# Dummy Variables



Dummy variables represent categories. Use dummy variables if you want to find out if being in a certain category makes a difference, compared with not being in that category.

It's called a dummy variable because its values are all either 0 or 1. "Dummy" is an adjective, not a noun. You give the dummy variable a value of 1 for each observation that is in some category that you have defined. You give the dummy variable a value of 0 for each observation that is not in the category. For any such category, either you're in or your out, so no values other than 0 or 1 are allowed. In a spreadsheet, a dummy variable looks like a column of 0's and 1's.

For example, if you are doing a study where your observations are of men and women, and you think gender matters, you can create a dummy variable that's 0 for men and 1 for women (or 0 for women and 1 for men – either way is OK).

When you do the regression and get your results, the estimated coefficient of a dummy variable shows how much difference it makes to be in the category for which the dummy variable is 1.

For example, suppose your equation is: Weight = a + b×Height + c×Gender, and Gender is 0 for men and 1 for women. Your estimate of c is how much more women weigh than men, given their height. Since women generally weigh less than men of the same height, c will be a negative number. c can be interpreted as how much less women weigh than men, with the height difference controlled for.

If you have more than two categories, and each observation in your data is in one and only one of those categories

- use a separate dummy variable for each category,
- but always use <u>one less</u> dummy variable than you have categories.

For example, suppose we divide S.C. into three regions: Piedmont, Midlands, Coast. We have data for each county for income and hospital bed-days. We expect that income affects the demand for bed-days. We want to see if region also matters. We can define a dummy variable called Piedmont that is 1 for Piedmont counties and 0 for other counties. We can define a Midlands dummy variable that is 1 for Midlands counties and 0 for other counties. We can also define a Coast dummy variable that is 1 for coastal counties and 0 for the others.

Our equation, though, must include only two of those dummy variables, not all three. For example, we might choose to estimate areas BedDays = a + b×Income + c×Piedmont + d×Midlands + error

Why not also include a dummy variable for Coast? If your categories are exhaustive (every observation is in one of your categories) and mutually exclusive (no observation is in more than one category), then you must not put dummies for all the categories in the same equation, unless you leave out the intercept. Violate this and the mathematical algorithm that least squares uses won't work.

The reason, and it's subtle, is that you have perfect multicollinearity among your dummies and the intercept. In the S.C. regions example, if Coast = 1 for Coast areas and 0 for elsewhere, and every observation is Coast or Piedmont or Midlands, then Coast = 1 - Piedmont - Midlands. This is a linear

relationship among your variables. That fits the definition of multicollinearity.

In your regression results, the coefficient of any particular dummy variable shows the difference between being in that category and being in the category whose dummy variable is not in the equation. In the S.C. regions example, if we leave out Coast, so that the equation is
BedDays = a + b×Income + c×Piedmont + d×Midlands + error ,
then the coefficient of Piedmont is the average difference in BedDays between Piedmont and Coast.

In a multiple-category situation like this, it doesn't matter which dummy you leave out. You will get the same predictions regardless.

The alternative to leaving out one dummy is to leave out the intercept and put dummies in for all your categories. In results you get with this approach, each dummy's coefficient is the intercept for observations in its category.

