Chapter 5
Week 12 – Tuesday
Applied Regression Analysis (STAT 757)

Paul J. Hurtado

5 Apr, 2016

Data

Merge and reshape data: `dplyr`, `tidyr`, `reshape2`, ...

```
head(dat1,4)
```

```
##   ID Species  Weight1  Weight2
## 1  1       A 11.37811 10.87522
## 2  2       B 12.01119 11.72802
## 3  3       C 12.60281 12.88633
## 4  4       A 11.32140 11.30118
```

```
dat2
```

```
##   Species Avg.Weight
## 1       A         11
## 2       B         12
## 3       C         13
```

```
dat3 = merge(dat1,dat2,by="Species",sort=FALSE)
head(dat3,3)
```

```
##   ID Species Avg.Weight  Weight1  Weight2
## 1  1       A         11 11.37811 10.87522
## 4  2       B         12 12.01119 11.72802
## 8  3       C         13 12.60281 12.88633
```

Data

Convert from Wide to Long format with `tidyr::gather()`

```
head(dat3,3)
```

```
##   ID Species Avg.Weight  Weight1  Weight2
## 1  1       A         11 11.37811 10.87522
## 4  2       B         12 12.01119 11.72802
## 8  3       C         13 12.60281 12.88633
```

```
dat=gather(dat3,Replicate,Weight,Weight1:Weight2)
dat$Replicate <- type.convert(gsub('Weight','',dat$Replicate))
head(dat,6)
```

```
##   ID Species Avg.Weight Replicate   Weight
## 1  1       A         11         1 11.37811
## 2  2       B         12         1 12.01119
## 3  3       C         13         1 12.60281
## 4  4       A         11         1 11.32140
## 5  5       B         12         1 11.92289
## 6  6       C         13         1 12.89738
```

More in data.R, at `www.statmethods.net`, and RStudio's Data Wrangling cheatsheet:

www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

ANCOVA Example

Consider the model with a "dummy" covariate $d \in \{0, 1\}$:

$$\begin{aligned}\mathbf{Y} &= \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{d} + \beta_3 \mathbf{X}\mathbf{d} + \mathbf{e} \\ &= (\beta_0 + \beta_2 \mathbf{d}) + (\beta_1 + \beta_3 \mathbf{d})\mathbf{X} + \mathbf{e}\end{aligned}$$

In **R**, we would specify this as either

```
lm(y~x+d+d:x)
# or equivalently (see ?formula for details)
lm(y~x*d)
```

Goal: Test null hypothesis H_0: $\beta_2 = 0 = \beta_3$

See the `travel.txt` example in the Ch. 5 R code at
http://www.stat.tamu.edu/~sheather/book/

Partial F-Test

Consider a null model which assumes $\beta_1, \ldots, \beta_k = 0$.

$$H_0\text{: } \beta_i = 0 \text{ for } i = 1, 2, \ldots, k$$

Call that null model the *reduced* model. Then we use the test statistic

$$\begin{aligned} F &= \frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{RSS_{\text{full}}/df_{\text{full}}} \\ &= \frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/k}{RSS_{\text{full}}/(n-1-p)} \end{aligned}$$

