Exploring relationships among variables

Dataset used for demo :

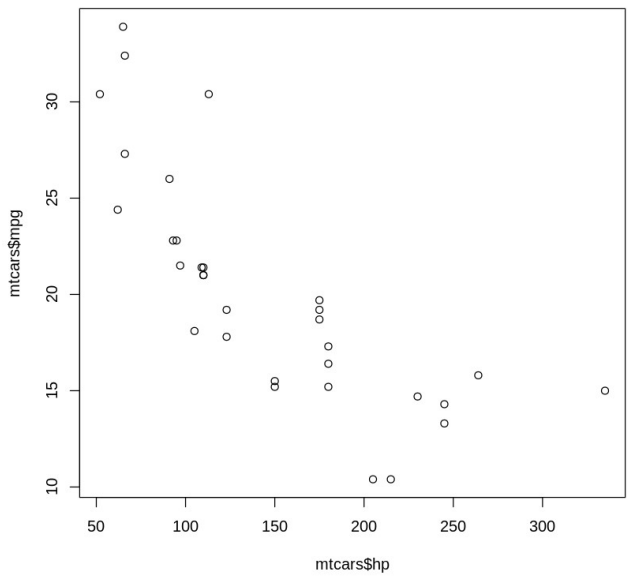
mtcars

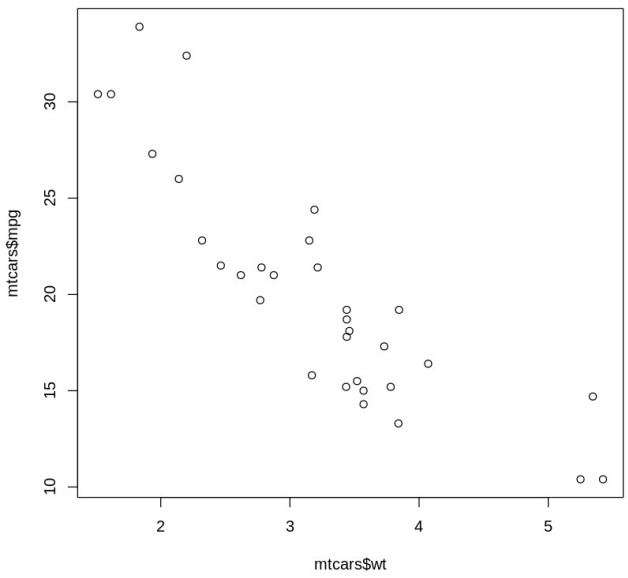
Variables in the dataset :

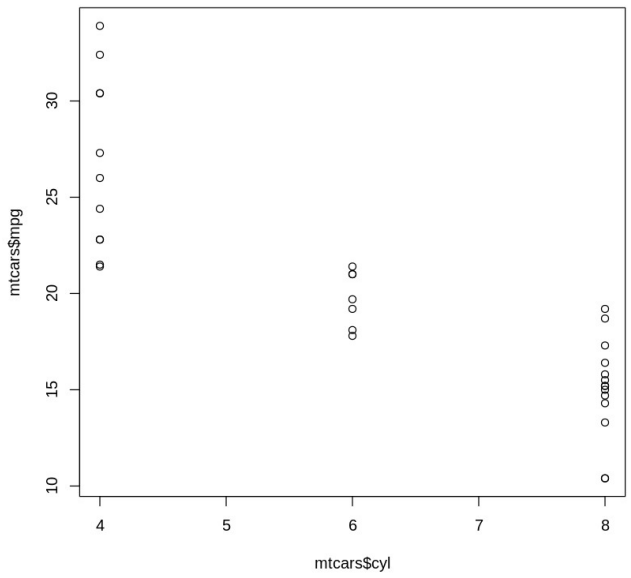
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

