- ANCOVA merges ANOVA with regression

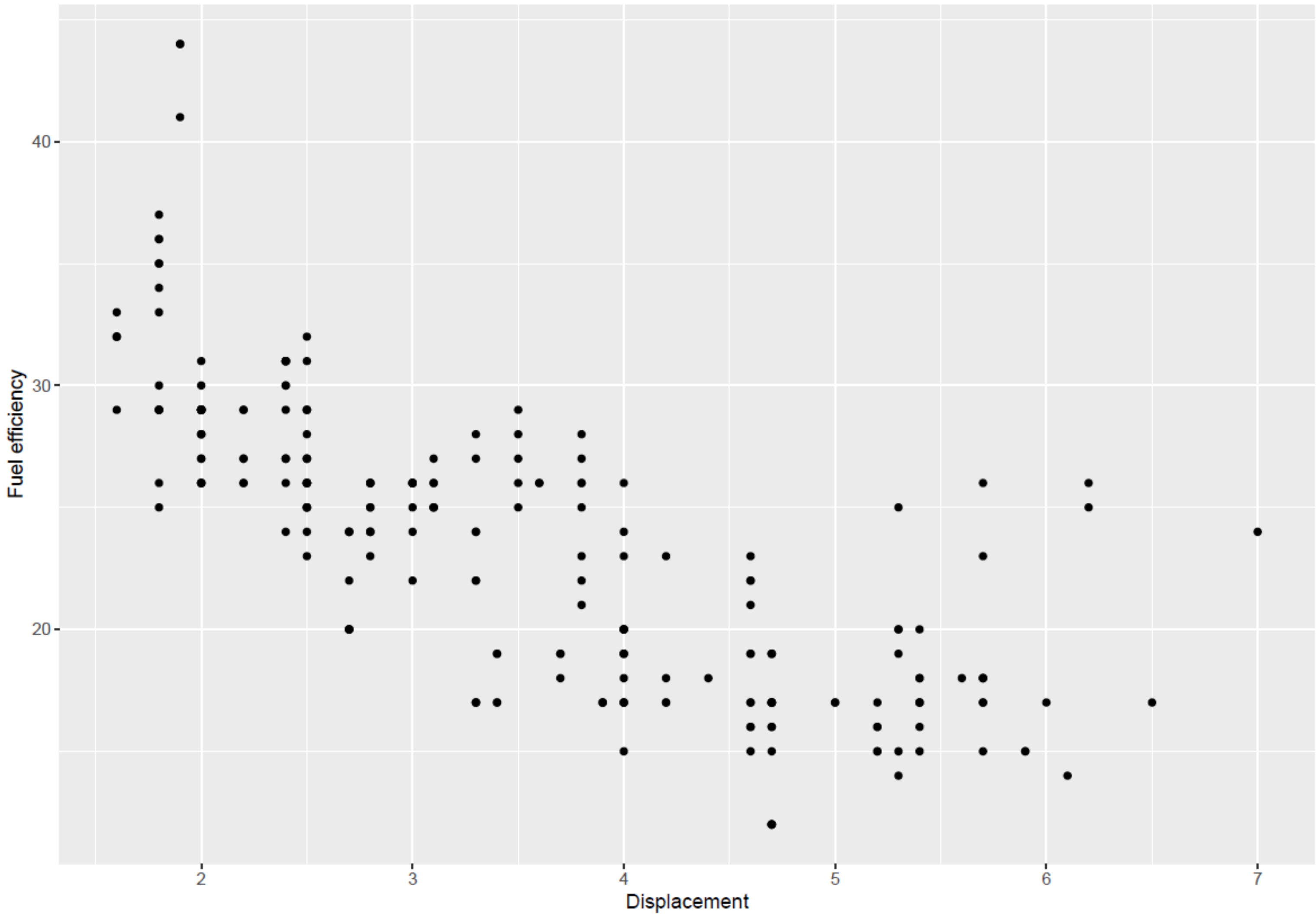
- ANOVA: categorical predictors (*factors*)
- Regression: continuous predictors
- ANCOVA: categorical predictors (*covariates*) and continuous predictors (*factors*)

