

ISyE 3103 Supply Chain Modeling: Transportation and Logistics

Regression Models with Indicator Variables

1 Categorical Variables

One of the advantages regression-based forecasting models have over time series extrapolation models (i.e. moving average, exponential smoothing, etc.) is that regression models can include other explanatory variables that are not necessarily related to time. It seems logical that demand (and other dependent variables) could be a function of factors such as advertising expenditures and price in addition to time series elements. Some of these factors could be *categorical* in nature as opposed to quantitative. Examples of categorical variables include gender, color, region of the country, and the political party currently in power. As we will see, categorical variables are also useful in modeling seasonal effects in regression models.

Regression models require that all of the variables included, both independent and dependent, be quantitative in nature. Consequently, we must devise a method of translating these categories into numbers. We will define the *levels* of a categorical variable as the number of values that variable can assume. It is not sufficient for us to haphazardly assign the levels to an integer (or any other) value in order to convert them. Creating integer values necessarily implies some kind of ordering on the levels, which does not make any sense in the context of a categorical variable.

Consider the following example. Suppose that we wanted to model the time it takes to deliver pallets to a warehouse. Clearly the number of pallets to be delivered is an important explanatory variable for estimating the time it takes to deliver the pallets. In addition, the time depends on the delivery conditions at the warehouse. Each warehouse provides its own forklift for loading and unloading pallets from the truck. We distinguish between hand-operated forklifts and motorized forklifts. Also, some warehouses have loading docks at the level of the bed of the truck, whereas other warehouses do not have loading docks, so that the pallets have to be removed from the truck by going up and down a metal ramp between the truck bed and the pavement outside. The difference in time between a hand-operated forklift and a motorized forklift depends on whether there is a loading dock or not — the difference is greater if pallets have to be moved up and down a ramp. That is, the time depends on the combination of forklift type and dock/ramp type. Therefore, we have a categorical variable with 4 levels: (1) hand-operated forklift with loading dock; (2) hand-operated forklift with ramp; (3) motorized forklift with loading dock; (4) motorized forklift with ramp. If we were to create an integer explanatory variable called `WAREHOUSE` where `WAREHOUSE = 0` denotes hand-operated forklift with loading dock; `WAREHOUSE = 1` denotes hand-operated forklift with ramp; `WAREHOUSE = 2` denotes motorized forklift with loading dock; and `WAREHOUSE = 3` denotes motorized forklift with ramp; we would be imposing an artificial ordering on the levels of the `WAREHOUSE` variable. This ordering renders the regression coefficient meaningless since we would be implying that motorized forklift with loading dock (`WAREHOUSE = 2`) has twice the effect of hand-operated forklift with ramp (`WAREHOUSE = 1`), whereas motorized forklift with ramp (`WAREHOUSE = 3`) has three times the effect of hand-operated forklift with ramp (`WAREHOUSE = 1`).

2 Using Indicator Variables

In order to incorporate categorical variables into a regression model, we must create indicator variables¹ that either take on a value of 0 or 1. To represent a categorical variable that has c levels,

¹Indicator variables are often called “dummy variables” or “binary variables.”

we require $c - 1$ indicator variables. If we created c indicator variables along with a constant term (β_0) in the regression model, we would get an error from our statistical software program because the data matrix (or more accurately, $X^T X$) is singular.² You must choose³ one of the levels to be the base case, and then each of the other levels has a corresponding indicator variable that is equal to 1 if the observation exhibits that level and 0 otherwise.

Continuing our pallet delivery example from above, suppose we choose hand-operated forklift with loading dock as the base case. Then we must create $4 - 1 = 3$ indicator variables⁴ HAND-RAMP, MOTORIZED-DOCK, and MOTORIZED-RAMP to represent the other categories. If we list a given observation's indicator variable values as an ordered triple (HAND-RAMP, MOTORIZED-DOCK, MOTORIZED-RAMP), then a hand-operated forklift with loading dock would be denoted by $(0, 0, 0)$, a hand-operated forklift with ramp would be denoted by $(1, 0, 0)$, a motorized forklift with loading dock would be denoted by $(0, 1, 0)$, and a motorized forklift with ramp would be denoted by $(0, 0, 1)$.

Suppose we wanted to include a second categorical variable with ℓ levels in the regression model. There are two important alternative models to consider here. In the more general, but more complicated, model, you capture the possibility that the dependent variable depends on the combinations of the levels of the two categorical variables, that is, the effect of the two categorical variables on the dependent variable is not additive but something more complicated. In statistics terminology they say there is "interaction" between the two categorical variables. This case requires the creation of $c\ell - 1$ indicator variables. In the more restrictive, but simpler, model, you assume that the dependent variable depends on the levels of the two categorical variables in an additive way, that is, the effect of one categorical variable on the dependent variable is not affected by the level of the other categorical variable. This case requires the creation of $(c - 1) + (\ell - 1)$ indicator variables, which is clearly less than the $c\ell - 1$ indicator variables required by the model that makes provision for interaction effects.

