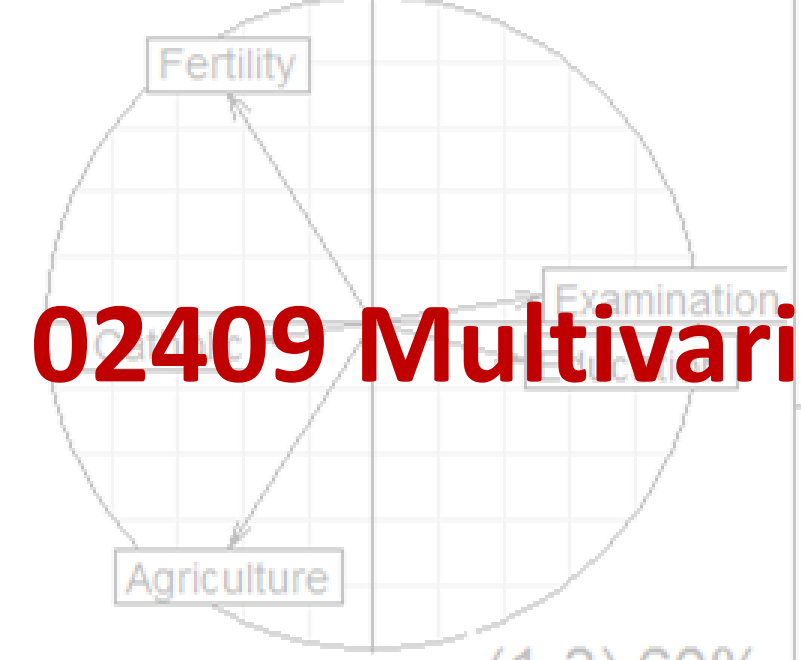
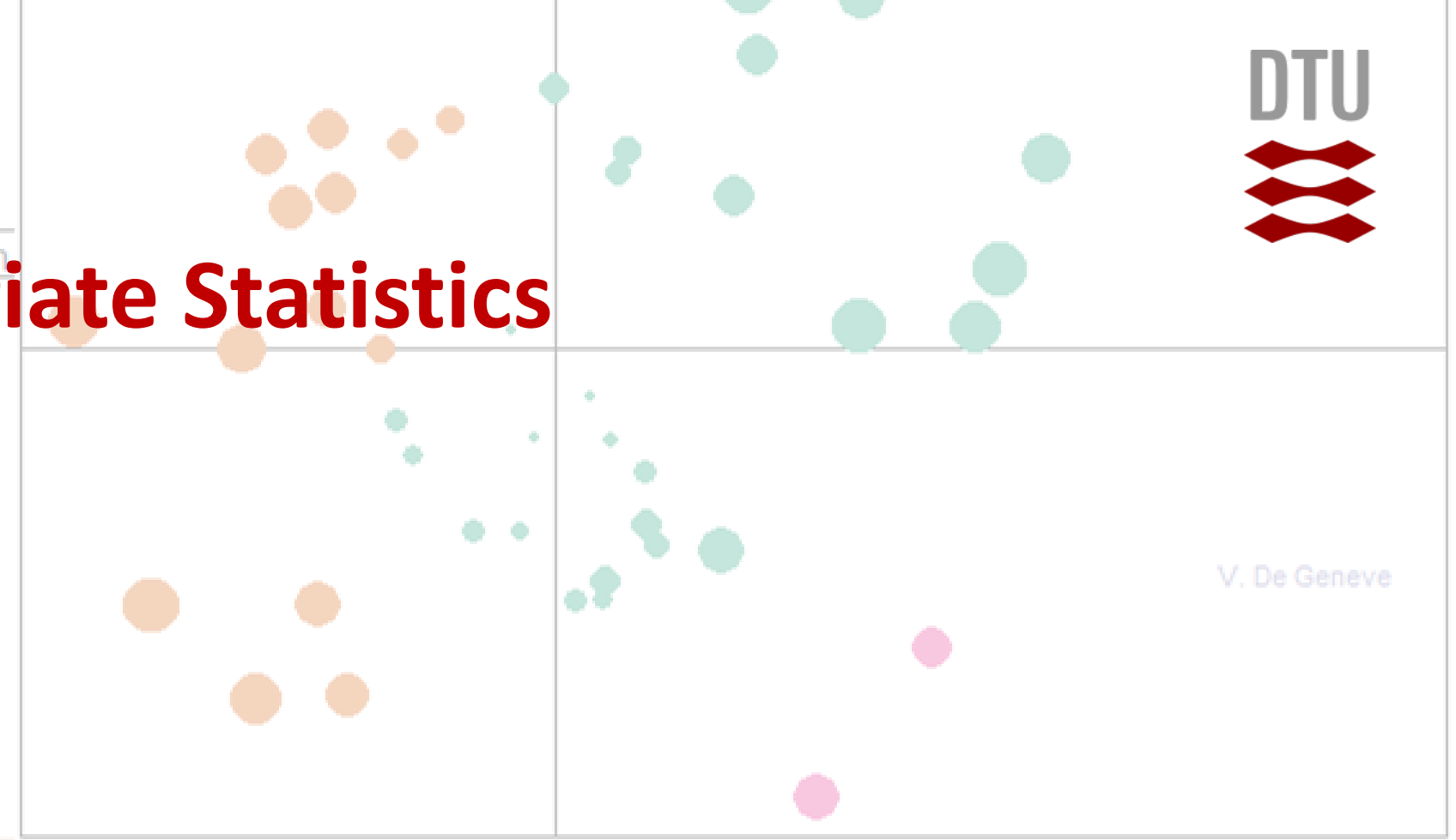