Analysis of covariance (ANCOVA)

- E.g. testing the effectiveness of different brands of detergents at removing marks on different types of fabric, controlling the effect that detergents may be more effective in warm water

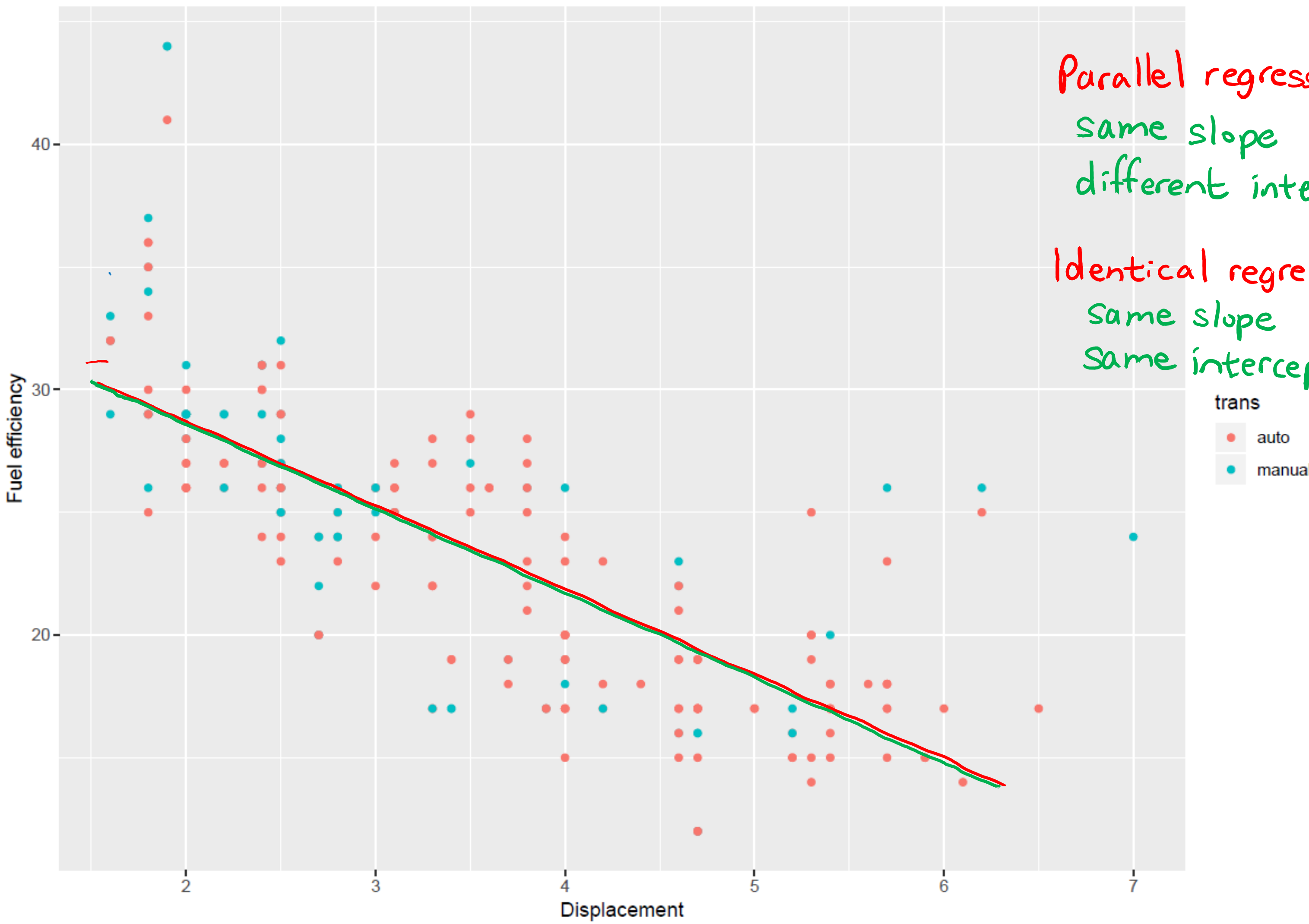
- MANOVA and MANCOVA is the multivariate version of ANOVA and ANCOVA (‘M’ stands for multivariate), which are used when there is more than one response variable

