Categorical Predictor Variables

* We often wish to use categorical (or qualitative) variables as covariates in a regression model.

* For binary variables (taking on only 2 values, e.g. sex), it is relatively easy to include them in the model.

* Usually one level is coded as 0 and the other as 1 and then the variable can be put into the model as normal.

* The interpretation of the estimate is slightly different, however.

A Single Binary Predictor

* Suppose that we have a response variable Y and a single binary variable Z coded as 0 and 1.

* If we fit a simple linear model $Y = \beta_0 + \gamma_1 X + \varepsilon$ to this data then we find that the least squares estimates are

$$\widehat{\beta}_0 \;=\; \overline{y}_0 \qquad \widehat{\gamma}_1 \;=\; \overline{y}_1 - \overline{y}_0$$

  where $\overline{y}_k$ is the mean of all the y values corresponding to observations with $x_i = k$, $k = 0, 1$.

* Furthermore we can write the residuals as

$$e_i \;=\; \begin{cases} y_i - \overline{y}_0 & \text{if } x_i = 0 \\ y_i - \overline{y}_1 & \text{if } x_i = 1 \end{cases}$$

A Single Binary Predictor

* The MSE, $\hat{\sigma}^2$, is then the pooled estimator of variance (s_p^2) in a two-sample situation with equal variances.
* We also see that we can write

$$\frac{1}{S_{xx}} = \frac{1}{n_0} + \frac{1}{n_1}$$

where n_k is the number of observations with $x_i = k$, $k = 0, 1$.

* Hence we have

$$t = \frac{\hat{\gamma}_1}{\text{se}(\hat{\gamma}_1)} = \frac{\overline{y}_1 - \overline{y}_0}{\sqrt{\left(\frac{1}{n_0} + \frac{1}{n_1}\right) s_p^2}}$$

which is the two-sample t-statistic for comparing the means of two groups and has a t_{n-2} distribution when the two population means are equal.

A Single Factor with $k > 2$ levels

* Many categorical variables have more than two levels.

* Even though they may be coded as consecutive integers we generally cannot treat them as such.

* To do so would be to assume that
  1. There is a unique ordering of the levels of the variable.
  2. The change in the response variable is the same for any change of one level in the covariate.

* These assumptions are very restrictive and generally not true.

A Single Factor with $k > 2$ levels

* Instead we need to create Dummy Variables.

* A dummy variable is a binary variable coded as 0's and 1's.

* A dummy variable for level j of a categorical variable is defined as

$$U_{i,j} = \begin{cases} 1 & \text{if } z_i = j \\ 0 & \text{if } z_i \neq j \end{cases}$$

A Single Factor with $k > 2$ levels

* Note that we only need $k - 1$ dummy variables to uniquely code for a categorical variable with k levels.

* The level for which we do not produce a dummy variable is called the base category or control group. It is the level to which all other levels are compared.

* It does not matter which level we designate as the base category.

* Effectively all other categories will be compared to this base category.

A Single Factor with $k > 2$ levels

* Suppose we select $Z = k$ as the base category then we fit the model

$$Y_i = \beta_0 + \gamma_1 U_{i,1} + \gamma_2 U_{i,2} + \cdots + \gamma_{k-1} U_{k-1,i} + \varepsilon_i$$

* It is easy to show that the least squares estimates are

$$\hat{\beta}_0 = \overline{y}_k \qquad \hat{\gamma}_j = \overline{y}_j - \overline{y}_k \quad j = 1, \ldots, k-1.$$

* Usually we are interested in testing if the mean of the response is the same for each group defined by the categories of Z.

* This is equivalent to testing $H_0 : \gamma_1 = \cdots = \gamma_{k-1} = 0$ and can be done from the F ratio in the ANOVA Table.

* This technique is usually called One-Way ANOVA.

One Continuous and One Categorical Covariate

* In most applications there are both continuous and categorical predictors.

* Consider the case in which there is a single binary covariate Z and a single continuous covariate X.

* The usual linear model is then

$$\begin{aligned} Y_i &= \beta_0 + \gamma_1 Z_i + \beta_1 X_i + \varepsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 X_i + \varepsilon_i & \text{if } Z_i = 0 \\ \beta_0 + \gamma_1 + \beta_1 X_i + \varepsilon_i & \text{if } Z_i = 1 \end{cases} \end{aligned}$$

One Continuous and One Categorical Covariate

* So the linear model can be thought of as two linear models with different intercepts but the same slope of the quantitative variable.

* The interpretation of β_1 is the slope of the linear relationship between Y and X as before.

* The interpretation of γ_1 is the change in the intercept of the line when comparing the group with $Z_i = 1$ to those with $Z_i = 0$.

* Thus the test of $\gamma_1 = 0$ is a test of whether the same line relating Y to X can be used for the two groups.

One Continuous and One Categorical Covariate

* If the Categorical Variable has more than 2 levels then a similar thing occurs.

* The coefficient of the dummy variable for level j is then the change in the intercept of the line for observations with $Z_i = j$ compared to the base category.

* A test of whether the same line is appropriate for all levels of Z can be done by fitting the reduced model with X alone and comparing the residual sums of squares as we did on slide 3-29.

* The linear model with one continuous and one categorical covariate is often called Analysis of Covariance (ANCOVA).

