5.4 Multi-level Categorical Variables

Consider a case where we model the response Y using two predictors X and D, where X is a numerical variable and D is categorical with k levels. As we discussed before, we need to generate $k-1$ dummy variables $D_2, \ldots, D_k$ where:

$$D_i = \begin{cases} 0, & \text{if not level } i \\ 1, & \text{if level } i \end{cases}$$

In this case, Level 1 is the *reference* level. `R` will do this process *automatically* for us. Let's see what we have in the following example:

Fruit Flies Example

The `fruitfly` data frame has 9 rows and 3 columns. 125 fruitflies were divided randomly into 5 groups of 25 each. The response was the longevity of the fruitfly in days.

- One group was kept solitary (`isolated`)
- One group was kept with a virgin female each day (`low`)
- One group was kept with 8 virgin females per day (`high`)
- One group was kept with one pregnant female per day (`one`)
- One group was kept with eight pregnant female per day (`many`)

Pregnant fruitflies will not mate. The thorax length of each male was measured as this was known to affect longevity. One observation in the many group has been lost. So the total sample size is **124**.

```
library(faraway)
data(fruitfly)
head(fruitfly)
```

```
##   thorax longevity activity
## 1   0.68        37     many
## 2   0.68        49     many
## 3   0.72        46     many
## 4   0.72        63     many
## 5   0.76        39     many
## 6   0.76        46     many
```

The levels of the fruitfly variable are shown below:

```
levels(fruitfly$activity)
```

```
## [1] "isolated" "one"      "low"      "many"     "high"
```

So, using our notation, the categorical variable of interest here has the following levels:

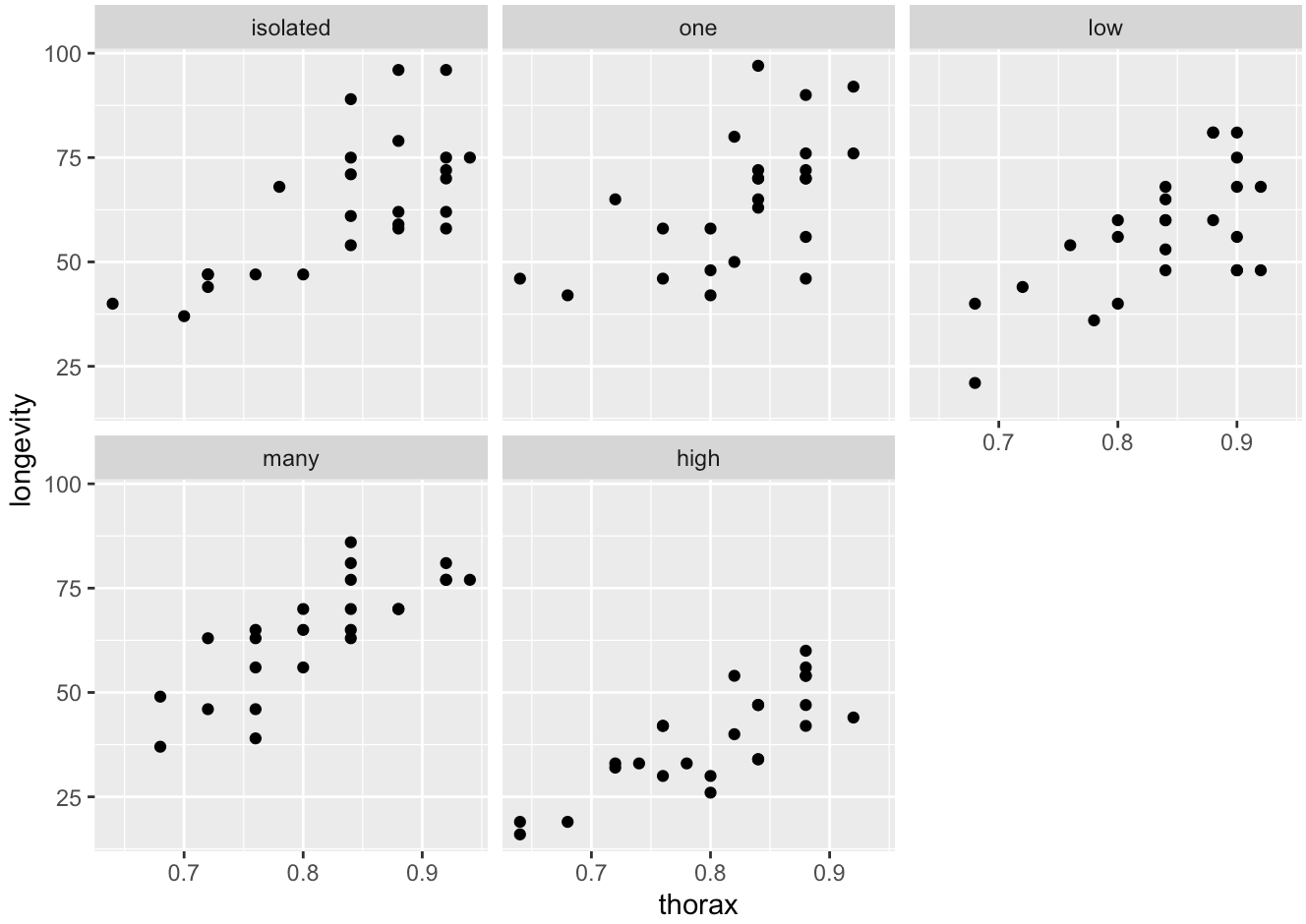
$$D = \begin{cases} \text{isolated} \\ \text{one} \\ \text{low} \\ \text{many} \\ \text{high} \end{cases}$$