02409 Multivariate Statistics

Lecture J, November 10 2025

Anders Stockmarr

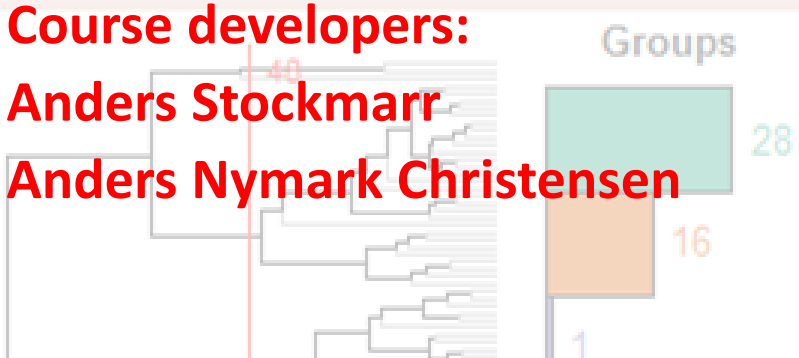
anst@dtu.dk

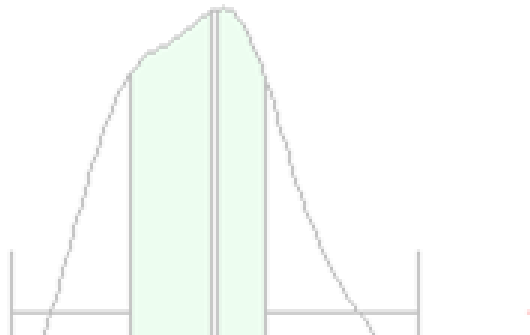


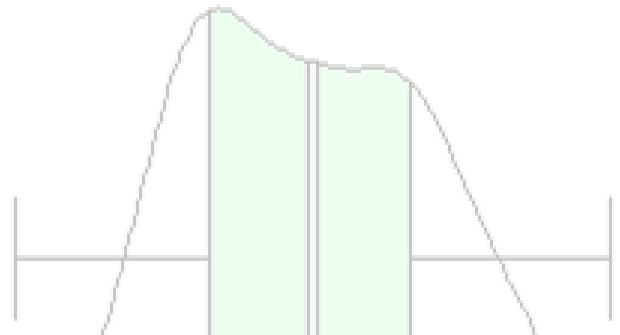
Course developers:

Anders Stockmarr

Anders Nymark Christensen







Agenda

- Analysis in practice in the ANOVA and ANCOVA situation
- Sequential hypothesis testing in practise
- Revisiting influence diagnostics
- Introduction to the Multivariate General Linear Model
  - Hotellings T^2

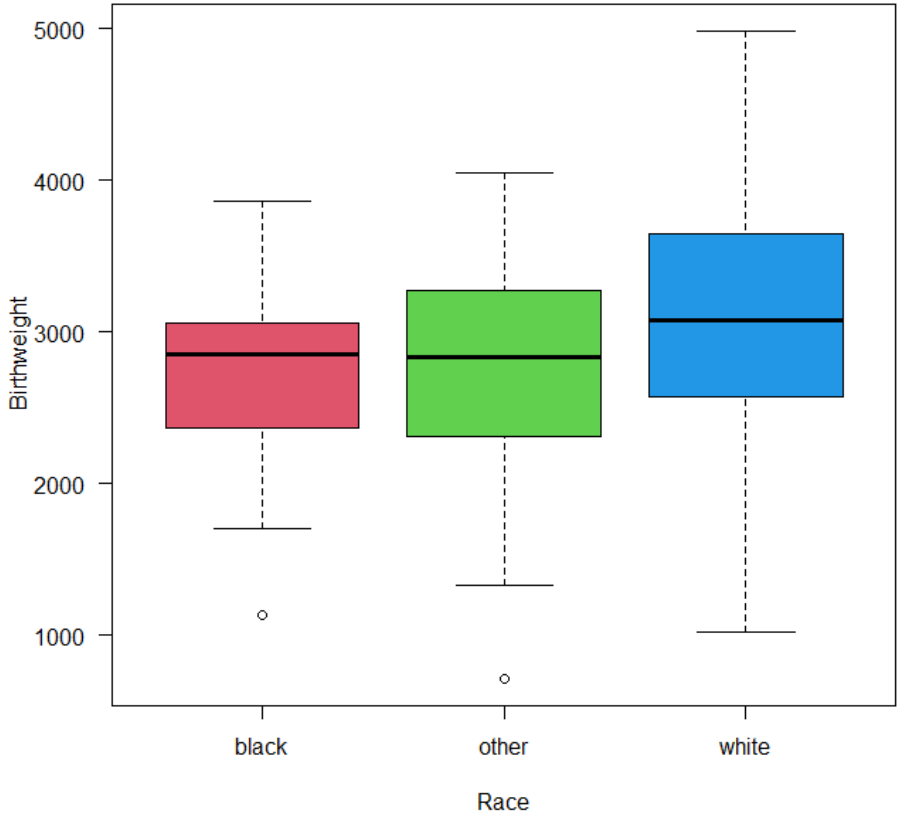
Example: Low Birth Weight

Data for 189 child births. `bwt` is the birth weight in grams. Contains additional information, eg. racial group `race`:

