5.3 Qualitative Predictor with 2-Levels

In this part, we focus on variables with only two levels 0 or 1, and we consider that we have one continuous variable X and one categorical D in the model.

General ANCOVA Model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + \varepsilon_i$$

where $\varepsilon_i \sim^{IID} \mathcal{N}(0, \sigma^2)$. The term $x_i d_i$ is the product of the two random variables X and D and is called the **interaction** term.

5.3.1 Interpretation of Regression Coefficients

Depending on the significance of the coefficients in the general model above, we may have several different models. So, let us start by discussing what are the possible regression models that derive from the general one:

1. Coincident regression lines
   This is the case where the categorical variable D has **no effect** on the response Y, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } d = 0 \text{ or } 1$$

   Hence, the regression line is the same for both groups of the variable D.

1'. Two-mean model
   This is the case where the continuous variable X has **no effect** on the response Y, i.e.

$$y_i = \beta_0 + \beta_2 d_i + \varepsilon_i = \begin{cases} \beta_0 + \varepsilon_i, & \text{if } d = 0 \\ (\beta_0 + \beta_2) + \varepsilon_i, & \text{if } d = 1 \end{cases}$$

2. Parallel regression lines

   This is the case where the categorical variable D **only** changes the intercept, or in other words it produces an **additive effect**:

$$y_i = \beta_0 + \beta_2 d_i + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i, & \text{if } d = 0 \\ (\beta_0 + \beta_2) + \beta_1 x + \varepsilon, & \text{if } d = 1 \end{cases}$$

   In this case, the design matrix becomes:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 1 & x_3 \\ 1 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

   The coefficient β_2 measures the change of the additive effect, i.e., difference of the intercept.

3. Regression lines with equal intercepts but different slopes

   This is the case where the categorical variable D only changes the effect of X on Y:

$$y_i = \beta_0 + \beta_1 x_i + \beta_3 x - i d_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 x + \varepsilon, & d = 0 \\ \beta_0 + (\beta_1 + \beta_3) x + \varepsilon, & d = 1 \end{cases}$$

   The design matrix in this case is written as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ 1 & x_3 & x_3 \\ 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

   The coefficient β_3 measures the change of the slope.

4. Unrelated regression lines

   In this case, the categorical variable D produces an *additive* change in Y and also *changes the effect of* X on Y.

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x.d) + \varepsilon = \begin{cases} \beta_0 + \beta_1 x + \varepsilon, & d = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x + \varepsilon, & d = 1 \end{cases}$$

The design matrix becomes

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 & 0 \\ 1 & 0 & x_2 & 0 \\ 1 & 1 & x_3 & x_3 \\ 1 & 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

Here, β_1 measures the effect of x on y when other predictors are held unchanged. This interpretation does not make much sense for models with interactions, since we cannot change x while holding d and $x * d$ unchanged.

Remark: *Why not just divide the data into two sets and run two separate regressions?*
There are two reasons for this. First, the sample size used to fit the general model will be larger than the sample size for each of the levels of the factor. This affects all inferences, such as the estimation of βs which can be made more precise since more degrees of freedom will be then associated with $MSE = RSS/n - p$. Secondly, we are able to explore the significant of interactions, which is not possible when we consider two separate models. In addition, even when an additive model is under consideration, i.e. a model that assumes equal slopes and the same constant error term variance for each level, the common slope β_1 can be estimated by pooling the two levels together.

Birthweight Example

In this example we are going to focus on three variables: `Birthweight` as the response, and `Headcirc` and `smoker` (smoker/non-smoker) as predictors. The `smoker` variable is a **categorical** variable that takes `0` if the mother is a non-smoker and `1` if it is a smoker.

We start our analysis by fitting the *full* model, i.e. a model that contains both predictors (`Headcirc` , `smoker`), as well as their interaction (`Headcirc:smoker`). Essentially, the full model is a model with *different intercepts* and *different slopes* for each of the groups.

```
birthweight.full = lm(Birthweight ~ Headcirc +  smoker + Headcirc:smoker, data=bi
summary(birthweight.full)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc + smoker + Headcirc:smoker,
##     data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76807 -0.29123 -0.05411  0.19927  1.18861
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.43447    1.69271  -0.847  0.40206
## Headcirc         0.14105    0.04821   2.926  0.00577 **
## smoker          -1.44641    2.10352  -0.688  0.49587
## Headcirc:smoker  0.03492    0.06043   0.578  0.56682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4391 on 38 degrees of freedom
## Multiple R-squared:   0.51,  Adjusted R-squared:  0.4713
## F-statistic: 13.18 on 3 and 38 DF,  p-value: 4.767e-06


# same as lm(Birthweight ~ Headcirc*smoker , data=birthweight2)
```