Motivating example



Motivating example (cont.)

Separate regression lines
different slopes
different intercepts



Parallel regression lines
same slope
different intercepts

Identical regression lines
same slope
same intercept

Setup

					from
Group 1	y_{11}, x_{11}	y_{12}, x_{12}	$\cdots$	y_{1n}, x_{1n}	$N(\mu_1, \sigma^2)$
Group 2	y_{21}, x_{21}	y_{22}, x_{22}	$\cdots$	y_{2n}, x_{2n}	$N(\mu_2, \sigma^2)$
$\vdots$					$\vdots$
Group k	y_{k1}, x_{k1}	y_{k2}, x_{k2}	$\cdots$	y_{kn}, x_{kn}	$N(\mu_k, \sigma^2)$

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

$$Y_{ij} = \mu_i + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim \text{iid } N(0, \sigma^2) \quad \text{for } \begin{array}{l} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n_i \end{array}$$

Three cases for μ_i:

$\mu_i = \beta_{i0} + \beta_{i1} x_{ij}$ separate regression lines

$\mu_i = \beta_{i0} + \beta_1 x_{ij}$ parallel regression lines

$\mu_i = \beta_0 + \beta_1 x_{ij}$ identical regression lines

The ANCOVA (with separate regression lines) can be set up as a MLR as follows:

covariate x_i

$$
\boldsymbol{Y}=\begin{bmatrix} y_{11}\\ y_{12}\\ \vdots\\ y_{1n_1}\\ y_{21}\\ y_{22}\\ \vdots\\ y_{2n_2}\\ \vdots\\ y_{k1}\\ y_{k2}\\ \vdots\\ y_{kn_k}\end{bmatrix},
\boldsymbol{X}=\begin{bmatrix}
1 & x_{11} & 0 & \cdots & 0 & 0 & \cdots & 0\\
1 & x_{12} & 0 & \cdots & 0 & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots\\
1 & x_{1n_1} & 0 & \cdots & 0 & 0 & \cdots & 0\\
1 & x_{21} & 1 & \cdots & 0 & x_{21} & \cdots & 0\\
1 & x_{22} & 1 & \cdots & 0 & x_{22} & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots\\
1 & x_{2n_2} & 1 & \cdots & 0 & x_{2n_2} & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots\\
1 & x_{k1} & 0 & \cdots & 1 & 0 & \cdots & x_{k1}\\
1 & x_{k2} & 0 & \cdots & 1 & 0 & \cdots & x_{k2}\\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots\\
1 & x_{kn_k} & 0 & \cdots & 1 & 0 & \cdots & x_{kn_k}
\end{bmatrix},
\boldsymbol{\beta}=\begin{bmatrix} \beta_{10}\\ \beta_{11}\\ \beta_{20}-\beta_{10}\\ \vdots\\ \beta_{k0}-\beta_{10}\\ \beta_{21}-\beta_{11}\\ \vdots\\ \beta_{k1}-\beta_{kn_k}\end{bmatrix}
$$

set to 0

set to 0

Then the parallel and identical regression lines can be specified by, respectively,

parallel H_1: the last $k-1$ elements of $\boldsymbol{\beta}$ are zero. The last k-1 columns of X are zero.

identical H_0: the last $2(k-1)$ elements of $\boldsymbol{\beta}$ are zero. The last 2(k-1) columns of X are zero.