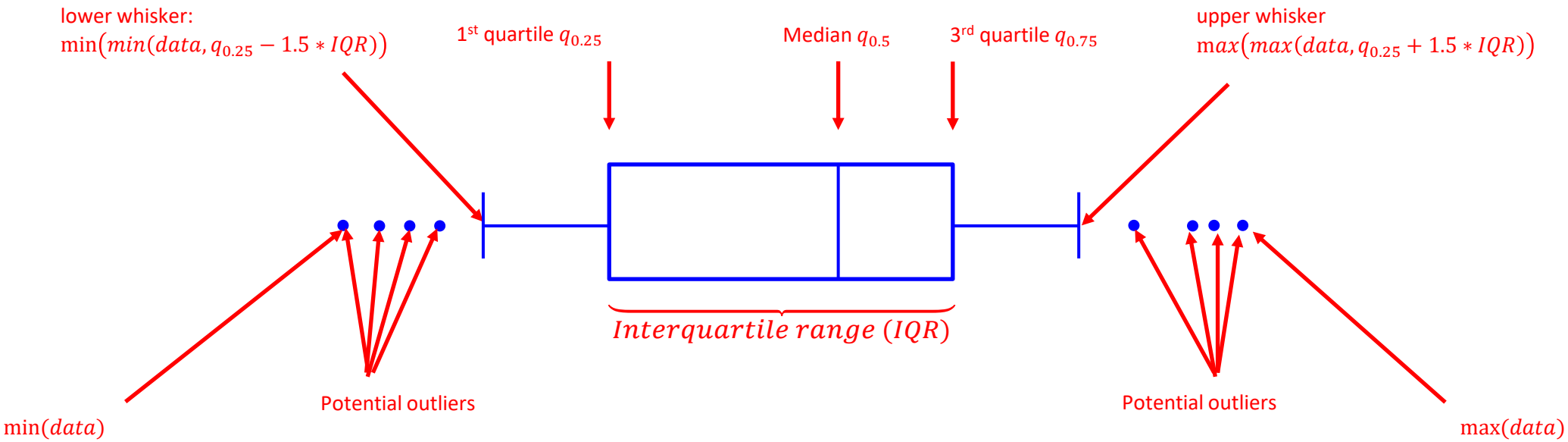
```
> lbw <- read.delim("Data/lbw.txt")
> by(lbw$bwt, lbw$race, summary)
lbw$race: black
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1135    2370    2849    2720    3057    3860
------------------------------------------------------------
lbw$race: other
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    709    2313    2835    2804    3274    4054
------------------------------------------------------------
lbw$race: white
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1021    2585    3076    3103    3651    4990
```

Example: Low Birth Weight

```
boxplot(bwt ~ race, data = lbw, xlab = 'Race', ylab = 'Birthweight',
las = 1, col = 2:4)
```



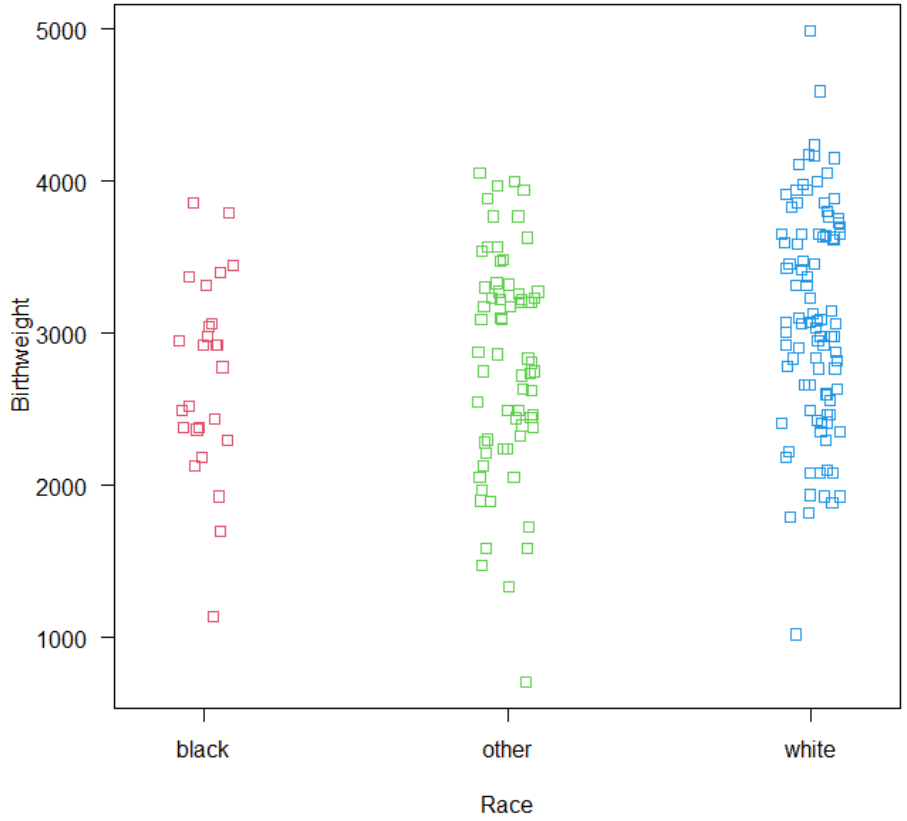
The Box Plot



Displays locality, spread and skewness of data

Example: Low Birth Weight

```
stripchart(bwt ~ race, data = lbw, vertical = TRUE, xlab = "Race",
ylab = "Birthweight", method = "jitter", las = 1, col = 2:4)
```



Example: Low Birth Weight

- **Statistical model:**

Let Y_{gi} be the birthweight of observation i within racial group g:

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

- One-way **AN**alysis **O**f **VA**riance (ANOVA).