and since the input of this variable is `text`, it is automatically understood as a factor by `R` and thus the following 4 (= # levels - 1) dummy/indicator variables will be created. The first level `isolated` will be the reference level when `R` creates the indicator variables.

$$D_2 = \begin{cases} 1, \text{if in level one} \\ 0, \text{otherwise} \end{cases}, \quad D_3 = \begin{cases} 1, \text{if in level low} \\ 0, \text{otherwise} \end{cases}, \quad D_4 = \begin{cases} 1, \text{if in level many} \\ 0, \text{otherwise} \end{cases}, \quad D_5 = \begin{cases} 1 \\ 0 \end{cases}$$

Remark: If the variable was coded using numbers (not text), then we would need to use `as.factor` to alert `R` that it should be treated as a categorical variable, not a continuous one.

```
library(ggplot2)
ggplot(aes(x = thorax, y=longevity), data=fruitfly) + geom_point() +
  facet_wrap(~activity)
```



We start by fitting the most general model, i.e., the model including both variables and their interaction.

```
fruitfly.full = lm(longevity ~ thorax * activity, fruitfly)
```

We can look at the dummy variables created by `R` by obtaining the design matrix of the fitted full model:

```
head(model.matrix(fruitfly.full))
```

```
##   (Intercept) thorax activityone activitylow activitymany activityhigh
## 1           1   0.68           0           0            1            0
## 2           1   0.68           0           0            1            0
## 3           1   0.72           0           0            1            0
## 4           1   0.72           0           0            1            0
## 5           1   0.76           0           0            1            0
## 6           1   0.76           0           0            1            0
##   thorax:activityone thorax:activitylow thorax:activitymany thorax:activityhig
## 1                  0                 0               0.68
## 2                  0                 0               0.68
## 3                  0                 0               0.72
## 4                  0                 0               0.72
## 5                  0                 0               0.76
## 6                  0                 0               0.76
```

