QUALITATIVE EXPLANATORY VARIABLES

QUANTITATIVE VS QUALITATIVE VARIABLES

A quantitative variable is a variable that can be numerically measured on some well defined scale (e.g., income, price, output, age, height, weight, and family size). A qualitative variable is a variable that indicates the presence or absence of a quality or a characteristic. It has as many categories as possible characteristics.
Examples of qualitative variables are as follows.
1) Gender: male or female.
2) Homeownership: own a house or don't own a house
3) Smoker: smoke or don't smoke.
4) Education: high school, college, graduate school.
To quantify a qualitative variable, we construct one or more artificial variables called *dummy variables*. A dummy variable can take two values: 0 or 1. The variable takes the value 1 if the characteristic is present and a value of 0 if the characteristic is absent.

ANALYSIS OF VARIANCE MODELS

A model for which the dependent variable is a quantitative variable, and all explanatory variables are qualitative variables is called an analysis of variance model. An analysis of variance model is a particular type of classical linear regression model.

Analysis of Variance Model with One Qualitative Variable with Two Categories

Suppose that we have information on the monthly wage of 49 workers. Suppose that we postulate that an individual's monthly wage depends upon his or her gender. To quantify the qualitative variable gender, we create a dummy variable, designated G_t . It is defined as follows.

$$G_t = 1 \text{ if male}$$
$$G_t = 0 \text{ if female}$$

We can specify this statistical model of wage determination as follows

$$Y_t = a + bG_t + \varepsilon_t$$

The error term ε_t satisfies all of the assumptions of the classical linear regression model. Y_t is the monthly wage of the *t*th worker, a quantitative variable. G_t measures the qualitative variable gender. The equation tells us that the monthly wage for the *t*th male (G_t =1) is given by

$$Y_t = a + b + \varepsilon_t$$

The monthly wage for the *t*th female (G_t = 0) is given by

$$Y_t = a + \varepsilon_t$$

This means that the intercept of the population regression line measures the average monthly wage of female workers (the control group). The slope of the population regression line measures the difference between the average salary of male workers and the average salary of female workers. Thus, if b > 0, then the average salary of male workers is greater than the average salary of female workers. If b< 0, then the average salary of male workers is less than the average salary of female workers.

Estimation of the Statistical Model

To obtain estimates of the population parameters a and b, we can regress Y_t on a constant term and the dummy variable G_t using the OLS estimator. It can be shown that the estimate of *a* is the sample mean salary for females, and the estimate of a + b is the sample mean salary for males. In this case, if we were to divide the 49 workers into 2 groups, male and female, and calculate the sample mean wage for each group, we would get the same result as that obtained from the regression model.

Test of the Hypothesis that the Two Population Means Are Equal

Suppose that we want to test the following hypothesis: "The population mean salary of females is equal to the population mean salary of males." The null and alternative hypotheses can be expressed as the following restriction on the parameters of the statistical model.

$$H_o: b_1 = 0$$
$$H_1: b_1 \neq 0$$

To test this null hypothesis, we can use a t-test. Note that the t-test of the null hypothesis b = 0 in the regression model is exactly the same as the t-test of the hypothesis that two population means are equal.

<u>Analysis of Variance Model with One Qualitative Variable with More Than Two Categories</u>

Suppose that we postulate that an individual's monthly wage depends upon his/her job type. Suppose that there are 4 different job types: professional, clerical, crafts, maintenance. This is an example of a qualitative variable that has 4 categories. To quantify the qualitative variable *job type*, we can define 4 dummy variables, one for each of the four categories. Define the following.

