GENERAL LINEAR MODELS

Multiple linear regression model

The separate-means model is a special case of the multiple linear regression model,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

where the β's are parameters and the x's are explanatory variables.

The separate-means model is a special case of the multiple linear regression model, where the beta's are parameters and the x's are explanatory variables.

Example. The regression model corresponding to the blood pressure example

Suppose y is the change in blood pressure, and x_1 and x_2 indicate which dose group the subject is in:

$$x_1 = \begin{cases} 1 & \text{if the observation is from group 1} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the observation is from group 2} \\ 0 & \text{otherwise} \end{cases}$$

Variables like x_1, x_2 are called **indicator variables**: their values indicate whether an observation is a member of group 1 or group 2 respectively.

In this example, if both $x_1 = 0$ and $x_2 = 0$, it indicates the observation is from group 3. (We do not need a x_3! That will be redundant.)

In general, we need $t - 1$ indicator variables to code t group memberships.

The general linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e.$$

Let's use the blood pressure data set as example to show the connection between the separate-means model and a general multiple linear regression model.

Suppose y is the response variable—the change in blood pressure.

We can use two indicator variables (x 1) and (x 2) to code the the factor levels. In this example, we let (x 1) equals 1 if the observation is from group 1, and 0 otherwise; we let (x 2) equals 2 if the observation is from group 2, and 0 otherwise.

Variables like (x 1), (x 2) are called indicator variables: their values indicate whether an observation is a member of group 1 or group 2 respectively.

In this example, when both (x 1) and (x 2) are 0, the observation must be from group 3. Note that we do not need a (x 3): that will be redundant. In fact, if you include redundant variables in a regression model, it will create numerical issues.

In general, we need (t -1) indicator variables to code t group memberships.

With these two indicator variables, we can define a regression model

y equals (beta 0) plus (beta 1)(x 1) plus (beta 2)(x 2) plus e.

The correspondence between the regression model and the cell-means model

	Value of		Mean response	
Group	x_1	x_2	Cell means model	General model
1	1	0	μ_1	$\beta_0 + \beta_1$
2	0	1	μ_2	$\beta_0 + \beta_2$
3	0	0	μ_3	β_0

So, $\beta_0 = \mu_3$, $\beta_1 = \mu_1 - \beta_0 = \mu_1 - \mu_3$, and $\beta_2 = \mu_2 - \beta_0 = \mu_2 - \mu_3$.

Compare the multiple-linear regression model to the separate-means model. We see that they are simply two different parameterizations of the same model.

This table gives the correspondence between the parameters under the two models or the two parameterizations.

For an observation from group 1, the indicator variable (x 1) will take the value 1 and the indicator variable (x 2) will take the value 0. For this observation, the regression model becomes

y equal beta0 + beta 1 + (an error term)

For an observation from group 2, (x 1) equals 0 and (x 2) equals 1, and the regression model becomes

y equal beta0 + beta 2 + (an error term)

For an observation from group 3, both (x 1) and (x 2) are 0, and the regression model becomes

y equal beta0 + (an error term)

Comparing with the separate-means model, we see that

(beta 0 = mu 3), (beta 1 = mu 1 - mu3) and (beta 2 = mu 2 - mu 3).

In other words, the intercept (beta 0) corresponds to the mean of the one of the groups (we can call that group the reference group). (beta 1) and (beta 2) correspond to group mean difference between group 1, group 2 and the reference group respectively.

Note that the regression model considered here differs from Kuehl, pp. 45-47, who uses k indicators for k levels of a factor.

Understanding regression output

Example. Blood pressure data (cont'd). See `script1` for instructions on obtaining regression output.

Suppose y is the change in blood pressure, dose20 is an indicator for the 20-unit dose group, and dose50 is an indicator for the 50-unit dose group.

Regression output from R:

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.000      2.914   1.029   0.3236
dose20            -7.200      4.121  -1.747   0.1062
dose50           -12.800      4.121  -3.106   0.0091

