Chapter 7. Factors and the ANOVA F-test

- A **factor** is an explanatory variable with discrete levels.
- Factors are also called **categorical 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` distinguishes 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. 24 mice are randomized to one of two diets and are weighed after two weeks.

- First, set up notation. Let y_{ij} be the weight of the jth mouse on treatment i, where $i = 1, 2$ corresponds to the normal and high fat diet respectively and $j = 1, \ldots, 12$ enumerates the replicates for each treatment group.

- A probability model for this experiment is

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

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

- Here, we have written the model in **double 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 $\mathbf{x}_1 = (x_{1,1}, \dots, x_{24,1}) = (1, \dots, 1, 0, \dots, 0)$ be a vector with 1 in the first 12 places and 0 in the remaining 12 places.
- Let $\mathbf{x}_2 = (x_{1,2}, \dots, x_{24,2}) = (0, \dots, 0, 1, \dots, 1)$ be a vector with 0 in the first 12 places and 1 in the remaining 12 places.
- Let $\mathbf{y} = (y_1, \dots, y_{24}) = (y_{1,1}, \dots, y_{1,12}, y_{2,1}, \dots, y_{2,12})$ be the mouse weights concatenated into a single vector.
- Let $\mathbf{e} = (e_1, \dots, e_{24}) = (e_{1,1}, \dots, e_{1,12}, e_{2,1}, \dots, e_{2,12})$ 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 $y_{ij} = \overset{M_i}{\cancel{\mu_i}} + e_{ij}$ in the usual single-subscript linear model form,

$$y_k = \overset{M_1}{\cancel{\mu_1}} x_{k,1} + \overset{M_2}{\cancel{\mu_2}} x_{k,2} + e_k.$$

Convince yourself that these equations are equivalent.

1) For the first 12, the $x_{k,2}$ term disappears, for the next 12, the $x_{k,1}$ term disappears

2) Check that for mouse $k=1$ and mouse $k=13$ the equations correspond.

Two things to notice about models with factors

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

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

We have used i and j in the double subscript version. In principle, we could also use i here — it is a dummy subscript — but it is clearer not to repeat.

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

1. Exactly one of $x_{k,1}$ and $x_{k,2}$ is 1 for each k, so the average weight is accounted for by either the $x_{k,1}$ coefficient or the $x_{k,2}$ coefficient.

2. Alternatively, there are 2 different means in the model — for 2 treatment groups. We already have 2 parameters, μ_1 and μ_2. A third parameter is unnecessary.

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, $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

$$\mathbb{X} = \left[X_{kl}\right]_{24\times 2}$$

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.

$$\begin{bmatrix} Y_1\\ Y_2\\ \vdots\\ Y_{12}\\ Y_{13}\\ \vdots\\ Y_{24}\end{bmatrix} = \underbrace{\begin{bmatrix}1&0\\1&0\\ \vdots& \\ 1&0\\0&1\\ \vdots&\vdots\\ 0&1\end{bmatrix}}_{\mathbb{X}} \underbrace{\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}}_{\underset{\sim}{\beta}} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_{12}\\ \varepsilon_{13}\\ \vdots\\ \varepsilon_{24}\end{bmatrix}, \quad \underbrace{\underset{\sim}{\varepsilon} \sim MVN\left(\underset{\sim}{0}, \sigma^2 \mathbb{I}\right)}_{\text{independent normal measurement error model.}}$$

Alternative representations of factors

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

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

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

(M2):

$$X\underset{\sim}{\beta} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \end{pmatrix}$$

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

Set $\mu = \mu_1$ and $\alpha_2 = \mu_2 - \mu_1$ and both models give the same means.

We can think of (M1) and (M2) as two different ways to write the same model, because from the point of view of Y_{ij} it doesn't matter which way you write it, so (M1) and (M2) give the same fitted values for data: The data can't say which of (M1) and (M2) is "right".

