36-617: Applied Linear Models

SS, F, interactions, dummies

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Reading, HW, Quiz

- Quiz 01
  - Available around 5pm today
  - Due around 5pm tomorrow
  - Open book/notes/etc
- Reading for next week:
  - Sheather Ch 6
  - Supplemental: ISLR 3.3.3, G&H Ch4
- HW 01 due tonight 1159pm
- HW 02 out later today – due Mon 1159pm

Outline

- SS Decompositions and F Statistics
  - Some Comments
- Interactions
  - The Hierarchy Principle
- Categorical and Dummy Variables
- ANOVA models
- ANCOVA models

Review: Matrix Form of Regression

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Usually $x_{i0} \equiv 1$, so we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Review: Distributional Properties: $\hat{\beta}$

$$\boxed{\text{Fact: } Y \sim N(\mu, \Sigma) \Rightarrow AY \sim N(A\mu, A\Sigma A^T)}$$

$$y \sim N(X\beta, \sigma^2 I)$$

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T y \\
E[\hat{\beta}] &= E[(X^T X)^{-1} X^T y] \;=\; (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T X\beta = \beta \\
\text{Var}\,(\hat{\beta}) &= \text{Var}\,((X^T X)^{-1} X^T y) \\
&= (X^T X)^{-1} X^T \text{Var}\,(y) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T X (X^T X)^{-1} \sigma^2 \;=\; (X^T X)^{-1} \sigma^2
\end{aligned}$$

$$\Rightarrow \boxed{\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)}$$

Review: H, & Distribution of $\hat{y}$ and $\hat{e}$

$$\hat{y} = X\hat{\beta} = X\left[(X^TX)^{-1}X^Ty\right] = \underbrace{\left[X(X^TX)^{-1}X^T\right]}_{\text{Hat matrix, H}}y = Hy$$

$$HX = X(X^TX)^{-1}X^TX = X$$

$$\Rightarrow \forall\beta^*, HX\beta^* = X\beta^*$$

$$H^T = H \text{ and } (I-H)^T = (I-H)$$

$$HH = H \text{ and } (I-H)(I-H) = (I-H)$$

$$E[\hat{y}] = E[Hy] = HE[y] = HX\beta = X\beta$$

$$\text{Var}(\hat{y}) = \text{Var}(Hy) = H\text{Var}(y)H^T = HH\sigma^2 = H\sigma^2$$

$$\Rightarrow \boxed{\hat{y} \equiv Hy \sim N(X\beta, H\sigma^2)}$$

Similarly, $\boxed{\hat{e} = y - \hat{y} \equiv (I-H)y \sim N(0, (I-H)\sigma^2)}$

SS Decompositions and F Statistics

$$\text{Fact: } y^T A y + y^T B y \;=\; (y^T A + y^T B) y \;=\; y^T (A+B) y$$

$$\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \;=\; (y-\bar{y})^T (y-\bar{y}) \\
&= y^T (I-H_1)^T (I-H_1) y \;=\; y^T (I-H_1) y \\
SS_{reg} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \;=\; (\hat{y}-\bar{y})^T (\hat{y}-\bar{y}) \;=\; y^T (H-H_1) y \\
RSS &= \sum_{i=1}^{n} (y_i - \hat{y})^2 \;=\; (y-\hat{y})^T (y-\hat{y}) \;=\; y^T (I-H) y
\end{aligned}$$

$$\Rightarrow SS_{reg} + RSS \;=\; y^T (H-H_1) y + y^T (I-H) y \;=\; y^T (I-H_1) y \;=\; SST$$

- *Cochran's Theorem* implies: SS_{reg}/σ^2 and RSS/σ^2 are indep. χ^2's under $H_0: \ \beta_1 = \cdots = \beta_p = 0$
- The df for each χ^2 will be the rank, or equivalently the trace, of each defining matrix. Using $tr(AB) = tr(BA)$: $tr(H) = p+1$, $tr(H_1) = 1$, $tr(I) = n$, so

$$df(SS_{reg}/\sigma^2) = p, \ df(RSS/\sigma^2) = n-p-1, \ df(SST/\sigma^2) = n-1$$

https://en.wikipedia.org/wiki/Cochran's_theorem

$H_1 = hat\ matrix\ for\ y = \beta_0 + \epsilon$

SS Decompositions and F Statistics

- The foregoing lead to the traditional Analysis of Variance Table

Source of variation	Degrees of freedom (df)	Sums of squares (SS)	Mean square (MS)	F
Regression	p	$SSreg$	$SSreg/p$	$F = \frac{SSreg/p}{RSS/(n-p-1)}$
Residual	$n-p-1$	RSS	$RSS/(n-p-1)$	
Total	$n-1$	SST		