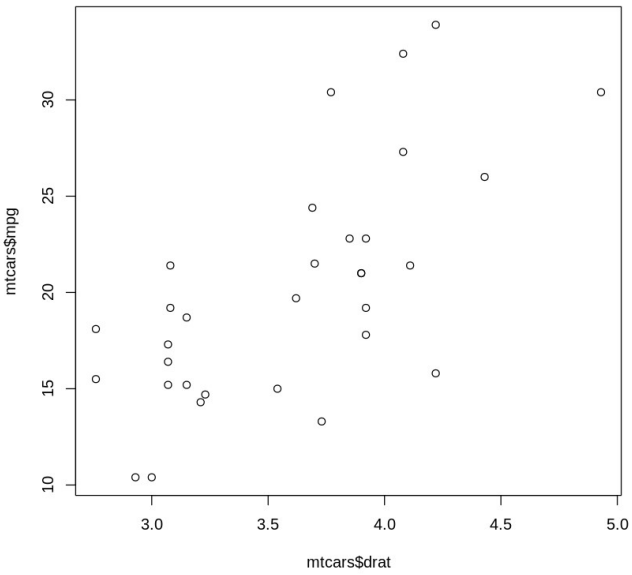
- Visual exploration

```
data(mtcars)
plot(mtcars$hp, mtcars$mpg)
plot(mtcars$wt, mtcars$mpg)
plot(mtcars$cyl, mtcars$mpg)
plot(mtcars$drat, mtcars$mpg)
```

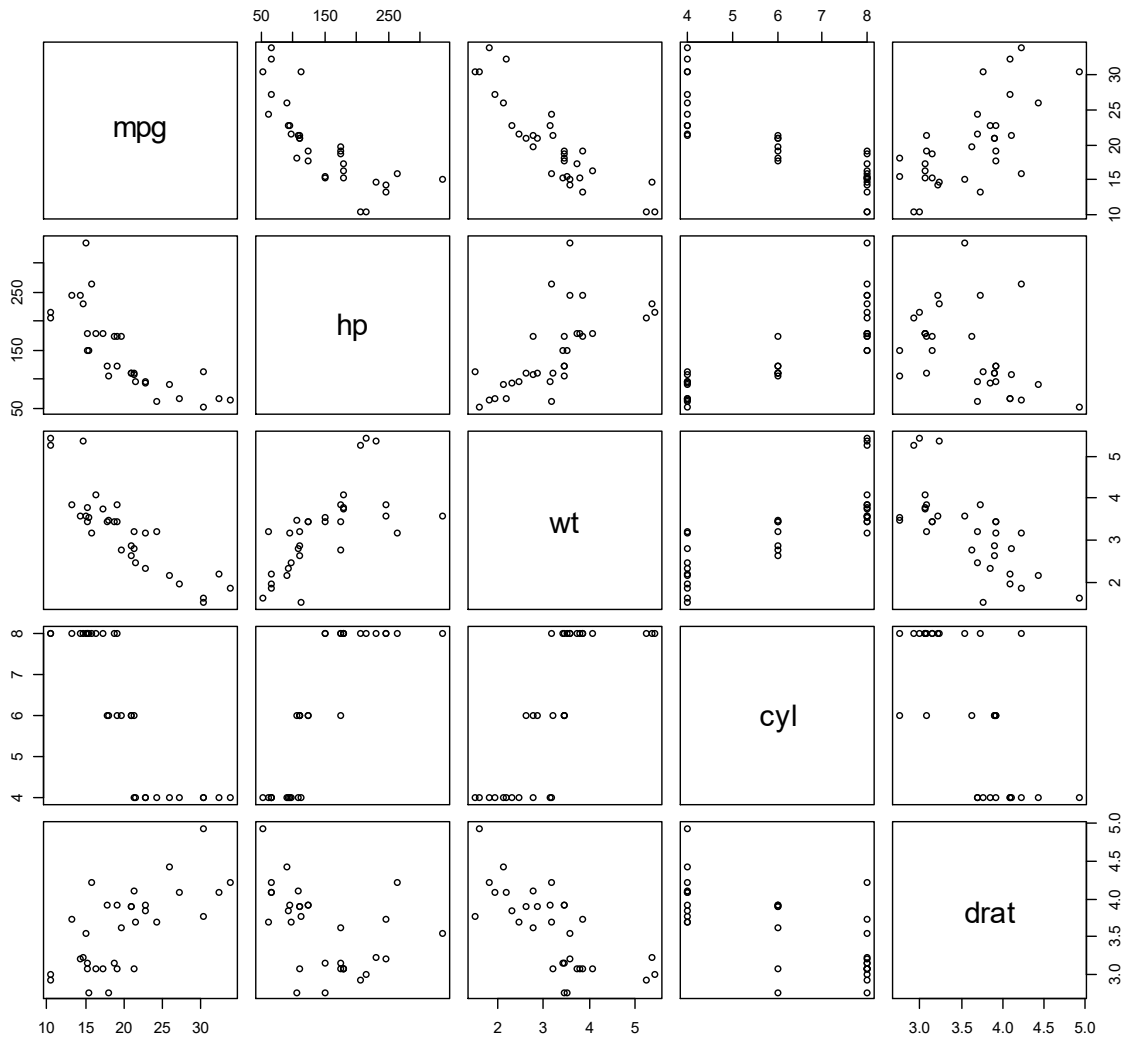









```
pairs(mtcars[,c('mpg','hp','wt','cyl','drat')])
```