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

The parameter α_2 is called a contrast. In (M2) we model using an overall mean & a contrast.

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)?

This gives multiple ways to describe the data which are all "correct". For example, we can get a mean of 20 for the 1st treatment group & 25 for the 2nd in 2 ways.

$(\mu = 10, \alpha_1 = 10, \alpha_2 = 15)$ or $(\mu = 15, \alpha_1 = 5, \alpha_2 = 10)$ or infinitely many other ways to get the same means.

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 hypothesis 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 $H_0 : \alpha_2 = 0$ 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/femaleMiceWeights.csv"
)
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
```

no parameter for chow. R is using the parametrization with $\alpha_1 = 0$.

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
```

this is R for

$$\mathbb{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\\vdots\\1\end{matrix}}\right\} 1\text{-}12 \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\\1\end{matrix}}\right\} 13\text{-}24 \end{matrix}$$

```
##   (Intercept) Diethf
## 1           1      0
## 2           1      0
## 3           1      0
```

$\mathbb{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ \vdots & 0 \\ & 1 \\ & \vdots \\ 1 & 1 \end{pmatrix}$, $(\mathbb{X}^T\mathbb{X}) = \begin{pmatrix} 24 & 12 \\ 12 & 12 \end{pmatrix}$

then we could continue to work out $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underset{\sim}{y}$ to check that m is the mean of treatment group 1.

Question 7.8. Write down the sample model R has fitted, in double subscript form, and interpret the parameters.

$y_{ij} = m + a_i + e_{ij}$, $a_1 = 0$ (Sample version)

the sample contrast

definition: the contrast is the difference btw the group mean & a reference group mean

sample mean for the 1st treatment group. This should follow from $\underset{\sim}{b} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underset{\sim}{y}$

for the purposes of a linear model, think of these stacked as a 24×1 column vector.

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $\alpha_1 = 0$, $\varepsilon_{ij} \sim MNV[0, \sigma^2\mathbb{I}]$ (probability model).

~~mean of~~ Expected value for the 1st treatment group.

the model contrast.

note: the t-test and z-test have the same test statistic. The t distribution allows for uncertainty in estimating σ, so has fatter tails.

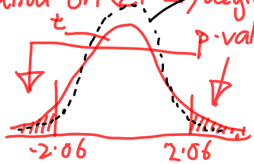
```
summary(lm1)$coef

##               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
```

Question 7.9. Consider the null hypothesis that the two diets are equivalent, so the observed difference in mouse weights is chance variation in the sample. Make both a t-test and a normal approximation test (also known as a z-test) of this hypothesis. Which test do you prefer, and why? A test at the 5% level is appropriate for this fairly small sample.

we are testing $H_0: \alpha_2 = 0$, the contrast between hf diet relative to the chow reference group diet is zero.

The test statistic is $\boxed{\frac{a_2}{SE(a_2)}} = 2.056$ from the R output.

The t-test compares this with the t distribution on (24-2) degrees of freedom. The p-value is 0.052 from R output. We don't reject H_0 at 5% level. The z-test compares with a normal distribution. Since $2.056 > 1.96$, we do reject H_0.



A linear model vs a two sample test

- The linear model test above is equivalent to a **two sample t-test with pooled variance**.

```
t.test(mice$Bodyweight[1:12],mice$Bodyweight[13:24],
  var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  mice$Bodyweight[1:12] and mice$Bodyweight[13:24]
## t = -2.0552, df = 22, p-value = 0.05192
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -6.06915183  0.02748516
## sample estimates:
## mean of x mean of y
##  23.81333  26.83417
```

- Check that the p-values are the same in both cases.

Why does the above linear model test match the two sample t-test with pooled variance?

