Module 10

2023-10-29

R Script

```
##                  PSYC 8100: Module 10
##                    Model Comparison
##               in General Linear Models

# Model comparisons can be conducted using an F statistic.
# The entire textbook by MDK is based on using this general approach
# for conducting statistical tests by comparing statistical models.
# It is called a partial F test or a test on the change in R squared.
# Essentially, we are determining whether the R squared has
# increased significantly when we add a variable (predictor)
# to the model.

# An equivalent way to conceptualize the model comparison approach is
# to think of it as a reduction in sum-of-squared errors. That is,
# when we add predictors to an existing model, do these additional
# predictors result in a statistically significant decrease in the
# sum-of-squared errors (SSE). If it results in a statistically significant
# reduction in the SSE from a REDUCED to FULL model, then this suggests
# that the predictors we added should be retained in the model.

# One sample mean (Module 5) ----

# Remember our one sample test on a mean. In this example, depression
# was the dependent variable.

depression <- c(35,35,40,10,6,20,35,35,35,30)
mean(depression)
```

```
## [1] 28.1
```

```
# Recall that we tested whether the students came from a population
# where the mean was 15. From the t test, we learned that they
# did not come from such a population.

# The sum-of-squared errors in prediction from the REDUCED model
# can be thought of as how much the observed scores differ from the
# hypothesized value, in our original example, 15. Recall that our
# null hypothesis was that the students came from a population
# where the mean was 15. Think of the sum-of-squares as how much
# the observed scores differ from the predicted score which in this
```

```r
# case is 15. The SSE for the REDUCED model is below.

sse.r1 <- sum((depression-15)^2)
sse.r1
```

```
## [1] 2981
```

```r
# The sum-of-squared errors in prediction from the FULL model can
# be thought of as how much the observed scores differ from the
# sample mean, in our example, 28.1. That is, 28.1 is the predicted score
# for every person from the FULL model. Think of the sum-of-squares error as
# how much the observed scores differ from the predicted score which in this
# case is 28.1. The SSE for the FULL model is below.

sse.f1 <- sum((depression-mean(depression))^2)
sse.f1
```

```
## [1] 1264.9
```

```r
# The degrees of freedom in the Numerator can be calculated by
# determining how many terms were added to the REDUCED model to
# form the FULL model. In this case, the REDUCED model has no terms
# in the model while the FULL model has one term (an intercept).
# Thus, the degrees of freedom in the numerator is 1.

df.11 <- 1

# The degrees of freedom in the Denominator can be calculated by
# subtracting the total number of terms (estimated parameters) in
# the FULL model from the total sample size (i.e., # of observations).
# Because we have 10 observations and only one term in the FULL
# model, we have 10 - 1 = 9 degrees of freedom in the denominator.

df.21 <- length(depression) - 1

# The F ratio comparing the FULL model versus REDUCED model is below.
# This equation is the same as Chapter 3, Equation 22 in MDK.

F.ratio1 <- ((sse.r1-sse.f1)/df.11)/(sse.f1/df.21)
F.ratio1
```

```
## [1] 12.21037
```

```r
# Because we have only 1 degree of freedom in the numerator, we
# can take the square root of the F statistic to obtain the
# equivalent t statistic.

t.ratio1 <- sqrt(F.ratio1)
t.ratio1
```

```
## [1] 3.494334
```

```r
# Now let's compare the result from our t statistic when we used
# the t.test() function. They are the same!

t.test(depression, mu=15)
```

```
##
##  One Sample t-test
##
## data:  depression
## t = 3.4943, df = 9, p-value = 0.006784
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##  19.61934 36.58066
## sample estimates:
## mean of x
##      28.1
```

```r
# Thus, when we conduct the one sample test on a mean, we are actually
# COMPARING TWO MODELS (REDUCED model vs. FULL model) using a
# partial F test.

# We can apply the same logic when comparing means from two
# independent samples. This was covered in Module 6.

# Two sample means (Module 6) ----

# When we had two independent samples, we compared females and
# males on the Wonderlic. We had 6 males and 5 females. The
# data are below.

wonderlic <- c(22,14,24,17,19,20,25,28,20,26,29)
sex <- rep(c(1,2), c(6,5))
sex <- factor(sex, labels = c("male", "female"))

wonderlic.df <- data.frame(wonderlic,sex)
rm(wonderlic, sex)

str(wonderlic.df)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ wonderlic: num  22 14 24 17 19 20 25 28 20 26 ...