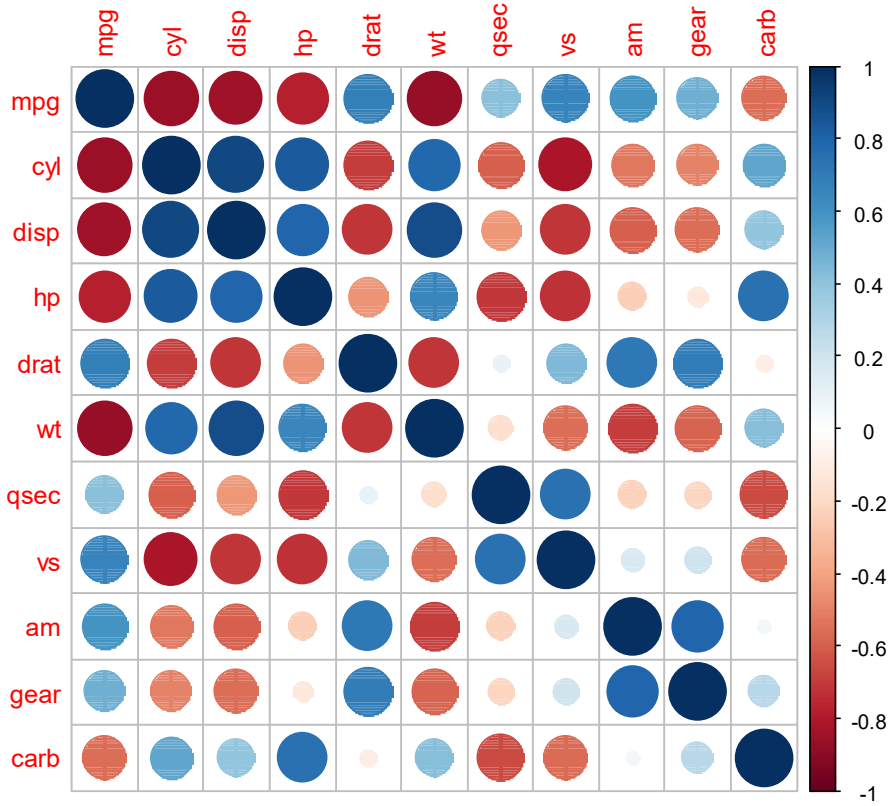


- Correlation analysis

```
round(cor(mtcars), digits = 2)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

```
library(corrplot)
corrplot(cor(mtcars), method = "circle")
```



- Exploring relationships with simple regression models

```
model1 <- lm(mpg ~ hp, data = mtcars)
summary(model1)
```

```
Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

How to interpret the results?

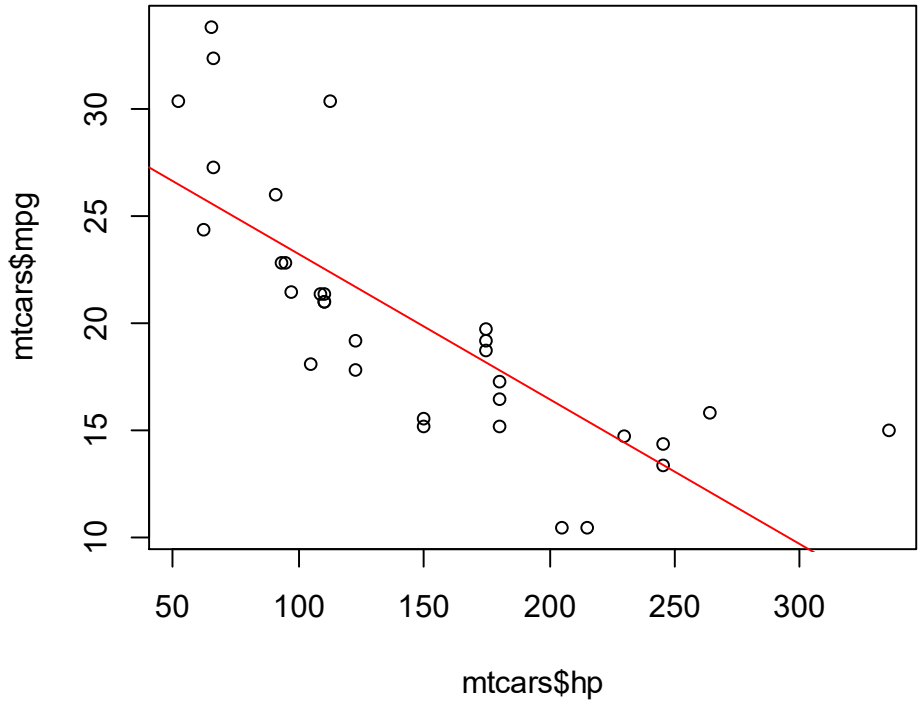
a. Intercept Term:
   - Represents the predicted value of mpg when hp is 0 (not meaningful in this context).
   - In this model, the intercept is 30.099, suggesting a car with 0 horsepower would have a predicted fuel efficiency of 30.099 mpg (hypothetical).

