STA302 METHODS OF DATA ANALYSIS 1

MODULE 6: DECOMPOSING THE VARIANCE

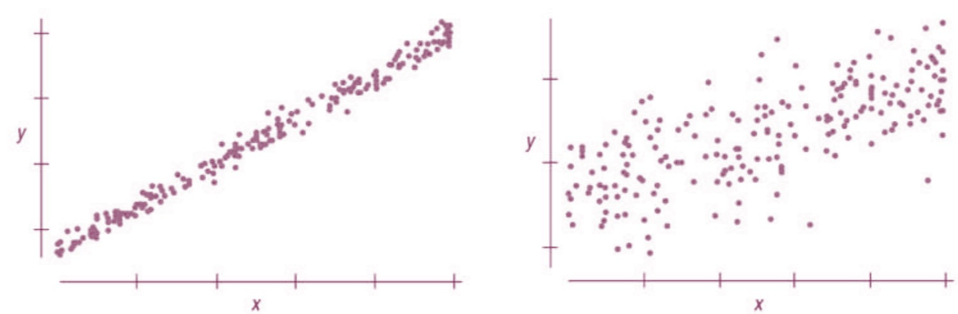
PROF. KATHERINE DAIGNAULT

MODULE 6 OUTLINE

1. Decomposition of the Sum of Squares
2. ANOVA Test for Overall Significance
3. Partial F Test

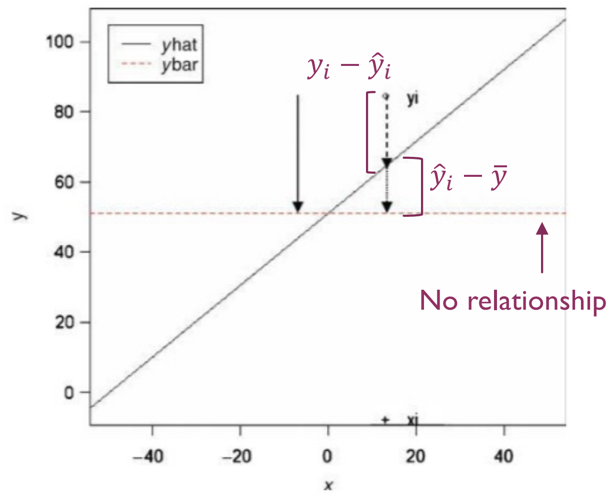
VARIATION IN LINEAR REGRESSION

- Fitting a linear model minimizes the RSS (i.e. variation around the line)
- Intuitively, we could use the idea of variation to decide which plot shows a stronger linear relationship
  - Left plot has less variation around the trend so trend is more prominent
  - Can also think of it as the predictor explains the spread better on the left than on the right
    - We have more leftover (i.e. residual) variation on the right
- The hypothesis test on slope in SLR also uses variation to decide on presence of significant linear relationship
  - If $\hat{\beta}_1 - 0$ is much bigger than variation expected in estimating $\hat{\beta}_1$, then 0 is not plausible value



