General Linear Statistical Models

Statistics 135

Autumn 2005

General Linear Statistical Models

This framework includes

- Linear Regression
- Analysis of Variance (ANOVA)
- Analysis of Covariance (ANCOVA)

These models can all be analyzed with the function `lm`.

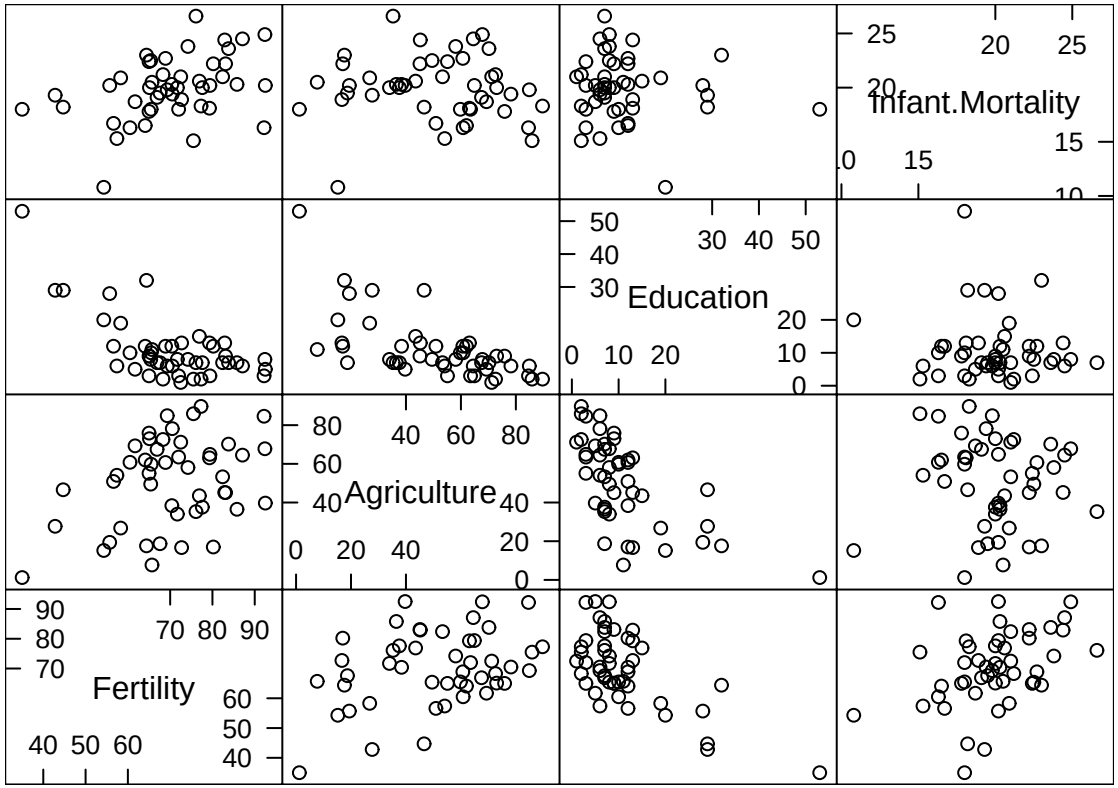
Note that much of what I plan to discuss will also extend to Generalized Linear Models (`glm`), Nonlinear Least Squares (`nls`), Generalized Additive Models (`gam`), and Regression Trees - Recursive Partitioning (`rpart`). Though not surprisingly, extensions will be required for some of these.

References:

- Ramsey FL and Schafer DW. The Statistical Sleuth, 2nd edition
- Neter J, Kutner MH, Nachtsheim CJ, and Wasserman W. Applied Linear Statistical Models, 4th edition.
- Montgomery DC and Peck EA. Introduction to Linear Regression Analysis, 2nd edition.
- Draper NR and Smith H. Applied Regression Analysis, 3rd edition.