- Only one hypothesis to test in one-way ANOVA:

$$H_0: \mu_1 = \cdots = \mu_k$$

Test sequence:

H
$\downarrow$
H_0

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \dots, n_g, \qquad g = 1, \dots, k.$$

- Factor dummy-coding:

$$X_{black,i} = \begin{cases} 1 & if\ race_i = "black" \\ 0 & if\ race_i \neq "black" \end{cases}; \quad X_{other,i} = \begin{cases} 1 & if\ race_i = "other" \\ 0 & if\ race_i \neq "other" \end{cases}; \quad X_{white,i} = \begin{cases} 1 & if\ race_i = "white" \\ 0 & if\ race_i \neq "white" \end{cases}$$

- $X = (X_{black}; X_{other}; X_{white}), \theta = \begin{pmatrix} \mu_{black} \\ \mu_{other} \\ \mu_{white} \end{pmatrix}$:

$$Y_i = \mu_{black} X_{black,i} + \mu_{other} X_{other,i} + \mu_{white} X_{white,i} + \varepsilon_i, \qquad i = 1, \dots 189$$

$$Y = X\theta + \varepsilon$$

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim N(0, \sigma^2), \quad i = 1, \ldots, n_g, \quad g = 1, \ldots, k.$$

- Note that $X_{black} + X_{other} + X_{white} = \mathbf{1}$.
- Alternative dummy-coding: $Z_{other} = X_{other} - X_{black}; Z_{white} = X_{white} - X_{black};$
- $\tilde{X} = (\mathbf{1}; Z_{other}; Z_{white})$ spans the same subspace $M \subset \mathbb{R}^{189}$ as X.

For testing, use $\tilde{X}$ rather than X.

- $Y_i = \alpha \mathbf{1}_i + \delta_{other} Z_{other,i} + \delta_{white} Z_{white,i} + \varepsilon_i, \ i = 1, \ldots 189$

$$Y = \tilde{X}\delta + \varepsilon$$

H_0 translates into

$$H_0' : \delta_{other} = \delta_{white} = 0$$

$$M_0 = \vee(\mathbf{1}) \subset M = \vee(X) = \vee(\tilde{X})$$

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

```
> model1<-lm(bwt ~ race, data = lbw)
> anova(model1)
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value   Pr(>F)
race        2  5048361 2524181   4.949 0.008052 **
Residuals 186 94866938  510037
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Data supports association of birth weight with racial group (p=0.008).

Comparisons of nested models with the anova function

```
> model2<-lm(bwt ~ 1, data = lbw)
> anova(model2,model1)
Analysis of Variance Table

Model 1: bwt ~ 1
Model 2: bwt ~ race
  Res.Df      RSS Df Sum of Sq     F   Pr(>F)
1    188 99915299
2    186 94866938  2   5048361 4.949 0.008052 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Compare to the previous slide. Note that `anova` labels the simplest model (1st argument) ‘model 1’.

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim N(0, \sigma^2), \quad i = 1, \ldots, n_g, \quad g = 1, \ldots, k.$$

```
> summary(model1)

Call:
lm(formula = bwt ~ race, data = lbw)

Residuals:
     Min       1Q   Median       3Q      Max
-2095.01  -503.01   -13.01   526.99  1886.99

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2719.69     140.06  19.418   <2e-16 ***
raceother      84.32     165.01   0.511   0.6100
racewhite     383.32     157.89   2.428   0.0161 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 714.2 on 186 degrees of freedom
Multiple R-squared:  0.05053,   Adjusted R-squared:  0.04032
F-statistic: 4.949 on 2 and 186 DF,  p-value: 0.008052
```

Individual p-values for δ_{other}, δ_{white}

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

Confidence intervals:

```
> tab <- cbind(coef(summary(model1))[ , 1:2], "Lower" = confint(model1)[ , 1],
+              "Upper" = confint(model1)[ , 2])
> data.frame(round(tab, 2),
+            "p-value" = format.pval(coef(summary(model1))[ , 4],
+            digits = 3, eps = 1e-3))
            Estimate Std..Error   Lower   Upper p.value
(Intercept)  2719.69     140.06 2443.38 2996.00  <0.001
raceother      84.32     165.01 -241.22  409.86  0.6100
racewhite     383.32     157.89   71.83  694.81  0.0161
```

- Good for testing, less good for illustrating group means.