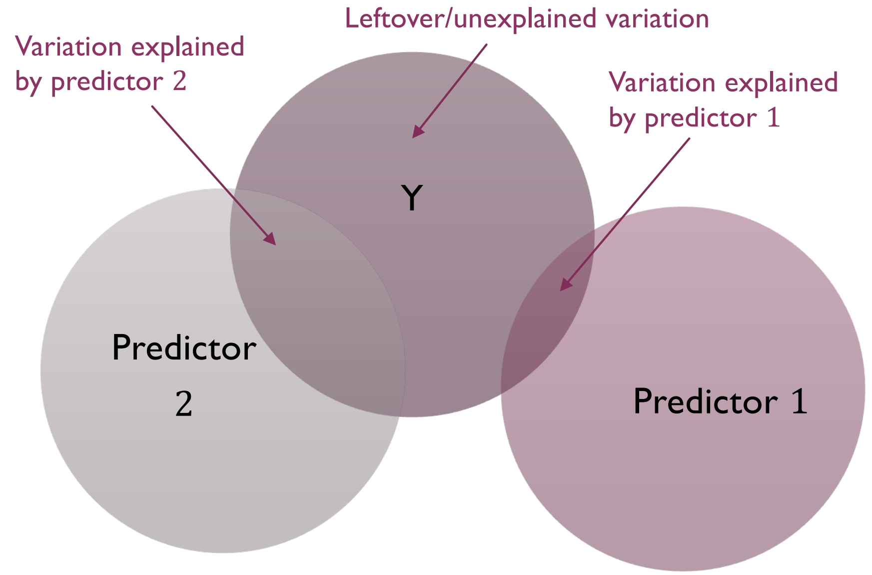
T-TEST IN SLR & VARIATION

- If we have no linear relationship, then $\beta_1 = 0$ which means $\hat{\beta}_0 = \bar{y}$
  - This also means $y_i = \beta_0 + \varepsilon$, a horizontal line
- When we have a linear relationship, there is a difference between $\bar{y}$ (no relationship) and $\hat{y}_i$
  - The farther from 0 the slope is, the bigger this difference, and the more significant the linear relationship
- But test statistic also compares $\hat{\beta}_1 - 0$ to variation in estimating $\hat{\beta}_1$
  - This variation includes $s^2 = {}^{\sum(y_i - \hat{y}_i)^2}/_{n-2}$
- For t-test to indicate significant linear relationship, $\hat{y}_i - \bar{y}$ across all observations (i.e. total) would need to overall be larger than residual differences $y_i - \hat{y}_i$ overall (i.e. total)



SOURCES OF VARIATION IN REGRESSION

- We aim to explain the variation/variability we see in our response
  - Quantified by $SST = \sum(y_i - \bar{y})^2$ (i.e. numerator of sample variance of y)
- Each predictor added to a model for y captures/explains a portion of this variation
  - Each is (hopefully) related to y and so explains the trend
  - Amount of overlap coincides with magnitude of difference $\hat{\beta}_j - 0$
- As more predictors are added, more variation in y is explained, but some will always be leftover



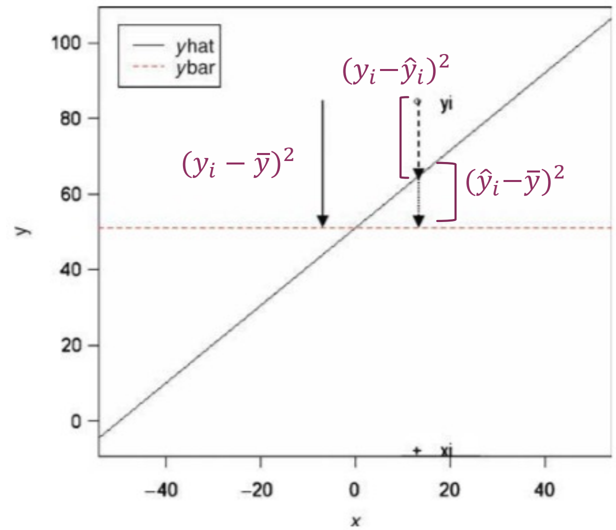
DECOMPOSITION OF SUM OF SQUARES

- There are three sources of variation in linear regression models:
  - Total amount of variation prior to fitting a model (ignoring predictors): $SST = \sum(y_i - \bar{y})^2$, called Total Sum of Squares
  - Overall variation explained by the predictors/the model:

    $SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$, called Regression Sum of Squares
  - Leftover/Unexplained variation leftover from fitting the model

    $RSS = \sum(y_i - \hat{y}_i)^2$, called the Residual Sum of Squares
- The decomposition relates these three sources of variation together:

$$SST = SS_{reg} + RSS$$



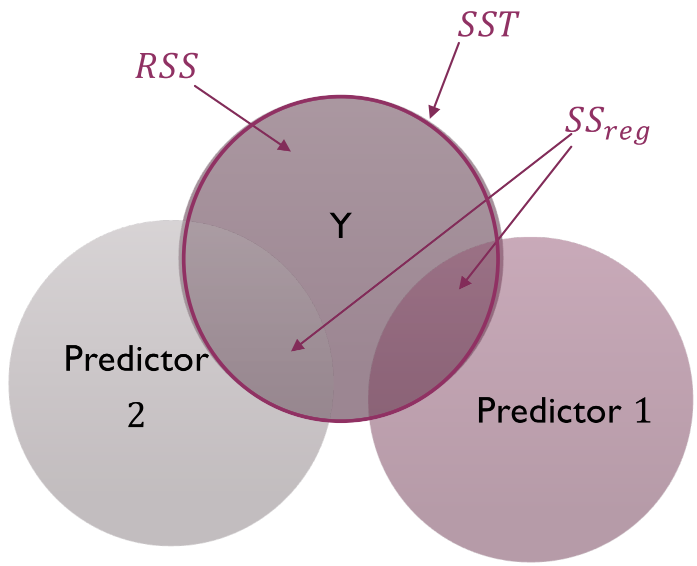
MORE ABOUT THE DECOMPOSITION

- The decomposition

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

can be shown to be true by expanding squares and manipulating terms

- Sum of squares are measures of distance, like numerator of sample variances
- Each term also has a degree of freedom (like denominator of sample variance)
  - SST has degree of freedom (df) $n-1$
  - RSS has degree of freedom (df) $n-p-1$ where p is number of predictors
  - SS_{reg} has degree of freedom (df) p
- DFs decompose too: $n-1=(n-p-1)+(p)$

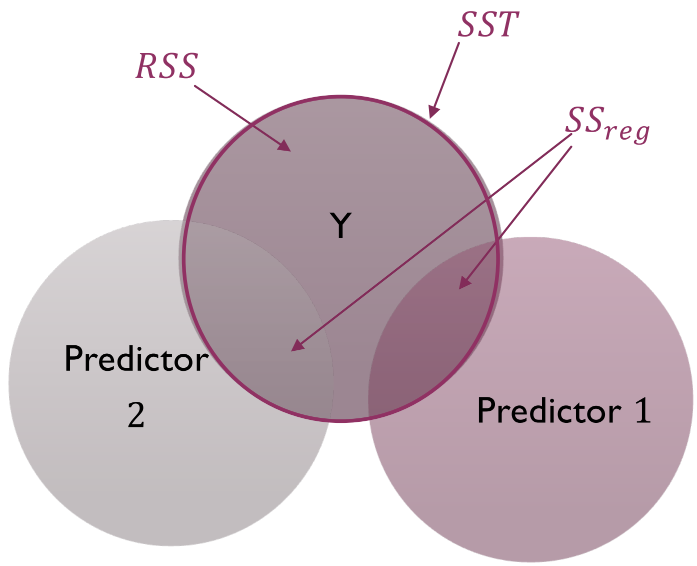


MODULE 6 OUTLINE

1. Decomposition of the Sum of Squares
2. ANOVA Test for Overall Significance
3. Partial F Test

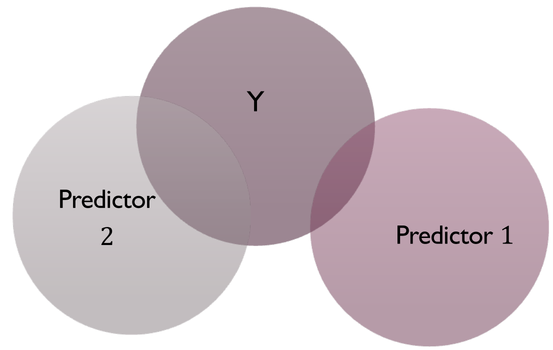
USING THE DECOMPOSITION TO BUILD A TEST

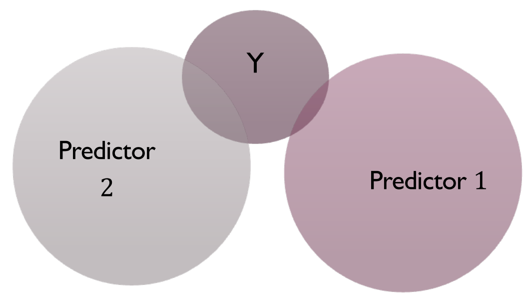
- T-test on slope in SLR tells us if the predictor is significantly related to response
  - Equivalent to saying a linear relationship exists
- T-tests on slopes in MLR can only say if this predictor is linearly related to response in presence of others.
  - Cannot give overall notion of whether a linear relationship exists in model
- Can use decomposition to create such a test
  - Indicates if at least one predictor overlaps Y enough to result in a significant linear relationship
  - Does this by checking if a significant amount of variation is explained