We can write down the two regression lines corresponding to smoker/non-smoker mothers:

* Non-Smoker

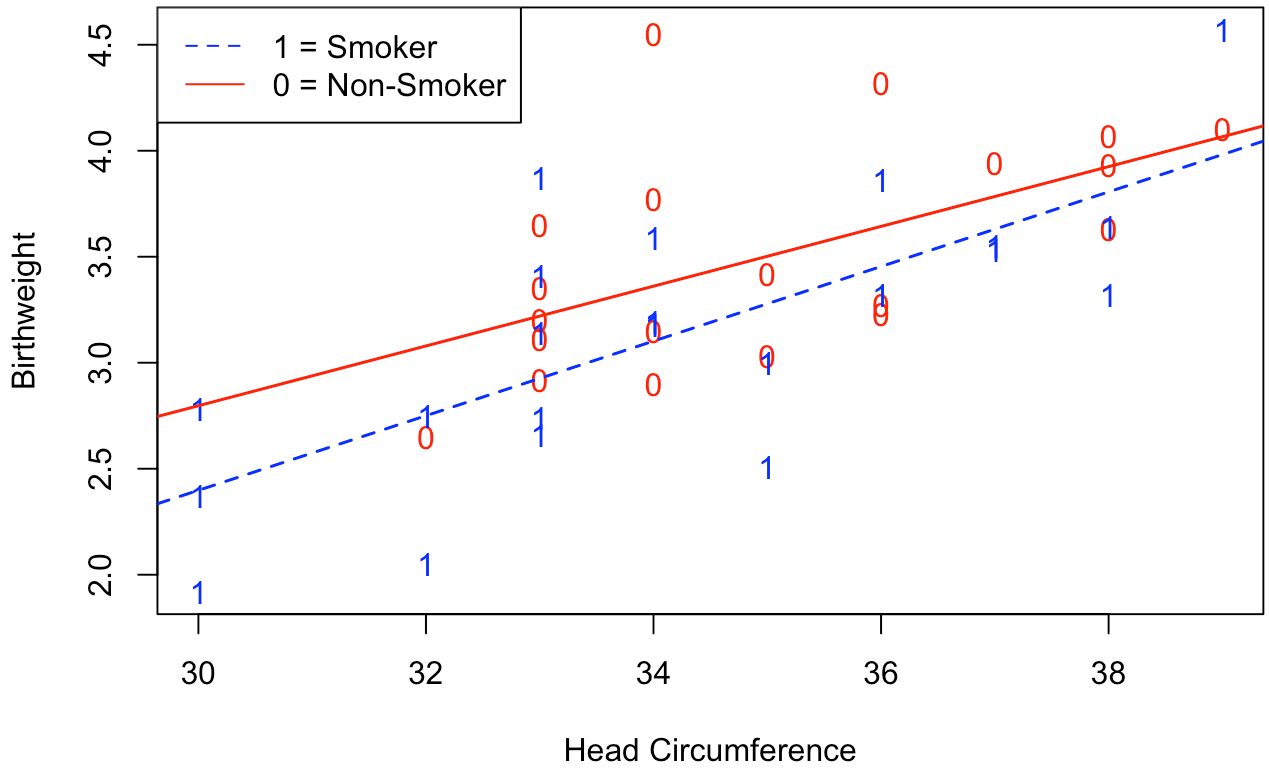
$$\hat{y}_i^n = -1.43447 + 0.14105 \cdot x$$

- Smoker

$$\hat{y}_i^s = (-1.43447 - 1.44641) + (0.14105 + 0.03492) \cdot x$$

We can also plot the two lines:

```
plot(birthweight2$Headcirc, birthweight2$Birthweight,type="n",xlab="Head Circumfe
text(birthweight2$Headcirc[birthweight2$smoker=='0'],birthweight2$Birthweight[bir
text(birthweight2$Headcirc[birthweight2$smoker=='1'],birthweight2$Birthweight[bir
abline(birthweight.full$coef[1], birthweight.full$coef[2], col="red", lty=1, lwd=
abline(sum(birthweight.full$coef[c(1,3)]), sum(birthweight.full$coef[c(2,4)]), co
legend('topleft', c("1 = Smoker","0 = Non-Smoker"), lty=c(2,1), col=c("blue","red
```