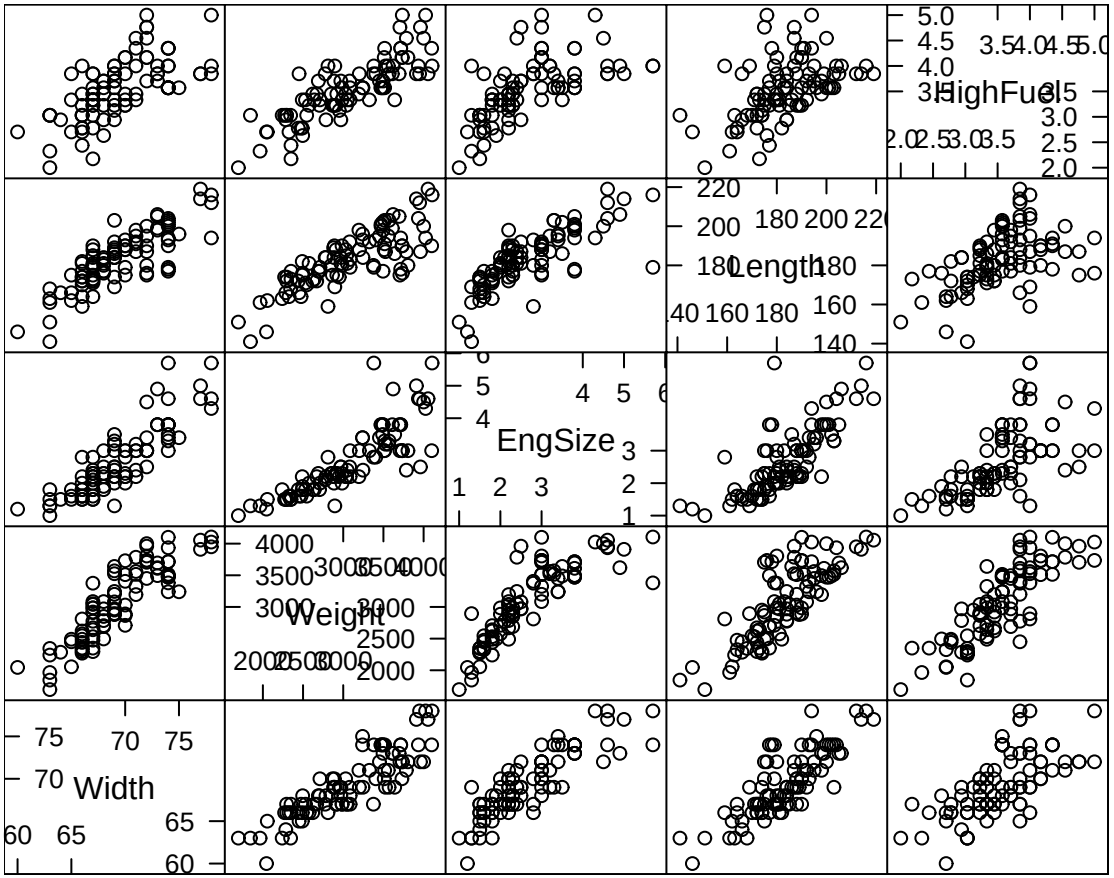
Linear Regression (Quantitative Predictors)

1. Model infant mortality (`Infant.Mortality`) in Switzerland by `Education`, `Agriculture`, and `Fertility` in the dataset `swiss`.



Scatter Plot Matrix

2. Model EPA highway fuel use (`HighFuel`) by `Weight`, engine size (`EngSize`), `Length`, and `Width` in the `cars93` dataset.



