STA302/STA1001, Weeks 10-11

Mark Ebden, 16–23 November (Section 1) and 23 November (Section 2)

With grateful acknowledgment to Alison Gibbs

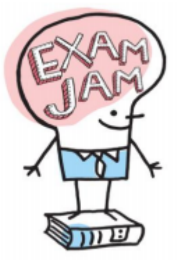
Overview

Multiple-regression ANOVA:

- The F-test
- R^2 and Adjusted R^2
- Interaction terms
- A first look at ANCOVA

Exam Jam

The STA302 review session will occur in SS 2135 from 10-11:30 am on 8 December. Please submit your requests for review topics closer to the time: there's a Piazza thread for this, under the 'Exam' topic.



In addition to our session: from 11 am to 3 pm there will be crafts, therapy dogs, a Photobooth, and other activities in the Sid Smith lobby. There will also be free coffee, juice, fruit, and granola bars there.

http://www.artsci.utoronto.ca/current/exam_jam

Recap of Regression ANOVA *(Week 3)*

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} b_1^2(x_i - \bar{x})^2}_{\text{SSReg}} + \underbrace{\sum_{i=1}^{n} \hat{e}_i^2}_{\text{RSS}}$$

Source	SS	d.f.	MS = SS/df
Regression line	$b_1^2 S_{xx} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$b_1^2 S_{xx}$
Error	$\sum_{i=1}^{n} \hat{e}_i^2$	$n-2$	S^2
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	–

The coefficient of determination is $R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}, \quad 0 \leq R^2 \leq 1.$

In Weeks 9–10 we showed that the ANOVA identity can be rewritten as:

$$\underbrace{\mathbf{Y}'\,(\mathbf{I} - \frac{1}{n}\mathbf{J})\,\mathbf{Y}}_{\text{SST}} = \underbrace{\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\text{SSReg}} + \underbrace{\mathbf{Y}'\,(\mathbf{I} - \mathbf{H})\,\mathbf{Y}}_{\text{RSS}}$$

Introducing Multiple-Regression ANOVA

In multiple regression, the ANOVA identity is the same as before, albeit with a different $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$:

$$\text{SST} = \text{SSReg} + \text{RSS}$$

$$\underbrace{\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\text{SST}} = \underbrace{\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\text{SSReg}} + \underbrace{\mathbf{Y}'\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}}_{\text{RSS}}$$

The MLR ANOVA table is similar to before, but the degrees of freedom have changed:

Source	SS	d.f.	MS = SS/df
Regression line	SSReg	p	SSReg$/p$
Error	RSS	$n - p - 1$	S^2
Total	SST	$n - 1$	–

The F-test in an MLR ANOVA table

The test hypotheses are:

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_a : At least one of the β_j's isn't 0

The test statistic is:

$$F_{\text{obs}} = \frac{\text{MSReg}}{\text{MSE}}$$

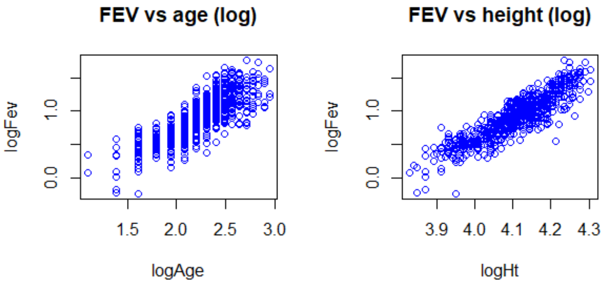
If H_0 is true, F_{obs} is an observation from an F distribution with $(p, n-p-1)$ degrees of freedom.

- Numerator d.f.: the # of β's being tested
- Denominator d.f.: the d.f. for the error

So in MLR ANOVA, we use the F-test to check for linear association between Y and *any* of the p predictors. If the F-test is significant, then we might ask, for *which* predictor(s) is there evidence of a linear association with Y? Some pitfalls in answering this question are investigated in Chapter 7.