$J_1 = 1$ if professional	$J_2 = 1$ if clerical	$J_3 = 1$ if crafts	$J_4 = 1$ if maintenance
$J_1 = 0$ otherwise	$J_2 = 0$ otherwise	$J_3 = 0$ otherwise	$J_4 = 0$ otherwise

Note that if for a particular worker $J_1 = 1$ then $J_2 = 0$, $J_3 = 0$, and $J_4 = 0$ for this worker. If $J_2 = 1$ then $J_1 = 0$, $J_3 = 0$, and $J_4 = 0$ for this worker, etc. The statistical model of wage determination can be specified as follows

$$Y_t = a + b_1J_{t1} + b_2J_{t2} + b_3J_{t3} + \varepsilon_t.$$

Note that we excluded one of the dummy variables. We excluded the dummy variable for maintenance, J_4. By doing this, we have selected maintenance as the control group. It is represented by the constant term *a*.
In general, when we represent a qualitative variable with *k* categories with *k* dummy variables, we can only include *k – 1* of those dummy variables in the regression model. This is because if we included all *k* dummy variables in the model, we would have perfect multicollinearity and we could not estimate the model.
The category for which we do not include a dummy variable is represented by the constant and is interpreted as the control group or reference group. Thus, it was not necessary to create a dummy variable for the category maintenance.

Interpretation

The monthly wage for the *t*th maintenance worker is given by

$$Y_t = a + \varepsilon_t$$

The monthly wage for the *t*th professional worker is given by

$$Y_t = a + b_1 + \varepsilon_t$$

The monthly wage for the *t*th clerical worker is given by

$$Y_t = a + b_2 + \varepsilon_t$$

The monthly wage for the *t*th crafts worker is given by

$$Y_t = a + b_3 + \varepsilon_t$$

It follows that b_1 is the difference between the average salary of a professional and a maintenance workers; b_2 is the difference between the average salary of a clerical worker and a maintenance worker; b_3 is the difference between the average salary of a crafts worker and a maintenance worker.

Estimation of the Statistical Model

To obtain estimates of the population parameters, a, b_1, b_2 and b_3, we can regress Y_t on a constant term and the dummy variables J_{t1}, J_{t2}, J_{t3} using the OLS estimator. It can be shown that *a* is the sample mean salary for maintenance workers, $a + b_1$ is the sample mean salary for professionals, $a + b_2$ is the sample mean salary for clerical workers, and $a + b_3$ is the sample mean salary for crafts workers. In this case, if we were to divide the 49 workers into 4 groups, maintenance, professional, clerical, and crafts, and calculate the sample mean wage for each group, we would get the same result as that obtained from the regression model.

Tests of Hypotheses

There are a number of alternative hypotheses that we can test.

Analysis of Variance Model with Two Qualitative Variables

Suppose that we postulate that an individual's monthly wage depends on his/her job type and his/her gender. Job-type is represented by the 4 dummy variables defined above: J_1, J_2, J_3, J_4. Gender is represented by the dummy variable *G* defined above. This is an example of a model with two qualitative variables. The qualitative variable job type has 4 categories. The qualitative variable gender has two categories. The statistical model of wage determination can be specified as follows

$$Y_t = a + b_1 J_{t1} + b_2 J_{t2} + b_3 J_{t3} + b_4 G_t + \varepsilon_t$$

Note that, once again, we have excluded one of the dummy variables for the qualitative variable *job type*.
We excluded the dummy variable for maintenance. By doing this we have selected maintenance for the control group.
For the gender dummy variable, $G = 0$ indicates female, and therefore we have selected female for the control group.
Thus, in this model, with two qualitative variables, the control group is female maintenance workers.
The group "female maintenance workers" is represented by the constant.

Interpretation

The monthly wage for the *t*th worker is given by

*t*th female maintenance worker: $Y_t = a + \varepsilon_t$
*t*th male maintenance worker. $Y_t = a + b_4 + \varepsilon_t$

*t*th female professional: $Y_t = a + b_1 + \varepsilon_t$
*t*th male professional: $Y_t = a + b_1 + b_4 + \varepsilon_t$

*t*th female clerical worker: $Y_t = a + b_2 + \varepsilon_t$
*t*th male clerical worker: $Y_t = a + b_2 + b_4 + \varepsilon_t$

*t*th female crafts worker: $Y_t = a + b_3 + \varepsilon_t$
*t*th male crafts worker: $Y_t = a + b_3 + b_4 + \varepsilon_t$

It is important to understand that this model imposes the restriction that the difference between the average salary of a male and a female is the same regardless of the job type; that is, the male/female wage differential is the same for maintenance workers, professionals, clerical workers, and crafts workers. This difference is given by the parameter attached to the dummy variable for gender (*G*), which is b_4.