INTUITION BEHIND THE TEST

- Significant linear relationship = significant amount of variation in response was explained by model
  - So SS_{reg} would be large relative to SST
- Every dataset is different, so every data's SST is different
  - Cannot look at ratio of SS_{reg} to SST
- Due to decomposition, a model where ${}^{SS_{reg}}/_{SST}$ is large necessarily means ${}^{RSS}/_{SST}$ is small.
  - A good model therefore should have $SS_{reg} > RSS$
- Test of Overall Significance checks if SS_{reg} is significantly bigger than RSS





ANOVA TEST OF OVERALL SIGNIFICANCE

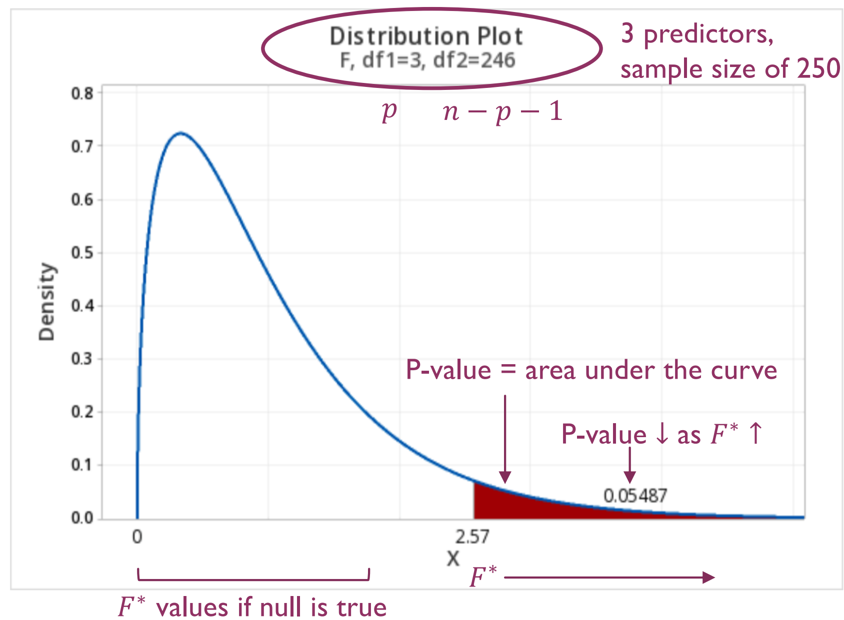
- Analysis of Variance Test of Overall Significance compares RSS and SS_{reg} to identify existence of a linear relationship
- Hypothesis: H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ versus H_a: $\boldsymbol{\beta}_1 \neq \mathbf{0}$ where we split our vector of coefficients $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)^T$
  - Testing the null "all slopes are zero" versus "at least one slope is not zero"
- Assuming the null is true, we test this using the statistic $F^* = \frac{SS_{reg}/p}{RSS/_{n-p-1}} \sim F(p, n-p-1)$
  - Dividing each SS by its df standardizes the quantity, giving us the Mean Squares Regression and Mean Squares Residual
- Test components sometimes summarized in an ANOVA table:

Source	DF	Sum Squares	Mean Squares	F value
Regression	p	SS_{reg}	$MS_{reg} = {SS_{reg}}/{p}$	${MS_{reg}}/{MSR}$
Residual	$n-p-1$	RSS	$MSR = {RSS}/{n-p-1}$	$-$
Total	$n-1$	SST	$-$	$-$

CONCLUDING ANOVA TEST

- F distribution assumes the null hypothesis is true.
  - Displays values we expect if no linear relationship exists
  - Expect to see small values of $F^* = \frac{SS_{reg}/p}{RSS/_{n-p-1}}$
- More extreme F^* values will be in the right tail
  - Bigger values are due to $MS_{reg} > MSR$
  - The larger this ratio, the larger the test statistic will be, and therefore the smaller the p-value
- If p-value $<\ \alpha$, or $F^* > F_{(1-\alpha),(p,n-p-1)}$, then reject H_0 and conclude a statistically significant linear relationship exists for at least one predictor

https://online.stat.psu.edu/stat200/book/export/html/213



EXAMPLE BY HAND & USING R

A model involving 3 predictors (X_1, X_2, X_3) is fit to a response Y using a sample of 30.

The response has sample variance $s_y^2 = 376.6853$.

The model yields an estimated error variance of $s^2 = 50.555$.