- Tests are the same if they use the same probability model for the null hypothesis and an equivalent test statistic.
- If you wrote out the probability model justifying the two sample t-test with pooled variance it would be exactly the model (M1) or (M2).
- Here we focus on tests via a linear model, but you might like to review two sample tests at `https://open.umich.edu/find/open-educational-resources/statistics/statistics-250-introduction-statistics-data-analysis`
- Your previous course in statistics likely did not explain the probability model behind the two sample t-test with pooled variance. Viewing this test as a special case of the linear model gives us a way to do that.
- The linear model also lets us analyze many more complex models.

A factor with many levels: Kicking field goals

- If an athlete has a good season, is the next one likely to be good? Or bad? Or does the previous season have no predictive skill?
- We consider field goal kicking success for the 19 National Football League (NFL) kickers who played every season during 2002-2006.

```
download.file(destfile="FieldGoals.csv",
  url="https://ionides.github.io/401f18/07/FieldGoals.csv")
```

```
goals <- read.table("FieldGoals.csv",header=TRUE,sep=",")
head(goals[,1:8])

##              Name Yeart Teamt FGAt  FGt Teamt1 FGAt1 FGt1
## 1 Adam Vinatieri  2003    NE   34 73.5     NE    30 90.0
## 2 Adam Vinatieri  2004    NE   33 93.9     NE    34 73.5
## 3 Adam Vinatieri  2005    NE   25 80.0     NE    33 93.9
## 4 Adam Vinatieri  2006   IND   19 89.4     NE    25 80.0
## 5    David Akers  2003   PHI   29 82.7    PHI    34 88.2
## 6    David Akers  2004   PHI   32 84.3    PHI    29 82.7
```

Understanding the field goal data

```
goals[1,1:8]

##               Name Yeart Teamt FGAt  FGt Teamt1 FGAt1 FGt1
## 1 Adam Vinatieri  2003    NE   34 73.5     NE    30   90
```

- Each record has the player Name and Year followed by
  Teamt: team that year.
  FGAt: number of field goal attempts in that year.
  FGt: percentage of field goal attempts which were successful that year.
  Teamt1: Team the previous year.
  FGAt1 and FGt1: Field goal attempts and percentage the previous year.

Question 7.10. Is there additional background on football that we need to understand the data and the question?

Note: field goals can be attempted from various distances. The closer the attempt, the easier the goal. There could be weather effects on open stadiums. The rest of the team has some role.

Brainstorming for a model

(we could write down many more models ...)

- If an NFL kicker has a good season, is the next one likely to be good? Or bad? Or does the previous season have no predictive skill?

Question 7.11. (1) Set up notation; (2) propose models in the context of our data; (3) write down hypotheses relevant to our question.

Let y_{ij} be goal percentage for kicker i in year j, $i=1,\dots,19$. $j=1,\dots,4$. Let x_{ij} be the percentage in the previous year, so $x_{ij}=y_{i,j-1}$, with y_{i0} being 2002. Using double subscript notation,

(M1) $y_{ij} = m + b x_{ij} + e_{ij}$ ← 2 parameters, shared intercept & slope

(M2) $y_{ij} = m_i + b_i x_{ij} + e_{ij}$ ← individuals have their own intercepts & slopes. (38 parameters)

(M3) $y_{ij} = m_i + b x_{ij} + e_{ij}$ (individual intercept & shared slope)

(M4) $y_{ij} = m + a_i + b x_{ij} + e_{ij}$, $a_1 = 0$.

(M4) is a different way of writing (M3).

Brainstorming continued

(3). To write a hypothesis formally, we need a probability model, for example

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij} \quad , \quad \alpha_1 = 0.$$

$\varepsilon_{ij} \sim$ iid normal$(0, \sigma)$

$H_0 : \alpha_i = 0$ for all $i = 2, \dots, 19$. Kickers are the same.

$H_0 : \beta = 0$. Previous year has no predictive skill for this year.

An alternative hypothesis is

$H_a : \beta \neq 0$.

A linear model for field goals

```
goals.lm <- lm(FGt~FGt1+Name,data=goals)
X <- model.matrix(goals.lm)
```

- Here, Name has R class factor. The levels are the kicker names.

