average daily census is associated with a mean decrease in the daily patients discharged of 0.6%.
- How about the variable xray ? It is defined as $100(\text{no.x-rays})/(\text{no. asymptomatic patients})$. Now, $(\text{no.x-rays})/(\text{no. asymptomatic patients})$ is the average number of x-rays per asymptomatic patient, which takes on an average of 0.816 in this data set. In the data set, the values range from 0.396 x-rays/patient to 1.335 x-rays/patient. An increase of one unit means one more x-ray per patient, which is enormous and covers the entire range of the data. If we multiply by 100, then a one-unit increase becomes an increase in average x-rays/patient of 0.01. This is a much more reasonable increment to work with. Hence the motivation for scaling this variable in this manner, as $100(\text{no.x-rays})/(\text{no. asymptomatic patients})$.

Just for curiosity's sake, what would we get if we used $\text{xray}/100 = (\text{no.x-rays})/(\text{no. asymptomatic patients})$ in the model?

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	22.349	2.032	11.00	<.0001
age	1	-0.082	0.027	-3.01	0.0033
risk	1	-0.400	0.119	-3.36	0.0011
xray100	1	-2.159	0.707	-3.05	0.0029
logcensus	1	-0.615	0.196	-3.14	0.0022
occupancy	1	-2.209	1.333	-1.66	0.1004
npratio	1	0.937	0.433	2.16	0.0328

Note that the t and p-values are the same; this is a linear transformation. For interpretation, we could say that an increase of one more x-ray/patient on average is associated with a decrease in the daily percentage of patients discharged of 2.2%. This model could also be reasonable to present, but an increase of one more x-ray per patient really is enormous, so the smaller increment might be more meaningful.

How about the variable occupancy? Occupancy is the average number of patients per bed, and ranges from 0.36 to 0.98. Since we cannot have more than one patient per bed or negative patients per bed (i.e., we cannot have occupancy outside the range of 0 to 1, or at least it is reasonable to assume so), we can think of this as a proportion of beds that are occupied. So this variable is a proportion. For proportions, it is not meaningful to think of a one-unit increase; this would be the equivalent of going from 0 to 1. It is more meaningful to think in terms of an increase of 0.01 or 0.10, or maybe 0.05. According to the coefficient in the model, a one-unit increase in occupancy is associated with a decrease in the mean percentage of patients daily discharged of 2.2%. It is perhaps more meaningful to think of an increase of 0.10; an increase of 0.10 in the proportion of beds occupied is associated with a decrease of 0.22%, controlling for the other variables in the model. We could speculate why this might happen. Perhaps busier hospitals also tend to have sicker patients who stay longer, or are so busy they don't heal the patients as fast. However, the variable is not significant ($P = 0.10$), so it is better not to draw conclusions.

Now let's look at the npratio, which is the ratio of the number of nurses to the daily number of patients. An increase of one unit in this variable means one additional nurse per patient. So we interpret the regression coefficient as indicating that one additional nurse per patient is associated with an increase in mean patients discharged of almost 1%. So having more nurses per patient is associated with discharging the patients more quickly.

What happens if we center all of the variables by subtracting the mean?

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	10.711	0.119	89.31	<.0001
age_cen	1	-0.082	0.027	-3.01	0.0033
risk_cen	1	-0.400	0.119	-3.36	0.0011
xray_cen	1	-0.021	0.007	-3.05	0.0029
logcensus_cen	1	-0.614	0.196	-3.14	0.0022
occup_cen	1	-2.209	1.333	-1.66	0.1004
npratio_cen	1	0.937	0.433	2.16	0.0328

All of the coefficients and standard errors are exactly the same except for the intercept. But now the intercept has a meaningful interpretation. It is the estimated mean of the dependent variable (average percentage of patients discharged each day) when all of the predictors are zero, i.e., when they are at their average values. Here, we see that when the other predictors are at their average value, the mean percentage of patients discharged per day is 10.7%. (Compare this with the sample mean in Table 1; we are right on target.) This is rather nice, because now we have a “benchmark” of what a typical discharge rate is, and we can get a better idea of how the other variables impact this. We can read this right off the table.

