Categorical independent variables with R*

```
> kars = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/mcars4.data")
> kars[1:4,]
  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
>
> attach(kars) # Variables are now available by name
> n = length(length); n
[1] 100
> # Make indicator dummy variables for Cntry. Just use 2 for now.
> # U.S. will be the reference category
> c1 = numeric(n); c1[Cntry=='Europ'] = 1
> table(c1,Cntry)
   Cntry
c1  Europ Japan US
  0     0    13 73
  1    14     0  0
> c2 = numeric(n); c2[Cntry=='Japan'] = 1
> table(c2,Cntry)
   Cntry
c2  Europ Japan US
  0    14     0 73
  1     0    13  0
>
> c3 = numeric(n); c3[Cntry=='US'] = 1
> table(c3,Cntry)
   Cntry
c3  Europ Japan US
  0    14    13  0
  1     0     0 73
```

```
> # Take a look at mean fuel consumption for each country
> aggregate(lper100k,by=list(Cntry),FUN=mean)
  Group.1        x
1   Europ 10.17857
2   Japan 10.68462
3      US 12.96438
> # Must specify a LIST of grouping factors
```

On average, the U.S. cars seem to be using more fuel. Back it up with a hypothesis test.

Origin	c1	c2	$E(Y\|\mathbf{X}=\mathbf{x}) = \beta_0 + \beta_1 c_1 + \beta_2 c_2$
Europe	1	0	$\beta_0 + \beta_1$
Japan	0	1	$\beta_0 + \beta_2$
U.S.	0	0	β_0

```
> # H0: mu1=mu2=mu3
> justcountry = lm(lper100k ~ c1+c2)
> summary(justcountry)

Call:
lm(formula = lper100k ~ c1 + c2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0644 -2.1644 -0.4644  2.5154  6.8356 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.9644     0.3651  35.511  < 2e-16 ***
c1           -2.7858     0.9101  -3.061  0.00285 ** 
c2           -2.2798     0.9390  -2.428  0.01703 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.119 on 97 degrees of freedom
Multiple R-squared: 0.1203,	Adjusted R-squared: 0.1022 
F-statistic: 6.634 on 2 and 97 DF,  p-value: 0.001993 
```

```
>
> # Which means are different?
> Have t-tests. What about Europe vs. Japan?

> # Repeating ...
> summary(justcountry)$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 12.964384  0.3650854 35.510547 2.167687e-57
c1          -2.785812  0.9101021 -3.060989 2.853779e-03
c2          -2.279768  0.9390140 -2.427832 1.703327e-02
>
```

$$T = \frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}}{s\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$$

```
> # First replicate test of H0: beta1=0
> betahat = justcountry$coefficients; betahat
(Intercept)          c1          c2
  12.964384   -2.785812   -2.279768
> a1 = rbind(0,1,0); a1
     [,1]
[1,]    0
[2,]    1
[3,]    0
> V = vcov(justcountry) # MSE * (X'X)-inverse
> T1 = t(a1) %*% betahat / sqrt(t(a1) %*% V %*% a1)
> T1 = as.numeric(T1)
> T1; 2*(1-pt(abs(T1),97)) # 2-tailed p-value
[1] -3.060989
[1] 0.002853779
>
> # Now test H0: beta1 = beta2
> a = rbind(0,1,-1)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> pval = 2*(1-pt(abs(T),97))
> T; pval
[1] -0.4211978
[1] 0.6745425
```

Conclusion: American cars are getting fewer kilometers per litre on average than Japanese and European cars.

```
> # Repeating the test H0: beta1 = beta2 (Europe vs. Japan)
> a = rbind(0,1,-1)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> pval = 2*(1-pt(abs(T),97))
> T; pval
[1] -0.4211978
[1] 0.6745425


> # R can make the dummy variables for you
> is.factor(Cntry)
[1] TRUE
> # The factor Cntry has dummy vars built in. What are they?
> contrasts(Cntry) # Note alphabetical order
      Japan US
Europ     0  0
Japan     1  0
US        0  1
>
> jc2 = lm(lper100k~Cntry); summary(jc2)

Call:
lm(formula = lper100k ~ Cntry)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0644 -2.1644 -0.4644  2.5154  6.8356 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.1786     0.8337  12.209  < 2e-16 ***
CntryJapan    0.5060     1.2014   0.421  0.67454    
CntryUS       2.7858     0.9101   3.061  0.00285 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.119 on 97 degrees of freedom
Multiple R-squared: 0.1203,	Adjusted R-squared: 0.1022 
F-statistic: 6.634 on 2 and 97 DF,  p-value: 0.001993
```

```
> # You can select the dummy variable coding scheme.
> contr.treatment(3,base=2) # Category 2 is the reference category
  1 3
1 1 0
2 0 0
3 0 1

> # U.S. as reference category again
> Country = Cntry
> contrasts(Country) = contr.treatment(3,base=3)
> summary(lm(lper100k~Country))

Call:
lm(formula = lper100k ~ Country)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0644 -2.1644 -0.4644  2.5154  6.8356 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.9644     0.3651  35.511  < 2e-16 ***
Country1     -2.7858     0.9101  -3.061  0.00285 ** 
Country2     -2.2798     0.9390  -2.428  0.01703 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 3.119 on 97 degrees of freedom
Multiple R-squared: 0.1203,	Adjusted R-squared: 0.1022 
F-statistic: 6.634 on 2 and 97 DF,  p-value: 0.001993
```