Estimation of the Model

To obtain estimates of the population parameters a, b_1, b_2, b_3, and b_4, we can regress Y_t on a constant term and the dummy variables J_1, J_2, J_3, J_4 and G using the OLS estimator

Tests of Hypotheses

Once again, we can test hypotheses about differences between or among population mean salaries for the various groups using the appropriate t-test or F-test.

Regression Model Vs Analysis of Variance

This regression model with two qualitative variables (job type and gender) represented by 4 dummy variables as regressors describes the same phenomenon and leads to the same test results as a two-way analysis of variance model.

Analysis of Variance Model with Two Qualitative Variables and Interaction Terms

The regression model with two or more qualitative variables can be generalized further by including interaction terms. Once again, suppose that we postulate that an individual's monthly wage depends upon his/her job type and his/her gender. The previous model imposes the restriction that the difference between the monthly wage of a male and a female is the same regardless of their job type. This male/female wage difference is given by the parameter attached to the dummy variable for gender (G), which is b_4. Suppose that we don't want to impose this restriction. Suppose that we want to allow the difference between the monthly wage of a male and a female to differ by job type. To do this, we need to include interaction terms between gender and job type in the model. The statistical model of wage determination can be specified as follows

$$Y_t = a + b_1J_{t1} + b_2J_{t2} + b_3J_{t3} + b_4G_t + b_5J_{t1}G_t + b_6J_{t2}G_t + b_7J_{t3}G_t + \varepsilon_t$$

To obtain an interaction term between gender and a particular job type category, we simply multiply the dummy variable for the gender by the dummy variable for that job type category. Notice that once again the control group is female maintenance workers.

Interpretation

The monthly wage for the tth worker is given by

tth female maintenance worker: $Y_t = a$
tth male maintenance worker. $Y_t = a + b_4$

tth female professional: $Y_t = a + b_1$
tth male professional: $Y_t = a + b_1 + b_4 + b_5$

tth female clerical worker: $Y_t = a + b_2$
tth male clerical worker: $Y_t = a + b_2 + b_4 + b_6$

tth female crafts worker: $Y_t = a + b_3$
tth male crafts worker: $Y_t = a + b_3 + b_4 + b_7$

Notice that the difference between the average salary for a male and female can be different for maintenance workers, professionals, clerical workers, and crafts workers. The male/female wage differential will differ for these different job types if the coefficients of the interaction terms, b_5, b_6, and b_7, are non-zero and have different magnitudes.

Estimation of the Model

To obtain estimates of the population parameters a, b_1, b_2, b_3, b_4, b_5, b_6, and b_7, we can regress Y_t on a constant term and the dummy variables and interaction variables (which are also dummy variables) J_1, J_2, J_3, J_4, G, J_1G, J_2G, and J_3G using the OLS estimator. It can be shown that the above means are identical to the sample means for 8 separate groups of workers: female maintenance workers, male maintenance workers, female professionals, male professionals, female clerical workers, male clerical workers, female crafts workers, male craft workers. Thus, if we were to divide the 49 workers into 8 groups, female maintenance workers, male maintenance workers, female professionals, male professionals, female clerical workers, male clerical workers, female crafts workers, male craft workers, and calculate the sample mean for each of the groups we would get the same results.

Tests of Hypotheses

Once again, we can test hypotheses about differences between or among population mean salaries for the various groups using the appropriate t-test or F-test.

Regression Model Vs Analysis of Variance

This regression model with two qualitative variables (job type and gender) and interaction terms between the two qualitative variables, represented by 7 dummy variables as regressors, describes the same phenomenon and leads to the same test results as a two-way analysis of variance model with interactions.