Scatter Plot Matrix

Want to fit models of the form

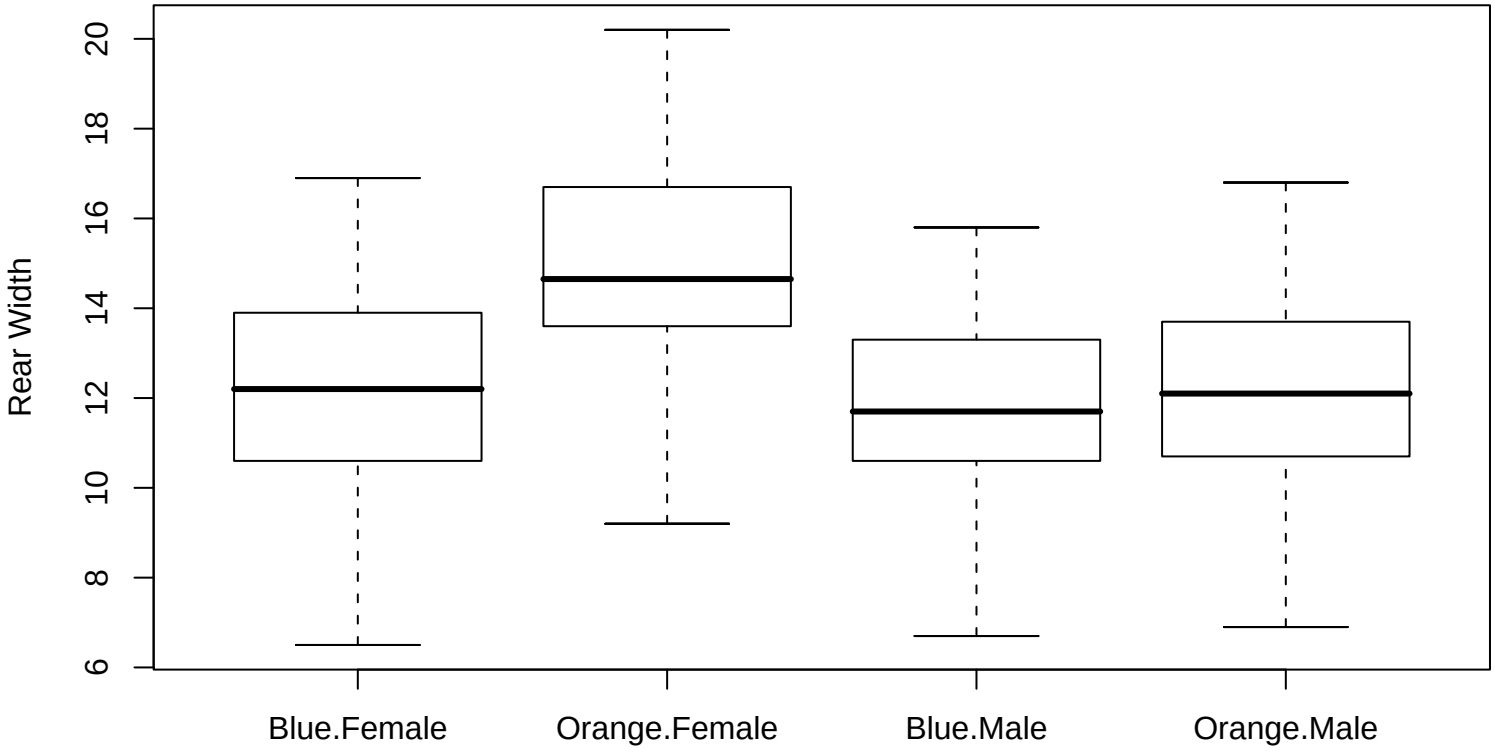
$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i; \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

This model also includes polynomial regression, as for example, could have $x_{ki} = x_{ji}^2$.