b. Slope Term (hp):
   - Represents the change in mpg for a one-unit increase in hp.
   - The slope is -0.068, indicating that for each additional horsepower, fuel efficiency decreases by 0.068 mpg.

c. R-squared:
   - A measure of how well the model fits the data (0 to 1).
   - In this model, R-squared is 0.602, meaning 60.2% of the variation in mpg can be explained by the linear relationship with hp.
   - Higher the value better the model.

d. P-values [Pr(>|t|)]:
   - The p-value represents the probability that the coefficient is zero (has no effect).
   - In social science, in general, p-value < 0.05 indicates statistically significant relationship between variables.
   - Exploring relationships with regression models

```
plot(mtcars$hp, mtcars$mpg)
abline(model1, col='red')
```



- Exploring relationships with multiple regression models.

```
model2 <- lm(mpg ~ hp + drat, data = mtcars)
summary(model2)
```

```
Call:
lm(formula = "mpg ~ hp + drat", data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0369 -2.3487 -0.6034  1.1897  7.7500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.789861   5.077752   2.125 0.042238 *
hp          -0.051787   0.009293  -5.573 5.17e-06 ***
drat         4.698158   1.191633   3.943 0.000467 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.17 on 29 degrees of freedom
Multiple R-squared:  0.7412,    Adjusted R-squared:  0.7233
F-statistic: 41.52 on 2 and 29 DF,  p-value: 3.081e-09
```

Let’s interpret the results of model2.

```
model3 <- lm(mpg ~ hp + drat + cyl + wt, data = mtcars)
summary(model3)
```

```
Call:
lm(formula = "mpg ~ hp + drat + cyl + wt", data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6171 -1.5663 -0.6058  1.2612  5.8161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.49588    7.44101   4.636  8.1e-05 ***
hp          -0.02089    0.01295  -1.613  0.11845
drat         0.81771    1.38684   0.590  0.56034
cyl         -0.76229    0.63502  -1.200  0.24040
wt          -2.97331    0.81818  -3.634  0.00116 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.541 on 27 degrees of freedom
Multiple R-squared:  0.8451,    Adjusted R-squared:  0.8222
F-statistic: 36.84 on 4 and 27 DF,  p-value: 1.438e-10
```

Let’s interpret the results of model3.

- Why **hp, drat**, and **cyl** have insignificant coefficient estimates? : Because all the independent variables are highly correlated to each other. This leads to Multicollinearity problem in the estimation.

```
model4 <- lm(hp ~ disp + carb, data = mtcars)
summary(model4)
```

```
Call:
lm(formula = "hp ~ disp + carb", data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-46.06 -17.14  -0.98  14.87  52.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.98848   11.61391   0.860    0.397
disp         0.32431    0.04298   7.546 2.56e-08 ***
carb        21.99928    3.29762   6.671 2.57e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.24 on 29 degrees of freedom
Multiple R-squared:  0.8523,    Adjusted R-squared:  0.8421
F-statistic: 83.66 on 2 and 29 DF,  p-value: 9.047e-13
```

Let’s interpret the results of model4.