Example of an F-test: the `fev` database

```
a2 = read.table("DataA2.txt",sep=" ",header=T) # Load the data set
logFev <- log(a2$fev); logAge <- log(a2$age); logHt <- log(a2$ht)
par(mfrow=c(1,2))
plot(logAge,logFev,type="p",col="blue",pch=21, main="FEV vs age (log)")
plot(logHt,logFev,type="p",col="blue",pch=21, main="FEV vs ht (log)")
mod1 = lm(logFev~logAge+logHt)
```



SLR in the `fev` database

```
##
## Call:
## lm(formula = logFev ~ logAge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60857 -0.13532  0.00227  0.14329  0.56348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.98772    0.05756  -17.16   <2e-16 ***
## logAge       0.84615    0.02535   33.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2026 on 652 degrees of freedom
## Multiple R-squared:  0.6309, Adjusted R-squared:  0.6303
## F-statistic:  1114 on 1 and 652 DF,  p-value: < 2.2e-16
```

SLR in the `fev` database

```
##
## Call:
## lm(formula = logFev ~ logHt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69369 -0.09122  0.01145  0.09832  0.44965
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.92110    0.25577  -46.61   <2e-16 ***
## logHt         3.12418    0.06223   50.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1512 on 652 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7941
## F-statistic:  2520 on 1 and 652 DF,  p-value: < 2.2e-16
```

MLR in the fev database

```
##
## Call:
## lm(formula = logFev ~ logAge + logHt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62020 -0.08894  0.01166  0.09807  0.46645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.29520    0.39196 -26.266  < 2e-16 ***
## logAge        0.18045    0.03346   5.392 9.74e-08 ***
## logHt         2.62968    0.11010  23.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1481 on 651 degrees of freedom
## Multiple R-squared:  0.8033, Adjusted R-squared:  0.8026
## F-statistic:  1329 on 2 and 651 DF,  p-value: < 2.2e-16
```

R^2 for MLR ANOVA

Let's consider the coefficient of determination for MLR ANOVA, a.k.a. the "coefficient of **multiple** determination":

$$R^2 = \frac{\text{SSReg}}{\text{SST}} = \frac{\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}{\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}$$

It's not the square of correlation r anymore! Correlation is between two variables, whereas we have potentially many variables now.

However, as before, it's the proportion of the total sample variability in the Y's explained by the regression model.

Question: What happens to R^2 when you add more predictor variables?

The effect on R^2 of additional predictors

Each time a predictor variable is added, SST stays the same because it depends on $\mathbf{Y}$ only.

However, adding a new predictor variable often improves (decreases) RSS: a richer model will often lead to a better fit, i.e. less error. Recall that RSS $= \hat{\mathbf{e}}'\hat{\mathbf{e}}$. A least-squares minimization of RSS, with additional predictors now, is minimizing over a larger-dimensional space. This guarantees that the minimum is at least as small. So, at worst, RSS will stay the same (if you add a predictor that's ignored by fitting $\hat{\beta}_j = 0$), and usually it will get better.

If SST is constant and RSS decreases, SSReg must increase. Therefore R^2 will increase. (Put another way, the $\mathbf{H}$ in the numerator will have changed.)

Adjusted R^2

Because R^2 generally increases with the number of predictors, how do we compare the R^2 for a simple model to the R^2 for a many-variable model?

We can use the **Adjusted $\mathbf{R}^2$**, a better measure of the model fit. It is adjusted for the number of predictors in the model.

$$\text{Adj } R^2 = 1 - (n-1)\,\frac{\text{MSE}}{\text{SST}} = 1 - \frac{n-1}{n-p-1}\,\frac{\text{RSS}}{\text{SST}}$$

With additional predictor variables, the Adjusted R^2 will only increase if MSE decreases.

Adjusted R^2 in action: First, reviewing regression ANOVA

For the `fev` vs `age` SLR dataset (HW2, question 1), $n = 654$ and $p = 1$.

From Weeks 9–10 slide 18, $R^2 \approx 0.5722$ and Adj $R^2 \approx 0.5716 \approx R^2$, a difference of approximately only 0.1%.

Taking logs, and rerunning the analysis, today we got $R^2 \approx 0.6309$ and Adj $R^2 \approx 0.6303 \approx R^2$.

Adjusted R^2 in action: MLR ANOVA

Let's compare the (adjusted) coefficients of determination for a small dataset, with and without an extra predictor.

Consider just the first ten points in the `fev` database (A = abridged):

```
set.seed(1)
N<-10; u <- sample(length(logFev),N)
logFevA<-logFev[u]; logAgeA<-logAge[u]
rA<-rnorm(N) # A new potential predictor

mod2 = lm(logFevA~logAgeA)
mod3 = lm(logFevA~logAgeA+rA)
summary(mod2) # SLR ANOVA
summary(mod3) # MLR ANOVA
```

Note that `rA` is noise, but adding it still increases the R^2.