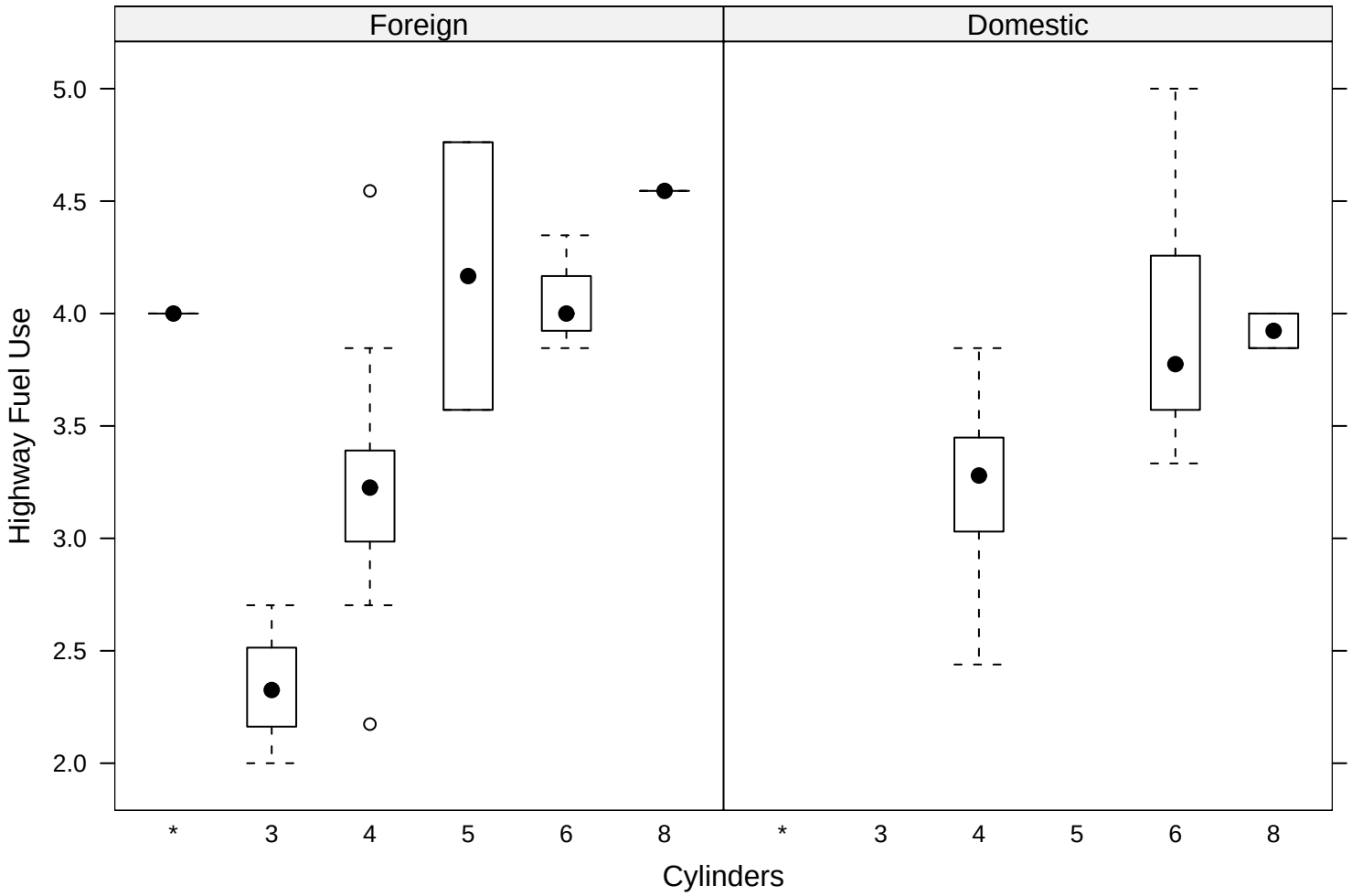
Note that linear regression refers to being linear in the parameters β, not the predictors. For example, polynomial or log transformations of the predictors is fine.