An example: These data, from Wonnacott and Wonnacott, show jewelry sales by quarter of the year (1957_1 means January-March 1957) in millions of dollars (what E6 means) for Canada in the C column. The B column has a time trend variable. Columns D, E, and F have dummy variables for the 1st, 3rd, and 4th quarters respectively. The second quarter is left out, so the coefficients of these dummy variables will measure the difference between each of the other quarters and the second quarter. In columns D, E, and F, the value is 1 for observations in the quarter that the column represents. The value is 0 in the other quarters.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | Time | SalesE6 | 1st Qtr | 3rd Qtr | 4th Qtr |
| 2 | 1957_1 | 0 | 2.4 | 1 | 0 | 0 |
| 3 | 1957_2 | 1 | 2.9 | 0 | 0 | 0 |
| 4 | 1957_3 | 2 | 2.9 | 0 | 1 | 0 |
| 5 | 1957_4 | 3 | 5 | 0 | 0 | 1 |
| 6 | 1958_1 | 4 | 2.4 | 1 | 0 | 0 |
| 7 | 1958_2 | 5 | 3 | 0 | 0 | 0 |
| 8 | 1958_3 | 6 | 2.9 | 0 | 1 | 0 |
| 9 | 1958_4 | 7 | 5.1 | 0 | 0 | 1 |
| 10 | 1959_1 | 8 | 2.6 | 1 | 0 | 0 |
| 11 | 1959_2 | 9 | 2.9 | 0 | 0 | 0 |
| 12 | 1959_3 | 10 | 3 | 0 | 1 | 0 |
| 13 | 1959_4 | 11 | 5.2 | 0 | 0 | 1 |
| 14 | 1960_1 | 12 | 2.5 | 1 | 0 | 0 |
| 15 | 1960_2 | 13 | 3 | 0 | 0 | 0 |
| 16 | 1960_3 | 14 | 2.9 | 0 | 1 | 0 |
| 17 | 1960_4 | 15 | 5 | 0 | 0 | 1 |

The regression equation is SalesE6 = a + $b_1$×Time + $b_2$×1stQtr + $b_3$×3rdQtr + $b_4$×4thQtr .

Dependent Variable: SalesE6 Observations: 16

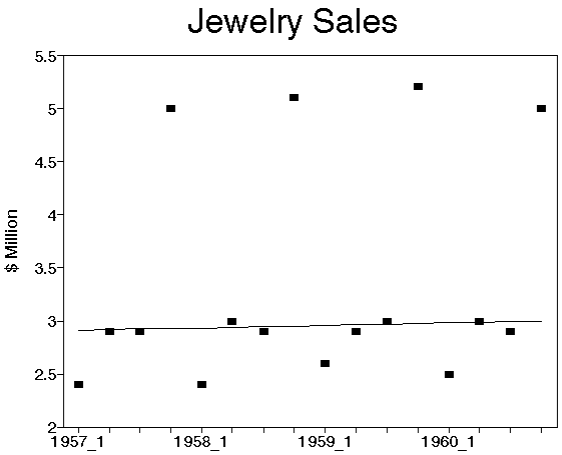
| Variable | Coefficient | Std Error | T-statistic | P-Value |
|---|---|---|---|---|
| Time | .00562500 | .00420953 | 1.3362526 | 0.2084511 |
| 1st Qtr | -.46937500 | .05341299 | -8.7876564 | 0.0000026 |
| 3rd Qtr | -.03062500 | .05341299 | -.57336240 | 0.5779328 |
| 4th Qtr | 2.1137500 | .05390833 | 39.210082 | 0.0000000 |
| Intercept | 2.9106250 | .04781111 | 60.877590 | 0.0000000 |

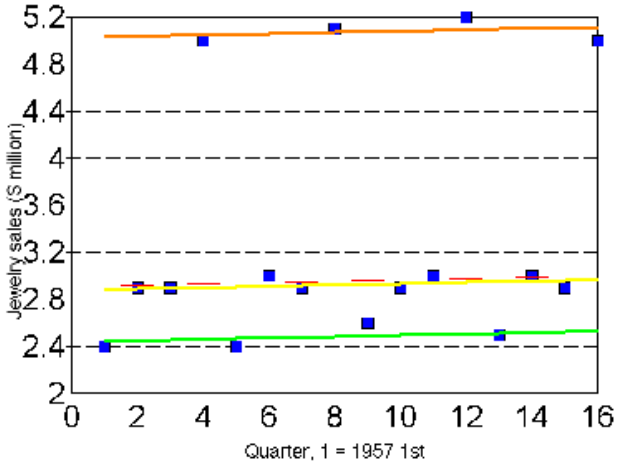
These test hypothesis that coefficient is 0.

| | Sum of Squares | Degrees of Freedom | R-Squared | F for Equation |
|---|---|---|---|---|
| Regression | 16.3370000 | 4 | 0.996197 | 720.269 |
| Residual | .062375000 | 11 | Adjust R-Sq | P Value |
| TOTAL | 16.3993750 | | 0.989647 | 0.00000 |

The results show an overall upward trend of $0.005625 million per quarter, based on the estimated coefficient of the Time variable. The 1st Qtr coefficient says that 1st quarter sales average $0.47 million below an overall trend line at the 2nd quarter level. The 3rd quarter isn't significantly different from the 2nd quarter. The 4th quarter averages $2.11 million higher than an overall trend line at the 2nd quarter level.