Results of SLR ANOVA

```
##
## Call:
## lm(formula = logFevA ~ logAgeA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34977 -0.04767 -0.00790  0.10280  0.26091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6288     0.5944  -2.740  0.02544 *
## logAgeA       1.1232     0.2523   4.452  0.00213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1747 on 8 degrees of freedom
## Multiple R-squared:  0.7125, Adjusted R-squared:  0.6765
## F-statistic: 19.82 on 1 and 8 DF,  p-value: 0.002132
```

Results of MLR ANOVA

```
##
## Call:
## lm(formula = logFevA ~ logAgeA + rA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32561 -0.05576 -0.01012  0.05902  0.29785
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.72678    0.64144  -2.692  0.03099 *
## logAgeA      1.16367    0.27176   4.282  0.00365 **
## rA           0.03408    0.05727   0.595  0.57055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1822 on 7 degrees of freedom
## Multiple R-squared:  0.7263, Adjusted R-squared:  0.6481
## F-statistic: 9.289 on 2 and 7 DF,  p-value: 0.01072
```

Overview

Multiple-regression ANOVA:

- The F-test
- R^2 and Adjusted R^2
- **Interaction terms**
- A first look at ANCOVA

Regression model with interaction

An *additive* model (no interaction):

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{ht} + e$$

A model that is *not* additive (has an interaction term):

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{ht} + \beta_3 \,\text{age} \times \text{ht} + e$$

It can help us answer the question, "Does the relationship of `fev` with `age` depend on `height`?"

Two explanatory variables are said to *interact* if the effect that one of them has on the response depends on the value of the other.

How can we quantitatively assess this?

MLR ANOVA without interaction

```
##
## Call:
## lm(formula = logFev ~ logAge + logHt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62020 -0.08894  0.01166  0.09807  0.46645
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.29520    0.39196 -26.266  < 2e-16 ***
## logAge        0.18045    0.03346   5.392 9.74e-08 ***
## logHt         2.62968    0.11010  23.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1481 on 651 degrees of freedom
## Multiple R-squared:  0.8033, Adjusted R-squared:  0.8026
## F-statistic:  1329 on 2 and 651 DF,  p-value: < 2.2e-16
```

MLR ANOVA with interaction

```
##
## Call:
## lm(formula = logFev ~ logAge * logHt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64913 -0.08337  0.01099  0.09729  0.42260
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.5057     1.5322  -2.941 0.003392 **
## logAge        -2.4648     0.6781  -3.635 0.000300 ***
## logHt          1.2039     0.3809   3.160 0.001649 **
## logAge:logHt   0.6495     0.1663   3.906 0.000104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1465 on 650 degrees of freedom
## Multiple R-squared:  0.8078, Adjusted R-squared:  0.8069
## F-statistic: 910.4 on 3 and 650 DF,  p-value: < 2.2e-16
```

Considering the t-test result

We called `lm(logFev~logAge*logHt)`, which is equivalent to calling `lm(logFev~logAge+logHt+logAge:logHt)`

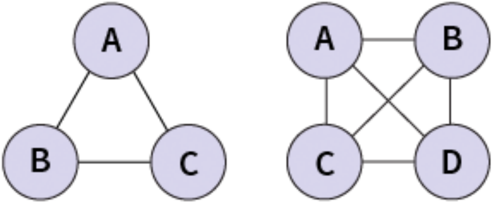
From the t-test regarding `logAge:logHt`, we can conclude that we have evidence that the coefficient of age $\times$ ht is statistically significantly different from 0, given that the other terms are in the model.

Note that this model has a slightly smaller MSE and larger Adj R^2 than the additive model.

We can conclude that adding the interaction term is worthwhile.

Should we routinely add interaction terms? (Hint: consider combinatorics.)

When to add interaction terms



When to add them can also be considered a research question.

However, a standard practice is that if an interaction term is in the model, we also include the individual terms for the predictor variables, even if their coefficients are not statistically significantly different from 0.

Next steps

- Try Chapter 5's **question 2**
- Remember that on Tuesday **21 November** we'll start at 11:10 am
- Solutions to Chapter 5's question 1 will be uploaded by 23 November