Residual standard error: 6.517 on 12 degrees of freedom
```

The regression equation is: $\hat{y} = 3.0 - 7.2\,(\text{dose20}) - 12.8\,(\text{dose50})$. Therefore, the predicted change in blood pressure is $3.0 - 7.2 = -4.2$ for dose 20; $3.0 - 12.8 = -9.8$ for dose 50; and 3.0 for the placebo group. These are the cell means that we saw earlier.

The 'residual standard error' is $\sqrt{s^2} = \sqrt{\text{MSE}}$. So MSE $= 6.517^2 = 42.47$, which agrees with the ANOVA table from earlier. □

Here is the regression output from R after fitting the regression model to the blood pressure data.

See script1.html for details on how to fit this regression model, summarize the regression output, and interpret the results.

Learning to analyze data using R is an integral part of this class. Throughout this class, I will provide you with many R scripts. It is important that you study these R scripts. Experiment with them, try them on a different data set, add your own notes, and so on.

You can use these R scripts as a reference when doing homework problems. More importantly, you want to organize these R scripts into your own collection. So in future, when you need to analyze your experiment, you have arsenal of tools that you can use.

Identify the regression equation

From the regression output, we can identify the regression equation presented as

$$\hat{y} = 3.0 - 7.2(\text{dose20}) - 12.8(\text{dose50}).$$

where (dose 20) and (dose 50) are two indicator variables for group 2 and group 3.

Note that in the fitted regression model, the reference group is group 1, the no-dose control group (not group 3). When fitting the regression model, you can choose which group to use as reference group.

Based on the fitted regression model, the predicted change in blood pressure is $3.0 - 7.2 = -4.2$ for dose 20; $3.0 - 12.8 = -9.8$ for dose 50; and 3.0 for the placebo group. These are the cell means that we saw earlier.

From the regression output, we can identify the regression equation as

(y hat equals 3.0 - 7.2 (dose 20) - 12.8 (dose 50))

where (dose 20) and (dose 50) are two indicator variables for group 2 and group 3.

Note that in the fitted regression model, the reference group is group 1, the no-dose control group (not group 3). **When fitting the regression model, you can choose which group to use as reference group.**

Based on the fitted regression model, the predicted change in blood pressure is (3.0 - 7.2 = -4.2) for dose 20; (3.0 - 12.8 = -9.8) for dose 50; and 3.0 for the placebo group. These are the cell means that we saw earlier.

The residual standard error and MSE

In the regression output, you will see

Residual standard error: 6.517 on 12 degrees of freedom

The relationship between the residual standard error (RSE) from the regression output and the mean square error (MSE) from the ANOVA output:

$$MSE = RSE^2$$

We have seen from the ANOVA output that the MSE from this fitted model is 42.47, which is 6.517^2.

Note: Once you have fit the linear regression model for a one-factor design, you can summarize the results either with the regression output using the R command `summary` or with the ANOVA output using the R command `anova`. (See `script1.html` for details.)

In the regression output, you will see

Residual standard error: 6.517 on 12 degrees of freedom

The relationship between the residual standard error (RSE) from the regression output and the mean square error (MSE) from the ANOVA output is that the mean square error is the square of the residual standard error.

We have seen from the ANOVA output that the MSE from this fitted model is 42.47, which is 6.517 squared.

Note that once you have fit the linear regression model for a one-factor design, you can summarize the results either with the regression output using the R command "summary" or with the ANOVA output using the R command "anova". See "script1.html" for details.)

t-test for regression coefficient

Note that regression coefficient for (dose 20) correspond to the mean difference in response between the group (does 20) and the reference group.

So the t-test for that coefficient is for testing the group mean difference between the group (does 20) and the reference group (no-dose).

From the regression output, we see that the p-value for this test is 0.1062. This is equivalent to the corresponding test we did in the previous lecture "Treatment comparison". We conclude that there is no significant difference in the mean blood pressure changes in these two groups.

Note that regression coefficient for (dose 20) correspond to the mean difference in response between the group (does 20) and the reference group.

So the t-test for that coefficient is for testing the group mean difference between the group (does 20) and the reference group (no-dose).

From the regression output, we see that the p-value for this test is 0.1062. This is equivalent to the corresponding test we did in the previous lecture "Treatment Comparison". We conclude that there is no significant difference in the mean blood pressure changes in these two groups.

Summary

The connection between the regression model and the separate-means (cell-means) model:

1. How to use indicator variables to code factor levels.
2. Study the R code and notes in script1.html.
3. From the R regression output:
   - how to identify the fitted regression model
   - how to interpret the regression coefficients and corresponding t tests
4. Compare the regression output to the ANOVA output. Both outputs are useful—they answer different questions.

In this lecture, we discussed the connection between the general regression model and the separate-means (cell-means) model:

In particular, we discussed how to use indicator variables to code factor levels.

It is important that you study script1.html and learn how to fit the regression model in R.

From the regression output, you should be able to identify the regression equations and know how to correctly interpret each regression coefficient.

For a regression model fitted to a data set from one-factor experiment, you can summarize the results either using the regression output or using the ANOVA output. Both output are useful. They help answer different questions.