28.3: The t-test as a Linear Model

The t-test is often presented as a specialized tool for comparing means, but it can also be viewed as an application of the general linear model. In this case, the model would look like this:

$$\widehat{BP} = \widehat{\beta_1} * Marijuana + \widehat{\beta_0}$$

However, smoking is a binary variable, so we treat it as a *dummy variable* like we discussed in the previous chapter, setting it to a value of 1 for smokers and zero for nonsmokers. In that case, $\widehat{\beta_1}$ is simply the difference in means between the two groups, and $\widehat{\beta_0}$ is the mean for the group that was coded as zero. We can fit this model using the `lm()` function, and see that it gives the same t statistic as the t-test above:

```
##
## Call:
## lm(formula = TVHrsNum ~ RegularMarij, data = NHANES_sample)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.293 -1.133 -0.133  0.867  2.867
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.133      0.119   17.87   <2e-16 ***
## RegularMarijYes    0.660      0.249    2.65   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.5 on 198 degrees of freedom
## Multiple R-squared:  0.0343, Adjusted R-squared:  0.0295
## F-statistic: 7.04 on 1 and 198 DF,  p-value: 0.00861
```

We can also view the linear model results graphically (see the right panel of Figure 28.1). In this case, the predicted value for nonsmokers is $\widehat{\beta_0}$ (2.13) and the predicted value for smokers is $\widehat{\beta_0} + \widehat{\beta_1}$ (2.79).

To compute the standard errors for this analysis, we can use exactly the same equations that we used for linear regression – since this really is just another example of linear regression. In fact, if you compare the p-value from the t-test above with the p-value in the linear regression analysis for the marijuana use variable, you will see that the one from the linear regression analysis is exactly twice the one from the t-test, because the linear regression analysis is performing a two-tailed test.

28.3.1 Effect sizes for comparing two means

The most commonly used effect size for a comparison between two means is Cohen's d, which (as you may remember from Chapter 18) is an expression of the effect in terms of standard error units. For the t-test estimated using the general linear model outlined above (i.e. with a single dummy-coded variable), this is expressed as:

$$d = \frac{\widehat{beta_1}}{SE_{residual}}$$

We can obtain these values from the analysis output above, giving us a d = 0.45, which we would generally interpret as a medium sized effect.

We can also compute R^2 for this analysis, which tells us how much variance in TV watching is accounted for. This value (which is reported in the summary of the lm() analysis) is 0.03, which tells us that while the effect may be statistically significant, it accounts for relatively little of the variance in TV watching.