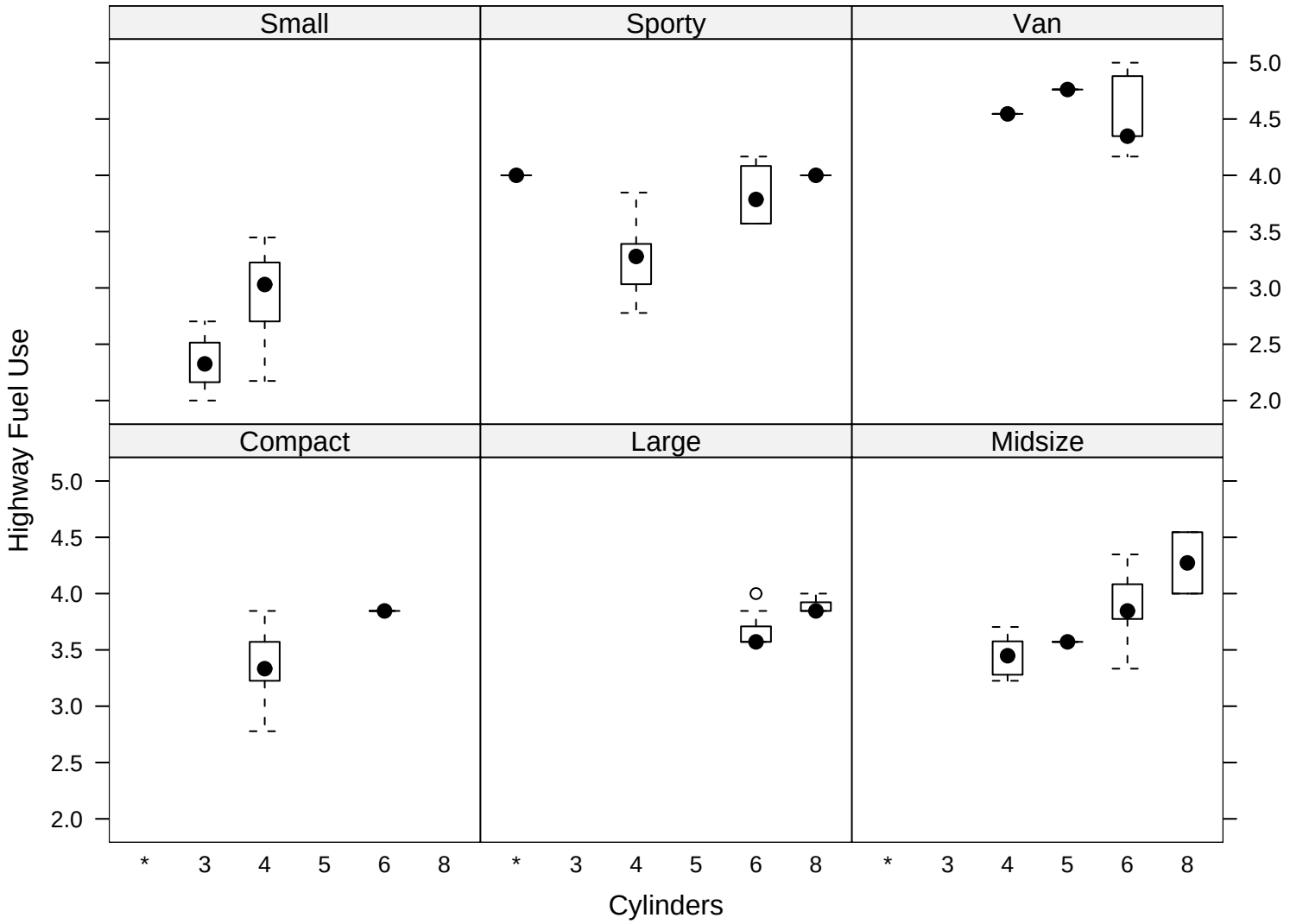
ANOVA (Qualitative Predictors)

1. Model rear width (`RW`) of *Leptograpsus variegates* by `sex` and `species` in the `crabs` dataset.



2. Model EPA highway fuel use (`HighFuel`) by car type (`Type`), number of cylinders (`Cylinders`), and where made (`Domestic`) in the `cars93` dataset.