Appendix

What happens when we add a Height^2 term and all interactions?

```
##
## Call:
## lm(formula = logFev ~ logAge * logHt * logH2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66838 -0.08213  0.00931  0.09914  0.41712
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1960.63    1760.15  -1.114    0.266
## logAge                979.32     744.44   1.316    0.189
## logHt                1425.69    1310.11   1.088    0.277
## logH2                -345.49     324.95  -1.063    0.288
## logAge:logHt         -714.54     550.66  -1.298    0.195
## logAge:logH2          173.53     135.77   1.278    0.202
## logHt:logH2            27.91      26.86   1.039    0.299
## logAge:logHt:logH2    -14.02      11.16  -1.257    0.209
##
## Residual standard error: 0.1465 on 646 degrees of freedom
## Multiple R-squared:  0.8088, Adjusted R-squared:  0.8067
## F-statistic:  390.4 on 7 and 646 DF,  p-value: < 2.2e-16
```

no significant p values

pretty high confidence for F-statistic

Analysis of the multi-parameter model

The F-test had a null hypothesis that $\beta_1 = \cdots = \beta_7 = 0$, which was rejected with a p-value below 10^{-15}.

However, the p-values for the seven tests were all well above 0.05. Namely, $H_0 : \beta_j = 0$ was never rejected.

Do we conclude that all of the β_j's should be zero?

Answer: No, each t-test is for the effect of one explanatory variable given that the others are in the model.

Second example: A meadowfoam experiment

Meadowfoam is a flower found on the West Coast, which produces an oil of use to the cosmetics and hair-care industries.

A randomized experiment was conducted to explore the effect of growing conditions on the number of flower blooms per plant.

Example 2: dataset overview

There were $6 \times 2 = 12$ unique treatments, for:

- Six light intensities: `Intensity` $\in \{150, 300, 450, 600, 750, 900\}$, measured in μmol/m^2/s
- Two timings at which light began: `Time` is 1 if early, 2 if late

Each treatment was applied in two trials, so there were 24 trials in total.

The response variable, Y, known as `Flowers` in the dataset, was the average number of flowers observed per plant (across ten plants in a single pot).

Two questions of interest: What's the effect on the number of flowers per plant, of:

- Timing
- Light intensity

A scientific paper with background, as optional reading:
`http://agris.fao.org/agris-search/search.do?recordID=US9500398`

The data

```
library(Sleuth3)
print(case0901)
```

```
##    Flowers Time Intensity
## 1     62.3    1       150
## 2     77.4    1       150
## 3     55.3    1       300
## 4     54.2    1       300
## 5     49.6    1       450
## 6     61.9    1       450
## 7     39.4    1       600
## 8     45.7    1       600
## 9     31.3    1       750
## 10    44.9    1       750
## 11    36.8    1       900
## 12    41.9    1       900
## 13    77.8    2       150
## 14    75.6    2       150
## 15    69.1    2       300
## 16    78.0    2       300
## 17    57.0    2       450
## 18    71.1    2       450
## 19    62.9    2       600
```

Strategy

keeping categorical since relationship can be nonlinear

We'll set categorical variable t to be 0 or 1 for late and early, respectively.

We'll also treat `Intensity` as a *categorical* variable (!)

We can do this because `Intensity` has a small number of values with multiple observations for each. This approach may be useful for learning which intensity leads to the highest value of response variable, without imposing a particular form of relationship on `Intensity` versus `Flowers`. It may be linear, quadratic, etc.

Shall we define six new indicator variables?

$$i150 = \begin{cases} 1 & \text{if Intensity} = 150 \\ 0 & \text{otherwise} \end{cases} \qquad \cdots \qquad i900 = \begin{cases} 1 & \text{if Intensity} = 900 \\ 0 & \text{otherwise} \end{cases}$$

1 hot encoding

problem: more complex, no notion of one indicator higher than another (which is OK).

Economical representation

Using all six indicator variables is redundant: e.g. if five variables are zero, you know that the sixth is 1. In the 24×8 design matrix, the columns for $i150, \ldots i900$ contain a linear dependence.

This will lead to an error in R when running the `lm` command.

In general: For a categorical variable with k categories, you need $k - 1$ indicator variables.

The model will be:

$$Y = \beta_0 + \beta_1 i150 + \beta_2 i300 + \beta_3 i450 + \beta_4 i600 + \beta_5 i750 + \beta_6 t + e$$

i900 is a reference default level, when all others zero, i900=1

Results

This code ensures that `Intensity` $= 900$ will be the reference level, as it's listed first:

```
i <- factor(case0901$Intensity, levels=c(900,150,300,450,600,750))
myFit <- lm(Flowers ~ i + as.factor(Time), data=case0901)
summary(myFit)
```

treat it as categories

The fitted model is:

$$\hat{Y} = 37.8 + 29.4i150 + 20.2i300 + 16.0i450 + 6.1i600 + 1.6i750 + 12.2t$$

When `Intensity` is 150, and `Time` is early, what's the estimate of the mean number of flowers per plant? 37.8 + 29.4 * 1 + 12.2 * 1