How about if we also divide each predictor by its standard deviation? Then all of the predictors are in units of standard deviations from their means. This can also be very useful, particularly because it puts all of the predictors into a common metric. This makes comparisons of their relative magnitudes easier.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	10.711	0.120	89.31	<.0001
age_s	1	-0.370	0.123	-3.01	0.0033
risk_s	1	-0.520	0.155	-3.36	0.0011
xray_s	1	-0.419	0.137	-3.05	0.0029
logcensus_s	1	-0.491	0.157	-3.14	0.0022
occup_s	1	-0.242	0.147	-1.66	0.1004
npratio_s	1	0.300	0.139	2.16	0.0328

Note that the t and p values have not changed; this is a linear transformation. Also note that the intercept did not change. Now, for each predictor, a one-unit change is a one standard deviation change. This enables us to get new insights. For example, we see that the relative magnitudes of the coefficients do not differ that much. There appears to be no dominant predictor. In a way, the first model (page 1) might be misleading about the relative importance of the predictors; it looked like occupancy had the biggest effect in that model, whereas it has the smallest effect in this standardized model.

If we want to figure out how much the daily discharged percentage changes due to a one-SD increase in average age (for example), we can get this as $(-0.370)(4.5) = -17\%$ (4.5 is the SD of age).

Another remark: you would want to make sure the distributions of the predictors are fairly symmetric before performing this kind of standardization. It does not make sense to standardize a highly skewed variable this way.

Interpreting dummy variables and interactions, illustrated using the CDI data

For illustration purposes, let's work with just a few of the variables from the CDI data set. The unit of analysis is county. We will use the following variables and region (West, North Central, Northeast, South):

Variable	Mean (SD)	Min	Max
crmp1000 (no. serious crimes/1000 people)	56.8 (24.9)	4.6	161.6
natural log of area in sq miles	6.5 (0.9)	2.7	9.9
natural log of population	12.5 (0.8)	11.5	16.0

Let's run a model with only the dummy variables for region:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	60.878	2.479	24.56	<.0001
ne	1	-21.594	3.277	-6.59	<.0001
nc	1	-9.772	3.244	-3.01	0.0027
s	1	9.860	3.043	3.24	0.0013

This model gives us estimates of the mean crime rate per 1000 people in each region. The reference region is the West. The estimated mean crime rate is 61 per 1000 for counties in the West. Other coefficients give an offset from the West. For counties in the Northeast, we have an estimated mean of $60.9 - 21.6 = 39$ crimes per 1000 people. Means for other regions are determined similarly. In this situation, we might think about changing the reference group to be the region with either the lowest crime rate (NE) or the highest (S). Let's make it the lowest (Northeast):

Variable	DF	Estimate	Error	t Value	Pr > t
Intercept	1	39.284	2.144	18.33	<.0001
nc	1	11.822	2.996	3.95	<.0001
s	1	31.454	2.776	11.33	<.0001
w	1	21.594	3.277	6.59	<.0001

Now we easily see that the reference region is the lowest, at a mean of 39 crimes per 1000 people, and the other regions are higher. Also, we see that there are large regional differences, so it would make sense to control for region in our models.

Let's add a continuous variable, $\log(\text{pop})$:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-125.130	14.950	-8.37	<.0001
nc	1	14.860	2.662	5.58	<.0001
s	1	33.929	2.464	13.77	<.0001
w	1	19.844	2.901	6.84	<.0001
logpop	1	13.075	1.179	11.09	<.0001

We see that when controlling for region, a 1% increase in population is associated with a mean increase in the crime rate of about 0.13 per 1000 people. Maybe a 10% increase in population is more meaningful; that would be associated with an increase of 1.3 crimes per 1000 people. But now the intercept is meaningless, because $\log(\text{pop})$ is never zero. Let us center logpop and rerun the model:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	38.304	1.897	20.19	<.0001
nc	1	14.860	2.663	5.58	<.0001
s	1	33.929	2.464	13.77	<.0001
w	1	19.844	2.901	6.84	<.0001
logpop_cen	1	13.074	1.179	11.09	<.0001

Now we can interpret the intercept as the estimated mean crime rate per 1000 for counties in the Northeast, when $\log(\text{pop})$ is at its mean. The estimated mean for counties in the North Central region when $\log(\text{pop})$ is at its mean is $38.3 + 14.9 = 53$ crimes per 1000 people. The coefficient 14.9 gives the difference in mean for North Central vs. Northeast when controlling for $\log(\text{pop})$, i.e., for two counties with the same population.