- We can define the "multiple R^2" as:

$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST} \quad (\;=\; \mathrm{Corr}\,(y, \hat{y})^2 \;)$$

- "Adjusted R^2": mean-squares instead of sums of squares, to account for capitalization on chance

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

SS Decompositions and F Statistics

Since $SST = y^T(I - H_1)y = RSS_{H_1}$, the residual sum-of-square from the smallest model (intercept-only), the F statistic from the Anova table can be written as

$$F = \frac{SS_{reg}/p}{RSS/n-p-1} = \frac{(RSS_{H_1} - RSS_H)/(df_{H_1} - df_H)}{RSS_H/df_H} \qquad (*)$$

This idea, and the sum-of-square calculations we did earlier, can be generalized so that, if H_{full} and $H_{reduced}$ are hat matrices from a "full" model, and from a "reduced" model obtained by linear restrictions on the "full" model, then the ***partial F statistic***

$$F = \frac{(RSS_{H_{reduced}} - RSS_{H_{full}})/(df_{H_{reduced}} - df_{H_{full}})}{RSS_{H_{full}}/df_{H_{full}}}$$

will have an F distribution under the null hypothesis that the linear restrictions are true.

(*) The df throughout are ***residual*** df, that is, tr(I-H)

Some Comments

- It's good to know the "canonical" theory of the linear model and the Analysis of Variance table
  - Distribution assumptions and multiple testing matters
  - We will more fully discuss later in the course
- The "linear restrictions" for the partial F statistic usually amount to just setting some β's = 0 . This is especially useful when a regressor is categorical, since a categorical X is recoded as a set of dummy variables, one for each level of X
- The partial F test brings us into "variable selection"
  - We will more fully discuss variable selection later as well!

Interactions

- An interaction is nothing more than the product of two (or more) X-variables.
  - If 2 X-variables, it is a 2-way interaction
  - If 3 X-variables, it is a 3-way interaction
  - Etc.
- The Hierarchy Principle
  - If a k-way interaction is included in the model, then all lower-order interactions,$(k-1)$-way, $(k-2)$-way, etc., should usually also be included.
  - Helps with <u>*flexibility*</u> and <u>*interpretability*</u> of the model!

Interpretation of Interactions

- Often people say "an interaction shows how two X-variables affect each other". **This is incorrect.**
- An interaction is appropriate when *the influence of one X-variable on y is affected by the value of one or more other X-variables.*

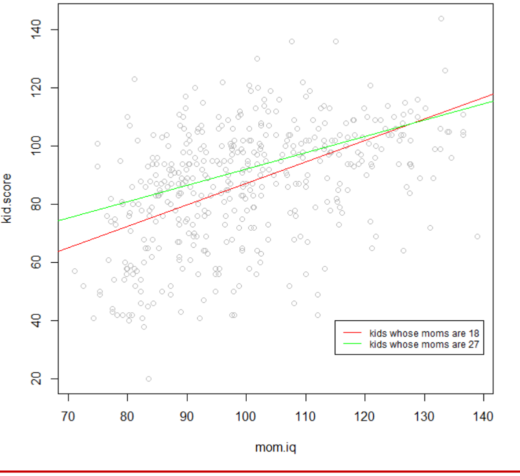
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$
$$y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon$$
$$y = \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1 + \epsilon$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \epsilon$$
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_1 X_2 + (\beta_3 + \beta_5 X_1 + \beta_6 X_2 + \beta_7 X_1 X_2) X_3 + \epsilon$$