This graph shows the data, and a trend line at the average 2nd quarter level, based on the regression results' time coefficient and intercept. The line's slope is $.005625 million (or $5625) per quarter and its intercept is $2.91 million. The 2.1137500 coefficient of the 4th quarter dummy variable tells us that 4th quarter sales average $2,113,750 above that line. The 1960 4th quarter was actually $2,005,000 above that line, so that quarter's sales were $108,750 below the trend.




This graph shows the regression lines for each quarter. Fitting an equation with dummy variables is like fitting a bunch of parallel lines, with one line for each category. The 2.11375 coefficient for the fourth quarter dummy means that the 4th quarter observations' line is$2,113,750 above the 2nd quarter observations' line.

The observed value for the 4th quarter of 1960, in the upper right corner of the graph, is below the 4th quarter line by $108,750. Jewelry sales during the 4th quarter of 1960 were lower than expected.

# Time Series Theory

Time series analysis is looking at data gathered over time. Time series analysis involves a time trend variable and dummy variables that the researcher constructs.

A time series can be decomposed into

- Trend
- Seasonal fluctuation that repeats
- Shifts in trend
- Tracking of errors ("autocorrelation")

Let's look at these in turn:

1. Trend

Trend can be represented by a variable that goes 1,2,3, ..., etc.
A linear equation would be $Y = \alpha + \beta t$ + other terms
A constant growth curve would be $\ln(Y) = \alpha + \beta t$ + other terms

2. Seasonal fluctuation

Provides seasonal predictions.
"Season" can be day, week, month, or season, etc., anything that recurs regularly.

You may not need to bother about seasonal adjustment if all you care about is estimating the time trend. You may be able to get by with estimating $\beta$ in the simple equation $Y = \alpha + \beta t$. However, as Wonnacott and Wonnacott's jewelry data show, sometimes leaving out seasonal adjustment can give you a biased estimate of $\beta$.

To calculate seasonal fluctuation, create a dummy variable for each season When doing the regression, leave one season out of your independent variables, or leave the intercept out.

3. Shift in trend.

If there are a block of observations that are out of line, and you can identify something special about that time period, use a dummy variable whose value is 1 during the period and 0 otherwise. This dummy is in addition to any seasonal dummies.

4. Tracking of errors, or "autocorrelation" of the errors

If errors "track," meaning that the errors follow each other, so that if one error is positive then the next error is likely to also be positive, this violates the assumption that the covariances of the errors are zero. Least squares is no longer the best way to analyze the data. Other techniques are available, however. One is to fit a curve, by transforming the data. If the autocorrelation persists, you can use the difference between one time period's Y value and the previous time period's Y value as the dependent variable. That technique is beyond the scope of this course.

**How to interpret the coefficient of a time or dummy variable in an equation with a dependent variable that is a logarithm.**

I'll use an equation with only one dummy variable. This simplifies the math. The principles apply the

same to equations with several variables.

Suppose that the equation is:

$Ln(Y) = \alpha + \beta \times Time + \gamma \times Dummy$ Time goes 1, 2, 3, etc. Dummy is 0 for some time periods and 1 for the others.

Let's first consider β, the coefficient of Time. As an approximation, you can say that Y is growing at the rate of 100%×β per year. For example, if β is .01 and Time is in years, then you can say Y is growing at about 1% per year.

