Chapter 7. Factors and F-tests

- A **factor** is an explanatory variable with discrete levels.
- Factors are also called **categorial variables**.
- The different values the variable can take are called **levels** of the factor.
- If we tested growth of a plant in three different soil types we might model this using a soil type factor with 3 levels, `clay`, `sand` and `loam`.
- A factor with 2 levels is a **binary factor**.
- In linear models, factors can describe different classes of units. For example, in HW7, the binary factor `Competition` distinquishes certain types of newspaper.
- We could have a different mean and/or different slope for each level of the factor.

Comparing two sample means via a model with a factor

- Recall the mouse weight experiment. 12 mice are given each of two diets and are then weighed.
- First, set up notation. Let y_{ij} be the weight of the jth mouse on treatent i, where $i = 1, 2$ corresponds to the normal and high fat diet respectively and j=1,. . . ,12 enumerates the replicates for each treatment group.
- A probability model for this experiment is

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{for } i = 1, 2 \text{ and } j = 1, \ldots, 12$$

where $\epsilon_{ij} \sim \mathrm{iid\ normal}(0, \sigma)$.

- Here, we have written the model in subscript form. We have a mean for each level of the treatment group factor.
- This looks superficially different from the way we have written linear models. There is an extra subscript.
- We can rewrite it to make it fit into our linear model framework by putting all the (i, j) values in a single column.

Dummy variables to code levels of factors

- Let $\mathbf{x}_1 = (x_{1,1}, \ldots, x_{24,1}) = (1, \ldots, 1, 0, \ldots, 0)$ be a vector with 1 in the first 12 places and 0 in the remaining 12 places.
- Let $\mathbf{x}_2 = (x_{1,2}, \ldots, x_{24,2}) = (0, \ldots, 0, 1, \ldots, 1)$ be a vector with 0 in the first 12 places and 1 in the remaining 12 places.
- Let $\mathbf{y} = (y_1, \ldots, y_{24}) = (y_{1,1}, \ldots, y_{1,12}, y_{2,1}, \ldots, y_{2,12})$ be the mouse weights concatenated into a single vector.
- Let $\mathbf{e} = (e_1, \ldots, e_{24}) = (e_{1,1}, \ldots, e_{1,12}, e_{2,1}, \ldots, e_{2,12})$ be residual error terms concatenated into a single vector.

Question 7.1. $\mathbf{x}_1$ and $\mathbf{x}_1$ are called **dummy variables** since they are built to allow us to write $y_{ij} = \mu_i + e_{ij}$ in the usual linear model form,

$$y_k = \mu_1 x_{k,1} + \mu_2 x_{k,2} + e_k.$$

Convince yourself that these equations are equivalent.

Two things to notice about models with factors

We consider the sample linear model $y_k = \mu_1 x_{k,1} + \mu_2 x_{k,2} + e_k$, $k = 1, \ldots, 24$.

Question 7.2. Usually we use i as a subscript when writing a linear model in subscript form. Why do we use k here?

Question 7.3. Notice there is no intercept term in this linear model. Why?

Question 7.4. Write the probability model $Y_{ij} = \mu_i + \epsilon_{ij}$ for $i = 1, 2$ and $j = 1, \ldots, 12$ in the matrix form for the probability model of a linear model, $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

This asks you to write down a choice of $\mathbf{Y}$, $\mathbb{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ so that the two equations are equivalent.

Alternative representations of factors

- Consider the following two models in subscript form, with $\epsilon_{ij} \sim \text{iid normal}(0, \sigma)$ for $i = 1, 2$ and $j = 1, \ldots, 12$.

$$\text{(M1).} \qquad Y_{ij} = \mu_i + \epsilon_{ij}$$

$$\text{(M2).} \qquad Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{with } \mu_1 = 0$$

Question 7.5. Why are (M1) and (M2) equivalent?

Question 7.6. What is the difference in the interpretation of the parameters between (M1) and (M2)?