ANALYSIS OF COVARIANCE MODELS

In an analysis of variance model, all of the explanatory variables are qualitative variables measured by dummy variables. These types of models are not used very often in economics. This is because the dependent variable usually depends on one or more quantitative variables as well as one or more qualitative variables. A statistical model that includes both qualitative and quantitative explanatory variables is often times called an analysis of covariance model in the statistics literature.

Varying Intercept Parameter Models

A varying intercept model allows the intercept to differ for two or more categories, but requires the slope to be the same for all categories.

Example Suppose that we postulate that an individual's monthly wage depends upon his/her gender and his/her years of work experience. That is,

Wage = *f*(gender, experience).
Gender is a qualitative variable that has two categories: male and female. Experience is a quantitative variable that is measured in years. We can specify this statistical model of wage determination as follows

$$Y_t = a + bG_t + b_2E_t + \varepsilon_t$$

Where *G* is the dummy variable for gender (*G* = 1 if male; *G* = 0 if female), and *E* is years of work experience. The error term satisfies all of the assumptions of the classical linear regression model. By including the quantitative variable *experience*, we can now address the question: "Is there a wage differential between male and female workers who have the same number of years of work experience?"

Interpretation

The monthly wage for the *t*th male worker with E_t years of experience, and the *t*th female worker with E_t years of experience are given by

*t*th male worker with E_t years of experience: $Y_t = a + b_1 + b_2E_t + \varepsilon_t$
*t*th female worker with E_t years of experience: $Y_t = a + b_2E_t + \varepsilon_t$

Thus, we are postulating that the intercept for the wage function differs for males and females. The intercept for males is given by $(a + b_1)$ and the intercept for females is given by *a*. Thus the difference between the intercepts is given by *b*. If $b_1 > 0$, then for any given number of years of experience males make higher wages than females. If $b_1 < 0$, then for any given number of years of experience males make lower wages than females.

Estimation of the Statistical Model

To obtain estimates of the population parameters, we can regress the monthly wage on the dummy variable for gender and the number of years of experience, using OLS.

Hypothesis Tests

To test the hypothesis of no gender discrimination, we can test the following restriction on the parameters of the statistical model.

$$H_o: b1 = 0$$
$$H_1: b \neq 0$$

Extension of Model

We can include more than one quantitative variable in the model if we so desire. For example, we could include both years of experience and years of schooling. The conditional means would then give us information on the average wages of males and females for any given number of years of experience and years of schooling. When testing for wage discrimination between males and females, it is important to control for all important factors that influence wages other than

gender. If we don't, then the coefficient attached to gender will be biased and the test of wage discrimination will not be valid.

Varying Slope Parameter Models

A varying slope parameter model allows the slope to differ for two or more categories, but requires the intercept to be the same for all categories.

Example

Suppose that we postulate that an individual's monthly wage depends his/her years of work **experience**. That is,
Wage = f(experience).
However, we also postulate that the amount by which the wage changes for each additional year of work experience differs for males and females. We can specify this statistical model of wage determination as follows

$$Y_t = a + b_1E_t + b_2E_tG_t + \varepsilon_t$$

Notice that we have included an interaction term between experience and gender.

Interpretation

The monthly wage for the tth male worker with E_t years of experience, and the tth female worker with E_t years of experience are given by

tth male worker with E_t years of experience: $Y_t = a + (b_1 + b_2)E_t + \varepsilon_t$
tth female worker with E_t years of experience: $Y_t = a + b_1E_t + \varepsilon_t$

Thus, we are postulating that the slope for the wage function differs for males and females. The slope for males is given by $(b_1 + b_2)$ and the slope for females is given by b_1. Thus the difference between the slopes is given by b_2, which is the coefficient of the interaction term between experience and gender. If $b_2 > 0$, then an additional year of experience for males results in a bigger wage increase than for males. If $b_2 < 0$, then an additional year of experience for males results in a smaller wage increase for males than for females.