Foreign
Domestic
Highway Fuel Use
5.0
4.5
4.0
3.5
3.0
2.5
2.0
*
3
4
5
6
8
Cylinders


Small
Sporty
Van
Compact
Large
Midsize
Highway Fuel Use
Cylinders
2.0
2.5
3.0
3.5
4.0
4.5
5.0
*
3
4
5
6
8

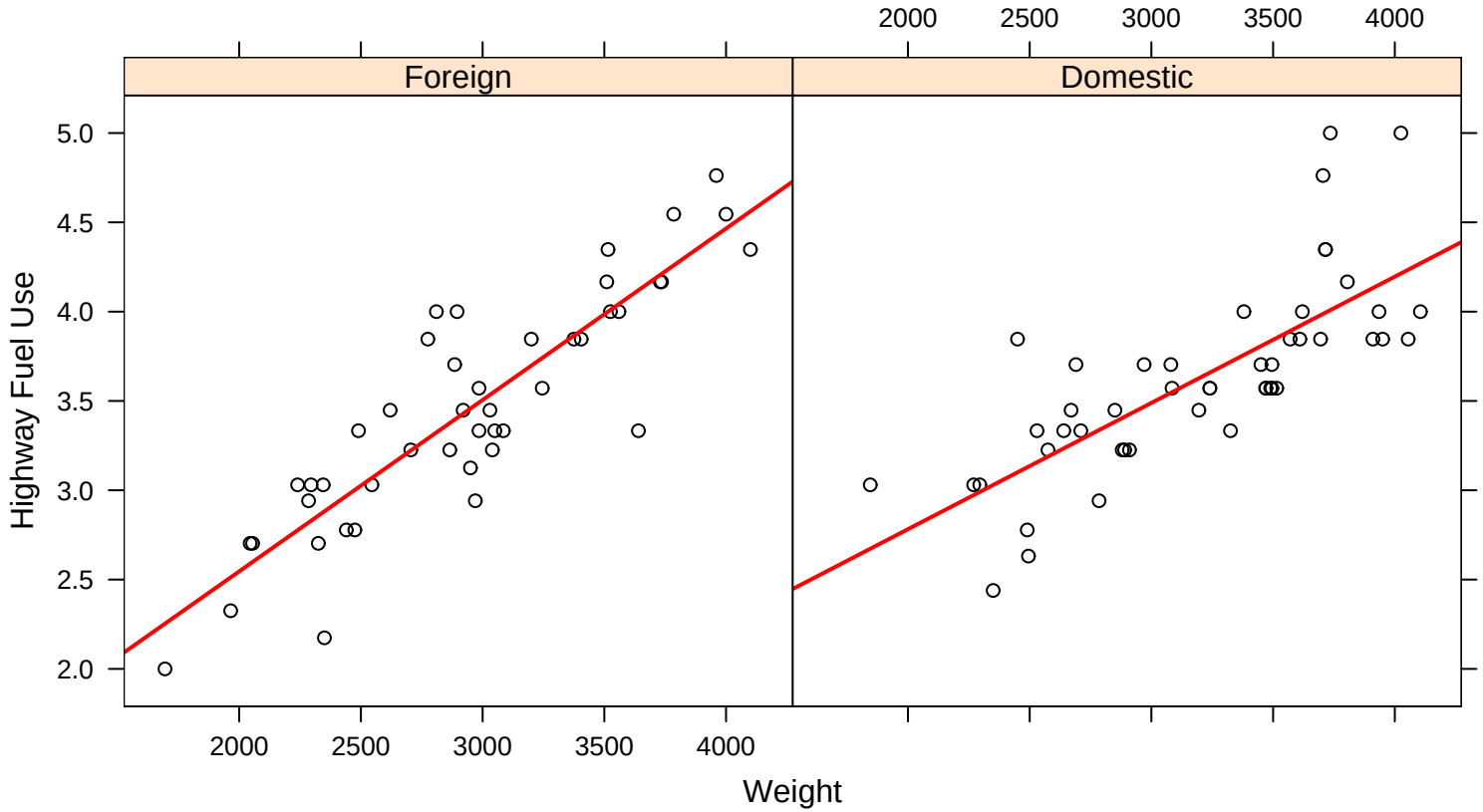
Want to fit models of the form

$$y_{ijkl} = \mu + (\alpha\beta\gamma)_{jkl} + \epsilon_{ijkl}; \qquad \epsilon_{ijkl} \overset{iid}{\sim} N(0, \sigma^2)$$

Have a potentially different mean for each combination of the factor levels.

