STA302/STA1001, Weeks 10-11

Mark Ebden, 16 November (Section 1) and 23 November (Section 2), 2017

With grateful acknowledgment to Alison Gibbs

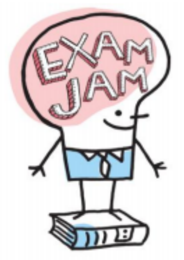
Overview

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 $R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}, \quad 0 \le R^2 \le 1.$

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

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

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

Introducing Multiple-Regression ANOVA

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

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

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

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:

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

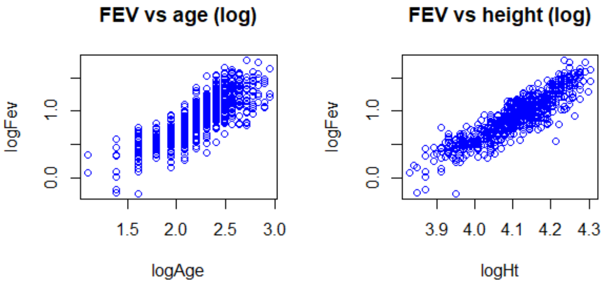
If H_0 is true, F_{obs} is an observation from an F distribution with $(p,\, n - p - 1)$ degrees of freedom.

- 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)`

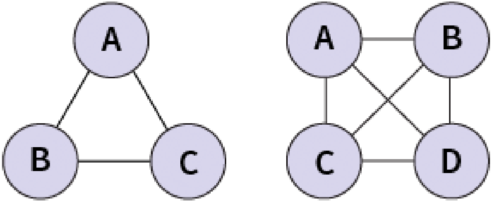
From the *t*-test regarding `logAge:logHt`, we can conclude that we have evidence that the coefficient of age $\times$ ht is statistically significantly different from 0, given that the other terms are in the model.

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.

Analysis of Covariance (ANCOVA)

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

Parallel regression lines:

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

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 case is an example of introducing an *interaction* as before.

In the next lecture we'll consider the analysis for ANCOVA in more depth.

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