What does the intercept estimate?

when t=0 (late) and i900

```
##
## Call:
## lm(formula = Flowers ~ i + as.factor(Time), data = case0901)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.979 -4.308 -1.342  5.204 10.204
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         37.846      3.629  10.430 8.33e-09 ***
## i150                29.350      4.751   6.178 1.01e-05 ***
## i300                20.225      4.751   4.257 0.000532 ***
## i450                15.975      4.751   3.362 0.003697 **
## i600                 6.125      4.751   1.289 0.214601
## i750                 1.600      4.751   0.337 0.740415
## as.factor(Time)2    12.158      2.743   4.432 0.000365 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.719 on 17 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.7606
## F-statistic: 13.18 on 6 and 17 DF,  p-value: 1.427e-05
```

intercept significant is reasonable, because with predictor=0 for all, we'd expect some flowers grown

closer to i900, less significant

Is timing important?

Let's consider $H_0 : \beta_6 = 0$ versus $H_a : \beta_6 \neq 0$. The test statistic is about 4.43, with a p-value calculated from a t_{17} distribution of about 0.0004.

Yes, timing is important. After accounting for the effect of intensity, there is strong evidence that the mean of the number of flowers per plant differs between the early and late timings.

Holding intensity constant, we get on average 12.2 flowers per plant more with early timing.

Is intensity important?

Let's consider $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. The p-value is less than 0.0001, so there is strong evidence that $\beta_1 \neq 0$ given other variables in the model.

If $\beta_1 = 0$, the model would be the same for intensities of 150 and 900.

So **yes** intensity is important. We conclude that a light intensity of 150 gives, on average, a different number of flowers per plant than an intensity of 900.

Individual tests for $\beta_1, \ldots \beta_5$ compare the mean response at a certain intensity to that at an intensity of 900.

Is intensity important?

What we really want to test is $H_0 : \beta_1 = \cdots = \beta_5 = 0$ versus H_a : at least one of $\beta_1, \ldots \beta_5$ isn't zero.

We should run a **partial *F*-test**. This tests whether a subset of β's are zero simultaneously.

The approach is:

1. Fit the model with all predictor variables (known as the *full model*), and calculate RSS, known as RSS(full)
2. Fit the model without the predictor variables whose coefficients we're testing (known as the *reduced model*), and calculate RSS, known as RSS(reduced)
3. Calculate the observed F:

$$F = \frac{(\mathrm{RSS(reduced)} - \mathrm{RSS(full)}) \,/\, (\mathrm{df}_{\mathrm{reduced}} - \mathrm{df}_{\mathrm{full}})}{\mathrm{RSS(full)} \,/\, \mathrm{df}_{\mathrm{full}}}$$

The partial F-test

We know that:

- RSS in reduced model $\geq$ RSS in full model
- SSReg in reduced model $\leq$ SSReg in full model
- SST in reduced model $=$ SST in full model

Note that df_{full} is the number of degrees of freedom in the error for the full model. The difference $\text{df}_{\text{reduced}} - \text{df}_{\text{full}}$ is the number of parameters that you're testing in the partial F-test.

It can be shown that, under H_0, F_{obs} has an F distribution with $(\text{df}_{\text{reduced}} - \text{df}_{\text{full}}, \text{df}_{\text{full}})$ degrees of freedom.

The **intuition** behind the test is: Did RSS go down by a statistically significant amount when new predictors were added to the model?

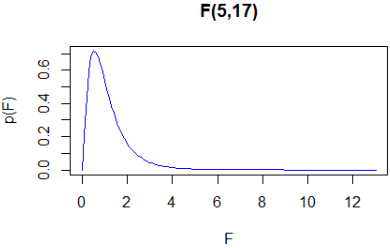
- Equivalently: Did R^2 increase by a statistically significant amount?

Back to our example: Is intensity important?

$H_0 : \beta_1 = \cdots = \beta_5 = 0$ versus H_a : at least one of $\beta_1, \ldots \beta_5$ isn't zero.

We obtain a test statistic of

$$F_{\text{obs}} \approx \frac{(3451 - 767)/5}{767/17} \approx 11.9$$



There is strong evidence that not all of $\beta_1, \ldots \beta_5$ are zero, given that time is in the model. So we have reconfirmed that **yes** intensity is important.

