HANDOUT : Stat 230 - MLR - nested F-tests and Kitchen Sink

P.B. Matheson adapted from A.S. Wagaman

Blood Pressure and Boston Housing Prices

MLR: Nested F-Tests (3.6)

EXAMPLE 1 - Back to Blood Pressure example from previous lecture

We considered several models to predict blood pressure (BP). Nested F-tests allow us to formally check for significance of the explanatory variables that we add/drop using a nested F-test.

The models are provided here:

```
BP <- read.csv("https://pmatheson.people.amherst.edu/stat230/bloodpress.csv")
#just SLR predicting BP from weight
fm1 <- lm(BP ~ Weight, data = BP)

 #full model - using all predictors
fm.full <- lm(BP ~ ., data = BP)

#3 preds(age, dur -duration of hypertension, stress)
fm2 <- lm(BP ~ Weight +  Age + Dur + Stress, data = BP)

#2 predictors (weight and age)
fm3 <- lm(BP ~ Weight +  Age, data = BP) #2 predictors (weight and age)
```

Do a pairwise comparison between full model (fm.full) with all six predictors (had pulse and BSA [body surface area] too) and the model with 4 predictors (fm2)

```
anova(fm2,fm.full)
```

```
## Analysis of Variance Table
##
## Model 1: BP ~ Weight + Age + Dur + Stress
## Model 2: BP ~ Age + Weight + BSA + Dur + Pulse + Stress
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     15 4.5451
## 2     13 2.1559  2    2.3893 7.2037 0.007843 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You are testing significance of the GROUP of predictors different in the 2 models.
A significant test result means you need the more complicated model (at least one of the predictors in the more complex model has an additional coefficient that is non-zero - but it won't tell you which one!)
1) What can we say from this output?

```
# Add the models in ascending order of complexity.
anova(fm1, fm3, fm2, fm.full)
```

We can run all the nested F tests in one step.

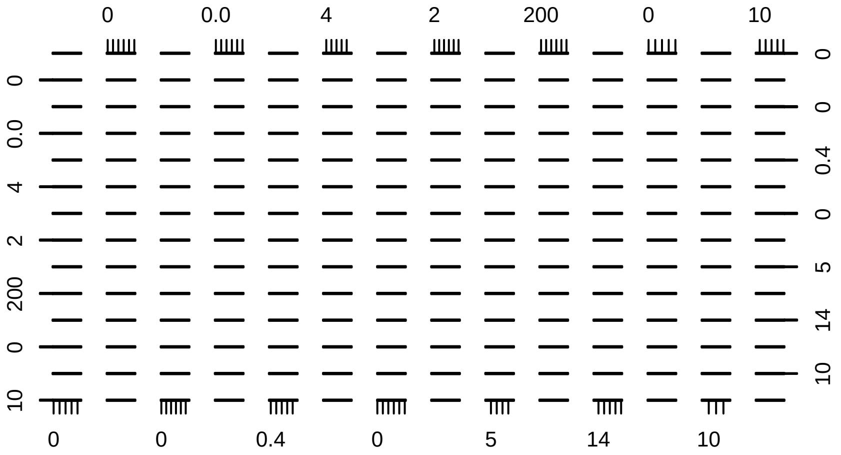
```
## Analysis of Variance Table
##
## Model 1: BP ~ Weight
## Model 2: BP ~ Weight + Age
## Model 3: BP ~ Weight + Age + Dur + Stress
## Model 4: BP ~ Age + Weight + BSA + Dur + Pulse + Stress
##   Res.Df    RSS Df Sum of Sq        F    Pr(>F)
## 1     18 54.528
## 2     17  4.824  1    49.704 299.7198 2.327e-10 ***
## 3     15  4.545  2     0.279   0.8406  0.453611
## 4     13  2.156  2     2.389   7.2037  0.007843 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. What does this tell us?

Leads us to question which variable should we add BSA or Pulse or both? More about that with Added Variable Plots (in 4.1)

