Cell means coding, Quantitative-Categorical interaction with R: Cars data*

```
> kars = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/mcars4.data")
> head(kars)
  Cntry lper100k weight length
1    US     19.8   2178   5.92
2 Japan      9.9   1026   4.32
3    US     10.8   1188   4.27
4    US     12.5   1444   5.11
5    US     12.5   1485   5.03
6    US     12.5   1485   5.03
>
> # Make dummy variables and interaction terms
> cars = within(kars,{
+ # Make indicator dummy variables for Cntry.
+ c1 = c2 = c3 = numeric(length(weight))
+ c1[Cntry=='Europ'] = 1
+ c2[Cntry=='Japan'] = 1
+ c3[Cntry=='US'] = 1
+ # Interaction terms: Just Weight by Country
+ wc1 = weight*c1; wc2 = weight*c2
+ }) # End within kars
> head(cars)
  Cntry lper100k weight length  wc2 wc1 c1 c2 c3
1    US     19.8   2178   5.92    0   0  0  0  1
2 Japan      9.9   1026   4.32 1026   0  0  1  0
3    US     10.8   1188   4.27    0   0  0  0  1
4    US     12.5   1444   5.11    0   0  0  0  1
5    US     12.5   1485   5.03    0   0  0  0  1
6    US     12.5   1485   5.03    0   0  0  0  1
>
> # Basic model with intercept and no interaction
> model1 = lm(lper100k ~ weight + c1 + c2, data = cars)
> summary(model1)

Call:
lm(formula = lper100k ~ weight + c1 + c2, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0550 -0.4890  0.0138  1.2755  2.8316 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.4241768  0.9376017  -0.452  0.65200    
weight       0.0086939  0.0005942  14.631  < 2e-16 ***
c1           1.2127472  0.5777671   2.099  0.03844 *  
c2           1.8932896  0.5976631   3.168  0.00206 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.745 on 96 degrees of freedom
Multiple R-squared:  0.7276, Adjusted R-squared:  0.7191 
F-statistic: 85.49 on 3 and 96 DF,  p-value: < 2.2e-16

>
```

* See last page for copyright information

```
> # Test Country controlling for weight with full-restricted approach
> justweight = lm(lper100k ~ weight, data = cars)
> anova(justweight,model1)
Analysis of Variance Table

Model 1: lper100k ~ weight
Model 2: lper100k ~ weight + c1 + c2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     98 326.23
2     96 292.22  2    34.015 5.5873 0.005065 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Now with general linear test: H0: L beta = h
> # Define ftest = function(model,L,h=0)
> source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/ftest.txt")
> L1 = rbind(c(0,0,1,0),
+            c(0,0,0,1))
> ftest(model1,L1)
          F          df1          df2      p-value
 5.587336363  2.000000000 96.000000000  0.005065184
>
```

```
> # Cell means model: No intercept and all 3 dummy variables
> model2 = lm(lper100k ~ 0 + c1 + c2 + c3 + weight, data = cars)
> summary(model2)

Call:
lm(formula = lper100k ~ 0 + c1 + c2 + c3 + weight, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0550 -0.4890  0.0138  1.2755  2.8316 

Coefficients:
         Estimate Std. Error t value Pr(>|t|)    
c1      0.7885704  0.7933027   0.994   0.3227    
c2      1.4691128  0.7942841   1.850   0.0674 .  
c3     -0.4241768  0.9376017  -0.452   0.6520    
weight  0.0086939  0.0005942  14.631   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.745 on 96 degrees of freedom
Multiple R-squared:  0.9819,	Adjusted R-squared:  0.9811 
F-statistic:  1302 on 4 and 96 DF,  p-value: < 2.2e-16

> summary(model1) # For comparison

Call:
lm(formula = lper100k ~ weight + c1 + c2, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0550 -0.4890  0.0138  1.2755  2.8316 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.4241768  0.9376017  -0.452  0.65200    
weight       0.0086939  0.0005942  14.631  < 2e-16 ***
c1           1.2127472  0.5777671   2.099  0.03844 *  
c2           1.8932896  0.5976631   3.168  0.00206 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.745 on 96 degrees of freedom
Multiple R-squared:  0.7276,	Adjusted R-squared:  0.7191 
F-statistic: 85.49 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
>
> # Test country controlling for weight with a general linear test.
> # Compare F = 5.587
> coefficients(model2)
         c1          c2          c3      weight 
 0.78857044  1.46911279 -0.42417681  0.00869387 
> L2 = rbind(c(1,-1, 0, 0),
+            c(0, 1,-1, 0))
> ftest(model2,L2)
          F         df1         df2     p-value 
 5.587336363  2.000000000 96.000000000  0.005065184 

>
> # Test pairwise differences between countries controlling for weight
> # c1 = Europe, c2 = Japan, c3 = US

> L3 = rbind(c(1,-1, 0, 0)) # Europe versus Japan
> ftest(model2,L3)
        F       df1       df2   p-value 
 1.0252817  1.0000000 96.0000000  0.3138147 

> L4 = rbind(c(0, 1,-1, 0)) # Japan versus US
> ftest(model2,L4) # Compare t = 3.168, p = 0.00206
          F         df1         df2     p-value 
10.035087779  1.000000000 96.000000000  0.002059955 

> L5 = rbind(c(1, 0,-1, 0)) # Europe versus US
> ftest(model2,L5) # Compare t = 2.099, p = 0.03844
         F        df1        df2    p-value 
 4.40590356  1.00000000 96.00000000  0.03843886 
```

```
> # Model with Interaction of country by weight, based on model1
```