We can also obtain the `summary` of the `fruitfly.full` model

```
summary(fruitfly.full)
```

```
##
## Call:
## lm(formula = longevity ~ thorax * activity, data = fruitfly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9509  -6.7296  -0.9103   6.1854  30.3071
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -50.2420    21.8012  -2.305    0.023 *
## thorax              136.1268    25.9517   5.245 7.27e-07 ***
## activityone           6.5172    33.8708   0.192    0.848
## activitylow          -7.7501    33.9690  -0.228    0.820
## activitymany         -1.1394    32.5298  -0.035    0.972
## activityhigh        -11.0380    31.2866  -0.353    0.725
## thorax:activityone   -4.6771    40.6518  -0.115    0.909
## thorax:activitylow    0.8743    40.4253   0.022    0.983
## thorax:activitymany   6.5478    39.3600   0.166    0.868
## thorax:activityhigh -11.1268    38.1200  -0.292    0.771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.71 on 114 degrees of freedom
## Multiple R-squared:  0.6534, Adjusted R-squared:  0.626
## F-statistic: 23.88 on 9 and 114 DF,  p-value: < 2.2e-16
```

The coefficients that are shown in this output corresponds to the following model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 d_2 + \hat{\beta}_3 d_3 + \hat{\beta}_4 d_4 + \hat{\beta}_5 x d_5$$
$$\hat{\beta}_6 x d_2 + \hat{\beta}_7 x d_3 + \hat{\beta}_8 x d_4 + \hat{\beta}_9 x d_5$$

where

$$\hat{\beta}_0 = -50.2420,\ \hat{\beta}_1 = 136.1268,\ \ldots,\ \hat{\beta}_9 = -11.1268$$

We can also write one regression line corresponding to each of the levels of the categorical predictor as follows:

$$\begin{aligned}
\text{Solitary } \hat{y} &= -50.2420 + 136.1268x \\
\text{One } \hat{y} &= (-50.2420 + 6.5172) + (136.1268 - 4.6771)x \\
\text{Low } \hat{y} &= (-50.2420 - 7.7501) + (136.1268 + 0.8743)x \\
\text{Many } \hat{y} &= (-50.2420 - 1.1394) + (136.1268 + 6.5478)x \\
\text{High } \hat{y} &= (-50.2420 - 11.0380) + (136.1268 - 11.1268)x
\end{aligned}$$

If we want to test the significance of the categorical and continuous predictors and their interaction in this case, we need to work with the `anova` table output instead of the summary output in `R`.

5.4.1 Model Selection

The main purpose of the analysis here is to decide which models fits the data. We have the following candidates:

- Model 1: $Y \sim 1$ (intercept-only model)
- Model 2: $Y \sim X$ (D has not effect on Y)
- Model 2': $Y \sim D$ (X has not effect on Y)
- Model 3: $Y \sim D + X$ (additive model)
- Model 4: $Y \sim D + X + D:X$ (full model)

The main tool we are going to use is the F-test. In fact, when the categorical variable D has *more than two levels*, the t-test may no longer be appropriate.

Multiplicative vs. Additive Model

$$\begin{cases} H_0: Y \sim X + D \text{ (Additive model)} \\ H_\alpha: Y \sim D + X + D:X \text{ (Full model)} \end{cases}$$

1. If the interaction $D:X$ is significant, **stop**. The *Full* model is appropriate.
2. If X is significant, keep X *or* if D is significant, keep D.
3. If neither X nor D are significant, report the intercept model $Y \sim 1$.

But, how do we test whether X or D is significant? And which variable should we test first?

Testing the significance of X and/or D.

Is X statistically significant?

There are two ways that we can test for the significance of the continuous predictor. We can either test for its marginal contribution to the model, i.e.

- Test the **marginal** contribution of X:

$$\begin{cases} H_0: Y \sim 1 \ \ \text{(intercept-only model)} \\ H_\alpha: Y \sim X \end{cases}$$

or we can test for the contribution of X in the model when D is already in the model, i.e.

- Test the contribution of X in **addition** to D:

$$\begin{cases} H_0: Y \sim D \\ H_\alpha: Y \sim X + D \end{cases}$$

Is D statistically significant?