The full model is **not** the same as if we were fitting a SLR model separately in each group. The coefficients are the same, but the t-statistics are different since $\hat{\sigma}$ is not the same. Indeed,

```
nonsmoker_model = lm(Birthweight~Headcirc, data=birthweight2[birthweight2$smoker=
summary(nonsmoker_model)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc, data = birthweight2[birthweight2$smoker =
##     "1", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76807 -0.24613 -0.05411  0.33981  0.94387
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.88088    1.25095  -2.303 0.032150 *
## Headcirc     0.17597    0.03649   4.822 0.000104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4399 on 20 degrees of freedom
## Multiple R-squared:  0.5376, Adjusted R-squared:  0.5145
## F-statistic: 23.25 on 1 and 20 DF,  p-value: 0.0001036
```

```
smoker_model=lm(Birthweight~Headcirc, data=birthweight2[birthweight2$smoker=='0',
summary(smoker_model)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc, data = birthweight2[birthweight2$smoker =
##     "0", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47245 -0.31863 -0.05139  0.14715  1.18861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.43447    1.68949  -0.849  0.40700
## Headcirc     0.14105    0.04812   2.931  0.00892 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4383 on 18 degrees of freedom
## Multiple R-squared:  0.3231, Adjusted R-squared:  0.2855
## F-statistic: 8.592 on 1 and 18 DF,  p-value: 0.008922
```

5.3.2 Model Selection

If we take a close look at the *full* model in the `Birthweight` example, we observe that the interaction term is **not** statistically significant. This means that we could consider removing it from the model.

In this section, we are going to discuss how to do model selection in an ANCOVA model with a categorical variable that has two levels.

As in previous cases, we will use partial F-test to select the appropriate model. In this process, we *always* start by testing the interaction term:

$$\begin{cases} H_0 : \text{Additive model } \ (\text{Model } 2) \\ H_\alpha : \text{Full model } \ (\text{Model } 4) \end{cases}$$

If we reject the null, then we **stop** and select the *full model* (Model 4). Otherwise, we select the *additive model* (Model 2), continue and test whether we can further reduce the additive model to a model with only one predictor, i.e. Model 1 or Model $1'$. But, what if β_3 (the interaction) is significant, but β_1 or β_2, is not significant? What about model 3? Can we get a model where the interaction is in the model, while the continuous predictor is not? To answer this questiom, we have the followig rule:

Hierarchical Rule for Interactions

An interaction term will be included in a model only if all its main effects have been included.

Due to this rule, we would include **both** β_1 and β_2 in the model, once β_3 is significant.

Remark: In practice we could test $\beta_1 = 0$ or $\beta_2 = 0$. We just need to understand what the model looks like when β_1 or β_2 equals zero.

- When $\beta_1 = 0$, it *does not mean* that X is not significant:

$$y_i = \begin{cases} \beta_0 + \varepsilon_i, \ \text{if } d_i = 0 \\ (\beta_0 + \beta_2) + \beta_3 x_i + \varepsilon_i, \ \text{if } d_i = 1 \end{cases}$$

- When $\beta_2 = 0$, it *does not mean* D is not significant:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i, \ \text{if } d_i = 0 \\ \beta_0 + (\beta_1 + \beta_3) x_i + \varepsilon_i, \text{if } d_i = 1 \end{cases}$$

Birthweight Example

Since the interaction term is not statistically significant, we remove it and fit the additive model. Indeed,

```
birthweight.additive = lm(Birthweight ~ Headcirc + smoker  , data=birthweight2)
summary(birthweight.additive)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc + smoker, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75768 -0.27776 -0.06271  0.19296  1.21194
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.21345    1.01472  -2.181   0.0353 *
## Headcirc     0.16328    0.02882   5.666 1.51e-06 ***
## smoker      -0.23365    0.13681  -1.708   0.0956 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4353 on 39 degrees of freedom
## Multiple R-squared:  0.5057, Adjusted R-squared:  0.4803
## F-statistic: 19.95 on 2 and 39 DF,  p-value: 1.08e-06
```

If the *additive* model is the final model, then we have two lines that have the same slope, but different intercepts. For example, the two regression lines corresponding to smoker/non-smoker mothers:

- Non-Smoker

$$\hat{y}_i^n = -2.21345 + 0.16328 \cdot x$$

- Smoker

$$\hat{y}_i^s = (-2.21345 - 0.23365) + 0.16328 \cdot x$$

However, the additive model is not the final model here since the p-value for the `smoker` variable is *higher* than 5% which means that we can remove it from the model. So, we have

```
birthweight.Headcirc = lm(Birthweight ~ Headcirc  , data=birthweight2)
summary(birthweight.Headcirc)
```

```
##
## Call:
## lm(formula = Birthweight ~ Headcirc, data = birthweight2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87259 -0.28101 -0.04531  0.24732  1.33969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.6472     1.0057  -2.632    0.012 *
## Headcirc      0.1723     0.0290   5.940 5.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4456 on 40 degrees of freedom
## Multiple R-squared:  0.4687, Adjusted R-squared:  0.4554
## F-statistic: 35.29 on 1 and 40 DF,  p-value: 5.735e-07
```

In this final model, the continuous variable `Headcirc` is statistically significant which means that we stop the model selection process and continue with diagnostics etc.