##  $ sex      : Factor w/ 2 levels "male","female": 1 1 1 1 1 1 2 2 2 2 ...
```

```r
summary(wonderlic.df)
```

```
##    wonderlic         sex
##  Min.   :14.00   male  :6
##  1st Qu.:19.50   female:5
##  Median :22.00
##  Mean   :22.18
##  3rd Qu.:25.50
##  Max.   :29.00
```

```r
aggregate(wonderlic ~ sex, wonderlic.df, mean)
```

```
##      sex wonderlic
## 1   male  19.33333
## 2 female  25.60000
```

```r
# What is the REDUCED model?
# The REDUCED model has only an intercept. Think of this as a common
# mean---the grand mean---for everyone. That is, the REDUCED model assumes
# that the means are the same in BOTH groups. Thus, we have only an
# intercept in the REDUCED model.

# What is the FULL model?
# The FULL model allows the means to differ between the two
# groups. The FULL model, in this case, has an intercept
# and one slope for the one dummy-variable. It does not matter
# whether the reference group is female or male. However,
# we should know how the coding is being done in statistical
# software. In R, the dummy-variable will use the smallest
# number as the reference group. Thus, our dummy-variable,
# technically, indicates whether an observation is female or not.
# The 1s were used to denote male and the 2s were for female.
# Thus, with an intercept and one slope (for the dummy-variable),
# we have two estimated parameters.

# The sum-of-squared errors in prediction from the REDUCED model
# can be thought of as how much the observed scores differ from the
# grand mean of the sample. Remember that we are ignoring group
# membership here because the REDUCED model assumes that the data
# is sampled from a single distribution with a common mean. Recall
# that our null hypothesis was that females and males have the same
# population mean. Think of the sum-of-squares error as how much the
# observed scores differ from the predicted score which is just the
# grand mean. The SSE for the REDUCED model is below.

sse.r2 <- sum((wonderlic.df$wonderlic - mean(wonderlic.df$wonderlic))^2)
sse.r2
```

```
## [1] 219.6364
```

```r
# The sum-of-squared errors in prediction from the FULL model can be
# thought of as how much the observed scores differ from the predicted
# scores from the FULL model. The best predicted score for each observation
# is the group mean. Thus, for males, the best predicted score according
# to the FULL model is 19.33333. For females, the best predicted score
# according to the FULL model is 25.6. The SSE for the FULL model is below.

grp.means2 <- rep(c(19.33333,25.6),c(6,5))
sse.f2 <- sum((wonderlic.df$wonderlic - grp.means2)^2)
sse.f2
```

```
## [1] 112.5333
```

```r
# The degrees of freedom in the Numerator can be calculated by
# determining how many terms were added to the REDUCED model to
# form the FULL model. In this case, the REDUCED model has one
# term (an intercept) while the FULL model has an intercept and
# a slope for the dummy-variable. Thus, the degrees of freedom in
# the numerator is 1. Only one term was added to the REDUCED model.

df.12 <- 1

# The degrees of freedom in the Denominator can be calculated by
# subtracting the total number of terms (estimated parameters) in
# the FULL model from the total sample size (i.e., # of observations).
# Because we have 11 observations and two terms in the FULL
# model, we have 11 - 2 = 9 degrees of freedom in the denominator.

df.22 <- dim(wonderlic.df)[1] - 2

# The F ratio comparing the FULL model versus REDUCED model is below.
# This equation is the same as Chapter 3, Equation 22 in MDK.

F.ratio2 <- ((sse.r2-sse.f2)/df.12)/(sse.f2/df.22)
F.ratio2
```

```
## [1] 8.565704
```

```r
# Because we have only 1 degree of freedom in the numerator, we
# can take the square root of the F statistic to obtain the
# equivalent t statistic.

t.ratio2 <- sqrt(F.ratio2)
t.ratio2
```

```
## [1] 2.926722
```

```r
# Now let's compare the result from our t statistic when we
# used the t.test() function. They are the same!

t.test(wonderlic ~ sex, data = wonderlic.df, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  wonderlic by sex
## t = -2.9267, df = 9, p-value = 0.01685
## alternative hypothesis: true difference in means between group male and group female is not equal to
## 95 percent confidence interval:
##  -11.11037  -1.42296
## sample estimates:
##   mean in group male mean in group female
##             19.33333             25.60000
```

```r
# Thus, when we conduct the two independent samples t test, we are
# COMPARING TWO MODELS (REDUCED model vs. FULL model) using a
# partial F test. The result is the same with the exception of
# the sign because we could calculate the mean difference as
# (females - males) or (males - females).

# We can apply the same logic when comparing means coming from
# three or more independent samples. This was also called
# one-way between-subjects analysis of variance. This was
# covered in Module 7.

# Comparing Three (or More) Independent Means (Module 7) ----