Example: Interaction of mom's age and iq

```
kidiq <- read.csv("kidiq.csv",header=T)
lm.0 <- lm(kid.score ~ mom.age *
   mom.iq,data=kidiq)
round(summary(lm.0)$coef,2)
##                    Est     SE      t  pval
## (Intercept)    -32.69  51.68 -0.63  0.53
## mom.age          2.55   2.21  1.15  0.25
## mom.iq           1.10   0.51  2.18  0.03
## mom.age:mom.iq -0.02   0.02 -0.99  0.32
##
## (kid.score)
##  = -32.69 + 2.55*(mom.age) + 1.10*(mom.iq)
##     - 0.02*(mom.age)*(mom.iq)
##  = -32.69 + 2.55*(mom.age)
##     + (1.10 - 0.02*(mom.age))*(mom.iq)
##
## For kids whose moms are 18:
##  (kid.score)
##   = -32.69 + 2.55*(18)
##      + (1.10 - 0.02*(18))*(mom.iq)
##   =  13.21 + 0.74*(mom.iq)
```

```
## For kids whose moms are 27:
##  (kid.score)
##   = -32.69 + 2.55*(27)
##      + (1.10 - 0.02*(27))*(mom.iq)
##   =  36.16 + 0.56*(mom.iq)
plot(kid.score ~
  mom.iq,data=kidiq,col="grey")
abline(13.21,0.74,col="red")
abline(36.16,0.56,col="green")
legend(115,40,lty=1,col=c("red","green"),
  cex=0.75,legend=
  c("kids whose moms are 18",
    "kids whose moms are 27"))
```



More on the Hierarchy Principle

- **Hierarchy Principle for Interactions:**

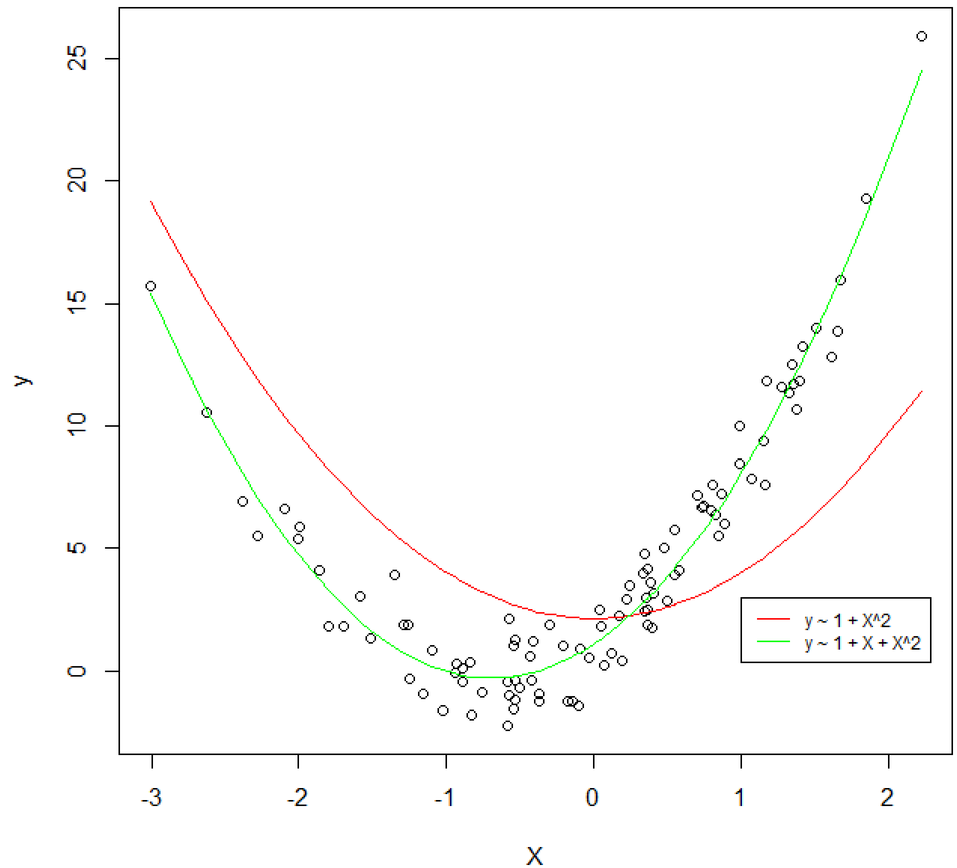
  Include all lower-order interactions with a k-way interaction

- Think of main effects as "1-way interactions" and the intercept as a "0-way interaction".
- We can think of a power of X as an interaction of X with itself. Then we have,
- **Hierarchy Principle for Polynomials:**

  Include all lower-order powers of X with the k^{th} power of X.