Presenting the model results using "stargazer"

- Descriptive statistics using "stargazer"

```
stargazer(mtcars, type = 'text')
```

```
==========================================
Statistic N   Mean   St. Dev.  Min     Max
------------------------------------------
mpg       32 20.091   6.027   10.400 33.900
cyl       32  6.188   1.786     4       8
disp      32 230.722 123.939  71.100 472.000
hp        32 146.688  68.563    52     335
drat      32  3.597   0.535   2.760   4.930
wt        32  3.217   0.978   1.513   5.424
qsec      32 17.849   1.787   14.500 22.900
vs        32  0.438   0.504     0       1
am        32  0.406   0.499     0       1
gear      32  3.688   0.738     3       5
carb      32  2.812   1.615     1       8
------------------------------------------
```

- Presenting the models results (model1, model2, model3, model4)

```
library(stargazer)

stargazer(model1, model2, model3, model4, type = "text",
          digits = 2,
          covariate.labels = c('Gross horsepower (hp)',
                               'Rear axle ratio (dart)',
                               'Number of cylinders (cyl)',
                               'Weight (1000 lbs) (wt)',
                               'Displacement (cu.in.) (disp)',
                               'Number of carburetors (carb)'),
          dep.var.labels = c("Miles/(US) gallon (mpg)", "Gross horsepower (hp)"),
          notes = "Standard errors are in parentheses.")
```

	Dependent variable:			
	Miles/(US) gallon (mpg)			Gross horsepower (hp)
	(1)	(2)	(3)	(4)
Gross horsepower (hp)	-0.07***	-0.05***	-0.02	
	(0.01)	(0.01)	(0.01)	
Rear axle ratio (dart)		4.70***	0.82	
		(1.19)	(1.39)	
Number of cylinders (cyl)			-0.76	
			(0.64)	
Weight (1000 lbs) (wt)			-2.97***	
			(0.82)	
Displacement (cu.in.) (disp)				0.32***
				(0.04)
Number of carburetors (carb)				22.00***
				(3.30)
Constant	30.10***	10.79**	34.50***	9.99
	(1.63)	(5.08)	(7.44)	(11.61)
Observations	32	32	32	32
R2	0.60	0.74	0.85	0.85
Adjusted R2	0.59	0.72	0.82	0.84
Residual Std. Error	3.86 (df = 30)	3.17 (df = 29)	2.54 (df = 27)	27.24 (df = 29)
F Statistic	45.46*** (df = 1; 30)	41.52*** (df = 2; 29)	36.84*** (df = 4; 27)	83.66*** (df = 2; 29)
Note:				*p<0.1; **p<0.05; ***p<0.01
				Standard errors are in parentheses.

Saving results to a html file named 'model_results.html'.

```
stargazer(model1, model2, model3, model4, type = "html", out = 'model_results.html',
          digits = 2,
          covariate.labels = c('Gross horsepower (hp)',
                               'Rear axle ratio (dart)',
                               'Number of cylinders (cyl)',
                               'Weight (1000 lbs) (wt)',
                               'Displacement (cu.in.) (disp)',
                               'Number of carburetors (carb)'),
          dep.var.labels = c("Miles/(US) gallon (mpg)", "Gross horsepower (hp)"),
          notes = "Standard errors are in parentheses.")
```

	Dependent variable:			
	Miles/(US) gallon (mpg)			Gross horsepower (hp)
	(1)	(2)	(3)	(4)
Gross horsepower (hp)	-0.07***	-0.05***	-0.02	
	(0.01)	(0.01)	(0.01)	
Rear axle ratio (dart)		4.70***	0.82	
		(1.19)	(1.39)	
Number of cylinders (cyl)			-0.76	
			(0.64)	
Weight (1000 lbs) (wt)			-2.97***	
			(0.82)	
Displacement (cu.in.) (disp)				0.32***
				(0.04)
Number of carburetors (carb)				22.00***
				(3.30)
Constant	30.10***	10.79**	34.50***	9.99
	(1.63)	(5.08)	(7.44)	(11.61)
Observations	32	32	32	32
R^2	0.60	0.74	0.85	0.85
Adjusted R^2	0.59	0.72	0.82	0.84
Residual Std. Error	3.86 (df = 30)	3.17 (df = 29)	2.54 (df = 27)	27.24 (df = 29)
F Statistic	45.46*** (df = 1; 30)	41.52*** (df = 2; 29)	36.84*** (df = 4; 27)	83.66*** (df = 2; 29)
Note:				*p<0.1; **p<0.05; ***p<0.01
				Standard errors are in parentheses.