Indicator linear models

Questions of interest

Suppose that we have two independent continuous random variables $Y_1, Y_2$ with random variable means $\mu(Y_1), \mu(Y_2)$. Common questions of interest:

- Are the random variable means of $Y_1$ and $Y_2$ equal?
- What is the difference $\mu(Y_2) - \mu(Y_1)$ in random variable means?

Examining the data

Fitting the model

Sampling Variability Assumptions

Statistical inference

Worked examples

Estimating a mean

As listed on page 95, two common types of statistical questions of interest for constant linear model analyses arise as follows:

- Are the random variable means of $Y_1$ and $Y_2$ equal?
- What is the difference $\mu(Y_2) - \mu(Y_1)$ in random variable means?
We traditionally address the first of these with what is called a **2-sample t test**. The method ordinarily used to address the second doesn’t really have a name but might be called computing a **2-sample t confidence interval**.

The next example illustrates both of these questions.

**Example: Masses of blue and orange Plain M and M candies.**

Do blue and orange Plain M&M candies have different average masses? Intuition would suggest that they don’t. Use the M&M candies data set to investigate this question statistically.

From the statement of the problem, we determine two closely related non-statistical questions of interest:

1. Does the average mass of a blue Plain M&M equal the average mass of an orange Plain M&M?

2. What is an estimate of the average mass of an orange Plain M&M minus the average mass of a blue Plain M&M?

We translate these into two corresponding statistical questions of interest. We state these questions in terms of the variables of interest: color, a categorical variable with levels Blue and Orange; and mass, a random variable that gives the mass of a randomly chosen M&M from all of the M&Ms in the world. We also use the notation that colorOrange is an indicator variable for the condition that color = Orange. Also, we use $\beta_1$ to denote the coefficient of colorOrange of the (unknown) true linear model of class

$$\text{mass} \sim \text{color}$$

with reference level color = Blue. With this notation, the statistical questions of interest are:

1. What are the results of a t test of:

$$H_{null} : \beta_1 = 0 \text{ grams} \quad \text{and} \quad H_{alt} : \beta_1 \neq 0 \text{ grams},$$

   evaluated at the usual significance level of 0.05?

2. What are an estimate and a 95% t confidence interval for $\beta_1$?

From the description of the data set, the source seems objective and reliable enough. The sample consists of blue and orange M&Ms from a
single bag of Plain M&Ms. For each level of color (Blue and Orange), this is certainly not a simple random sample of all the Plain M&Ms of that color in the world. However, for the purposes of our analysis, we make the simplifying assumption that our blue and orange samples are simple random samples of Plain M&Ms of those two colors. This means that for each color, the masses of these M&Ms constitute a sample of the two random variables mass|color (one for each level of color).

However, our samples of these two colors are not completely unrelated to each other, since both are from the same bag of M&Ms. For the purposes of our analysis though, we assume that they are unrelated, so that the random variables mass|(color = Blue) and mass|(color = Orange) are independent.

With these assumptions, we are ready to proceed. We next make density plots of our samples of mass|(color = Blue) and mass|(color = Blue), looking for anything unusual or noteworthy.

This plot reveals no unusual patterns or outliers. Perhaps its most noteworthy aspect is that, as we would expect, the distributions are quite similar and overlap heavily. Therefore our statistical inference procedures should not detect a difference between the average masses of the two types of M&Ms. We will soon verify this.
To get another view of how similar these two distributions are though, we can also look at side-by-side boxplots.

![Boxplots of masses for blue and orange objects](image)

As we saw with the densities, these two boxes and whiskers overlap heavily and are very similar, so we shouldn’t be able to detect a difference in the average masses using statistical inference.

To verify this, we first use a computer to fit the model of class

$$mass \sim color_{Orange}.$$  

To answer the first statistical question of interest, we should conduct a hypothesis test of whether the true coefficient $\beta_1$ of color_{Orange} equals 0. For this, we must first verify the Sampling Variability Assumptions for the model class. To do so, we first use a computer to fit the model of this class to our data.

The two Sampling Variability Assumptions for models of this class are:

- Normality of the conditional error terms and constant variance of the conditional error terms.