What can go wrong without the H.P.

```
X <- sort(rnorm(100))
eps <- rnorm(100)
X2 <- X^2
y <- 1 + 4*X + 3*X^2+ eps
mydata <- data.frame(y,X,X2)
plot(y ~ X,data=mydata)
fit1 <- lm(y ~ X2,data=mydata)
fit2 <- lm(y ~ X + X2,data=mydata)
p1 <- predict(fit1,mydata)
p2 <- predict(fit2, mydata)
lines(p1 ~ X,col="red")
lines(p2 ~ X,col="green")
legend(1,3,legend=c("y ~ 1 + X^2",
 "y ~ 1 + X + X^2"),lty=1,cex=0.75,
  col=c("red", "green"))
```



Dummy Variables and Categorical Variables

- A <u>continuous</u> variable conceivably takes on infinitely many values *and the values can be arbitrarily close together*
  - Age, weight, height, angle, distance of a planet to the Sun, etc.
  - These get one coefficient (slope) in an R model formula
- A <u>categorical</u> variable conceivably takes on finitely or infinitely many values, *either non-numerical, or if numerical then there is a minimum positive distance between the variables.*
  - Age, year in school, letter grade (ABCDF), name, count
  - R maps the labels of non-numeric variables onto integers ("factor" and "ordered factor" types)
  - R treats numeric categorical variables as continuous, unless you recode the variable in some way for R

Coding of Categorical Variables

- R usually uses ***two different representations*** of non-numeric categorical variables
- "factor" or "ordered factor" variables
  - R maps the values of the variable onto integers 1,2,3… and keeps track of which text label goes with which integer
- Dummy-coding for an lm() model formula
  - R treats numeric variables as continuous
  - R recodes factor variables using one dummy variable per value of the factor variable (aka 1-hot coding)

Example of R's two representations

```
X1 <- c("Tommy","Sue","Billy", "Sue","George","Tommy","Sue")
X1
## [1] "Tommy"  "Sue"    "Billy"  "Sue"    "George"
## [6] "Tommy"  "Sue"
X1 <-as.factor(X)
X1
## [1] Tommy  Sue    Billy  Sue    George Tommy  Sue
## Levels: Billy George Sue Tommy
y <- rnorm(7)
round(summary(lm(y ~ X1))$coef,2)
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.29       1.06   -0.27     0.80
## X1George       -0.11       1.49   -0.07     0.94
## X1Sue          -0.69       1.22   -0.57     0.61
## X1Tommy         0.24       1.29    0.19     0.86
```

X1 = c(4,3,1,3,2,4,3), and X keeps the mapping
"Billy" -> 1, "George" -> 2, "Sue" -> 3, "Tommy" -> 4

The X matrix:

(Int)	X1George	X1Sue	X1Tommy
1	0	0	1
1	0	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	0	0	1
1	0	1	0

Baseline Categories and the Intercept

- The matrix of predictors X must be of full column rank (i.e. linearly independent columns) so that $\left(X^T X\right)^{-1}$ exists.
- For a dummy-coded categorical variable, *R will delete the first level* of the variable to get X to be full rank.
- Sometimes you don't want this. Some common alternatives:
  - Omit the intercept
  - Sum-to-zero constraint

Examples of making X full-rank

```
round(summary(lm(y ~ X1))$coef,2) Let R delete the first category
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)Billy -0.29       1.06   -0.27     0.80  A.K.A. “baseline category”
## X1George         -0.11       1.49   -0.07     0.94
## X1Sue            -0.69       1.22   -0.57     0.61   Deviations from baseline category
## X1Tommy           0.24       1.29    0.19     0.86
round(summary(lm(y ~ X1 - 1))$coef,2) Remove intercept
            Estimate Std. Error t value Pr(>|t|)
## X1Billy     -0.29       1.06   -0.27     0.80
## X1George    -0.40       1.06   -0.38     0.73   The estimated cell means
## X1Sue       -0.98       0.61   -1.61     0.21
## X1Tommy     -0.05       0.75   -0.07     0.95
contrasts(X1) <- contr.sum(4)
round(summary(lm(y ~ X1))$coef,2) Sum-to-zero constraint
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.43       0.44   -0.97     0.40
## X11 Billy         0.14       0.87    0.16     0.88
## X12 George        0.03       0.87    0.03     0.98
## X13 Sue          -0.55       0.62   -0.89     0.44   Deviations from intercept/grand-mean
round(X14 <- -(0.14 + 0.03 - 0.55),2)
## [1] 0.38 Tommy
```

ANOVA Models with lm() or aov()

