ORIGIN ≡ 0

Analysis of Covariance (ANCOVA) for a single fixed factor

Analysis of Covariance (ANCOVA) is a technique that combines the ANOVA strategy comparing multiple levels for one or more test factor (T) with regression of one or more numeric "covariate" or "concomitant" variables (X). The objective may be to use the covariate (X) to control for variance in the dependent variable (Y) to determine factor effects (T) even though Y varies with X. Alternatively, ANCOVA tests can be used to test whether multiple regressions (Y_i with X_i) have equivalent slopes and intercepts for multiple measurements i within each of j levels of T. If both both slopes and intercepts match, then the several regressions are termed "coincident", and may be pooled to allow greater precision in estimates of regression parameters and prediction. The worked example here is drawn from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Example in R:

```
#ANCOVA SINGLE FACTOR
#KNNL TABLE 22.1
setwd("c:/DATA/Models")
K=read.table("CH22TA01.txt",header=TRUE)
K
attach(K)

plot(X,Y,pch=20)
points(X[treatment==1],Y[treatment==1],pch=20,col="red")
points(X[treatment==2],Y[treatment==2],pch=20,col="blue")
points(X[treatment==3],Y[treatment==3],pch=20,col="green")

T=factor(treatment)

# GLM TEST OF TREATMENT EFFECT (INTERCEPT):
FM=lm(Y~X+T)
summary(FM)
```

KNNL Table 22.1 Cracker Data

```
> K
    Y  X treatment store
1  38 21         1     1
2  39 26         1     2
3  36 22         1     3
4  45 28         1     4
5  33 19         1     5
6  43 34         2     1
7  38 26         2     2
8  38 29         2     3
9  27 18         2     4
10 34 25         2     5
11 24 23         3     1
12 32 29         3     2
13 31 30         3     3
14 21 16         3     4
15 28 29         3     5
```

```
> summary(FM)
Call:
lm(formula = Y ~ X + T)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4348 -1.2739 -0.3363  1.6710  2.4869 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.3534     2.5230   6.878 2.66e-05 ***
X             0.8986     0.1026   8.759 2.73e-06 ***
T2           -5.0754     1.2290  -4.130  0.00167 ** 
T3          -12.9768     1.2056 -10.764 3.53e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.873 on 11 degrees of freedom
Multiple R-squared: 0.9403,     Adjusted R-squared: 0.9241 
F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07
```

^ regression coefficients for FM (full model) are listed in column called Estimate. Note that because R's default dummy coding (also called contrast matrix) "contr. treatment" differs from that employed by KNNL, the estimates are calculated differently. Here (and in R), coefficients for T2 measure differences between the first & second classes in T, and T3 measures differences between first & third classes in T. The t-statistics and probabilities are marginal tests of each regression coefficient which generally are not useful.

GLM Test of Treatment Effect (Intercept):

Assumptions:

- Standard Linear Regression depends on specifying in advance which variable is to be considered 'dependent' and which 'independent'. This decision matters as changing roles for Y & X usually produces a different result.

- $Y_1, Y_2, Y_3, \dots, Y_n$ (dependent variable) is a random sample ~ $N(\mu,\sigma^2)$.

- $X_1, X_2, X_3, \dots, X_n$ (independent variable) with each value of X_i matched to Y_i

Within this setup, two models for the relationship between X and Y variables are explicitly compared:

Full Model:

$$Y_{ij} = \mu + \gamma X_{ij} + \tau_i + \varepsilon_{ij}$$

where: μ is the grand mean = intercept
γ is the regression coefficient (slope) for Y vs X
τ_i is the treatment effect T (as coded using contr.treatment
ε_{ij} are "within" errors compared to each mean μ_j ~$N(0,\sigma^2)$

Reduced Model:

$$Y_{ij} = \mu + \gamma X_{ij} + \varepsilon_{ij}$$

where: same as above but setting treatment effects τ_i to zero.

Hypotheses:

H_0: The Reduced Model is *sufficient* to describe the relationship between Y & X

H_1: The Full Model is *required* to describe relationship between Y & X

Degrees of Freedom:

$n := 15$

$c_f := 2 + 2$ **< 2 parameters for regression, 2 parameters for 3 classes of T**

$c_r := 2$ **< 2 parameters for regression**

$df_F := n - c_f$ $df_F = 11$ **< "full" model with c estimated parameters in model**

$df_R := n - c_r$ $df_R = 13$ **< "reduced" model with 2 parameters for regression coefficents**

GLM Test Statistic:

$SSE_F := 38.57$ $SSE_R := 455.72$ **< from anova tables in R for full & reduced models**

$$F := \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}}$$

$F = 59.4847$ **< results confirmed p. 929**

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$CV := qF(1 - \alpha, df_R - df_F, df_F)$ $CV = 3.9823$ **< note degrees of freedom utilized here! calculation for CV confirmed p. 929.**

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 59.4847$ $CV = 3.9823$

Probability Value:

$P := 1 - pF(F, df_R - df_F, df_F)$ $P = 1.2634 \times 10^{-6}$ **< results confirmed in R (slight difference in values due to rounding of SS_F & SS_R above).**

Prototype in R:

anova(FM)

```
> anova(FM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 190.68 190.678  54.379 1.405e-05 ***
T          2 417.15 208.575  59.483 1.264e-06 ***
Residuals 11  38.57   3.506
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note F-statistic & probability for T >

^ the ANOVA table provides SS for Regression (adding together rows for X & T) and SS Error (or Residuals with values matching KNNL p. 928 verified). Note that F-statistics and probabilities reported here represent serial "extra" SS with order of variables in the model specified from left to right. KNNL instead turn to the GLM reduced model (RM) versus full model (FM) approach, shown below, although results for factor T (the last variable entered into the model) are identical.

RM=lm(Y~X)
summary(RM)
anova(RM)

anova(FM,RM)

```
> summary(RM)
Call:
lm(formula = Y ~ X)

Residuals:
   Min     1Q Median     3Q    Max
-8.711 -5.481  1.289  3.975  9.017

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.6056     7.9497   1.963   0.0714 .
X             0.7278     0.3121   2.332   0.0364 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.921 on 13 degrees of freedom
Multiple R-squared: 0.295,     Adjusted R-squared: 0.2408
F-statistic: 5.439 on 1 and 13 DF,  p-value: 0.03641

> anova (RM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X          1 190.68 190.678  5.4393 0.03641 *
Residuals 13 455.72  35.056
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # GLM Test:
> anova (FM,RM)
Analysis of Variance Table

Model 1: Y ~ X + T
Model 2: Y ~ X
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     11  38.57
2     13 455.72 -2   -417.15 59.483 1.264e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

compare F & P with above >

GLM Test for Parallel Slope:

Assumptions:

- Same as above.

Full Model:

$$Y_{ij} = \mu + \gamma X_{ij} + \tau_i + \beta_i X_{ij}\tau_i + \varepsilon_{ij}$$

where: **μ is the grand mean = intercept**
γ is the regression coefficient (slope) for Y vs X
τ_i is the treatment effect T (as coded using contr.treatment
$\beta_i X_{ij}\tau_i$ is interaction between treatment and X_{ij}
ε_{ij} are "within" errors compared to each mean μ_j ~N(0,σ^2)

Reduced Model:

$$Y_{ij} = \mu + \gamma X_{ij} + \tau_i + \varepsilon_{ij}$$

where: same as above but setting interactions β_i to zero.

Hypotheses:

H_0: The Reduced Model is *sufficient* to describe the relationship between Y & X

H_1: The Full Model is *required* to describe relationship between Y & X

Degrees of Freedom:

$n := 15$ $c_f := 2 + 2 + 2$ $c_f = 6$ **< 2 regression, 2 for T, 2 for interactions**

$c_r := 2 + 2$ $c_r = 4$ **< 2 regression, 2 for T, as above**

$df_F := n - c_f$ $df_F = 9$ **< "full" model with c estimated parameters in model**

$df_R := n - c_r$ $df_R = 11$ **< "reduced" model with 2 parameters for regression coefficents**

GLM Test Statistic:

$SSE_F := 31.52$ $SSE_R := 38.57$ **< from anova tables in R for full & reduced models**

$$F := \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}}$$

$F = 1.0065$ **< results confirmed p. 933 with rounding error**

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$CV := qF(1 - \alpha, df_R - df_F, df_F)$ $CV = 4.2565$ **< note degrees of freedom utilized here! calculation for CV confirmed p. 933.**

Decision Rule:

IF F > CV, THEN REJECT H_0 OTHERWISE ACCEPT H_0

$F = 1.0065$ $CV = 4.2565$

Probability Value:

$P := 1 - pF(F, df_R - df_F, df_F)$ $P = 0.4032$ **< results confirmed in R (slight difference in values due to rounding of SS_F & SS_R above).**

Prototype in R:

```
#GLM TEST OF PARALLEL SLOPE:
FMs=lm(Y~X*T)
summary(FMs)
anova(FMs)

RMs=lm(Y~X+T) #same model as FM above but now used as the reduced model
summary(RMs)
anova(RMs)

anova(FMs,RMs)
```

```
> anova(FMs)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)    
X          1 190.68 190.678 54.4434 4.198e-05 ***
T          2 417.15 208.575 59.5536 6.457e-06 ***
X:T        2   7.05   3.525  1.0065    0.4032    
Residuals  9  31.52   3.502                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(RMs)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)    
X          1 190.68 190.678  54.379 1.405e-05 ***
T          2 417.15 208.575  59.483 1.264e-06 ***
Residuals 11  38.57   3.506                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(FMs,RMs)
Analysis of Variance Table

Model 1: Y ~ X * T
Model 2: Y ~ X + T
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 31.521                           
2     11 38.571 -2   -7.0505 1.0065 0.4032
```

^ residual SS from each model

^ test result - compare with above. This result does NOT reject H_0, thus preferring the simpler RM