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim N(0, \sigma^2), \quad i = 1, \ldots, n_g, \quad g = 1, \ldots, k.$$

Confidence intervals for group means:

For confidence intervals, use X rather than $\tilde{X}$.

```
> model1a<-lm(bwt ~ race-1, data = lbw)
>
> tab <- cbind(coef(summary(model1a))[ , 1:2], "Lower" = confint(model1a)[ , 1],
+              "Upper" = confint(model1a)[ , 2])
> data.frame(round(tab, 2))
           Estimate Std..Error   Lower   Upper
raceblack   2719.69     140.06 2443.38 2996.00
raceother   2804.01      87.25 2631.89 2976.14
racewhite   3103.01      72.89 2959.21 3246.81
```

- No p-values; they are not relevant here, as we are testing that group means are 0.
- Relations:

$$\begin{pmatrix} \alpha \\ \delta_o \\ \delta_w \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ -3 & 3 & 0 \\ -3 & 0 & 3 \end{pmatrix}\begin{pmatrix} \mu_b \\ \mu_o \\ \mu_w \end{pmatrix}; \qquad \begin{pmatrix} \mu_b \\ \mu_o \\ \mu_w \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 3 & -1 & 1 \\ 3 & 2 & -1 \\ 3 & -1 & 2 \end{pmatrix}\begin{pmatrix} \alpha \\ \delta_o \\ \delta_w \end{pmatrix}$$

- One could have retained the coding and used the formula $V(A\hat{\delta}) = AV(\hat{\delta})A^T$ instead with A as above.

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

- Model assumptions:
  - Normality of residuals;
  - Variance homogenity; same σ^2 in each group;
  - Independence;
  - Same mean within each group.

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

Indepence:

- Context. Here:
  - No siblings among children (same mothers);
  - No siblings among mothers (relqated mothers);

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

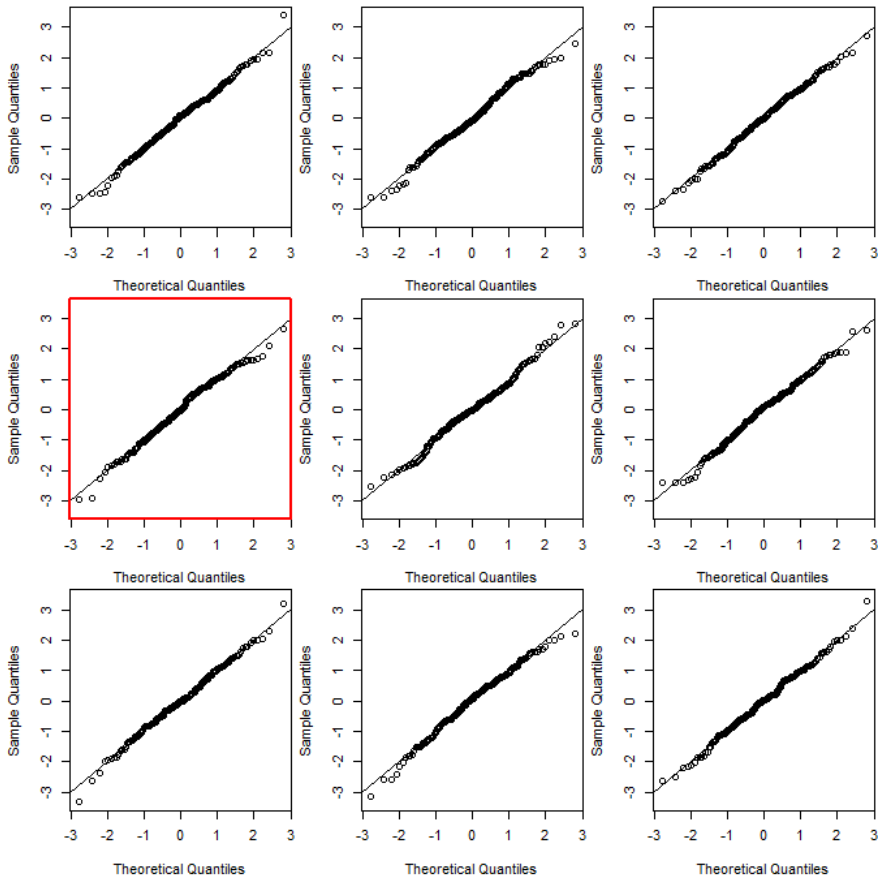
Normality: QQ-plot.

How does one recognize a good QQplot with the EBT?

```
qqwrap <- function(x, y, ...) {qqnorm(y,main="",...)
                                abline(a=0, b=1)}
wallyplot(model1, FUN=qqwrap)
```