Let's see an interaction between region and $\log(\text{pop})$:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-104.765	31.274	-3.35	0.0009
nc	1	-23.047	43.269	-0.53	0.5946
s	1	-37.513	41.115	-0.91	0.3621
w	1	61.333	43.417	1.41	0.1585
logpop	1	11.455	2.483	4.61	<.0001
ncXlogpop	1	3.041	3.466	0.88	0.3807
sXlogpop	1	5.743	3.285	1.75	0.0811
wXlogpop	1	-3.248	3.428	-0.95	0.3439

An F-test for the 3 interaction terms gives a p-value of 0.035. But the coefficients are hard to interpret. Let's use the centered version of logpop in the model:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	38.425	1.892	20.31	<.0001
nc	1	14.963	2.665	5.61	<.0001
s	1	34.279	2.458	13.95	<.0001
w	1	20.739	2.926	7.09	<.0001
logpop_cen	1	11.455	2.483	4.61	<.0001
ncXlogpop_cen	1	3.041	3.466	0.88	0.3807
sXlogpop_cen	1	5.743	3.285	1.75	0.0811
wXlogpop_cen	1	-3.248	3.428	-0.95	0.3439

The nice aspect of using the centered version of the variable is that the last four terms zero out when $\log(\text{pop})$ is at its mean, so the intercept and regional indicator variables are meaningful. Once again we can easily see that after controlling for log population, the mean crime rate in the Northeast is about 38 crimes per 1000 people when log of population is at its mean. Once again, we can also interpret the coefficients for the north-central, south and west as offsets from the rate in the northeast, when $\log(\text{pop})$ is at its mean.

How about the coefficients for the interaction terms? These indicate that the change in the mean crime rate associated with changes in $\log(\text{pop})$ is different in different regions. To get the slope for the Northeast (reference region), we use the coefficient for logpopcen; the regression model for counties in the Northeast is $38.4 + 11.5(\text{logpopcen})$. In the Northeast, a 1% increase in population is associated with an increase in the mean crime rate of about 0.11 crimes per 1000 people, and a 10% increase in population is associated with an increase of 1.1 per 1000. In the North Central region, a 10% increase in population is associated with an increase of about $1.15 + .30 = 1.45$ crimes per 1000 people, on average. In the West, a 10% increase in population is associated with an increase of about $1.14 - .32 = .82$ crimes per 1000 people, on average.

So that was an interaction between a categorical variable and a continuous variable. How about an interaction between two continuous variables? Let's look at the interaction between $\log(\text{area})$ and $\log(\text{pop})$, while controlled for region:

Here I have used centered versions of all continuous variables, to make interpretation easier (and reduce collinearity):

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	37.049	1.894	19.56	<.0001
nc	1	15.541	2.635	5.90	<.0001
s	1	35.016	2.449	14.30	<.0001
w	1	24.570	3.187	7.71	<.0001
logarea_cen	1	-3.141	1.219	-2.58	0.0103
logpop_cen	1	13.511	1.187	11.38	<.0001
logarea_cenXlogpop_cen	1	-2.720	1.111	-2.45	0.0147

The interaction is significant ($P = .015$). How do we interpret the interaction? It means that the change in the mean crime rate due to differences in $\log(\text{area})$ depend on the level of $\log(\text{pop})$, and that the change in the mean crime rate due to differences in $\log(\text{pop})$ depend on the level of $\log(\text{area})$.

Using centered variables makes it a little easier to interpret the “main effects” of the variables involved in the interaction. These coefficients give the change in the mean crime rate due to changes in $\log(\text{area})$ or $\log(\text{pop})$ when the other variable is at its mean (in which case, the last term is zero). So when $\log(\text{pop})$ is at its mean, a 1% increase in area is associated with a decrease in mean crime rate of about .03 per 1000 people (and a 10% increase is associated with a decrease of .3 per 1000). When $\log(\text{area})$ is at its mean, a 1% increase in population is associated with an increase in mean crime rate of about .14 per 1000 people. Perhaps this makes sense; it means lower population density is associated with a lower crime rate.

To find the regression equation for values of $\log(\text{pop})$ and $\log(\text{area})$ other than their means, we need to plug in the specific values of the predictors we are interested in.