# With the first example comparing three independent groups,
# we were interested in whether three types of Conditions
# (TrtA, TrtB, TrtC) influenced Knowledge. This was a one-way
# between-subjects analysis of variance. Each participant
# contributed only one score on the dependent variable.

knowl <- data.frame(knowledge = c(22,14,24,17,19,25,20,26,29,35,28,37,40),
                    condition = factor(rep(c(1,2,3),c(5,4,4)),
                                       labels=c('TrtA','TrtB','TrtC')))
summary(knowl)
```

```
##    knowledge     condition
##  Min.   :14.00   TrtA:5
##  1st Qu.:20.00   TrtB:4
##  Median :25.00   TrtC:4
##  Mean   :25.85
##  3rd Qu.:29.00
##  Max.   :40.00
```

```r
aggregate(knowledge ~ condition, knowl, mean)
```

```
##   condition knowledge
## 1      TrtA      19.2
## 2      TrtB      25.0
## 3      TrtC      35.0
```

```r
# What is the REDUCED model?
# The REDUCED model has only an intercept. Think of this as a common
# mean---the grand mean. The REDUCED model assumes that the means are
# the same in all three groups. Thus, we have only an intercept
# in the REDUCED model.

# What is the FULL model?
# The FULL model allows the means to differ between the three
# groups. The FULL model, in this case, has an intercept
# and two slopes. There are two slopes because there are two
# dummy-variables in the model. It does not matter which group
# serves as the reference group. However, we should know how the
# coding is being done in statistical software. In our current
```

```r
# example, the three levels were entered as 1s, 2s, and 3s and
# the labels were 'TrtA', 'TrtB', and 'TrtC'. In R, the group
# with the 1s will serve as the reference group because 1 is the
# lowest number used for our variable. Thus, one dummy-variable
# indicates whether an observation is 'TrtB' or not and the other
# dummy-variable indicates whether an observation is 'TrtC' or not.
# Thus, with an intercept and two slopes (one for each dummy-variable),
# we have three parameters to estimate.

# The sum-of-squared errors in prediction from the REDUCED model
# can be thought of as how much the observed scores differ from the
# grand mean of the sample. Remember that we are ignoring group
# membership here because the REDUCED model assumes that the data
# is sampled from a single distribution with a common mean. Recall
# that our null hypothesis was that all three conditions have the same
# population mean. Think of the sum-of-squares error as how much the
# observed scores differ from the predicted score which is just the
# grand mean. The SSE for the REDUCED model is below.

sse.r3 <- sum((knowl$knowledge - mean(knowl$knowledge))^2)
sse.r3
```

```
## [1] 741.6923
```

```r
# The sum-of-squared errors in prediction from the FULL model can be
# thought of as how much the observed scores differ from the predicted
# scores from the FULL model. According to the FULL model, the best
# predicted score for a given observation is the mean of the group to
# which the observation belongs. The group means were 19.2, 25, and 35.
# The SSE for the FULL model is below.

grp.means3 <- rep(c(19.2,25,35),c(5,4,4))
sse.f3 <- sum((knowl$knowledge - grp.means3)^2)
sse.f3
```

```
## [1] 182.8
```

```r
# The degrees of freedom in the Numerator can be calculated by
# determining how many terms were added to the REDUCED model to form
# the FULL model. In this case, the REDUCED model has one term
# (an intercept) while the Full model has an intercept and two
# slopes (for the two dummy-variables). Thus, the degrees of freedom
# in the numerator is 2. Two terms were added to the REDUCED model.

df.13 <- 2

# The degrees of freedom in the Denominator can be calculated by
# subtracting the total number of terms (estimated parameters) in
# the FULL model from the total sample size (i.e., # of observations).
# Because we have 13 observations and three terms in the FULL
# model, we have 13 - 3 = 10 degrees of freedom in the denominator.

df.23 <- dim(knowl)[1] - 3
```

```
# The F ratio comparing the FULL model versus REDUCED model is below.
# This equation is the same as Chapter 3, Equation 22 in MDK.

F.ratio3 <- ((sse.r3-sse.f3)/df.13)/(sse.f3/df.23)
F.ratio3
```

```
## [1] 15.28699
```

```
# Now let's compare the result from our F statistic when we used
# the lm() function. They are the same!

oneway.fit <- lm(knowledge ~ condition, data = knowl)
anova(oneway.fit)
```

```
## Analysis of Variance Table
##
## Response: knowledge
##           Df Sum Sq Mean Sq F value    Pr(>F)
## condition  2 558.89  279.45  15.287 0.0009094 ***
## Residuals 10 182.80   18.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Notice that the numerator does not have 1 degree of freedom. Thus,
# we cannot take the square root of the F ratio to compute an
# equivalent t statistic.

# Up until now, we have only encountered one independent variable.
# When we have two or more independent variables and the levels
# are crossed, this is known as a factorial design. The logic of
# model comparison using an F test extends to more complicated models
# including factorial designs which we will cover in Module 11.

# Simple Linear Regression (Module 4) ----

# What models were being compared when we conducted a simple linear regression
# in Module 4. Recall that we wanted to predict Eating Difficulties using Stress.
# Stress was a continuous predictor (independent variable). That is, Stress was
# not a categorical predictor. The data are below.

stress_eat <- data.frame(student = LETTERS[1:10],
                         stress = c(17,8,8,20,14,7,21,22,19,30),
                         eat_difficulties = c(9,13,7,18,11,2,5,15,26,28))

str(stress_eat)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ student         : chr  "A" "B" "C" "D" ...