In the previous example, we captured the possibility that the delivery time depends on the forklift type and the ramp type in a way that is more complicated than additive, because as we argued before, "the difference in time between a hand-operated forklift and a motorized forklift depends on whether there is a loading dock or not." Therefore we created $(2)(2) - 1 = 3$ indicator variables. If the difference in time between a hand-operated forklift and a motorized forklift did not depend on whether there is a loading dock, then we would have created only $(2 - 1) + (2 - 1) = 2$ indicator variables.

In both cases the definition of the base case must include the chosen base level for *each* of the categorical variables.

To illustrate this concept further, suppose we thought that whether payment for the delivered goods has to be processed with delivery was an important factor determining the delivery time. This variable has two levels: payment and no-payment. Let us select the no-payment level as our base case. If we think that the effect of forklift type and ramp type on delivery time depends on payment type, that is, there is interaction between (forklift type, ramp type) and payment type, then we will need $(4)(2) - 1 = 7$ indicator variables in total. On the other hand, if we think that the effect of forklift type and ramp type on delivery time does not depend on payment type, that is, there is no interaction between (forklift type, ramp type) and payment type, then we will need

²Recall that a singular matrix has no inverse, and we must be able to invert that matrix in order to use least-squares regression.

³The specific choice of the base level does not matter. Any of the levels is as good a base as any of the others.

⁴Note that these variables must be *created*. No data set that you receive will include all of these 0's and 1's. Once you choose an indicator variable mechanism to represent a categorical variable, you must go through the data and determine the appropriate values of each indicator variable for each data observation.

$(4 - 1) + (2 - 1) = 4$ indicator variables in total.

For the rest, we consider the second type of model, in which the effect of forklift type and ramp type on delivery time does not depend on payment type. Consequently, we will use the 3 indicator variables created before, and create one additional indicator variable, `PAYMENT`, that is equal to 1 if payment for the delivered goods has to be processed with delivery and zero otherwise. The base case of the model now corresponds to a delivery at a warehouse with a hand-operated forklift and a loading dock and without payment processing with the delivery. There are eight possible types of delivery, encoded with the indicator variables as follows.

Delivery Type	(HAND-RAMP, MOTORIZED-DOCK, MOTORIZED-RAMP, PAYMENT)
hand-operated forklift, loading dock, no payment processing	(0, 0, 0, 0)
hand-operated forklift, loading dock, payment processing	(0, 0, 0, 1)
motorized forklift, ramp, no payment processing	(0, 0, 1, 0)
motorized forklift, ramp, payment processing	(0, 0, 1, 1)
motorized forklift, loading dock, no payment processing	(0, 1, 0, 0)
motorized forklift, loading dock, payment processing	(0, 1, 0, 1)
hand-operated forklift, ramp, no payment processing	(1, 0, 0, 0)
hand-operated forklift, ramp, payment processing	(1, 0, 0, 1)

3 Interpreting Regression Coefficients for Indicator Variables

Regression coefficients (true β_i parameters) typically represent the amount of change in the expected value of the dependent variable as the corresponding explanatory variable increases by one unit. Estimated coefficients ($\hat{\beta}_i$ statistics) represent the forecasted change in the expected value of the dependent variable as the explanatory variable increases by one unit. These interpretations must be modified for indicator variables since they are binary and we have a base case. Regression coefficients for indicator variables signify the change in the dependent variable as our corresponding categorical variable changes from the base level to the level represented by that coefficient's indicator variable.

Suppose that we have the following estimated regression model

$$\begin{aligned} \text{DELIVERY_TIME} = & \hat{\beta}_0 + \hat{\beta}_1 \text{NUMBER_PALLET} \\ & + \hat{\beta}_2 \text{HAND-RAMP} \times \text{NUMBER_PALLET} + \hat{\beta}_3 \text{MOTORIZED-DOCK} \times \text{NUMBER_PALLET} \\ & + \hat{\beta}_4 \text{MOTORIZED-RAMP} \times \text{NUMBER_PALLET} + \hat{\beta}_5 \text{PAYMENT} \end{aligned}$$

where `NUMBER_PALLETS` is a quantitative variable⁵ representing the number of pallets to be delivered at a warehouse. Recall that we defined our base case as a delivery at a warehouse with a hand-operated forklift and a loading dock and without payment processing with the delivery. Consequently, the estimated delivery time for a delivery at a warehouse with a hand-operated forklift and a loading dock and without payment processing with the delivery is $\text{DELIVERY_TIME} = \hat{\beta}_0 + \hat{\beta}_1 \text{NUMBER_PALLET}$.