(we might have expected goals$Name to be a vector of type "character")

```
class(goals$Name)

## [1] "factor"

attributes(goals$Name)$levels[1:6]

## [1] "Adam Vinatieri" "David Akers"     "Jason Elam"
## [4] "Jason Hanson"   "Jay Feely"       "Jeff Reed"
```

- We want to find out what model we have fitted! Time to study the design matrix, X.

```
dim(X)

## [1] 76 20
```

- Working out the model (in double-subscript form) corresponding to a 76×20 matrix takes some thought.

This is the top left corner of X

this strips row & column names, letting us see more of the matrix.

```
unname(X[1:15,1:10])
```

x_{ij} stacked as a vector

intercept

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1 90.0    0    0    0    0    0    0    0     0
##  [2,]    1 73.5    0    0    0    0    0    0    0     0
##  [3,]    1 93.9    0    0    0    0    0    0    0     0
##  [4,]    1 80.0    0    0    0    0    0    0    0     0
##  [5,]    1 88.2    1    0    0    0    0    0    0     0
##  [6,]    1 82.7    1    0    0    0    0    0    0     0
##  [7,]    1 84.3    1    0    0    0    0    0    0     0
##  [8,]    1 72.7    1    0    0    0    0    0    0     0
##  [9,]    1 72.2    0    1    0    0    0    0    0     0
## [10,]    1 87.0    0    1    0    0    0    0    0     0
## [11,]    1 85.2    0    1    0    0    0    0    0     0
## [12,]    1 75.0    0    1    0    0    0    0    0     0
## [13,]    1 82.1    0    0    1    0    0    0    0     0
## [14,]    1 95.6    0    0    1    0    0    0    0     0
## [15,]    1 85.7    0    0    1    0    0    0    0     0
```

x_{12} $x_{\cdot 3}$ x_{14} x_{21} x_{22} x_{23} x_{24} x_{31}

kicker 1

kicker 2

there is no column with 1s for the first kicker: $a_1 = 0$.

Question 7.12. What is the probability model fitted by `lm(FGt~FGt1+Name,data=goals)`? Use double-subscript form.

$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$. $\alpha_1 = 0$,

$\varepsilon_{ij} \sim$ iid normal $(0, \sigma)$

$i = 1, \dots, 19$.

$j = 1, \dots, 4$.

Note: this one equation actually represents 76 = 4×19 equations, for each data point.

Note: $\varepsilon_{ij} \sim MVN(0, \sigma^2 \mathbb{I})$

is equivalent to $\varepsilon_{ij} \sim$ iid normal $(0, \sigma)$

preferred notation for matrix calculations.

simpler for most purposes.

Question 7.13. What are the terms in the sample linear model corresponding to the following R output?

```
coef(goals.lm)[1:6]

##      (Intercept)                FGt1  NameDavid Akers
##      126.6871588          -0.5037008       -4.6462893
##   NameJason Elam NameJason Hanson     NameJay Feely
##       -3.0166534           2.1172185      -10.3736848
```

$m = 126.69$, $b = -0.504$, $a_2 = -4.64$

$a_3 = -3.02$, $a_4 = 2.12$, $a_5 = -10.4$.

Question 7.14. Is there anything surprising about these results?

(i) b is negative! → a good season predicts an unsuccessful season !?

(ii) m is above 100%. → think about how m and b trade off.

Plotting the fitted model with a line for each factor level

the line for kicker 1

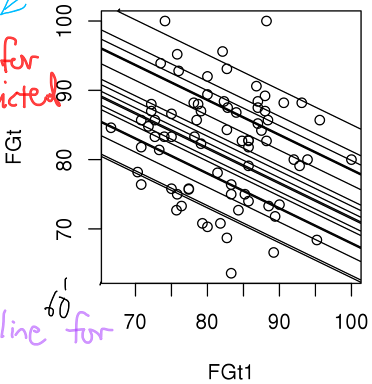
contrast for kicker i.