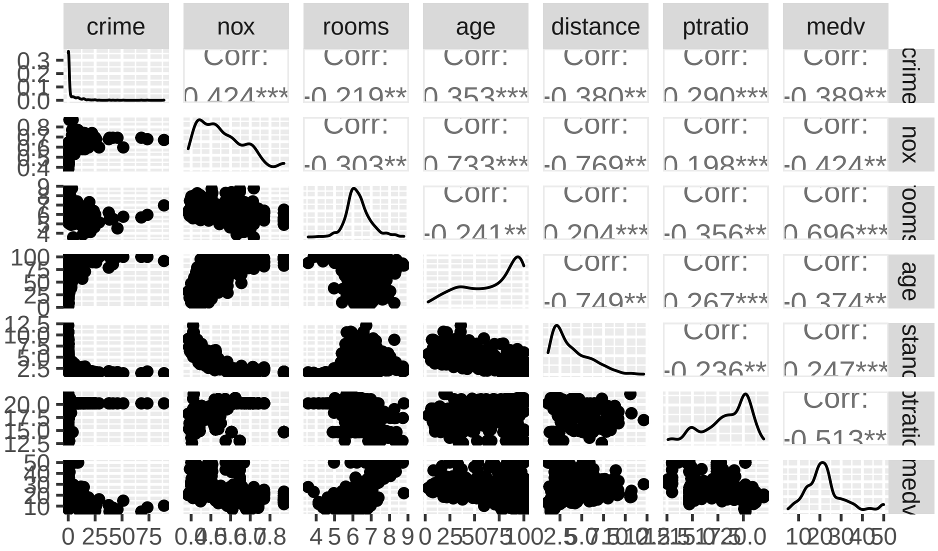
Example 2 - Boston Housing Prices

```
boston <- read.table("https://pmatheson.people.amherst.edu/stat230/boston.txt", header = T)
#usingbase R code below to produce a scatterplot matrix because ggpairs takes a while to render
plot(boston) #note this is *slightly* hard to read - take a peek
```

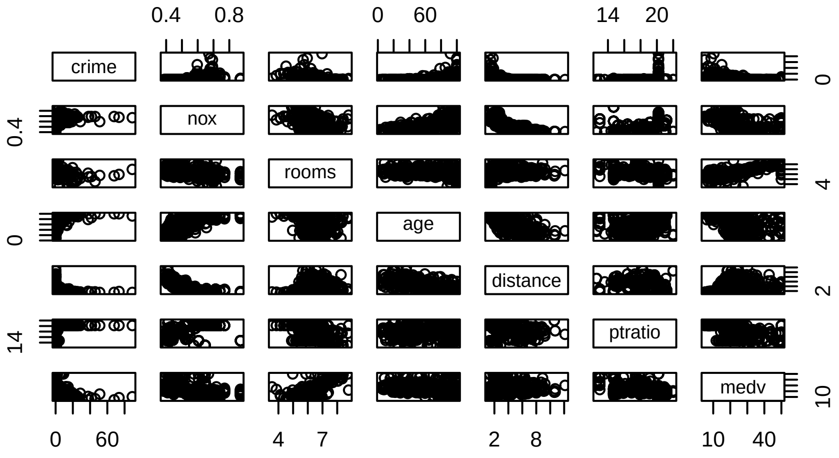


Let's work with the subset of predictors we've already identified to predict median value of the home - *medv*. Using the select statement we create a new datafile 'boston2' and now we can use ggpairs with fewer variables. ptratio is parent-teacher ration and distance is distance to 5 employment centers.

```
boston2 <- select(boston, "crime", "nox", "rooms", "age", "distance", "ptratio", "medv")
ggpairs(boston2)
```

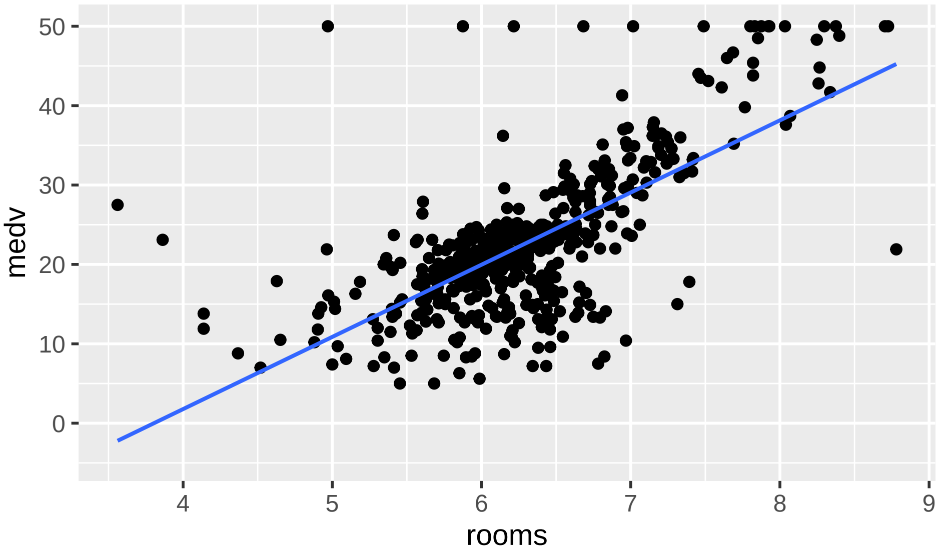


```
plot(boston2)
```



3. What do you see in these scatterplots? Which of the variables are most highly correlated with *medv*?

```
gf_point(medv ~ rooms, data = boston2) %>%
  gf_lm()
```