Find the elements of the ANOVA table and conclude the ANOVA test.

$$\boxed{H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs } H_a: \text{at least one } \beta_j \neq 0}$$

1. Degrees of freedom: $n = 30, p = 3$

$$df_{RSS} = 26, df_{SST} = 29, df_{SS_{reg}} = 3$$

2. Find the sums of squares:

$$SST = (n-1)s_y^2 = 29(376.6853) = 10923.87$$

$$RSS = (n-p-1)s^2 = 26(50.555) = 1314.43$$

$$SS_{reg} = SST - RSS = 10923.87 - 1314.43 = 9609.44$$

3. Find the mean squares:

$$MSR = {}^{RSS}/_{n-p-1} = {}^{1314.43}/_{26} = 50.555$$

$$MS_{reg} = {}^{SS_{reg}}/_{p} = {}^{9609.44}/_{3} = 3203.147$$

4. Compute test statistic: $F^* = \frac{MS_{reg}}{MSR} = \frac{3203.147}{50.555} = 63.36$

5. Conclude:

```
> qf(0.95, 3, 26)
[1] 2.975154
```

$\rightarrow 63.36 > 2.98$

Reject null and conclude significant linear relationship exists for at least one predictor.

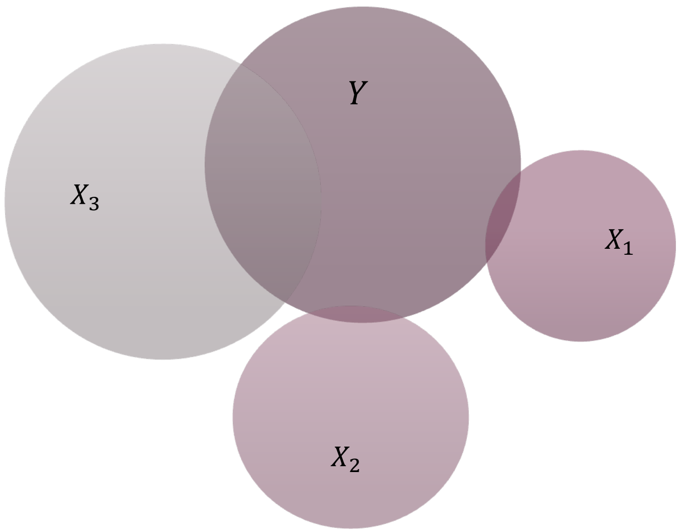
```
> data <- read.table("defects.txt", header=T)
>
> # fit a model using Defective as response, others are predictors
> model1 <- lm(Defective ~ Temperature + Density + Rate, data=data)
> summary(model1)

Call:
lm(formula = Defective ~ Temperature + Density + Rate, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7367  -4.1116  -0.5755   2.7617  16.3279

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3244    65.9265   0.157   0.8768
Temperature  16.0779     8.2941   1.938   0.0635 .
Density      -1.8273     1.4971  -1.221   0.2332
Rate          0.1167     0.1306   0.894   0.3797
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.11 on 26 degrees of freedom
Multiple R-squared:  0.8797,	Adjusted R-squared:  0.8658
F-statistic: 63.36 on 3 and 26 DF,  p-value: 4.371e-12
```

MODULE 6 OUTLINE

1. Decomposition of the Sum of Squares
2. ANOVA Test for Overall Significance
3. Partial F Test

TESTING A SUBSET OF PREDICTORS

- Suppose ANOVA test says at least one of X_1, X_2, X_3 is related to Y
  - Then t-test on each coefficient says only X_3 significantly related
- Is the simple model with only X_3 as good as the 3-predictor model?
  - Can we remove the two insignificant predictors at the same time?
    - Since each predictor is conditionally related to Y, t-test only says you can remove them one at a time.
- A Partial F Test can account for the conditionality to say if both X's can be removed simultaneously
  - If simple model as good as 3-predictor model, then SS_{reg} should be similar, and equivalently RSS of each model should be similar
  - If smaller model's RSS is too much bigger, then we can't remove all insignificant predictors at once.