Similarly, there are two ways we can test for the significance of D:

- Test the **marginal** contribution of D:

$$\begin{cases} H_0: Y \sim 1 \ \ \text{(intercept-only model)} \\ H_\alpha: Y \sim D \end{cases}$$

- Test the contribution of D in **addition** to X:

$$\begin{cases} H_0 : Y \sim X \\ H_\alpha : Y \sim X + D \end{cases}$$

Which test you are going to perform first

5.4.2 Sequential F Tests & ANOVA

In *Regression* there are two F tests that we have seen: The first one is the **partial F test** that is defined as

$$F = \frac{\frac{RSS_0 - RSS_\alpha}{df_0 - df_\alpha}}{\frac{RSS_\alpha}{df_\alpha}}$$

where RSS_0 corresponds to the model under the Null and RSS_α corresponds to the model under the alternative.

We have also seen F tests defined as follows:

$$F = \frac{MSRegression}{MSE} = \frac{\frac{FSS}{df_{FSS}}}{\frac{RSS}{n-p}}$$

We used this particular test to perform the *overall F test* of whether the set of all the predictors is more adequate than a model with just the intercept. This F tests is very common in ANOVA and ANCOVA models and they are the ones that are shown in the ANOVA Table.

In fact, the ANOVA Table is more detailed, because it further breaks down the "Regression" quantity to its components, i.e. variables X_1, X_2, etc. However, when we have an ANCOVA model the order with which we input the variables in the `lm` function in `R` affects the F tests and corresponding p-values we obtain.

In particular, if we consider one X and one D (with k levels), we typically have the following outputs depending on which variable we include in the model first:

X first, D second ANOVA Table

`anova(lm(Y ~ X + D + X:D))`

Source	df	SS	MS	F-value
X	df_X	SSX	$MS_X = \frac{SSX}{df_X}$	$F_X = \frac{MS_X}{MSE}$
D	df_D	$SS(D\|X)$	$MS_D = \frac{SS(D\|X)}{df_D}$	$F_D = \frac{MS_X}{MSE}$
$X:D$	df_{XD}	$SS(XD\|X,D)$	$MS_{XD} = \frac{SS(XD\|X,D)}{df_{XD}}$	$F_{XD} = \frac{MS_{XD}}{MSE}$
Residuals	df_{Res}	RSS	MSE	

The degrees of freedom of the variables in the ANOVA table are typically:

- df_X = 1, because it corresponds to one random variable and therefore we only need to estimate a single coefficient when fitting the regression model.
- df_D = $k-1$, because if D has k levels then we generate $k-1$ dummy variables that are associated to $k-1$ coefficients that need to be estimated when fitting the regression model.
- $df_{XD} = df_X \cdot df_D$, because it corresponds to the total number of interaction terms in the model - the product of each dummy variable with the continuous one.
- $df_{Residual} = n - p_A$, where $p_A = 2k$, 1 df for the intercept + 1 df for X + (k-1) df for D + (k-1) df XD.

In the above ANOVA Table, the following sequence of F -tests is given by

	H_0	H_α
Test 1:	$Y \sim 1$	$Y \sim X$
Test 2:	$Y \sim X$	$Y \sim X + D$
Test 3:	$Y \sim X + D$	$Y \sim X + D + X:D$

D first, X second ANOVA Table

`anova(lm(Y ~ D + X + X:D))`

Source	df	SS	MS	F-value
D	df_D	SSD	$MS_D = \frac{SSD}{df_D}$	$F_D = \frac{MS_X}{MSE}$
X	df_X	$SS(X\|D)$	$MS_X = \frac{SS(X\|D)}{df_X}$	$F_X = \frac{MS_X}{MSE}$
$X:D$	df_{XD}	$SS(XD\|X,D)$	$MS_{XD} = \frac{SS(XD\|X,D)}{df_{XD}}$	$F_{XD} = \frac{MS_{XD}}{MSE}$
Residuals	df_{Res}	RSS	MSE	

The degrees of freedom of the variables in the ANOVA table are the same as before, while the sequence of F -tests becomes

	H_0	H_α
Test 1:	$Y \sim 1$	$Y \sim D$
Test 2:	$Y \sim D$	$Y \sim D + X$
Test 3:	$Y \sim D + X$	$Y \sim D + X + X:D$

Remark: Some of the F-test statistics and p-values from the sequential ANOVA table are *(different)* from the ones we calculated based on the partial F-test for comparing two *nested* models.