As pointed out earlier, this is only an approximation. It is good only if β is close to 0. Let us look again at the above equation and figure out what happens to Y when Time goes up by 1.
If we replace Time with Time+1, we get $\alpha + \beta \times (Time+1) + \gamma \times Dummy$.
This equals $\alpha + \beta \times Time + \beta + \gamma \times Dummy$.
The difference between this and $\alpha + \beta \times Time + \gamma \times Dummy$ is just β.
So, when Time goes up by 1, Ln(Y) goes up by β.
That means that Y is *multiplied* by $e^{\beta}$. ( $e^{Ln(Y)+\beta} = e^{Ln(Y)} e^{\beta} = Y e^{\beta}$ )
In percentage terms, multiplying Y by $e^{\beta}$ means increasing Y by $100 \times ( e^{\beta} - 1 )$ percent.

Here is a table showing some β values and the percent equivalent of $e^{\beta} - 1$.

| β | $e^{\beta} - 1$ Percent equivalent |
|---|---|
| 1 | 1.718 171.8% |
| .3 | .350 35.0% |
| .1 | .105 10.5% |
| .05 | .051 5.1% |
| .01 | .010 1.0% |
| 0 | .000 .0% |
| -.01 | -.010 -1.0% |
| -.05 | -.049 -4.9% |
| -.1 | -.095 -9.5% |
| -.3 | -.259 -25.9% |
| -1 | -.632 -63.2% |

How b and $e^b$-1 differ

I would say that if β is between -0.05 and 0.1, then β is small enough so you can just multiply β by 100 and call that the monthly percent change in Y. For example, if β is .01, you can say that Y is changing by 1% per month. If β is bigger than 0.1 or more negative than -0.05, use the $e^{\beta} - 1$ formula to get the percent equivalent. For example, if b is 1 then the monthly rate of change is 171.8%. (If your numbers show that something is growing this fast, check for a mistake!)

Often, in write-ups of equations like this, β is said to be the estimated "rate of growth" in Y. Yet, the preceding paragraphs say that, on average, Y grows to more than $Y \times (1+\beta)$ in one unit of time. So, is the rate of growth really β or is it a little more than β?

The answer is that β is the *continuously compounded* rate of growth. Suppose you have a savings account whose nominal interest rate is 5% per year. If interest is paid and compounded continuously, money left in the bank for one full year will grow by 5.1% over the course of the year.

Now let's interpret γ, the coefficient of the γ dummy variable in our equation,
$Y=\alpha + \beta Time + \gamma Dummy$.
The interpretation this γ is similar to the interpretation of β. We also use the table above. If γ is between -0.05 and 0.1, then you can say that the time periods when the dummy is 1 have, on the average, $100 \times \gamma$ percent more surgeries than the time periods when the dummy is 0. If γ is not this small, use the

table above or a spreadsheet to calculate $100\% \times (e^{\gamma} - 1)$. For example, if $\gamma = 0.3$ then the time periods when the dummy is 1 have 35% more surgeries than the other time periods.

**Advanced stuff: A further refinement**

The interpretation suggested above is slightly biased. It tends to slightly overstate the effect of a unit change in time or in the dummy variable, because of the non-linearity in our equations.

A less-biased interpretation can be obtained by adding another term to the $e^{\beta\text{-hat}} - 1$ formula, like this:

The estimated effect on Y of a unit change in X,
in the equation $\ln(Y)=\alpha+\beta X$

$$= e^{\hat{\beta} - \frac{\hat{V}(\hat{\beta})}{2}} - 1$$

The V-hat of beta-hat term is the estimated variance of your estimate for $\beta$. The estimated variance is the **square** of the standard error of the coefficient, which is typically shown in your regression results.

If your estimated coefficient is statistically significant, and the estimated coefficient (beta-hat) is not much bigger than 1, then this correction will make little difference, compared with just using $e^{\beta\text{-hat}} - 1$, because the square of the standard error will be a small fraction. Still, we have run into journals that insist that this correction be made.

This correction was published in Peter E. Kennedy, "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *American Economic Review,* Volume 71, No. 4, Sept. 1981, p. 801. Kennedy is the author of *A Guide to Econometrics*, MIT Press, now in its fifth edition. If you are interested in learning more about regression, I highly recommend Kennedy's book as a readable survey of econometric theory. The 4th edition's General Notes, section 14.2, gives the $e^{\beta\text{-hat}} - 1$ formula. It does not mention the more complex formula shown above that Kennedy himself developed.