Origin	C1	C2	$E(y\|\mathbf{X}=\mathbf{x}) = \beta_0 + \beta_1\mathbf{X}_1 + \beta_3\mathbf{C}_1 + \beta_4\mathbf{C}_2 + \beta_5\mathbf{X}_1\mathbf{C}_1 + \beta_6\mathbf{X}_1\mathbf{C}_2$
Europe	1	0	$(\beta_0 + \beta_3) + (\beta_1+\beta_5)\mathbf{X}_1$
Japan	0	1	$(\beta_0 + \beta_4) + (\beta_1+\beta_6)\mathbf{X}_1$
U.S.	0	0	$\beta_0 + \beta_1 \mathbf{X}_1$

```
> model3 = lm(lper100k ~ weight + c1 + c2 + wc1 + wc2, data = cars)
> summary(model3)

Call:
lm(formula = lper100k ~ weight + c1 + c2 + wc1 + wc2, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8461 -0.5647 -0.1310  1.3273  2.6569 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.4005480  0.9545858   0.420   0.6757    
weight       0.0081583  0.0006065  13.452   <2e-16 ***
c1          -3.8072812  2.3485193  -1.621   0.1083    
c2          -8.7126778  5.0437692  -1.727   0.0874 .  
wc1          0.0044198  0.0020348   2.172   0.0324 *  
wc2          0.0097631  0.0046908   2.081   0.0401 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.687 on 94 degrees of freedom
Multiple R-squared:  0.7507,	Adjusted R-squared:  0.7375 
F-statistic: 56.63 on 5 and 94 DF,  p-value: < 2.2e-16 

> anova(model1,model3) # Test equality of slopes
Analysis of Variance Table

Model 1: lper100k ~ weight + c1 + c2
Model 2: lper100k ~ weight + c1 + c2 + wc1 + wc2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)  
1     96 292.22                             
2     94 267.43  2    24.793 4.3573 0.0155 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
```

The heavier the car, the greater the average fuel consumption. Rates of increase are greater for Japanese and European cars than for American cars.

```
> # How about European vs (minus) Japan slopes?
> EvsJ = rbind(c(0,0,0,0,1,-1))
> ftest(model3,EvsJ)
        F       df1        df2    p-value 
 1.1236914  1.0000000 94.0000000  0.2918412 
```

```
> # Plot the regression lines
> with(cars,{
+ yhat = model3$fitted.values
+ plot(weight,yhat,pch=' ',xlab='Weight in kg',
+ ylab='Fuel Consumption in lper100k')
+ title('Weight and Estimated Fuel Consumption')
+ lines(weight[Cntry=='US'],yhat[Cntry=='US'],lty=1)
+ lines(weight[Cntry=='Europ'],yhat[Cntry=='Europ'],lty=2)
+ lines(weight[Cntry=='Japan'],yhat[Cntry=='Japan'],lty=3)
+ x1 = c(1800,2000); y1 = c(8,8); lines(x1,y1,lty=1); text(1700,8,'US   ')
+ x2 = c(1800,2000); y2 = c(9,9); lines(x2,y2,lty=2); text(1700,9,'Japan')
+ x3 = c(1800,2000); y3 = c(10,10); lines(x3,y3,lty=3); text(1700,10,'Europ')
}) # End with cars
```