ANCOVA (Combination of quantitative variables and qualitative factors)

Model EPA highway fuel use (`HighFuel`) by `Weight` and where made (`Domestic`) in the `cars93` dataset.



Want to fit models of the form

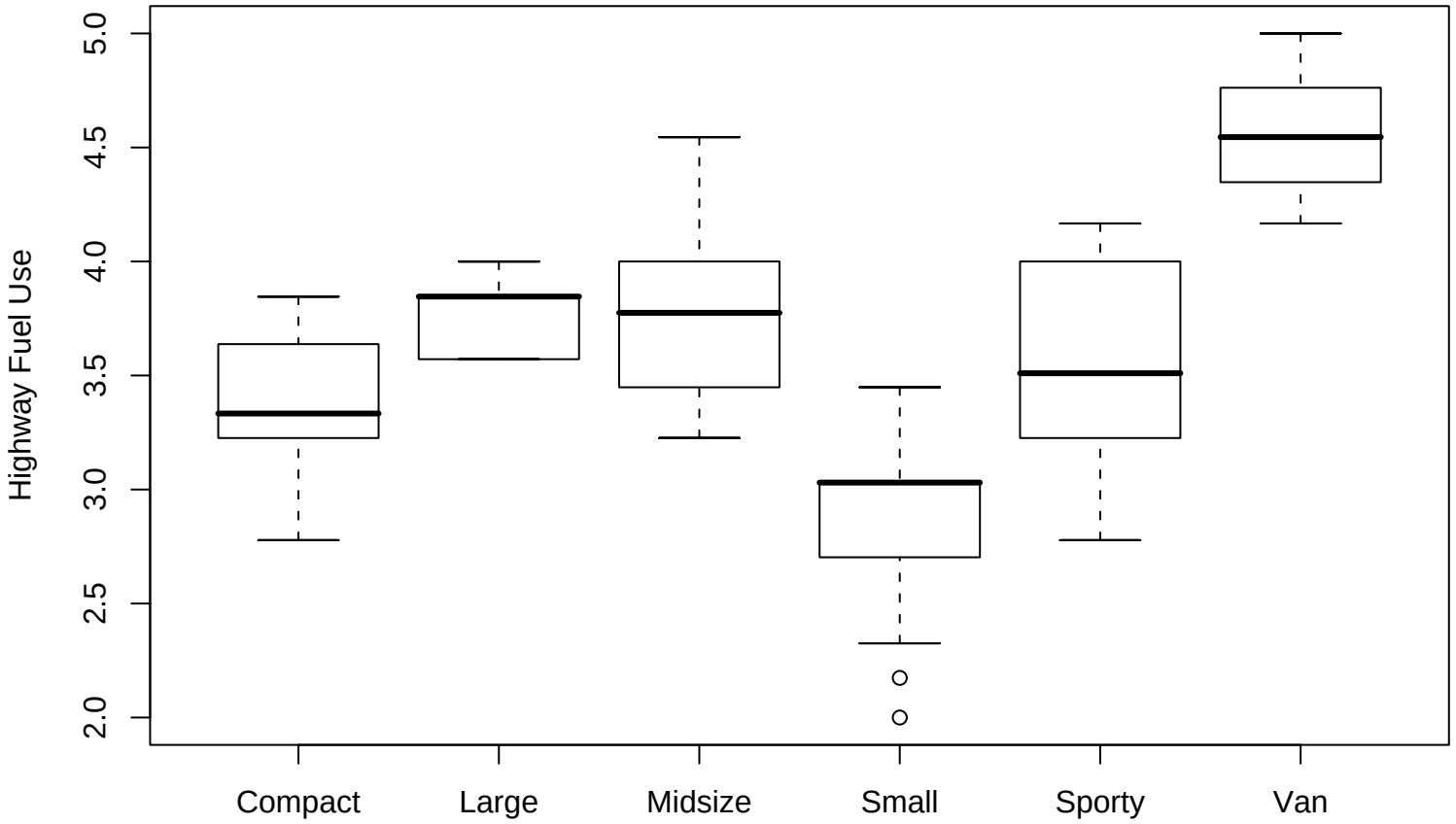
$$y_{ji} = \beta_{0j} + \beta_{1j}x_{1ji} + \ldots + \beta_{pj}x_{pji} + \epsilon_{ji}; \qquad \epsilon_{ji} \overset{iid}{\sim} N(0, \sigma^2)$$