More General Settings

* When we have one categorical variable and multiple continuous variables, the interpretation is the same. The only thing that changes for different levels of the categorical variable is the intercept in the model.
* The effect of the continuous covariates is assumed to be the same for all levels of the categorical variable.
* When we have more than one categorical variable there are more groups and more different intercepts.
* This is called an additive structure since the change in intercept for a combination of levels from the different factors is found by summing together the individual changes for each categorical variable.

Interactions

* We have made two major assumptions so far.
  1. Levels of a categorical variable only alter the intercept of the model, not its slope on any of the continuous covariates.
  2. The change in the intercept is additive over multiple categorical variables.

* In many cases one, or both, of these assumptions may not be reasonable for our data.

Interactions

* For example, it may be that one's initial salary depends on education level (categorical) and also that increases with years of experience (continuous) also depend on education level.

* In that case we may want different slopes for different education levels.

* The additive model assumes, for example, that the effect of management status on salary is the same (in $ amount) for each education level. This may also be untrue.

* Interaction terms can be used to expand the linear model to deal with these situations.

Interaction of 2 Categorical Variables

* Interaction terms are products of variables in the regression model.

* First consider the case of two categorical variables each with two levels.

* These are represented by two 0/1 variables and so their product is also a 0/1 variable which is 1 if, and only if, both of the categorical variables are 1.

* By including this product in the linear model, we allow the effect of one of the binary variables to be different depending on the level of the other binary variable.

Interaction of 2 Categorical Variables

* Suppose that X is continuous and Z_1 and Z_2 are binary and consider the model

$$Y_i = \beta_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \alpha_{1,2} Z_{1i} Z_{2i} + \beta_1 X_i + \varepsilon_i$$

* We can write the conditional expectation of Y in the following way

Z_1	Z_2	$\mathsf{E}\big(Y \mid X, Z_1, Z_2\big)$
0	0	$\beta_0 + \beta_1 X$
1	0	$\beta_0 + \gamma_1 + \beta_1 X$
0	1	$\beta_0 + \gamma_2 + \beta_1 X$
1	1	$\beta_0 + \gamma_1 + \gamma_2 + \alpha_{1,2} + \beta_1 X$

* The $\alpha_{1,2}$ parameter shows how much the intercept in the model deviates from the additive model for the two variables.

Interaction of 2 Categorical Variables

* When one, or both, of the categorical variables have more than two categories, then they are represented by dummy variables.

* There are then a number of interaction parameters, one for each product of a dummy variable for Z_1 and a dummy variable for Z_2.

* For example in the case where both variables have 3 categories, there will be $2 \times 2 = 4$ interaction terms.

* These four interaction terms allow the 9 possible intercepts to deviate from the additive structure which is present when only the main effects are modeled.

Interaction of 2 Categorical Variables

* Including interactions like this increases the number of parameters to be estimated and so reduces the error degrees of freedom.

* In many instances, the additive structure fits the data quite well.

* We can fit the model with and without the interaction terms and use the usual F test for comparing a reduced model to a full model to test this.

Interaction Between a Categorical and Continuous Variable

* In our discussion to date, the only thing that is affected by the categorical variables and their interactions is the intercept term.

* The slope for any continuous variable is assumed the same for any combination of levels of the categorical variables.

* In some cases, however, we may want to allow for the possibility that the slope of a continuous variable is different for different levels of a categorical variable.

* We can do this by allowing interactions between categorical and continuous variables in our model.

Interaction Between a Categorical and Continuous Variable

* Consider a single binary variable Z and a single continuous covariate X for response variable Y.

* The full model which allows for different lines to be fitted to the two levels of Z is

$$Y_i = \beta_0 + \gamma_1 Z_i + \beta_1 X_i + \delta_1 Z_i X_i + \varepsilon_i$$

* This corresponds to the pair of models

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i && \text{if } Z_i = 0 \\ Y_i &= (\beta_0 + \gamma_1) + (\beta_1 + \delta_1) X_i + \varepsilon_i && \text{if } Z_i = 1 \end{aligned}$$

Interaction Between a Categorical and Continuous Variable

* Testing if the interaction term in the previous model is different from 0, corresponds to a test of whether the two slopes are equal.

* When the categorical variable only has two categories, this is the usual t test but if it has more than 2 then an F-test comparing the full model with a reduced one is needed.

* We could also fit a simple linear regression model and test if Z, or its interaction, is needed in the model.

* Again this would be an F-test comparing the reduced model (without Z or $Z \cdot X$) and the full model (with both Z and $Z \cdot X$).

* Note that, in general, you should always include main effects for any variables that are used in an interaction.

Interaction Between a Categorical and Continuous Variable

* Note that the full model described above is not identical to the model which separately fits lines in each of the levels of Z.

* The estimates of intercept and slope would be identical with the separate regressions but the standard errors would be different.

* This is because the single model assumes that the variance is the same for all values of all covariates, including the categorical covariate.

Interaction Between a Categorical and Continuous Variable

* Fitting separate regressions would require the error variances to be homogeneous within levels of Z but allow them to be different for different levels of Z.

* Separate regressions, however, do no allow for an easy test of the reduced models.

* If we conclude from our test that there are different intercepts and slopes for each of the levels of Z then it is generally a good idea to fit the separate regressions at that point to see if different covariates are important in different levels of Z.