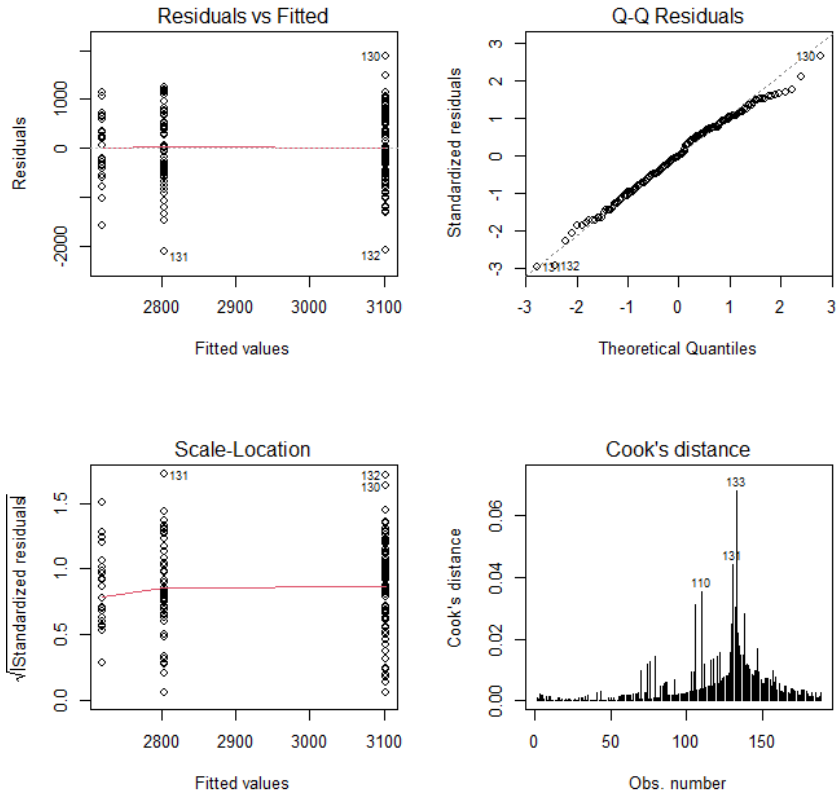
Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$



Example: Low Birth Weight

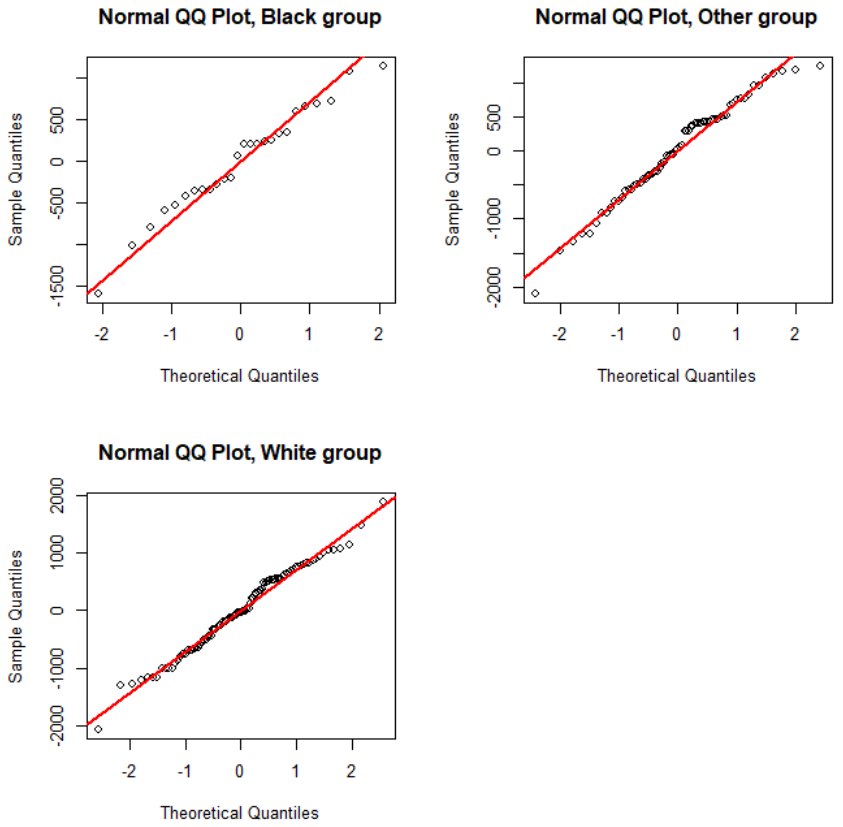
$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$



Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

- Same mean within each group:



Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \dots, n_g, \qquad g = 1, \dots, k.$$

Variance homogeneity:

- Testable with a numeric test!

$$Q = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{L(\hat{\mu}, \hat{\sigma}_1^2, \dots, \hat{\sigma}_k^2)} = \frac{(2\pi)^{-N/2} \hat{\sigma}^{2^{-N/2}} exp\left(-\frac{1}{2\hat{\sigma}^2}(N-k)\hat{\sigma}^2\right)}{\prod_{i=1}^{k} (2\pi)^{-n_i/2} \hat{\sigma}_i^{2^{-n_i/2}} exp\left(-\frac{1}{2\hat{\sigma}_i^2}(n_i - 1)\hat{\sigma}_i^2\right)} = \frac{\prod_{i=1}^{k} \hat{\sigma}_i^{2^{n_i/2}}}{\hat{\sigma}^{2^{N/2}}},$$

$$-2logQ = Nlog(\hat{\sigma}^2) - \sum_{k}^{k} n_i log\left(\hat{\sigma}_i^2\right).$$

- Bartlett's test:

$$\chi^2 = \frac{(N-k)log(\hat{\sigma}^2) - \sum_{i=1}^{k}(n_i - 1)log\left(\hat{\sigma}_i^2\right)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N-k}\right)} \underset{as}{\sim} \chi^2(k-1)$$

- Large values of $n_1, \dots, n_k$: $\chi^2 \approx -2logQ$.

In a multivariate setting: Theorem 4.32, apply it with $p = 1$.