We can interpret $\hat{\beta}_0$ as the fixed time (setup time) for delivery, irrespective of the number of pallets and payment processing. It may represent the time required for parking, paperwork, and

⁵We are allowed to use quantitative explanatory variables along with indicator variables in regression models. This will enable us to build models of trend and seasonality.

other overhead activities. We can interpret $\hat{\beta}_2$ as the incremental time *per pallet* for delivery at a warehouse with a hand-operated forklift and a ramp over delivery at a warehouse with a hand-operated forklift and a loading dock. (It is reasonable to expect that $\hat{\beta}_2 > 0$. Why?) Similarly, $\hat{\beta}_3$ represents the incremental time per pallet for delivery at a warehouse with a motorized forklift and a loading dock over delivery at a warehouse with a hand-operated forklift and a loading dock. (It is reasonable to expect that $\hat{\beta}_3 < 0$. Why?). Similarly, $\hat{\beta}_4$ represents the incremental time per pallet for delivery at a warehouse with a motorized forklift and a ramp over delivery at a warehouse with a hand-operated forklift and a loading dock. (What do you expect the sign of $\hat{\beta}_4$ to be?). Finally, $\hat{\beta}_5$ represents the incremental time for a delivery with payment processing over a delivery with no payment processing.

4 Time Series Example

The Belgian Trucking Company⁶ needs to determine the number of refrigerated⁷ trucks to satisfy the transportation demand between Antwerp and Brussels on a daily basis. The demand for refrigerated trucks is dependent on the daily temperature (because more refrigerated vans are needed when the outside temperature increases), and it also appears that there are trend and daily seasonality elements present in the data. Six weeks' worth of data are provided in Table 1.

Since the data exhibits daily seasonality within each week, we need to create appropriate indicator variables to model this seasonality in the regression model. There are five workdays in each week, so we need four indicator variables. Let us choose Friday as our base day and create variables MON, TUE, WED, and THU. Table 2 contains the first two weeks of data observations, including the indicator variable values. The subsequent weeks of data are modified in the same fashion to obtain the data set on which we will fit the regression model.

Regressing demand on all of the independent variables (2 quantitative and 4 indicator), we obtain the following estimated regression function.

$$\widehat{\text{DEMAND}} = 72.551 + 0.474\text{TIME} - 0.849\text{TEMP} + 5.476\text{MON} - 11.538\text{TUE} - 17.811\text{WED} - 22.143\text{THU}$$

All of the t-tests for the regression coefficients are significant at even the $\alpha = 0.01$ level, as is the overall F-test for the model as a whole. The measures of fit for this model are $R^2 = 0.9606$ and $R^2_{\text{ADJ}} = 0.9503$, which indicate that we have an extremely good fit.

As we expected, the daily temperature is inversely related to the reefer demand. We would estimate that on any given day a one-degree (Celsius) increase in temperature results in a 0.849 decrease in the expected demand. Similarly, the regression coefficients for the indicator variables denote the change in expected demand for that day of the week compared with our base day of Friday.

The interpretation of the coefficient for the TIME variable is somewhat complicated here. We would ordinarily say that demand should increase by an average of 0.474 as each day goes by. The problem here, though, is that an increase of one day would correspond to a different level in the daily seasonality categorical variable, so we are unable to keep the level of the seasonal variable constant if we increase by one day.

To deal with this problem, we should consider what happens to the expected demand as the time changes by 5 days.⁸ This allows us to keep the seasonal variable at a constant level. Consequently,

⁶This example is adapted from Ghiani, Laporte, and Musmanno's *Introduction to Logistics Systems Planning and Control* (2004; pg.72), published by John Wiley & Sons.

⁷These are also known as "reefer" trucks.

⁸Note that we only increase by 5 instead of 7 because the weekend days are not included in the model's trend formulation. We have modeled the situation as if weekends did not exist. (Sometimes that's the way it feels since they go by so quickly!)

Table 1: Data for example problem

Week	Day	t	D_t	T_t ($^{\circ}\text{C}$)
1	Monday	1	67	13
	Tuesday	2	54	10
	Wednesday	3	51	9
	Thursday	4	46	10
	Friday	5	62	10
2	Monday	6	65	15
	Tuesday	7	55	14
	Wednesday	8	47	14
	Thursday	9	45	15
	Friday	10	64	13
3	Monday	11	75	11
	Tuesday	12	58	10
	Wednesday	13	56	7
	Thursday	14	49	9
	Friday	15	71	10
4	Monday	16	76	14
	Tuesday	17	57	12
	Wednesday	18	53	11
	Thursday	19	50	11
	Friday	20	69	13
5	Monday	21	78	12
	Tuesday	22	60	12
	Wednesday	23	54	10
	Thursday	24	53	8
	Friday	25	80	9
6	Monday	26	81	11
	Tuesday	27	63	13
	Wednesday	28	58	12
	Thursday	29	52	11
	Friday	30	83	11

Table 2: First two weeks of data with indicator variable values

TIME	TEMP	MON	TUE	WED	THU	DEMAND
1	13	1	0	0	0	67
2	10	0	1	0	0	54
3	9	0	0	1	0	51
4	10	0	0	0	1	46
5	10	0	0	0	0	62
6	15	1	0	0	0	65
7	14	0	1	0	0	55
8	14	0	0	1	0	47
9	15	0	0	0	1	45
10	13	0	0	0	0	64

it makes sense for us to say that moving a week forward in time results in an average increase in demand of $0.474 \times 5 = 2.37$ reefer trucks (as long as the temperature stays constant).