

## Categorical independent variables and interactions 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

> # 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
> c2 = numeric(n); c2[Cntry=='Japan'] = 1
> c3 = numeric(n); c3[Cntry=='US'] = 1

> # Illustrate interactions in a model with just weight and country
> eqslope = lm(lper100k ~ weight+c1+c2)
> summary(eqslope)
```

Call:

```
lm(formula = lper100k ~ weight + c1 + c2)
```

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

---

\* Copyright information is on the last page.

Origin	C1	C2	$E(Y X=x) = \beta_0 + \beta_1 X_1 + \beta_3 C_1 + \beta_4 C_2 + \beta_5 X_1 C_1 + \beta_6 X_1 C_2$
Europe	1	0	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$
Japan	0	1	$(\beta_0 + \beta_4) + (\beta_1 + \beta_6) X_1$
U.S.	0	0	$\beta_0 + \beta_1 X_1$

```
> wc1 = weight*c1; wc2 = weight*c2
> uneqslope = lm(lper100k ~ weight+c1+c2+wc1+wc2)
> summary(uneqslope)
Call:
lm(formula = lper100k ~ weight + c1 + c2 + wc1 + wc2)
```

```
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(eqslope,uneqslope)
Analysis of Variance Table
```

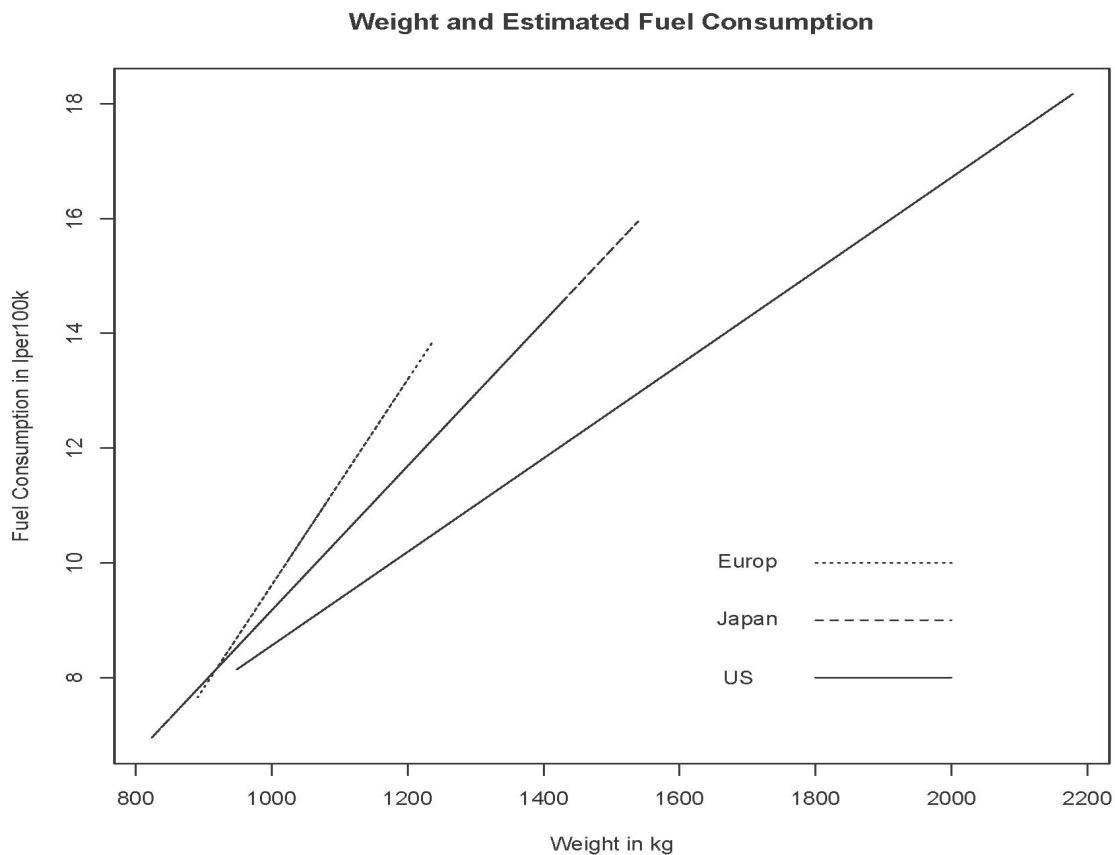
```
Model 1: lper100k ~ weight + c1 + c2
Model 2: lper100k ~ weight + c1 + c2 + wc1 + wc2
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     96 292.22
2     94 267.43  2    24.793 4.3573 0.0155 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```

> # Plot the regression lines
> yhat = uneqslope$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')

```



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/appliedf17>