As an example, suppose we want to test the significance of the categorical parameter D. The partial F test and the sequential ANOVA table F test may lead to different answers.

- **Partial F Test**:

This is the F test we have been performing since the beginning of the semester. So, to perform this test, we need to fit two models: the model under the H_0 (e.g. the model with X) and the model under the H_α (e.g. with X and D). Then, as usual we obtain the RSS and dfs for each model and perform the following F test:

$$F = \frac{(RSS_0 - RSS_\alpha)/(k-1)}{RSS_\alpha/(n-p_\alpha)} \sim F_{k-1,\,n-p_\alpha} \text{ (under the null)}$$

either *"by hand"* or using the `anova()` function of the full/reduced models. Here, k is the total number of categories of variable D and $(n-p_\alpha)$ are the degrees of freedom of the RSS_α, which are $(n-p_\alpha) = n-(k-1)-1-1$ (D corresponds to $k-1$ indicator variables, we have 1 X and 1 intercept).

Fruitfly Example

In this example we will discuss what are the different F tests that we can extract from the various `R` outputs.

We start by fitting a model with `thorax` (X) first and `activity` (D) second.

```
anova(lm(longevity ~ thorax*activity, fruitfly))
```

```
## Analysis of Variance Table
##
## Response: longevity
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax            1 15003.3 15003.3 130.733 < 2.2e-16 ***
## activity          4  9634.6  2408.6  20.988 5.503e-13 ***
## thorax:activity   4    24.3     6.1   0.053    0.9947
## Residuals       114 13083.0   114.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The various F tests in the ANOVA Table above are as follows:

- Line 1 (`thorax`): The test behind this F value is

$$\begin{cases} H_0 : Y \sim 1 \\ H_a : Y \sim X \end{cases}$$

(marginal contribution of `thorax`)
The F value and corresponding p-value are

$$F = \frac{15003.3}{114.8} = 130.733 \sim F_{1,114}$$

with p-value $< 2.210^{-16}$.

- Line 2 (`activity`): The test behind this F value is

$$\begin{cases} H_0: Y \sim X \\ H_\alpha: Y \sim X + D \end{cases}$$

(contribution of `activity` in addition to `thorax`)
The F value and corresponding p-value are

$$F = \frac{2408.6}{114.8} = 20.988 \sim F_{4,114}$$

with p-value $5.503e-13$.

- Line 3 (`thorax:activity`): The test behind this F value is

$$\begin{cases} H_0: Y \sim X + D \\ H_\alpha: Y \sim X + D + X:D \end{cases}$$

(contribution of `thorax:activity` in addition to `thorax` and `activity`)
The F value and corresponding p-value are

$$F = \frac{6.1}{114.8} = 0.053 \sim F_{4,114}$$

with p-value 0.9947.

Next, we a model with `activity` (D) first and `thorax` (X) second.

```
anova(lm(longevity ~ activity*thorax, fruitfly))
```

```
## Analysis of Variance Table
##
## Response: longevity
##                  Df  Sum Sq Mean Sq F value  Pr(>F)
## activity          4 12269.5  3067.4  26.728 1.2e-15 ***
## thorax            1 12368.4 12368.4 107.774 < 2e-16 ***
## activity:thorax   4    24.3     6.1   0.053  0.9947
## Residuals       114 13083.0   114.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The various F tests in the ANOVA Table above are as follows:

- Line 1 (`activity`): The test behind this F value is

$$\begin{cases} H_0: Y \sim 1 \\ H_\alpha: Y \sim D \end{cases}$$

  (marginal contribution of `activity`)
  The F value and corresponding p-value are

$$F = \frac{3067.4}{114.8} = 26.728 \sim F_{4,114}$$

  with p-value $< 1.210^{-15}$.