Types of models

- Identical regression lines (for all groups):

i is the index of the groups

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \quad i = 1,2,\ldots,I, \; j = 1,2,\ldots,n_i.$$

- Parallel regression lines:

$$Y_{ij} = \beta_{i0} + \beta_1 x_{ij} + \epsilon_{ij}, \quad i = 1,2,\ldots,I, \; j = 1,2,\ldots,n_i.$$

- Separate regression lines:

$$Y_{ij} = \beta_{i0} + \beta_{i1} x_{ij} + \epsilon_{ij}, \; i = 1,2,\ldots,I, \; j = 1,2,\ldots,n_i.$$

$\varepsilon_{ij} \sim \text{iid } N(0, \sigma^2)$

Identical regression lines

```r
model_matrix(mpg, hwy ~ displ)
```

```
## # A tibble: 234 x 2
##    `(Intercept)` displ
##            <dbl> <dbl>
##  1             1   1.8
##  2             1   1.8
##  3             1   2
##  4             1   2
##  5             1   2.8
##  6             1   2.8
##  7             1   3.1
##  8             1   1.8
##  9             1   1.8
## 10             1   2
## # ... with 224 more rows
```

Identical regression lines (cont.)

We can represent this as the linear model:

hwy displ

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where

$$\epsilon_i \sim i.i.d. N(0, \sigma^2), i = 1, 2, \ldots, n$$

Parallel regression lines

```r
model_matrix(mpg, hwy ~ displ + trans)
```

```
## # A tibble: 234 x 3
##    `(Intercept)` displ transmanual
##            <dbl> <dbl>       <dbl>
##  1             1   1.8           0
##  2             1   1.8           1
##  3             1   2             1
##  4             1   2             0
##  5             1   2.8           0
##  6             1   2.8           1
##  7             1   3.1           0
##  8             1   1.8           1
##  9             1   1.8           0
## 10             1   2             1
## # ... with 224 more rows
```

Parallel regression lines (cont.)

We can represent this as the linear model:

$$\overset{\text{hwy}}{Y_i} = \beta_0 + \beta_1 \overset{\text{displ}}{\boxed{x_{i1}}} + \beta_2 \overset{\text{trans}}{\boxed{x_{i2}}} + \epsilon_i,$$

where x_{i1} is the displacement, x_{i2} is 1 for manual and 0 for automatic, and

$$\epsilon_i \sim i.i.d.\, N(0, \sigma^2), i = 1, 2, \ldots, n$$

What is the model for a manual? for an automatic?

auto ($x_{i2} = 0$): $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

manual ($x_{i2} = 1$): $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \epsilon_i = (\beta_0 + \beta_2) + \beta_1 x_{i1} + \epsilon_i$

Separate regression lines

```
model_matrix(mpg, hwy ~ displ + trans + displ:trans)
```

```
## # A tibble: 234 x 4
##    `(Intercept)` displ transmanual `displ:transmanual`
##            <dbl> <dbl>       <dbl>               <dbl>
##  1             1   1.8           0                 0
##  2             1   1.8           1                 1.8
##  3             1   2             1                 2
##  4             1   2             0                 0
##  5             1   2.8           0                 0
##  6             1   2.8           1                 2.8
##  7             1   3.1           0                 0
##  8             1   1.8           1                 1.8
##  9             1   1.8           0                 0
## 10             1   2             1                 2
## # ... with 224 more rows
```

Separate regression lines

```r
model_matrix(mpg, hwy ~ displ * trans)
```

```
## # A tibble: 234 x 4
##    `(Intercept)` displ transmanual `displ:transmanual`
##            <dbl> <dbl>       <dbl>               <dbl>
##  1             1   1.8           0                 0
##  2             1   1.8           1                 1.8
##  3             1   2             1                 2
##  4             1   2             0                 0
##  5             1   2.8           0                 0
##  6             1   2.8           1                 2.8
##  7             1   3.1           0                 0
##  8             1   1.8           1                 1.8
##  9             1   1.8           0                 0
## 10             1   2             1                 2
## # ... with 224 more rows
```

Separate regression lines (cont.)