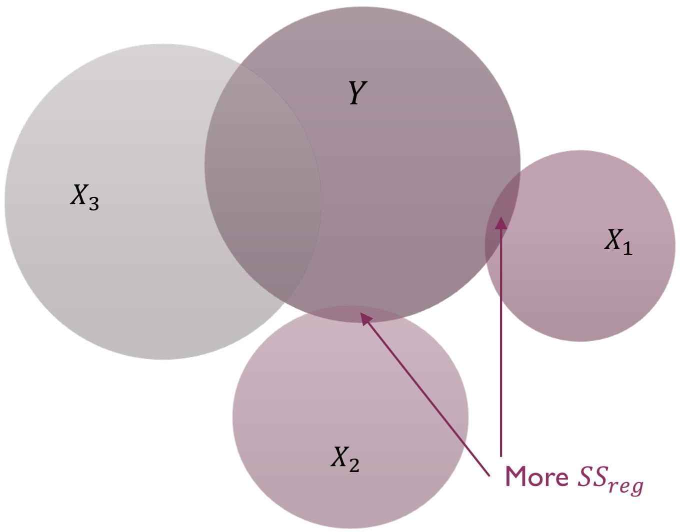


A TALE OF TWO MODELS

- A Partial F Test compares 2 models
  - Complete/full model (the bigger one) with p predictors, e.g. the 3-predictor model
  - Reduced model (the smaller one), e.g. the simple X_3 model
    - We say reduced model has $p-k$ predictors left (i.e. k were removed)
- Each model uses same data/response, so SST is the same.
- A model with more predictors (even non-significant ones) has smaller RSS so

$$RSS_{reduced} > RSS_{full} \Leftrightarrow SS_{reg,reduced} < SS_{reg,full}$$

- Adjust for additional predictors by standardizing with degrees of freedom of each model.



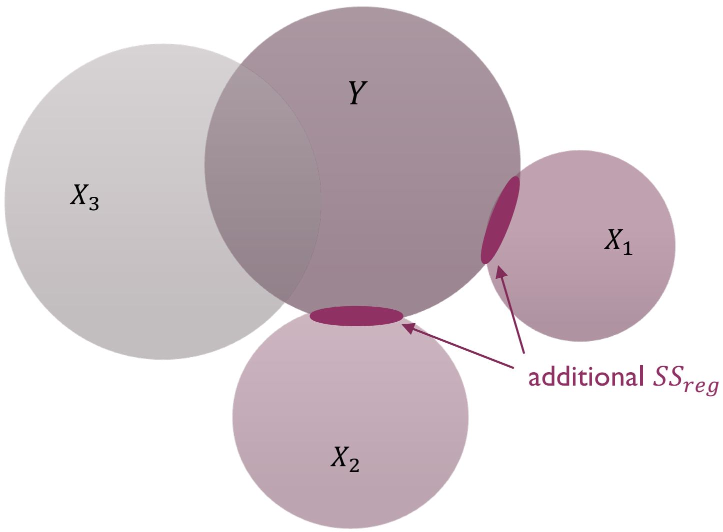
PARTIAL F TEST FOR SUBSETS OF MODELS

- Test is built assuming null hypothesis is true: $H_0\colon \boldsymbol{\beta}_2 = \mathbf{0}$ versus $H_a\colon\ \boldsymbol{\beta}_2 \neq \mathbf{0}$ where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$
  - $\boldsymbol{\beta}_2$ is a vector of k coefficients that were removed from the full model to make the reduced model.
    - i.e. reduced model has the $p - k$ predictors matching coefficients in $\boldsymbol{\beta}_1$, so testing if ok to drop k predictors because their β's are 0
- We quantify the difference between the RSS in the two models as $RSS_{drop} = RSS_{reduced} - RSS_{full}$
  - A substantially better full model gives a bigger RSS_{drop}, while similar full and reduced models give a smaller RSS_{drop}
- Like the ANOVA test, we look at a ratio of mean sums of squares, comparing the mean SS difference $\left({}^{RSS_{drop}}/_{k}\right)$ to the model with the smallest mean squares residual (i.e. the full model):

$$F^* = \frac{{}^{RSS_{drop}}/_{k}}{{}^{RSS_{full}}/_{n-p-1}} \sim F(k, n-p-1)$$

CONCLUDING THE PARTIAL F TEST

- The null assumes it is true that the smaller model is better (MSR_{drop} is small relative to MSR_{full})
- If $F^* < F_{(1-\alpha),(k,n-p-1)}$, we fail to reject the null
  - The additional SS_{reg} from k predictors not enough to keep the predictors $\leftarrow$ outcome we want if aiming for simpler/smaller models
  - Conclude there does not exist a significant linear relationship between Y and <u>any</u> of the k predictors.
- If $F^* > F_{(1-\alpha),(k,n-p-1)}$, or p-value $< \alpha$, we reject the null
  - Additional SS_{reg} from k predictors explains a lot of variation so we want to keep these predictors
  - Conclude there exists a significant linear relationship between Y and <u>at least one</u> of the k predictors.