A First Model

```
fm1 <- lm(medv ~ rooms, data = boston2)
msummary(fm1)
```

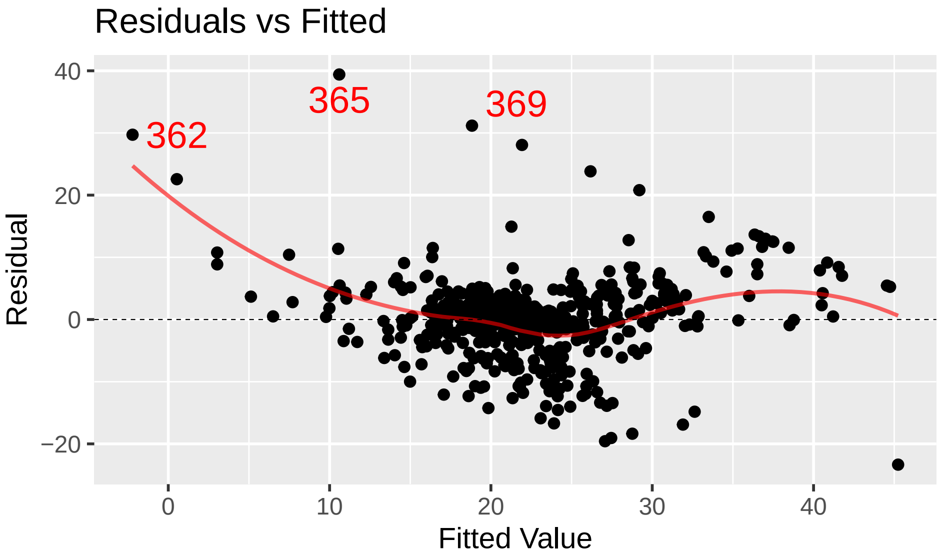
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.5943     2.6611  -13.00   <2e-16 ***
## rooms         9.0928     0.4207   21.61   <2e-16 ***
##
## Residual standard error: 6.638 on 498 degrees of freedom
## Multiple R-squared:  0.484,  Adjusted R-squared:  0.483
## F-statistic: 467.2 on 1 and 498 DF,  p-value: < 2.2e-16
```

4. How well does the SLR model for predicting median value using number of rooms work?

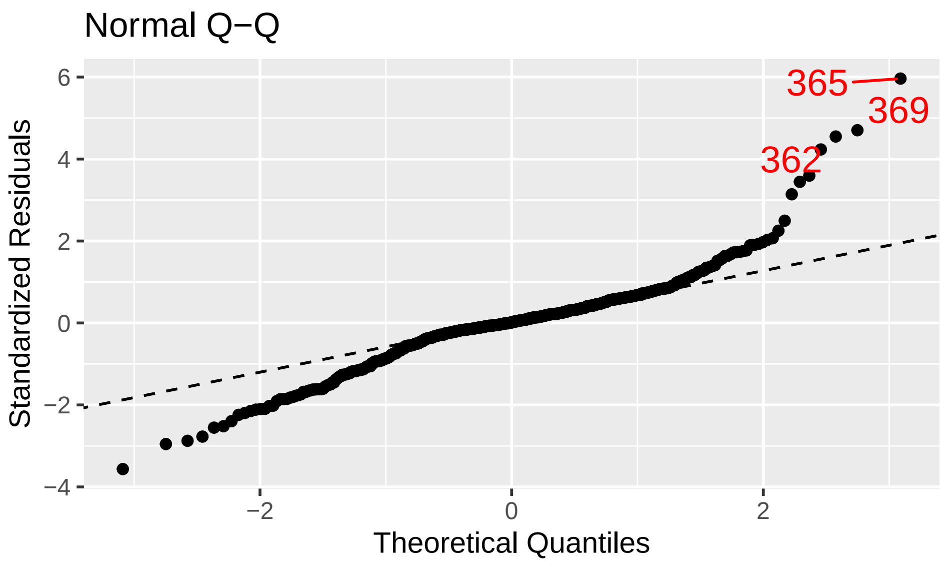
5. Check the assumptions for SLR. Are they met? (should ask before question #4)