Estimation of the Statistical Model

To obtain estimates of the population parameters, we can regress the monthly wage on the number of years of experience and the interaction term between experience and gender, using OLS.

Hypothesis Tests

To test the hypothesis of no gender discrimination in wage increases for additional years of experience, we can test the following restriction on the parameters of the statistical model:

$$H_o: b_2 = 0 \text{ against } H_1: b_2 \neq 0$$

Extension of Model

We can include more than one quantitative variable in the model if we so desire. For example, we could include both years of experience and years of schooling. We could allow the wage increase for an additional year of experience and the wage increase for an additional year of schooling to differ for males and females.

Varying Intercept and Slope Parameter Models

A varying intercept and slope parameter model allows both the intercept and slope to differ for two or more categories.

Example

Suppose that we postulate that an individual's monthly wage depends upon his/her gender and his/her years of work experience. That is,

Wage = f(gender, experience).

Suppose that we postulate that the wages of males and females will be different for any given number of years of experience, and wage increases of males and females will be different for an additional year of work experience. We can specify this statistical model of wage determination as follows

$$Y_t = a + b_1G_t + b_2E_t + b_3E_tG_t + \varepsilon_t$$

Interpretation

The monthly wage for the tth male worker with E_t years of experience, and the tth female worker with E_t years of experience are given by

tth male worker with E_t years of experience: $Y_t = (a + b_1) + (b_2 + b_3)E_t + \varepsilon_t$
tth female worker with E_t years of experience: $Y_t = a + b_2E_t + \varepsilon_t$

Thus, we are postulating that both the intercept and the slope for the wage function differs for males and females. The intercept for males is given by $(a + b_1)$ and the intercept for females is given by a. The slope for males is given by $(b_2 + b_3)$ and the slope for females is given by b_2. Thus the difference between the intercepts is given by b. If $b_1 > 0$, then for any given number of years of experience males make higher wages than females. If $b_1 < 0$, then for any given number of years of experience male make lower wages than females. The difference between the slopes is given by b_3, which is the coefficient of the interaction term between experience and gender. If $b_3 > 0$, then an additional year of experience for males results in a bigger wage increase than for males. If $b_3 < 0$, then an additional year of experience for males results in a smaller wage increase for males than for females

Estimation of the Statistical Model

To obtain estimates of the population parameters, we can regress the monthly wage on the dummy variable for gender and the number of years of experience, and the interaction term between experience and gender, using OLS.

Hypothesis Tests

To test the hypothesis of no gender discrimination, we can test the following restrictions on the parameters of the statistical model.

$$H_o: b_1 = b_3 = 0$$
$$H_1: \text{At least one is non-zero}$$

Alternative Specification of the Model

It can be shown that the coefficient estimates of the varying intercept and slope parameter model given by

$$Y_t = a + b_1 G_t + b_2 E_t + b_3 E_t G_t + \varepsilon_t$$

are exactly the same as the coefficient estimates that we would obtain if we estimated two separate wage equations: one wage equation for males, and one wage equation for females. Thus, we could specify this model equivalently as follows.

Wage equation for males: $Y_t = \alpha_o + \alpha_1 E_t + \varepsilon_t$, where $\alpha_o = (a + b_1)$, and $\alpha_1 = (b_2 + b_3)$.

Wage equation for female: $Y_t = a + b_2 E_t + \varepsilon_t$.

Thus, we have a sample of 49 workers: 25 male workers, and 24 female workers. We can regress the wage of males on years of experience for males for the subsample of 25 male workers. We can regress the wage of females on years of experience for females for the subsample of 24 workers.

Difference between the Two Models

The only difference between the two models involves the estimation of the error variance σ^2. If the error variance for males is the same as the error variance for females, then using the dummy variable model will give us a more efficient estimate of σ^2. This estimate is $ESS/(n-k)$, where ESS is the residual sum of squares and $n - k$ is the number of degrees of freedom for the dummy variable model.