EXAMPLE BY HAND & USING R

A model involving 3 predictors (X_1, X_2, X_3) is fit to a response Y using a sample of 30.

The response has sample variance $s_y^2 = 376.6853$.

The 3-predictor model yields an estimated error variance of $s^2 = 50.555$.

If a model with only X_1 yields an $RSS = 1497.03$, test whether this model is better.

H_0: $\beta_2 = \beta_3 = 0$ vs H_a: at least one $\beta_j \neq 0$

1. Define models and values:

Full model: $RSS = 1314.43,\ df = 26$

Reduced model: $RSS = 1497.03,\ df = 28$

2. Find RSS_{drop}:

$$RSS_{drop} = RSS_{reduced} - RSS_{full}$$
$$= 1497.03 - 1314.43 = 182.6$$

3. Compute test statistic:

$$F^* = \frac{RSS_{drop}/k}{RSS_{full}/n-p-1} = \frac{182.6/2}{1314.43/26} = 1.81$$

4. Conclude:

```
> qf(0.95, 2, 26)
[1] 3.369016
```

$1.81 < 3.37$

We fail to reject the null, so no significant linear relationship exists between Y and either X_2or X_3, so they can be removed from the full model.

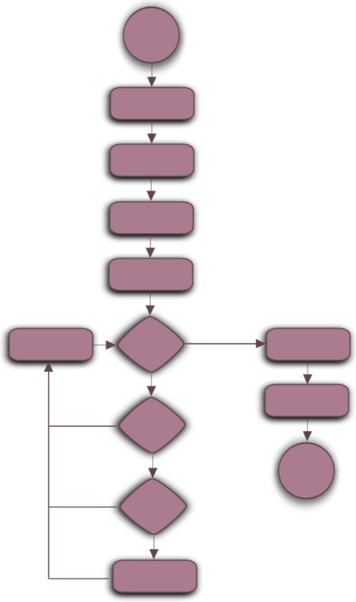
```
> # fit simple model using only Temperature
> model2 <- lm(Defective ~ Temperature, data=data)
> summary(model2)

Call:
lm(formula = Defective ~ Temperature, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.5952  -4.9203  -0.6253   4.2133  15.1861 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -40.938      5.298  -7.727 2.04e-08 ***
Temperature   30.904      2.327  13.279 1.32e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.312 on 28 degrees of freedom
Multiple R-squared:  0.863,	Adjusted R-squared:  0.8581 
F-statistic: 176.3 on 1 and 28 DF,  p-value: 1.317e-13

> # to conduct partial F test
> anova(model2, model1)
Analysis of Variance Table

Model 1: Defective ~ Temperature
Model 2: Defective ~ Temperature + Density + Rate
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 1496.8                           
2     26 1314.4  2    182.42 1.8042 0.1846
```

anova(reduced, full)

RSS_{drop}

CONSIDERATIONS FOR USING THESE TESTS

- General order of tests in an analysis:
  - ANOVA overall test → T-tests on individual coefficients → Partial F Test
  - T-tests can help inform what predictors you may consider wanting to drop
- Always check assumptions before using a hypothesis test of any kind
  - In partial F test, important to verify assumptions in both full and reduced models
- Caution when using partial F test:
  - Cannot compare models from different datasets
  - Cannot compare models that are not subsets of one another
    - E.g. model 1 uses predictors X_1, X_2, X_3, while model 2 uses predictors X_2, X_4 and model 3 uses X_2, X_3, partial F test can only be used on models 1 and 3



MODULE TAKE-AWAYS

1. What do the components of the sum of squares decomposition measure?
2. How are these components measured/computed?
3. What is the difference between the hypothesis for the ANOVA test and the Partial F test?
4. How is the decomposition used differently for the ANOVA test and the Partial F test?
5. How do we conduct each decomposition-based test?
6. What are the conclusions of each decomposition-based test?