```
mplot(fm1, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
mplot(fm1, which = 2)
```



The Kitchen Sink

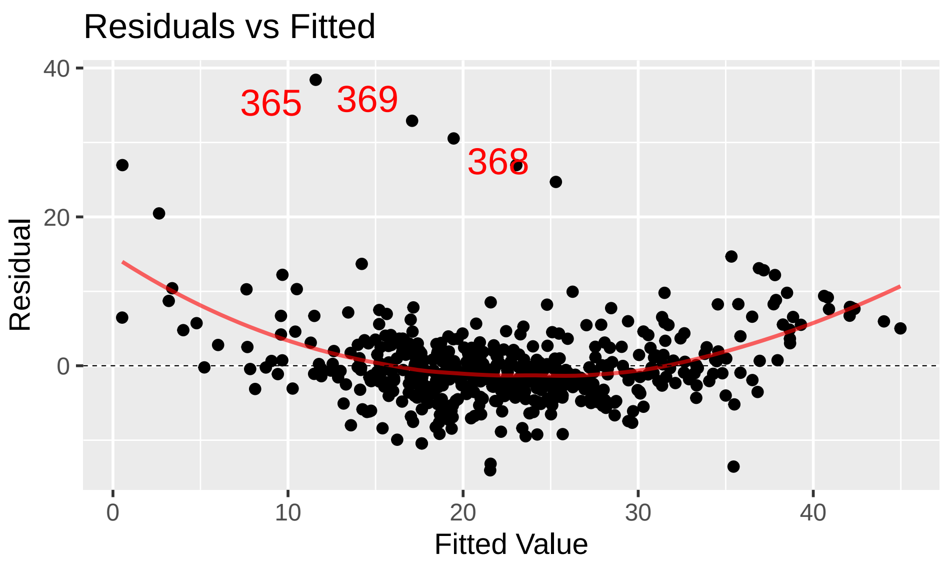
Without any intuition, one way to proceed is to simply throw all of the variables into our regression model. This is called our “kitchen sink” model, or just your “full” model. Note: the full model does NOT contain any interaction or polynomial terms.

```
# Using the . in the formula interface includes all non-response variables in the data frame
fm.full <- lm(medv ~ ., data = boston2)
msummary(fm.full)
```

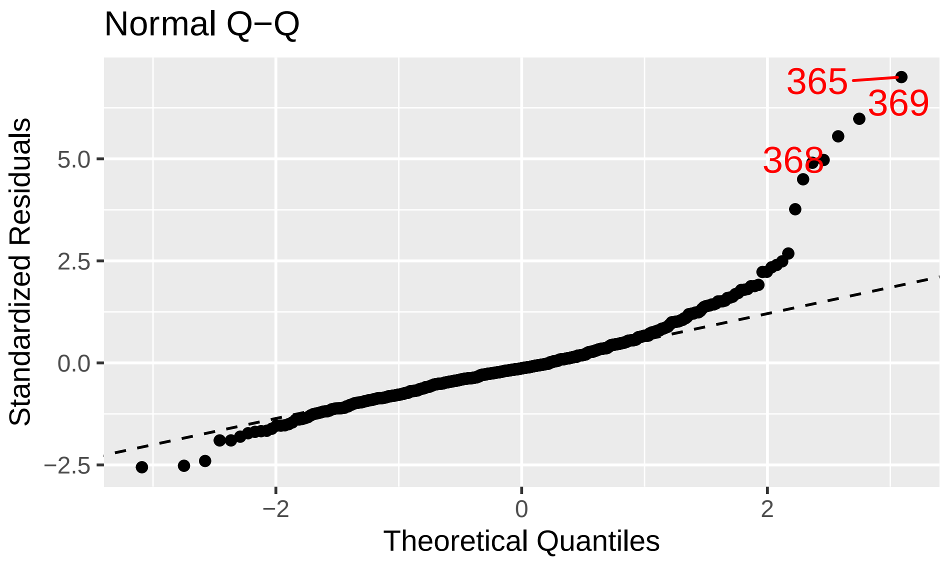
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.53068    5.00930   4.498 8.57e-06 ***
## crime        -0.15293    0.03253  -4.701 3.37e-06 ***
## nox         -22.04460    3.77024  -5.847 9.12e-09 ***
## rooms         6.68211    0.38910  17.173  < 2e-16 ***
## age          -0.05021    0.01439  -3.490 0.000526 ***
## distance     -1.31362    0.20402  -6.439 2.86e-10 ***
## ptratio      -1.12513    0.12769  -8.812  < 2e-16 ***
##
## Residual standard error: 5.52 on 493 degrees of freedom
## Multiple R-squared:  0.6468, Adjusted R-squared:  0.6425
## F-statistic: 150.5 on 6 and 493 DF,  p-value: < 2.2e-16
```

```
mplot(fm.full, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
mplot(fm.full, which = 2)
```