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

- Bartlett's test in **R**:

```
> bartlett.test(bwt~race,data=lbw)

        Bartlett test of homogeneity of variances

data:  bwt by race
Bartlett's K-squared = 0.65601, df = 2, p-value = 0.7204
```

- Data supports equal variances (p=0.72)

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

- Model assumptions fulfilled
- Significantly different means in the groups
- But we only know that the three means are significantly different, not how they differ.

Post hoc analysis: Pairwise comparisons

A multiple comparisons element: In general $\mathrm{m} = k(k-1)/2$ comparisons. Probability that one of them is significant by chance under $\mathrm{H_o} \approx 1 - (1-\alpha)^m$. For $k = 3$ we get 0.14.

One solution: lower the test level from α to α/m; Then $1 - \left(1 - \frac{\alpha}{m}\right)^m \approx 1 - \left(1 - m \cdot \frac{\alpha}{m}\right) = \alpha$.

Similar: Multiply the p-values by m, and use the original test level:

Bonferroni correction

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

Post hoc analysis: Pairwise comparisons

```
> pairwise.t.test(lbw$bwt, lbw$race, p.adj = "none")

        Pairwise comparisons using t tests with pooled SD

data:  lbw$bwt and lbw$race

      black  other
other 0.6100 -
white 0.0161 0.0093
```

Example: Low Birth Weight

$$Y_{gi} = \mu_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n_g, \qquad g = 1, \ldots, k.$$

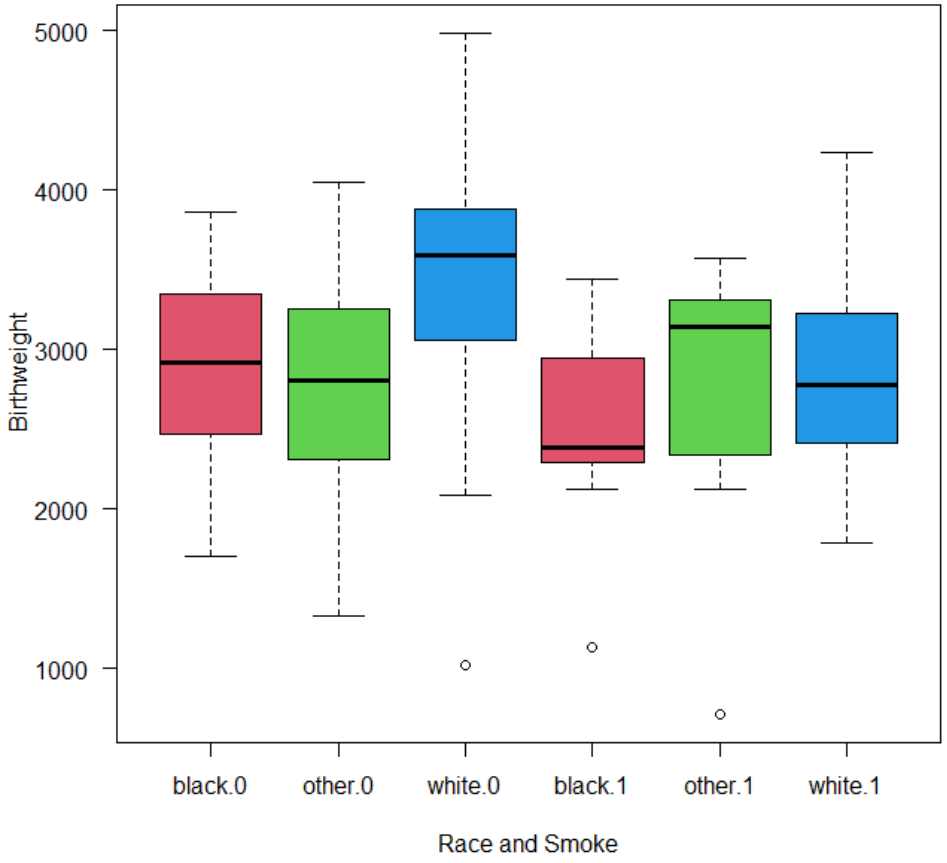
Post hoc analysis: Pairwise comparisons

```
> pairwise.t.test(lbw$bwt, lbw$race, p.adj = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  lbw$bwt and lbw$race

      black other
other 1.000 -
white 0.048 0.028

P value adjustment method: bonferroni
```

Example: Low Birth Weight

Let us now, on top of `race` (3 levels), introduce the factor `smoking` (2 levels), indicating whether the mother as been smoking during prenancy.

```
boxplot(bwt ~ race*smoke, data=lbw,
        xlab = 'Race and Smoke',
        ylab = 'Birthweight',
        las = 1, col = 2:4)
```



Example: Low Birth Weight

Model:

$$Y_{rsi} = \alpha_{rs} + \varepsilon_{rsi}, i = 1, \dots, n_{rs}$$

Two-sided analysis of variance.

Dimension of corresponding subspace: 6 (6 =3x2 groups).