We can represent this as the linear model:

interaction term

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \boxed{\beta_3 x_{i1} x_{i2}} + \epsilon_i,$$

where x_{i1} is the displacement, x_{i2} is 1 for manual and 0 for automatic, and

$$\epsilon_i \sim i.i.d. N(0, \sigma^2), i = 1, 2, \ldots, n$$

What is the model for a manual? for an automatic?

auto ($x_{i2} = 0$): $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

manual ($x_{i2} = 1$): $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1} + \epsilon_i$

$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} + \epsilon_i$$

Model selection

Start with the largest model – separate regression.

```
model_sep<- lm(hwy ~ displ * trans, data=mpg)
summary(model_sep)
```

```
##
## Call:
## lm(formula = hwy ~ displ * trans, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1441 -2.2946 -0.2436  2.1184 14.7553
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       35.39457    0.94674  37.386   <2e-16 ***
## displ             -3.52217    0.24090 -14.621   <2e-16 ***
## transmanual        0.02559    1.51343   0.017    0.987
## displ:transmanual  0.27194    0.44143   0.616    0.538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 230 degrees of freedom
## Multiple R-squared:  0.5921, Adjusted R-squared:  0.5868
## F-statistic: 111.3 on 3 and 230 DF,  p-value: < 2.2e-16
```

Parallel regression lines

```r
model_parallel <- update(model_sep, .~. - displ:trans)
summary(model_parallel)
```

```
##
## Call:
## lm(formula = hwy ~ displ + trans, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8130 -2.2109 -0.2639  2.0964 14.5517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.0933     0.8096  43.348   <2e-16 ***
## displ        -3.4412     0.2016 -17.070   <2e-16 ***
## transmanual   0.8933     0.5531   1.615    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.823 on 231 degrees of freedom
## Multiple R-squared:  0.5914, Adjusted R-squared:  0.5879
## F-statistic: 167.2 on 2 and 231 DF,  p-value: < 2.2e-16
```

Identical regression lines

```
model_identical <- update(model_parallel, .~. - trans)
summary(model_identical)
```

```
##
## Call:
## lm(formula = hwy ~ displ, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1039 -2.1646 -0.2242  2.0589 15.0105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.6977     0.7204   49.55   <2e-16 ***
## displ        -3.5306     0.1945  -18.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.836 on 232 degrees of freedom
## Multiple R-squared:  0.5868, Adjusted R-squared:  0.585
## F-statistic: 329.5 on 1 and 232 DF,  p-value: < 2.2e-16
```

all terms are significant

We will stop and use identical regression lines

ANOVA

```
anova(model_identical, model_parallel, model_sep)
```

```
Analysis of Variance Table

Model 1: hwy ~ displ
Model 2: hwy ~ displ + trans
Model 3: hwy ~ displ * trans
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    232 3413.8
2    231 3375.7  1    38.113 2.6011 0.1082
3    230 3370.2  1     5.561 0.3795 0.5385
```

comparing models 1 and 2

comparing models 2 and 3

We choose Model 1 in this case.

AIC

```
AIC(model_identical, model_parallel, model_sep)
```

```
##                  df      AIC
## model_identical   3 1297.246
## model_parallel    4 1296.619
## model_sep         5 1298.233
```

Parallel model has the lowest AIC, hence this model is preferred by AIC.

BIC

```
BIC(model_identical, model_parallel, model_sep)
```

```
##                 df      BIC
## model_identical  3 1307.612
## model_parallel   4 1310.440
## model_sep        5 1315.510
```

Identical model has the lowest BIC, hence this model is preferred by BIC.

Note that the *F*-tests, AIC, and BIC may led to different conclusions (as in this case). Always specify the selection criterion used to choose your final model.

Plot