- The previous example is an example of a *1-way Analysis of Variance (ANOVA)* model, since it regresses y on a single "factor" variable
- Regressing y on two "factor" variables is a *2-way ANOVA* model
- In traditional use of ANOVA models
  - Don't care so much about cell means
  - Care if "interaction" is present (F-tests!)
  - More efficient ways than lm() of doing calculation: aov() *[try aov on your own!]*

Example: Prehistoric Pottery Sherds

- Response: number of sherds (= shards of prehistoric pottery)
- Factor 1: region of archaeological excavation
- Factor 2: type of ceramic sherd (three circle red on white, Mogollon red on brown, Mimbres corrugated, bold face black on white)

Question: Does region have a uniform effect on number of sherds found, for each type of pottery, or are the number of sherds for different pottery types different in different regions?

Woosley, A. I., & McIntyre, A. J. (1995). *Mimbres Mogollon Archaeology: Charles C. Di Peso's Excavations at Wind Mountain* (No. 10). University of New Mexico Press.

Coefficient estimates are not useful to answer the question…

```
sherds <- read.csv("pottery-sherds-01.csv",
   header=T)
str(sherds)
## 'data.frame':   60 obs. of  3 vars:
##  $ Region: chr  "I" "I" "I" "II" ...
## $ Type  : chr  "Red on White" ...
##  $ Sherds: int  68 33 45 59 43 37 ...
```

```
summary(lm.1 <- lm(Sherds ~ Region + Type,
   data=sherds))$coef
```

```
##                        Est    SE      t pval
## (Intercept)          80.00  7.16  11.17 0.00
## RegionII             -5.92  8.00  -0.74 0.46
## RegionIII           -10.50  8.00  -1.31 0.20
## RegionIV            -12.83  8.00  -1.60 0.11
## RegionV             -32.75  8.00  -4.09 0.00
## TypeMimbres         -14.93  7.16  -2.09 0.04
## TypeMogollon        -13.27  7.16  -1.85 0.07
## TypeRed on White    -17.13  7.16  -2.39 0.02
```

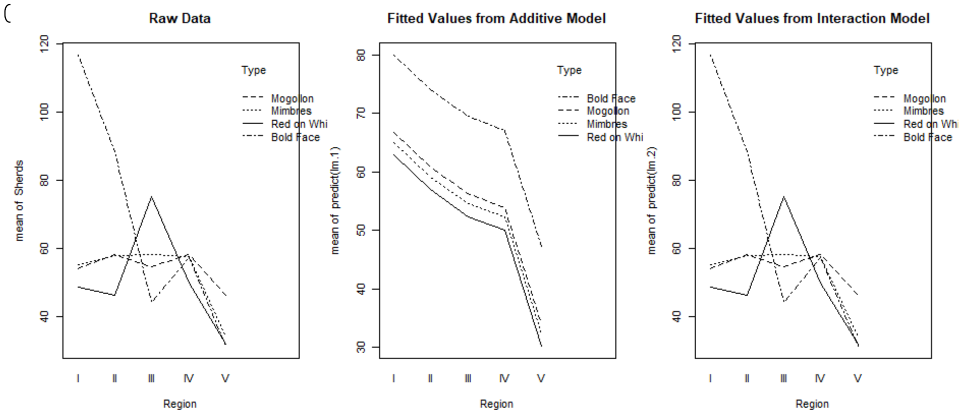
```
summary(lm.2 <- lm(Sherds ~ Region * Type,
   data=sherds))$coef
```

```
##                                Est    SE     t pval
## (Intercept)                 116.67  8.37 13.94 0.00
## RegionII                    -28.33 11.84 -2.39 0.02
## RegionIII                   -72.33 11.84 -6.11 0.00
## RegionIV                    -59.33 11.84 -5.01 0.00
## RegionV                     -85.33 11.84 -7.21 0.00
## TypeMimbres                 -61.33 11.84 -5.18 0.00
## TypeMogollon                -62.67 11.84 -5.29 0.00
## TypeRed on White            -68.00 11.84 -5.74 0.00
## RegionII:TypeMimbres         31.00 16.74  1.85 0.07
## RegionIII:TypeMimbres        75.33 16.74  4.50 0.00
## RegionIV:TypeMimbres         61.67 16.74  3.68 0.00
## RegionV:TypeMimbres          64.00 16.74  3.82 0.00
## RegionII:TypeMogollon        32.67 16.74  1.95 0.06
## RegionIII:TypeMogollon       73.00 16.74  4.36 0.00
## RegionIV:TypeMogollon        63.67 16.74  3.80 0.00
## RegionV:TypeMogollon         77.67 16.74  4.64 0.00
## RegionII:TypeRed on White    26.00 16.74  1.55 0.13
## RegionIII:TypeRed on White   99.00 16.74  5.91 0.00
## RegionIV:TypeRed on White    60.67 16.74  3.62 0.00
## RegionV:TypeRed on White     68.67 16.74  4.10 0.00
```