- Line 2 (`thorax`): The test behind this F value is

$$\begin{cases} H_0: Y \sim D \\ H_\alpha: Y \sim X + D \end{cases}$$

  (contribution of `thorax` in addition to `activity`)
  The F value and corresponding p-value are

$$F = \frac{12368.4}{114.8} = 107.774 \sim F_{1,114}$$

  with p-value $< 2.210^{-16}$.

- Line 3 (`activity:thorax`): The test behind this F value is

$$\begin{cases} H_0: Y \sim X + D \\ H_\alpha: Y \sim X + D + X:D \end{cases}$$

(contribution of `activity:thorax` in addition to `thorax` and `activity`)
The F value and corresponding p-value are

$$F = \frac{6.1}{114.8} = 0.053 \sim F_{4,114}$$

with p-value 0.9947.

Similarly, when we remove the interaction term and focus on fitting the additive model, the order matters.

Now, we are going to discuss how to use a **partia F test** to test for the interaction term. So, the two hypotheses we have are:

$$\begin{cases} H_0: Y \sim X + D \\ H_\alpha: Y \sim X + D + X:D \end{cases}$$

The partial F test is

$$F = \frac{(RSS_0 - RSS_\alpha)/(df_0 - df_\alpha)}{RSS_\alpha/df_\alpha}$$

So, we need to fit the full and reduced models

```
anova(lm(longevity ~ thorax*activity, fruitfly))
```

```
## Analysis of Variance Table
##
## Response: longevity
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax            1 15003.3 15003.3 130.733 < 2.2e-16 ***
## activity          4  9634.6  2408.6  20.988 5.503e-13 ***
## thorax:activity   4    24.3     6.1   0.053    0.9947
## Residuals       114 13083.0   114.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(longevity ~ thorax+activity, fruitfly))
```

```
## Analysis of Variance Table
##
## Response: longevity
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax      1 15003.3 15003.3 135.069 < 2.2e-16 ***
## activity    4  9634.6  2408.6  21.684 1.974e-13 ***
## Residuals 118 13107.3   111.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can either perform this test “by hand” by extracting the usual RSS and df values or we can directly use the `anova()` function

```
anova(lm(longevity ~ thorax*activity, fruitfly),
      lm(longevity ~ thorax+activity, fruitfly))
```

```
## Analysis of Variance Table
##
## Model 1: longevity ~ thorax * activity
## Model 2: longevity ~ thorax + activity
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1    114 13083
## 2    118 13107 -4   -24.314 0.053 0.9947
```

As we can see, the partial F test for the interaction term yields the same results as the other two tests.

However, this is not the case when we want to test for the significance of the factor D. Indeed, the hypothesis to test is

$$\begin{cases} H_0 : Y \sim X \\ H_\alpha : Y \sim X + D \end{cases}$$

where the partial F test above gives us:

```
anova(lm(longevity ~ thorax, fruitfly),
      lm(longevity ~ thorax+activity, fruitfly))
```

```
## Analysis of Variance Table
##
## Model 1: longevity ~ thorax
## Model 2: longevity ~ thorax + activity
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    122 22742
## 2    118 13107  4    9634.6 21.684 1.974e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

which is not the same as the ANOVA F test of the model including the interaction. This is expected, because for these two models the estimator for σ^2 that is used is different. In fact, the RSS in the ANOVA tables is computed after fitting the full model (with all terms) and its degrees of freedom are $n - 2k$ while the σ^2 estimator for the partial F test is based on the additive model under the H_α.