The ANOVA table for the Meadowfoam dataset

We have decomposed SSReg into two components: intensity and timing.

partial F test is square of t test for 1 variable (i.e. timing)

Source	df	SS	MS	F
Regr(timing)	1	887	887	$887/45.15 = 19.6$
Regr(intensity)	5	2684	538	$538/45.15 = 11.9$
Error	17	767	45.15	
Total	23	4338		

Note that $887/45.15 \approx 19.6 \approx (4.43)^2$.

Also note that we could carry out a partial F-test on $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, i.e. on one parameter. Of course, this assumes all other variables are in the model.

Exercise: Try this for β_5, i.e. for $i750$'s coefficient.

New question

Does the way light intensity affects the mean of the number of flowers per plant depend on timing?

In setting up a model to answer this question, we'll continue to model timing as a qualitative variable, but we'll begin to model intensity as a quantitative variable.

Analysis of Covariance (ANCOVA)

In ANCOVA, the predictors include both quantitative variables and qualitative variables, e.g. $d \in \{0, 1\}$.

literally 2 parallel lines

Parallel regression lines:

intensity, quantitative

$$Y = \beta_0 + \beta_1 x + \beta_2 d + e$$

timing, qualitative

Regression lines with equal intercepts but different slopes:

$$Y = \beta_0 + \beta_1 x + \beta_3 d\,x + e$$

Unrelated regression lines:

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 d\,x + e$$

The last cases are examples of introducing an *interaction*, as we saw earlier.

Using ANCOVA to answer our new question

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 d x + e$$

We'll test whether the resulting change in Y (`Flowers`) when x (`Intensity`) changes is the same for early- versus late timings ($d = 1$ or 0). In other words, $H_0 : \beta_3 = 0$.

This isn't the same as asking: "Is the relationship between Y and `Intensity` the same for early and late timings? Do they have the same line?" (What is the hypothesis test in that case?)

beta2 = beta3 = 0 ?

The R code for our test is:

```
myFit <- lm(Flowers ~ Intensity * as.factor(Time), data=case0901)
summary(myFit)
```

R output

```
##
## Call:
## lm(formula = Flowers ~ Intensity * as.factor(Time), data = case0901)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -9.516 -4.276 -1.422  5.473 11.938
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                71.623333   4.343305  16.491 4.14e-13 ***
## Intensity                  -0.041076   0.007435  -5.525 2.08e-05 ***
## as.factor(Time)2           11.523333   6.142360   1.876   0.0753 .
## Intensity:as.factor(Time)2  0.001210   0.010515   0.115   0.9096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.598 on 20 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:  0.7692
## F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```

not significant, so no interaction

Conclusions regarding the interaction

From the unrelated-regressions model, there is no evidence that the effect of light intensity on the number of flowers per plant differs with timing ($p \approx 0.91$).

If there were significant interactions (as we saw in the `fev` example), it would be difficult to talk about the effects of the individual predictor variables because they'd depend on the value of others.

Next step: Since the coefficient of interaction is not statistically significantly different from 0, remove it so that we can talk about the individual effects of timing and intensity.

R output

```
##
## Call:
## lm(formula = Flowers ~ Intensity + as.factor(Time), data = case0901)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.652 -4.139 -1.558  5.632 12.165
##
## Coefficients:
```

intercept too high?

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      71.305833   3.273772  21.781 6.77e-16 ***
## Intensity        -0.040471   0.005132  -7.886 1.04e-07 ***
## as.factor(Time)2 12.158333   2.629557   4.624 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.441 on 21 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:   0.78
## F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

Continued analysis

There is strong evidence ($p < 0.001$) that light intensity affects the number of flowers per plant over and above timing.

For a given timing, increasing the light intensity by 100 μmol/m^2/s decreases the number of flowers per plant on average by approximately 4.0.

Exercise: Show that the 95% CI for this decrease is $(-5.1, -3.0)$.

There is strong evidence ($p \approx 0.0001$) that timing affects the number of flowers per plant over and above light intensity.

For a given intensity, introducing early timing increases the number of flowers per plant on average by approximately 12.2.

Exercise: Show that the 95% confidence interval for this increase is (6.7, 17.7).

Continued analysis

We could have fit two separate regression lines by splitting the data into the twelve early observations and the twelve late observations.

Advantages of using ANCOVA included:

- We have tests for equal slopes and intercepts
- We have higher df_{error}, meaning the power increases and the CIs are narrower
- We get a better estimate of the error variance based on 24 observations rather than 12

A possible disadvantage of using ANCOVA was:

- An implicit assumption that both groups have the same error variance