The ANOVA table provides F tests that can help

```
anova(lm.1)
##            Df  Sum Sq Mean Sq     F  pval
## Region      4  7364.6 1841.14  4.79 0.002
## Type        3  2681.7  893.91  2.33 0.086
## Residuals 52 19991.4  384.45
anova(lm.2)
##               Df  Sum Sq Mean Sq     F  pval
## Region         4  7364.6 1841.14  8.76 0.000
## Type           3  2681.7  893.91  4.25 0.011
## Region:Type 12 11585.4  965.45  4.59 0.00(
## Residuals    40  8406.0  210.15
anova(lm.1,lm.2)
## Model 1: Sherds ~ Region + Type
## Model 2: Sherds ~ Region * Type
##   Res.Df   RSS Df Sum of Sq      F pval
## 1     52 19991
## 2     40  8406 12     11585 4.5941 0.000
```

```
par(mfrow=c(1,3))
attach(sherds)
interaction.plot(Region,Type,Sherds,
   main="Raw Data")
interaction.plot(Region,Type,
   predict(lm.1),
main="Fitted Values from Additive Model")
interaction.plot(Region,Type,
   predict(lm.2),
main="Fitted Values from Interaction Model")
detach(sherds)
```



So, clearly (by statistical test & by eye) the effect of region on number of sherds is not uniform in pottery type!

ANCOVA Models

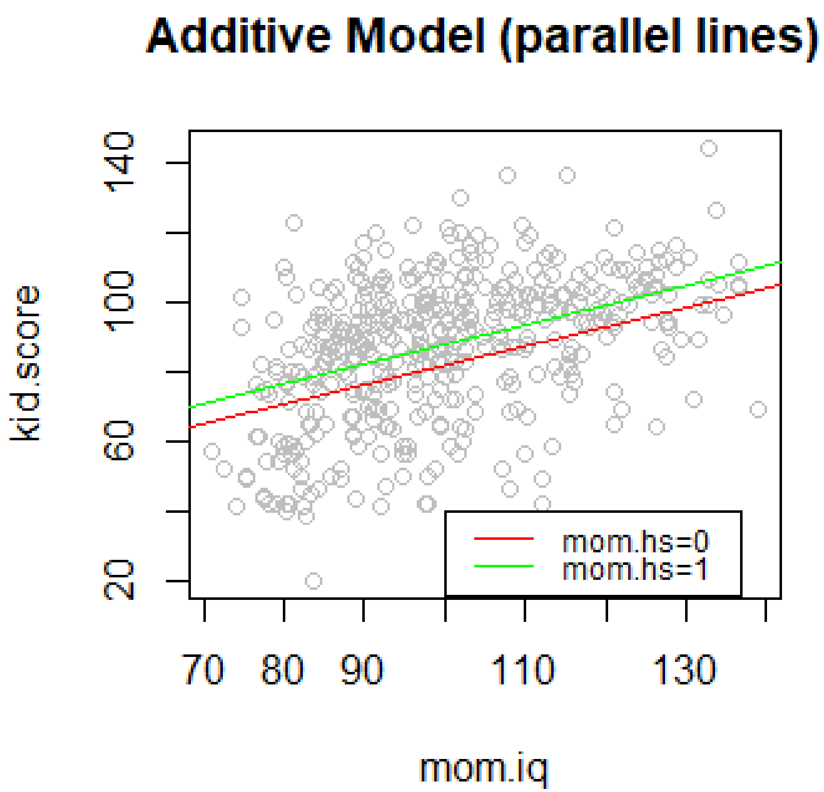
- ANCOVA (Analysis of Covariance) models are regression models with one continuous predictor and one discrete predictor
- The interesting models are
  - Additive model: parallel lines with different intercepts
  - Interaction model: lines with different slopes and intercepts
- The interesting question is usually which of these is a better model?