```
plot(FGt~FGt1,data=goals)
intercept <- coef(goals.lm)[1]
slope <- coef(goals.lm)[2]
kicker <- coef(goals.lm)[3:20]
abline(a=intercept,b=slope)
for(i in 1:18) abline(a=intercept+kicker[i],b=slope)
```

this is extrapolation. It also highlights a data selection issue: kickers with a bad season may lose their job & therefore not be in the dataset!

what is M=126.89 on this plot?
This is the intercept for kicker 1; the predicted percentage after a season with 0% success.

0 20 40 60



note: i=1 is the line for kicker 2.

e.g.
for (i in c("red","green"))

we can think of i taking values in a set. Here, 1:18 is the set {1,2, ..., 18}.

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

18. The intercept can be used for kicker 1 and then we have 18 contrasts. Or we need 1 degree of freedom for a shared intercept, 19 for separate intercepts, so the difference is 18.

- 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.

note: H_a is also a linear model, but with a larger design matrix.

- 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 7.16. It is always true that $\mathrm{RSS}_a \le \mathrm{RSS}_0$. Why?

equality when the extra predictors have zero coefficients.

The alternative hypothesis has additional predictors, so the errors are smaller. A useless or worse than useless predictor could always be given a coefficient of zero in the least squares fit.

- 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 $\epsilon_1, \dots, \epsilon_n$ as iid $N[0, \sigma]$.

- Consider the following sample test statistic:

$$f = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/d}{\mathrm{RSS}_a/\underbrace{(n-q)}_{\text{residual d.f. under } H_a}}.$$

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

$\approx$: approximately equal

$\gg$: much greater than

- 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.

this is true but not obvious!

- 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

This is also true but not obvious.

- 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 versus H_a using this p-value is called the F test.
- 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: Interpreting the ANOVA table

```
anova(goals.lm)
## Analysis of Variance Table
##
## Response: FGt
##           Df Sum Sq Mean Sq F value    Pr(>F)
## FGt1       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
```

anova( ) is a function whose argument is a fitted linear model of class "lm".

the sum of squares explained by "Name"

$RSS_0 - RSS_a$

degrees of freedom.

not the # kickers, but the # of extra d.f. if we have an intercept for each kicker.

$(RSS_0 - RSS_a)/18 = \frac{2252.5}{18}$

75

data points = 19×4 = 76

2161.0/56

RSS_a

Note: the intercept is not shown on the ANOVA table.

Question 7.17. Focus on the row labeled `Name`. Explain what is being tested, how it is being tested, and what you conclude. In other words, write out the hypothesis test corresponding to this line.

Probability model: $H_0: \alpha_i = 0$ for all i, $H_a: \alpha_1 = 0, \alpha_2, \ldots, \alpha_{19}$ unconstrained.

F statistic: $\frac{125.137}{38.589}$

p-value.

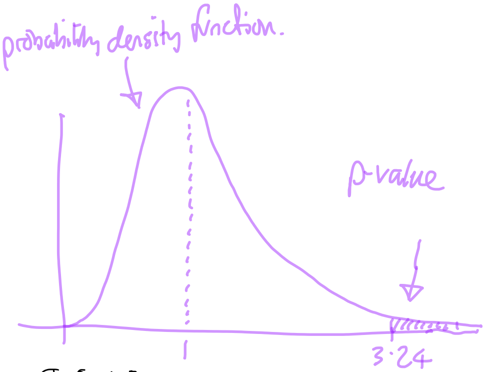
ANOVA hypothesis test, continued

The sample F statistic:

$$f = \frac{(RSS_0 - RSS_a)/18}{RSS_a/56} = 3.24$$ from the R output.

p-value for $f = 3.24$

F distribution on 18 and 56 degrees of freedom



Conclusion: p-value is 0.00039 so we reject H_0 at all usual testing levels. Kickers have different individual success percentages.