Have a different regression line (surface) for each combination of the qualitative factors.

In fact, all three situations are special cases of a common model. They can all be written in the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \epsilon_i; \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Consider 1-way ANOVA, where there is a single qualitative variable as a predictor. An example of this would be `HighFuel` modeled by `Type`



This data could be described by the model

$$y_{ji} = \mu + \alpha_j + \epsilon_{ji}; \qquad \epsilon_{ji} \overset{iid}{\sim} N(0, \sigma^2)$$

It can be converted to the other setting with

$$x_{1i} = \begin{cases} 1 & \text{car } i \text{ is Compact} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{car } i \text{ is Large} \\ 0 & \text{otherwise} \end{cases}$$

$$\dots$$

$$x_{5i} = \begin{cases} 1 & \text{car } i \text{ is Sporty} \\ 0 & \text{otherwise} \end{cases}$$

Note that we need one less x variable than the number of levels of the categorical factor.

This is only one possible way of defining x variables for the regression setting. The are other equally valid approaches. What is required that the different observed combinations of the xs describe the different levels of the categorical factor. How these variables are defined induces the relationship between the βs and μ and the αs.

In **S**, there are easy approaches of creating the x automatically from the factors (to come).

Defining a Model

The basic approach of defining a model is with the form

```
y ~ x1 + x2 + . . . + xk
```

where `xj` could be a quantitative variable, a qualitative factor, or a combination of variables.

For example, in the Infant Mortality example,

```
Infant.Mortality ~ Education + Agriculture + Fertility
```

describes the model

$$y_i = \beta_0 + \beta_1 x_{1i} + + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i; \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

To fit this model we can use the command

```
swiss.lm <- lm(Infant.Mortality ~ Education + Agriculture
                    + Fertility, data=swiss)
```

A description of the model fits can be given by the `summary` function.

```
> summary(swiss.lm)

Call:
lm(formula = Infant.Mortality ~ Education + Agriculture
    + Fertility, data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1086 -1.3820  0.1706  1.7167  5.8039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.14163    3.85882   2.628  0.01185 *
Education    0.06593    0.06602   0.999  0.32351
Agriculture -0.01755    0.02234  -0.785  0.43662
Fertility    0.14208    0.04176   3.403  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.625 on 43 degrees of freedom
Multiple R-Squared: 0.2405,     Adjusted R-squared: 0.1875
F-statistic:  4.54 on 3 and 43 DF,  p-value: 0.007508
```

Note that this only gives part of the standard regression output. To get the ANOVA table, use the `ANOVA` command.

```
> anova(swiss.lm)
Analysis of Variance Table

Response: Infant.Mortality
             Df  Sum Sq Mean Sq F value   Pr(>F)
Education     1   3.850   3.850  0.5585 0.458920
Agriculture   1  10.215  10.215  1.4820 0.230103
Fertility     1  79.804  79.804 11.5780 0.001454 **
Residuals    43 296.386   6.893
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```