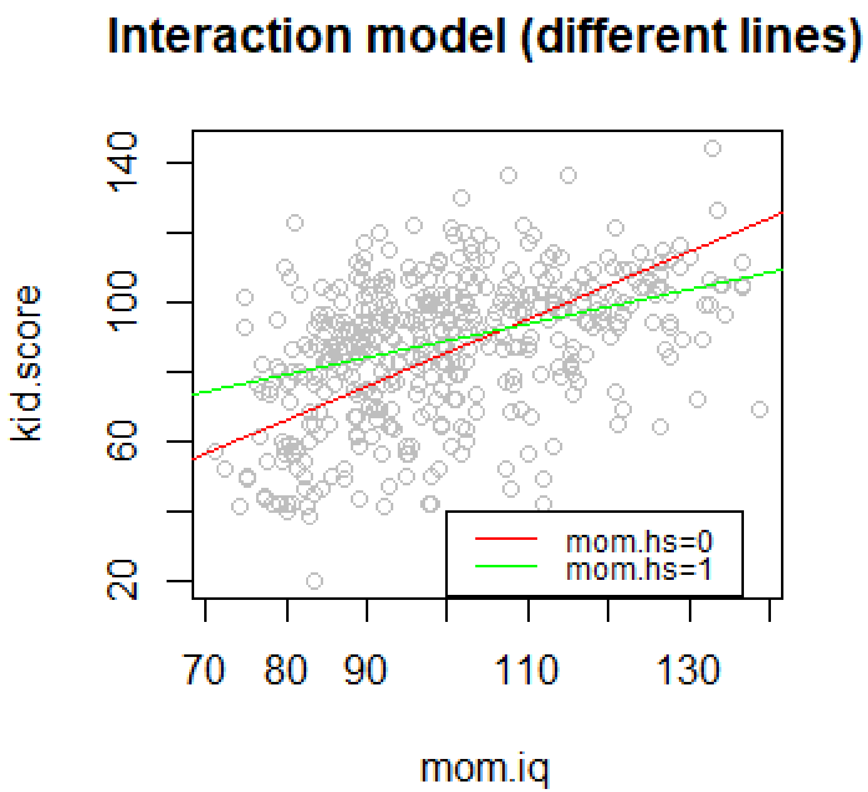
Example: ANCOVA for mom.hs and mom.iq vs kid.score

```
kidiq <- read.csv("kidiq.csv",header=T)
lm.3 <- lm(kid.score ~ mom.iq +
   mom.hs,data=kidiq)
round(summary(lm.3)$coef,2)
##                Est    SE     t   pval
## (Intercept)  25.73  5.88  4.38   0.00
## mom.iq        0.56  0.06  9.31   0.00
## mom.hs        5.95  2.21  2.69   0.01
lm.4 <- lm(kid.score ~ mom.iq *
   mom.hs,data=kidiq)
round(summary(lm.4)$coef,2)
##               Est    SE     t   pval
## (Intercept) 25.73  5.88  4.38   0.00
## mom.iq       0.56  0.06  9.31   0.00
## mom.hs       5.95  2.21  2.69   0.01
anova(lm.4)
##                Df  SumSq MeanSq      F  pval
## mom.iq          1  36249  36249 112.23 0.000
## mom.hs          1   2380   2380   7.37 0.007
## mom.iq:mom.hs   1   2878   2878   8.91 0.003
## Residuals     430 138879    323
```

```
par(mfrow=c(1,2))
plot(kid.score ~
mom.iq,data=kidiq,col="grey",
     main="Additive Model (parallel
lines)")
abline(25.73,0.56,col="red")
abline(25.73+5.95,0.56,col="green")
legend(100,40,lty=1,col=c("red","green"),
cex=0.75,
       legend=c("mom.hs=0",
                "mom.hs=1"))
plot(kid.score ~
mom.iq,data=kidiq,col="grey",
     main="Interaction model (different
lines)")
abline(-11.48,0.97,col="red")
abline(-11.48+51.27,0.97-
0.48,col="green")
legend(100,40,lty=1,col=c("red","green"),
cex=0.75,
       legend=c("mom.hs=0",
                "mom.hs=1"))
```

Example: ANCOVA for mom.hs and mom.iq vs kid.score





Summary

- SS Decompositions and F Statistics
  - Some Comments
- Interactions
  - The Hierarchy Principle
- Categorical and Dummy Variables
- ANOVA models
- ANCOVA models