

## Categorical independent variables with R\*

```
> kars = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/mcars4.data.txt")
> 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
>
> attach(kars) # Variables are now available by name
> n = length(Cntry); 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
```

---

\* Copyright information is on the last page.

```

> # 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	$E(Y X=x) = \beta_0 + \beta_1 C_1 + \beta_2 C_2$
Europe	1	0	$\beta_0 + \beta_1$
Japan	0	1	$\beta_0 + \beta_2$
U.S.	0	0	$\beta_0$

```

> # 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?
> # Test H0: beta1 = beta2

```

$$t = \frac{\ell' \mathbf{b} - \gamma}{s \sqrt{\ell' (X'X)^{-1} \ell}}$$

```

> betahat = justcountry$coefficients; betahat
(Intercept)          c1          c2
 12.964384    -2.785812
> V = vcov(justcountry) # MSE * (X'X)-inverse
> ell = rbind(0,1,-1); ell # It's a column vector.

```

```

      [,1]
[1,]    0
[2,]    1
[3,]   -1

```

```

> T = as.numeric( t(ell)%*%betahat/sqrt(t(ell) %*%V%*% ell) )
> 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. There is no evidence of different fuel efficiency for European and Japanese cars.

```

> # 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

```

```

> # Names of dummy variables 1=Europe, 2=Japan could be nicer
> colnames(contrasts(Country)) = c("Europe","Japan")
> contrasts(Country)

```

```

      Europe Japan
Europ    1     0
Japan    0     1
US       0     0

```

## Include covariates

Origin	c1	c2	$E(Y X=x) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3C_1 + \beta_4C_2$
Europe	1	0	$(\beta_0 + \beta_3) + \beta_1X_1 + \beta_2X_2$
Japan	0	1	$(\beta_0 + \beta_4) + \beta_1X_1 + \beta_2X_2$
U.S.	0	0	$\beta_0 + \beta_1X_1 + \beta_2X_2$

```
> # 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 *
CountryEurope 1.487722   0.575633   2.584 0.011274 *
CountryJapan  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

```

> # Test car size controlling for country
> anova(justcountry,fullmodel) # Full vs reduced
Analysis of Variance Table

Model 1: lper100k ~ c1 + c2
Model 2: lper100k ~ weight + length + Country
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     97 943.81
2     95 275.61  2     668.2 115.16 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # I advise using anova ONLY to compare full and reduced models
>
> # Might as well test country controlling for size too.
> justsize = lm(lper100k ~ weight+length); summary(justsize)

```

```

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

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-4.3857 -1.0684 -0.0556  1.3077  4.0429

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.617472   2.958472  -1.223  0.22439
weight       0.004949   0.001546   3.202  0.00185 **
length       1.835625   1.017349   1.804  0.07428 .
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 1.804 on 97 degrees of freedom
Multiple R-squared: 0.7058, Adjusted R-squared: 0.6997
F-statistic: 116.4 on 2 and 97 DF, p-value: < 2.2e-16

```

```

> anova(justsize,fullmodel)

```

```

Analysis of Variance Table

```

```

Model 1: lper100k ~ weight + length
Model 2: lper100k ~ weight + length + Country
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     97 315.64
2     95 275.61  2     40.035 6.8999 0.001592 **
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

## Cell means Coding

Origin	c1	c2	c3	$E(Y X=x) = \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 X_1 + \beta_5 X_2$
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+Cntry+weight+length)
> summary(cellmeans)
> # Beware! R-squared was 0.7431 for an equivalent model.
```

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
```

This handout was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. It is available in OpenOffice.org from the course website:

<http://www.utstat.toronto.edu/~brunner/oldclass/302f17>