6. What do you think of the fit of this model? Do you prefer this model over the SLR?

```
cor(boston2)# correlation between every pair of variables in the data set
```

```
##               crime        nox      rooms        age   distance    ptratio
## crime     1.0000000  0.4241157 -0.2190410  0.3532444 -0.3798094  0.2898353
## nox       0.4241157  1.0000000 -0.3025448  0.7327918 -0.7693710  0.1982093
## rooms    -0.2190410 -0.3025448  1.0000000 -0.2407848  0.2041556 -0.3564510
## age       0.3532444  0.7327918 -0.2407848  1.0000000 -0.7492671  0.2674005
## distance -0.3798094 -0.7693710  0.2041556 -0.7492671  1.0000000 -0.2356956
## ptratio   0.2898353  0.1982093 -0.3564510  0.2674005 -0.2356956  1.0000000
## medv     -0.3886253 -0.4243319  0.6957300 -0.3742982  0.2472432 -0.5127620
##                medv
## crime    -0.3886253
## nox      -0.4243319
## rooms     0.6957300
## age      -0.3742982
## distance  0.2472432
## ptratio  -0.5127620
## medv      1.0000000
```

Note that if not all variables were quantitative or if you just wanted the correlation matrix for a subset of variables, you can use select to isolate the variables you want.

Let's set up some more models so we can show more nested F testing.

```
fm2 <- lm(medv ~ crime + nox, data = boston2)
fm3 <- lm(medv ~ crime + nox + rooms, data = boston2)
fm4 <- lm(medv ~ crime + nox + age, data = boston2)
fm5 <- lm(medv ~ crime + nox + age + rooms, data = boston2)
```

Remember that model 1 (fm1) from before had only rooms in it: fm1 <- lm(medv ~ rooms, data = boston2)

and the full model (fm.full) had all 6 predictors: fm.full <- lm(medv ~ crime + nox + age + rooms + ptratio + distance, data = boston2)

7a. In which models is model 1 nested?

7b. In which models is model 2 nested?

7c. In which models is model 3 nested?

7d. In which models is model 4 nested?

7e. In which models is model 5 nested?

To use the anova command on several models, the entire chain needs to be nested, so here, if we want to include model 1 (fm1), which just had rooms as a predictor, certain models can't be compared (fm2 and fm4). Here is the anova for the comparisons we can do:

```
anova(fm1,fm3,fm5,fm.full)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ rooms
## Model 2: medv ~ crime + nox + rooms
## Model 3: medv ~ crime + nox + age + rooms
## Model 4: medv ~ crime + nox + rooms + age + distance + ptratio
##   Res.Df   RSS Df Sum of Sq       F  Pr(>F)
## 1    498 21945
## 2    496 18592  2    3352.5 55.0145 < 2e-16 ***
## 3    495 18415  1     177.4  5.8235 0.01618 *
## 4    493 15021  2    3393.9 55.6939 < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

8. All three p-values are significant. What does this indicate?

a. Row 2 gives an F statistic of 55.01 and the smallest p-value R can provide. What set of hypotheses are being tested by this row?

b. Row 3 gives an F statistics of 5.82 and a p-value of 0.01618. What set of hypotheses are being tested by this row?

c. The comparison between fm3 and fm5 could also have been accomplished with which procedure?

d. Row 4 gives an F statistic of 55.7 and the smallest p-value R can provide. What set of hypotheses are being tested by this row?

Finally, notice if you just do a series of pairwise comparisons, your F statistics and p-values will be slightly off from what is reported above for the whole chain, because the df above here are being adjusted for testing multiple models at once. This is usually not a large enough concern to necessitate doing all comparisons pairwise, however, the CONCEPT of multiple testing and adjusting significance levels is one you should work at understanding.

What this means is that if you were to do separate commands (as shown below), your F test statistics and p values might be different. anova(fm1, fm3) anova(fm2, fm4) anova(fm3, fm5) anova(fm5, fm.full)

OPTIONAL: For the first exercise on BP, you could directly compute the sum of the squares and compare the ANOVA tables for two of the regression models. If interested the R code below compares the full model (fm.full with all predictors) to the model with four predictors (fm2) The computations are here for your reference, but the anova command we just learned works well enough for these.

```r
SSR.full <- sum((fm.full$fitted.values - mean(BP$BP))^2)
# same thing
var(fm2$fitted.values) * (nrow(BP) - 1)
```

```
## [1] 377.57
```

```r
SSR.reduce <- sum((fm2$fitted.values - mean(BP$BP))^2)
SSR.full - SSR.reduce
```

```
## [1] 17594.83
```