##  $ stress          : num  17 8 8 20 14 7 21 22 19 30
##  $ eat_difficulties: num  9 13 7 18 11 2 5 15 26 28
```

```r
summary(stress_eat)
```

```
##    student              stress      eat_difficulties
##  Length:10          Min.   : 7.00   Min.   : 2.00
##  Class :character   1st Qu.: 9.50   1st Qu.: 7.50
##  Mode  :character   Median :18.00   Median :12.00
##                     Mean   :16.60   Mean   :13.40
##                     3rd Qu.:20.75   3rd Qu.:17.25
##                     Max.   :30.00   Max.   :28.00
```

```r
# What is the REDUCED model?
# The REDUCED model has only an intercept. Think of this as a common
# mean---the grand mean. There are no groups. Thus, we have only an
# intercept in the REDUCED model.

# What is the FULL model?
# The FULL model has an intercept and a slope. We have one slope
# because we have one continuous predictor (Stress). Notice that our
# FULL model has two parameters that we must estimate. There is
# an intercept and a slope.

# The sum-of-squared errors in prediction from the REDUCED model
# can be thought of as how much the observed scores differ from the
# grand mean of the sample. Remember that there are no groups. Recall
# the null hypothesis being tested is that the population slope for
# Stress is equal to 0. The alternative hypothesis is that the population
# slope for Stress is not equal to 0. Think of the sum-of-squares as how
# much an observed score differs from the predicted score (grand mean) in
# the REDUCED model. The SSE for the REDUCED model is below.

sse.r4 <- sum((stress_eat$eat_difficulties - mean(stress_eat$eat_difficulties))^2)
sse.r4
```

```
## [1] 662.4
```

```r
# The sum-of-squared errors in prediction from the FULL model can be
# thought of as how much the observed scores deviate from the predicted
# scores from the FULL model. Remember that the FULL model has an intercept
# and a slope. The predicted scores will fall on a straight line. The
# SSE for the FULL model is below.

slr.mod <- lm(eat_difficulties ~ stress, data = stress_eat)
sse.f4 <- sum((stress_eat$eat_difficulties - slr.mod$fitted.values)^2)
sse.f4
```

```
## [1] 360.4354
```

```r
# The degrees of freedom in the Numerator can be calculated by
# determining how many terms were added to the REDUCED model to form
# the FULL model. In this case, the REDUCED model has one term
# (an intercept) while the FULL model has an intercept and one
# slope. Thus, the degrees of freedom in the numerator is 1.
```

```r
# Only one term was added to the REDUCED model.

df.14 <- 1

# The degrees of freedom in the Denominator can be calculated by
# subtracting the total number of terms (estimated parameters) in
# the FULL model from the total sample size (i.e., # of observations).
# Because we have 10 observations and two terms in the FULL
# model, we have 10 - 2 = 8 degrees of freedom in the denominator.

df.24 <- dim(stress_eat)[1] - 2

# The F ratio comparing the FULL model versus REDUCED model is below.

F.ratio4 <- ((sse.r4-sse.f4)/df.14)/(sse.f4/df.24)
F.ratio4
```

```
## [1] 6.702218
```

```r
# The F ratio is the same as what we found when we used the lm() function.

summary(slr.mod)
```

```
##
## Call:
## lm(formula = eat_difficulties ~ stress, data = stress_eat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8457  -3.5688  -0.0146   3.5642  10.7206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4005     5.4515   0.073   0.9432
## stress        0.7831     0.3025   2.589   0.0322 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.712 on 8 degrees of freedom
## Multiple R-squared:  0.4559, Adjusted R-squared:  0.3878
## F-statistic: 6.702 on 1 and 8 DF,  p-value: 0.03217
```

```r
# We have seen how ALL of the analyses that we have encountered this semester
# can be thought of as part of a larger overall model (General Linear Model).
# Each test is, in fact, just a comparison of models. We compare a REDUCED model
# to a FULL model and we make a decision about which model to retain.
# Do we retain the REDUCED model or do we go with the FULL model?

# As noted above, the next module will involve TWO categorical independent
# variables. We can test whether each independent variable has a main effect.
# However, we can also test whether the two independent variables interact to
# influence the dependent variable.
```