This fitted model has two conditional error terms, one for each level of color: $\hat{e}(\text{mass}|(\text{color} = \text{blue}))$ and $\hat{e}(\text{mass}|(\text{color} = \text{orange}))$. To assess the normality of each of these, we examine normal quantile plots of the conditional residuals associated with each one.
If these conditional error terms are normally distributed, we expect about 95% of the points in each sample to be within the 95% confidence bands (shown by dashed lines). In these plots, we see at least that many points within the bands, so the conditional error terms do appear to be normally distributed.

In addition, a Shapiro-Wilk test of the normality of $\hat{e}(\text{mass} \mid \text{color} = \text{blue})$ gives a $p$-value of 0.256; one for the normality of $\hat{e}(\text{mass} \mid \text{color} = \text{orange})$ gives a $p$-value of 0.544. With the traditional significance level of 0.05, in neither case do we have statistically significant evidence that the conditional error term is non-normally distributed. This is in keeping with what we see in the normal quantile plots as well.

Both graphically and numerically, the two conditional error terms appear to be at least approximately normally distributed, satisfying the first Sampling Variability Assumption.

The second Sampling Variability Assumption is that the conditional error terms have equal variances. To verify this assumption, we examine a residuals versus fitted plot.
For indicator models, residuals versus fitted plots are not the best tools to assess equal variance of conditional residuals. However, for more complicated classes of models, such plots will prove to be useful, so we introduce them here.

We should note that in this plot there are only two fitted values. Fitted values for indicator models are just sample means for the different levels of the indicator variable, so there will always be exactly two of them. More complicated classes of models will generally have lots of fitted values though.

The above residuals versus fitted plot is not particularly informative because there is plenty of overplotting, so it is hard to see the distribution of each conditional residual. With what little information we can glean from this plot though, the variances of the two conditional error terms do at least seem to be similar, since the two vertical stripes seem to have approximately the same spread.

To be able to compare the variances of the conditional error terms better, we examine density plots of them. Such density plots may not be available to us for more complicated classes of models, but they are very useful for indicator models.
From this picture, we can see that the variances of the two distributions are quite similar. They are not exactly the same in the picture, but the density curves there are only estimates based on the data. We would expect some variation in these density curve estimates even for samples taken from the same random variable, so the plot above does seem to indicate equal or at least very similar variances for the two conditional error terms. In other words, the second Sampling Variability Assumption appears to be satisfied, so we can proceed to draw statistical inferences.

To answer the first statistical question of interest, we conduct the specified hypothesis test. For this, we first compute the value $t_1$ of the test statistic $T_1$ for our sample of mass. The theorem PUT REF IN gives us the formula for $t_1$ as

$$t_1 = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{0.00457 - 0}{0.00759} = 0.602.$$ 

Since we have a total of 162 observations, the theorem also tells us that the distribution of $T_1$ under the null hypothesis is $t_{160}$. Computing the area of the outer tails at 0.602 of $t_{160}$, we obtain a $p$-value of:

$$p = 0.548.$$
Since this is above our chosen significance level of 0.05, we do not have statistically significant evidence that the true coefficient $\beta_0$ is different from 0 grams.

We next address the second statistical question of interest. Using least squares as our optimality criterion, our best estimate of this true model coefficient is the fitted model coefficient $\hat{\beta}_1$, which equals 0.00457 grams.

To compute a 95% confidence interval for the true model coefficient $\beta_0$, we use the formula given in the theorem PUT REF HERE The computer gives us the values that we need for this formula:

$$\hat{\beta}_1 = 0.00457$$
$$\text{se}(\hat{\beta}_1) = 0.00759$$
$$t^* = 1.975,$$

where $t^*$ is the central .95 quantile for a $t_{160}$ distribution. Putting these numbers into the formula, we compute that a 95% confidence interval for the average orange Plain M&M mass minus the average blue Plain M&M mass is from $-0.0104$ grams to $0.0196$ grams.

In conclusion, with our hypothesis test we did not find statistically significant evidence that the average mass of an orange Plain M&M is different from the average mass of a blue Plain M&M. We estimate the average orange Plain M&M mass minus the average blue Plain M&M mass to be 0.0046 grams, with a 95% confidence interval from $-0.0104$ grams to $0.0196$ grams. We should remember, however, our samples came from only a single bag of Plain M&Ms and were therefore far from simple random samples of all the blue and orange Plain M&Ms in the world. To make our statistical inferences, we made the false but simplifying assumption that these were simple random samples, which could affect whether our results are correct.