An over-specified model

- Recall (M2) with $\epsilon_{ij} \sim \text{iid normal}(0, \sigma)$ for $i = 1, 2$ and $j = 1, \ldots, 12$.

$$\text{(M2).} \qquad Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{with } \alpha_1 = 0$$

- Suppose we modify model (M2) to omit the important detail that $\alpha_1 = 0$. This gives

$$\text{(M3).} \qquad Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- Many rude words are used to describe the problem with (M3) such as **over-specified**, **over-parameterized**, **unidentifiable**, **redundant**.

Question 7.7. Can you see and explain the concern about (M3)?

Using a linear model with factors to test equality of means

- A null hypothesis is that the mice weights for both treatment groups are drawn from the same distribution. Any difference is just chance variation in this particular sample. If the null hypotheis is a plausible description of our data, we don't want to spent too much time interpreting this experimental results.
- A natural way to write this null hypothesis is $H_0 : \mu_1 = \mu_2$ in the model representation (M1)
- **A USEFUL TRICK**. Using the equivalent model representation (M2), this becomes $H_0 : \alpha = 0$ which is the easiest type of null hypothesis for a linear model.

Factors in `lm()`

- If you give `lm()` an explanatory variable of class `character` it interprets the variable as levels of a factor.

```
mice <- read.csv("https://ionides.github.io/401f18/hw/hw01/femaleMi
head(mice,3)

##   Diet Bodyweight
## 1 chow      21.51
## 2 chow      28.14
## 3 chow      24.04

lm1 <- lm(Bodyweight~Diet,data=mice)
summary(lm1)$coef

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 23.813333   1.039353 22.911684 7.642256e-17
## Diethf       3.020833   1.469867  2.055174 5.192480e-02
```

What model has R actually fitted?

- It can be hard to figure out what R is actually doing when it fits models with a factor.
- If you can't correctly write the model R is fitting using subscript notation you may well interpret the results wrong.
- A good check is to look at R's design matrix

```
model.matrix(lm1)[c(1:2,12:13,23:24),]

##    (Intercept) Diethf
## 1            1      0
## 2            1      0
## 12           1      0
## 13           1      1
## 23           1      1
## 24           1      1
```

Hypothesis tests for groups of parameters

- We've seen how the least squares coefficient can be used as a test statistic for the null hypothesis that a parameter in a linear model is zero.
- Sometimes we want to test many parameters at the same time. For example, when analyzing the field goal kicking data, we must decide whether to have a separate intercept for each player.

Question 7.8. There are 19 kickers in the dataset. How many extra parameters are needed if we add an intercept for each player?

- This type of question is called **model selection**. Our test statistic should compare **goodness of fit** with and without the additional parameters.
- We need to know the distribution of the model-generated test statistic under the null hypothesis to find the p-value for the test.

Residual sum of squares to quantify goodness of fit

Let $\mathbf{y}$ be the data. Let H_0 be a linear model, $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Let H_a extend H_0 by adding d additional explanatory variables.

- Let RSS_0 be the residual sum of squares for H_0. The residual errors are $\mathbf{e} = \mathbf{y} - \mathbb{X}\mathbf{b}$ where $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$. So, $\mathrm{RSS}_0 = \sum_{i=1}^{n} e_i^2$.

- Let RSS_a be the residual sum of squares for H_a.

- Residual sum of squares is a measure of goodness of fit. A small residual sum of squares suggests a model that fits the data well.

Question 7.9. It is always true that $\mathrm{RSS}_a \leq \mathrm{RSS}_0$. Why?

- We want to know how much smaller RSS_a has to be than RSS_0 to give satisfactory evidence in support of adding the extra explanatory variables into our model. In other words, when should we reject H_0 in favor of H_a?

The f statistic for adding groups of parameters

Formally, we have $H_0 : \mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \epsilon$ and $H_a : \mathbf{Y} = \mathbb{X}_a\boldsymbol{\beta}_a + \epsilon$, where $\mathbb{X}$ is an $n \times p$ matrix and $\mathbb{X}_a = [\,\mathbb{X}\;\mathbb{Z}\,]$ is an $n \times q$ matrix with $q = p + d$. Here, $\mathbb{Z}$ is a $n \times d$ matrix of additional explanatory variables for H_a. As usual, we model $\epsilon_1, \ldots, \epsilon_n$ as iid $N[0, \sigma]$.