Include covariates

Origin	c1	c2	$E(Y\|\mathbf{X}=\mathbf{x}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 C_1 + \beta_4 C_2$
Europe	1	0	$(\beta_0 + \beta_3) + \beta_1 X_1 + \beta_2 X_2$
Japan	0	1	$(\beta_0 + \beta_4) + \beta_1 X_1 + \beta_2 X_2$
U.S.	0	0	$\beta_0 + \beta_1 X_1 + \beta_2 X_2$

```
> # Include covariates
> fullmodel = lm(lper100k ~ weight+length+Country)
> summary(fullmodel) # Look carefully at the signs!

Call:
lm(formula = lper100k ~ weight + length + Country)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5063 -0.8813  0.0147  1.3043  2.9432 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.276937   3.006354  -2.421 0.017399 *  
weight       0.005457   0.001472   3.707 0.000352 ***
length       2.345968   0.980329   2.393 0.018676 *  
Country1     1.487722   0.575633   2.584 0.011274 *  
Country2     1.994239   0.584995   3.409 0.000958 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.703 on 95 degrees of freedom
Multiple R-squared: 0.7431,	Adjusted R-squared: 0.7323 
F-statistic: 68.71 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
> ###### Predictions and prediction intervals ######
>
> # Predict litres per 100 km for a Japanese car weighing
> # 1295kg, 4.52m long

> # (1990 Toyota Camry)
> betahat = fullmodel$coefficients; betahat
 (Intercept)       weight       length     Country1     Country2
-7.276936526  0.005456609  2.345968436  1.487721833  1.994238863
> contrasts(Country)
      1 2
Europ 1 0
Japan 0 1
US    0 0
> x1 = c(1,1295,4.52,0,1)
> sum(x1*betahat)
[1] 12.38739
>
> # Use the predict function
> # help(predict.lm)
>
> camry1990 = data.frame(weight=1295,length=4.52,Country='Japan')
> camry1990
  weight length Country
1   1295   4.52   Japan
> predict(fullmodel,newdata=camry1990)
       1
12.38739
> # With 95 percent prediction interval (default)
> predict(fullmodel,newdata=camry1990, interval='prediction')
       fit      lwr      upr
1 12.38739 8.856608 15.91817
```

```
>
> # Multiple predictions
> cadillac1990 = data.frame(weight=1800,length=5.22,Country='US')
> volvo1990 = data.frame(weight=1371,length=4.823,Country='Europ')
> newcars = rbind(camry1990,cadillac1990,volvo1990); newcars
  weight length Country
1   1295  4.520   Japan
2   1800  5.220      US
3   1371  4.823   Europ

> is.data.frame(newcars)
[1] TRUE

> predict(fullmodel,newdata=newcars, interval='prediction')
       fit       lwr      upr
1 12.38739  8.856608 15.91817
2 14.79091 11.354379 18.22745
3 13.00640  9.481598 16.53121
>
```

Cell means Coding

Origin	c1	c2	c3	$E(Y\|\mathbf{X}=\mathbf{x}) = \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 X_1 + \beta_5 X_2$
Europe	1	0	0	$\beta_1 + \beta_4 X_1 + \beta_5 X_2$
Japan	0	1	0	$\beta_2 + \beta_4 X_1 + \beta_5 X_2$
U.S.	0	0	1	$\beta_3 + \beta_4 X_1 + \beta_5 X_2$

```
> cellmeans = lm(lper100k ~ 0+Country+weight+length)
> summary(cellmeans)

Call:
lm(formula = lper100k ~ 0 + Cntry + weight + length)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5063 -0.8813  0.0147  1.3043  2.9432 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
CntryEurop -5.789215   2.855736  -2.027 0.045441 *  
CntryJapan -5.282698   2.926052  -1.805 0.074179 .  
CntryUS    -7.276937   3.006354  -2.421 0.017399 *  
weight      0.005457   0.001472   3.707 0.000352 ***
length      2.345968   0.980329   2.393 0.018676 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 1.703 on 95 degrees of freedom
Multiple R-squared:  0.9829,  Adjusted R-squared:  0.982 
F-statistic:  1094 on 5 and 95 DF,  p-value: < 2.2e-16 

> # lm(lper100k ~ 0+c1+c2+c3+weight+length) gives the same results,
> # but the labels (c1 c2 c3) are not as nice.

> sum(cellmeans$residuals)
[1] 9.950374e-15
```

```
> # Repeating ...
> predict(fullmodel,newdata=newcars, interval='prediction')
       fit       lwr      upr
1 12.38739  8.856608 15.91817
2 14.79091 11.354379 18.22745
3 13.00640  9.481598 16.53121
>

> predict(cellmeans,newdata=newcars, interval='prediction')
       fit       lwr      upr
1 12.38739  8.856608 15.91817
2 14.79091 11.354379 18.22745
3 13.00640  9.481598 16.53121
>
```

Rule: All valid dummy variable coding schemes are equivalent and give identical results when there are no mistakes. The choice is based on convenience.

Origin	c1	c2	c3	$E(Y\vert\mathbf{X}=\mathbf{x}) = \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 X_1 + \beta_5 X_2$
Europe	1	0	0	$\beta_1 + \beta_4 X_1 + \beta_5 X_2$
Japan	0	1	0	$\beta_2 + \beta_4 X_1 + \beta_5 X_2$
U.S.	0	0	1	$\beta_3 + \beta_4 X_1 + \beta_5 X_2$

Need: Test country controlling for size H_0: $\beta_1 = \beta_2 = \beta_3$
Test size controlling for country H_0: $\beta_4 = \beta_5 = 0$