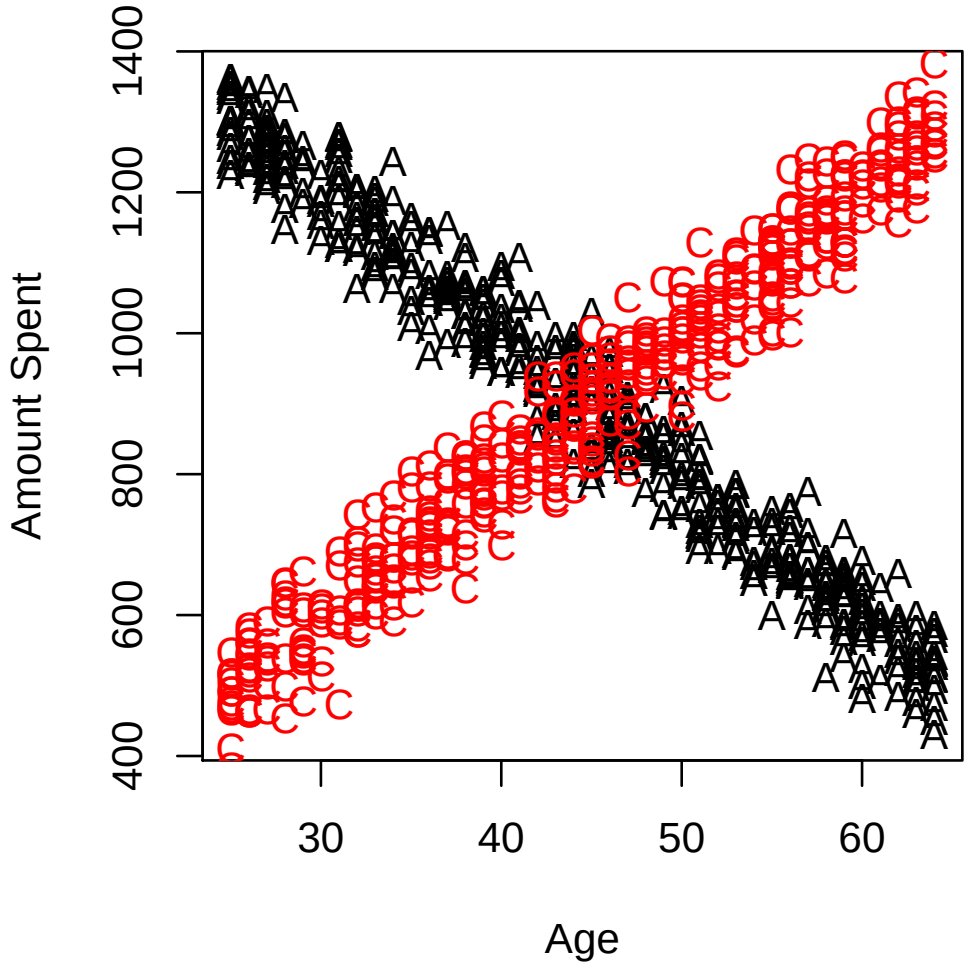
When the null hypothesis is true, the **partial F-statistic** is F distributed with parameters k and $df_{\text{full}} = n - 1 - p$.

ANCOVA Example

```
travel <- read.table("travel.txt",header=TRUE); attach(travel)
mfull <- lm(Amount~Age+C+C:Age)
summary(mfull)  #Regression output on page 141
```

```
##
## Call:
## lm(formula = Amount ~ Age + C + C:Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -143.298  -30.541   -0.034   31.108  130.743
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1814.5445     8.6011   211.0   <2e-16 ***
## Age           -20.3175     0.1878  -108.2   <2e-16 ***
## C           -1821.2337    12.5736  -144.8   <2e-16 ***
## Age:C          40.4461     0.2724   148.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.63 on 921 degrees of freedom
## Multiple R-squared:  0.9601, Adjusted R-squared:  0.9599
## F-statistic:  7379 on 3 and 921 DF,  p-value: < 2.2e-16
```

```
#Figure 5.7 on page 142
par(mfrow=c(1,1), oma=c(0,0,0,0),mar=c(4,4,2,1))
plot(Age[C==0],Amount[C==0],pch=c("A"),col=c("black"),ylab="Amount Spent",xlab=
points(Age[C==1],Amount[C==1],pch=c("C"),col=c("red"))
```



```
#Regression output on page 143
mreduced <- lm(Amount~Age)
summary(mreduced)
```

```
##
## Call:
## lm(formula = Amount ~ Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -545.06 -199.03    6.34  198.74  497.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 957.9103    31.3056  30.599   <2e-16 ***
## Age          -1.1140     0.6784  -1.642    0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 237.7 on 923 degrees of freedom
## Multiple R-squared:  0.002913, Adjusted R-squared:  0.001833
## F-statistic: 2.697 on 1 and 923 DF,  p-value: 0.1009
```

```
#Regression output on page 144
anova(mreduced,mfull)
```

```
## Analysis of Variance Table
##
## Model 1: Amount ~ Age
## Model 2: Amount ~ Age + C + C:Age
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1    923 52158945
## 2    921  2089377  2  50069568 11035 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
detach(travel)
```

Due Thursday!

Your **project proposal** should address the following:

1. **Questions** (include description of system, data)
2. **Modeling Framework?**
3. **Tests, Diagnostics, etc.** needed to answer question(s)

I'm looking for two things!

1. Is it feasible?
2. Is it non-trivial?

Diagnostics

(See Chapter 6)