- Consider the following sample test statistic:

$$f = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/d}{\mathrm{RSS}_a/(n-q)}.$$

- The denominator is an estimate of σ^2 under H_a. Using this denominator **standardizes** the test statistic.

- The numerator $(\mathrm{RSS}_0 - \mathrm{RSS}_a)/d$ is the **change in RSS per degree of freedom**. Parameters in linear models are often interpreted as degrees of freedom of the model.

- Let F be a model-generated version of f, with the data $\mathbf{y}$ replaced by a random vector $\mathbf{Y}$. If H_0 is true, then the RSS per degree of freedom should be about the same on the numerator and the denominator, so $F \approx 1$. Large values, $f \gg 1$, are therefore evidence against H_0.

The F test for model selection

- Under H_0, the model-generated F statistic has an F distribution on d and $n-q$ degrees of freedom.
- Because of the way we constructed the F statistic, its distribution under H_0 doesn't depend on σ. It only depends on the dimension of $\mathbb{X}$ and $\mathbb{X}_a$.
- We can obtain p-values for the F distribution in R using `pf()`. Try `?pf`.
- Testing H_0 verus H_a using this p-value is called the F test.
- When we add a single parameter, so $d=1$ and $q=p+1$, the F test is equivalent to carrying out Student's `t` test using the estimated coefficient as the test statistic. As homework, you are asked to check this using `pt()` and `pf()` in R.
- Degrees of freedom are mysterious. The mathematics for how they work involves matrix algebra beyond this course. An intuition is that fitting a parameter that is not in the model "explains" a share of the residual sum of squares; in an extreme case, fitting n parameters to n data points may give a perfect fit (residual sum of squares = zero) even if none of these parameters are in the true model.

The F test is called “analysis of variance”

- The F test was invented before computers existed.
- Working out the sums of squares efficiently, by hand, was a big deal!
- Sums of squares of residuals are relevant for estimating variance.
- Building F tests is historically called **analysis of variance** or abbreviated to **ANOVA**.
- The sums of squares and corresponding F tests are presented in an **ANOVA table**. We will see one in the following data analysis.

An F test for kickers. (i) Reviewing the data

```
goals <- read.table("FieldGoals2003to2006.csv",header=T,sep=",")
goals[1:5,c("Name","Teamt","FGt","FGtM1")]

##              Name Teamt  FGt FGtM1
## 1 Adam Vinatieri    NE 73.5  90.0
## 2 Adam Vinatieri    NE 93.9  73.5
## 3 Adam Vinatieri    NE 80.0  93.9
## 4 Adam Vinatieri   IND 89.4  80.0
## 5    David Akers   PHI 82.7  88.2

lm0 <- lm(FGt~FGtM1+Name,data=goals)
```

- This is model syntax we have not seen before.
- `Name` is a **factor**

```
class(goals$Name)

## [1] "factor"
```

- A factor is a vector with **levels**. Here, the levels are the kicker names.

An F test for kickers. (ii) Checking the design matrix

```
X <- model.matrix(lm0)
dim(X)

## [1] 76 20

unname(X[c(1,5,9,13,17),1:8])

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1 90.0    0    0    0    0    0    0
## [2,]    1 88.2    1    0    0    0    0    0
## [3,]    1 72.2    0    1    0    0    0    0
## [4,]    1 82.1    0    0    1    0    0    0
## [5,]    1 80.0    0    0    0    1    0    0
```

Question 7.10. Is this the design matrix that you want? Can we use our experience working with design matrices to understand what R is doing?

An F test for kickers. (ii) Interpreting the ANOVA table

```
anova(lm0)

## Analysis of Variance Table
##
## Response: FGt
##           Df Sum Sq Mean Sq F value    Pr(>F)
## FGtM1      1   87.2  87.199  2.2597 0.1383978
## Name      18 2252.5 125.137  3.2429 0.0003858 ***
## Residuals 56 2161.0  38.589
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 7.11. Focus on the row labeled `Name`. Explain what is being tested, how it is being tested, and what you conclude.