Formulation through dummy coding appropriate for testing:

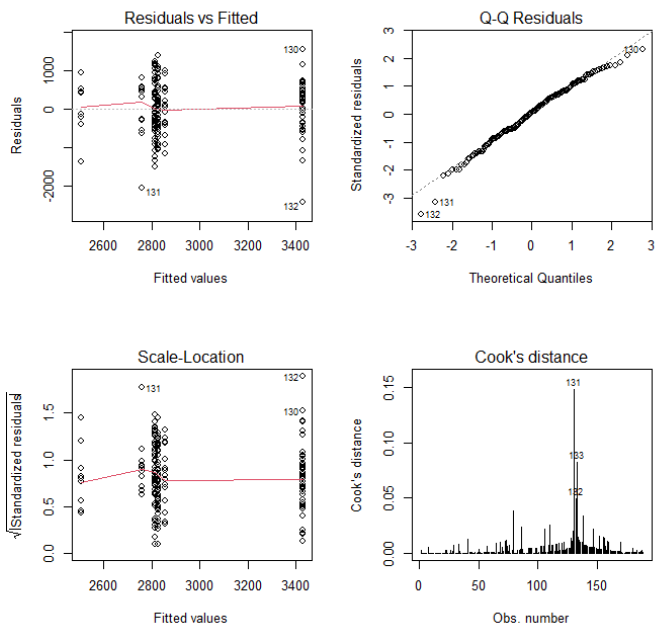
$$Y_{rsi} = \alpha + \beta_r + \gamma_s + \delta_{rs} + \varepsilon_{rsi}, i = 1, \dots, n_{rs}$$

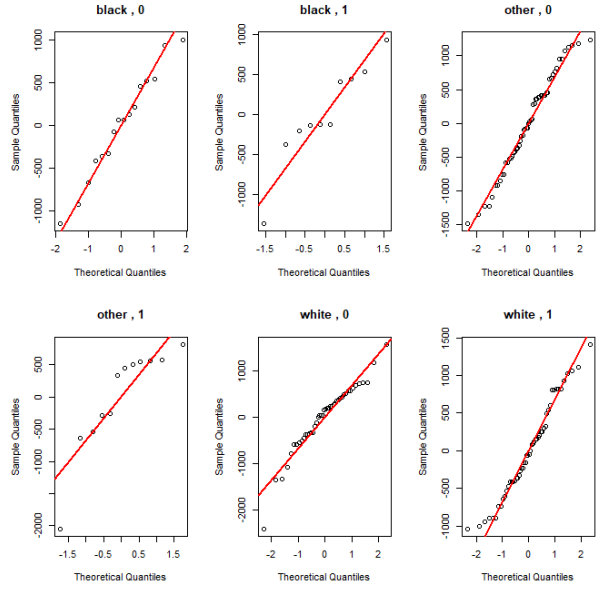
```
> lbw$smoke <- as.factor(lbw$smoke)
> model1 <- lm(bwt ~ race + smoke + race:smoke, data = lbw)
```

Example: Low Birth Weight

Model control:

```
> lbw2<-lbw; lbw2$group<-paste(lbw$race,",",lbw$smoke)
> bartlett.test(bwt ~ group, data=lbw2)

        Bartlett test of homogeneity of variances

data:  bwt by group
Bartlett's K-squared = 2.0061, df = 5, p-value = 0.8483
```





Formula objects in R

- Used as argument for analytic routines (eg. `lm`), plotting functions and more.
- Of the form $x \sim y$; corresponds to

$$x = \begin{cases} \alpha + \beta y + \varepsilon & if\ y\ is\ numeric \\ \alpha_y & if\ y\ is\ a\ factor \end{cases}$$

- Dummy-coding of factors equates this.

Formula objects in R

- Operators and relations in formulas: $\sim$, $+$, $:$, $*$, $-$, ^, I.

- "$\sim$" separates response from explanatory variables
- "$+$" indicates additive effects; $bwt \sim race + smoke$
- " $:$" indicates interaction effects; the *crossing* of factors:

$$x{:}\,y = \begin{cases} a\ factor\ with\ a\ level\ for\ each\ combination\ of\ x\ and\ y & if\ x, y\ factors \\ a\ numeric\ factor\ for\ each\ level\ of\ x\ regressing\ data\ with\ that\ level\ on\ y & if\ x\ factor, y\ numeric \\ the\ numeric\ variable\ xy & if\ x, y\ numeric \end{cases}$$

- "$*$" indicates interaction and main effects; ie. $x * y = x + y + x{:}\,y$

- " ^" indicates adding main effects and lower order interaction terms:

$$(x + y + z)^2 = x + x + z + x{:}\,y + y{:}\,z, \qquad (x + y + z)^3 = x + x + z + x{:}\,y + y{:}\,z + x{:}\,y{:}\,z$$

- "$-$" indicates negative selection;$(x + y + z)^2 - y{:}\,z = x + x + z + x{:}\,y$, $y \sim x - 1$ removes constant term from the model.

- " I" overrides the formula calculus, and allows you to do calculations within the formula: $y \sim x + I(x\text{^}2)$.

Example: Low Birth Weight

- 3 equivalent model formulations:

```
model1 <- lm(bwt ~ race + smoke + race:smoke, data = lbw)
model1 <- lm(bwt ~ race*smoke, data = lbw)
model1 <- lm(bwt ~ (race+smoke)^2, data = lbw)
```

```
> summary(model1)

Call:
lm(formula = bwt ~ race + smoke + race:smoke, data = lbw)

Residuals:
     Min       1Q   Median       3Q      Max
-2407.75  -417.38    31.25   462.50  1561.25

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2854.50     170.84  16.708  < 2e-16 ***
raceother          -40.26     194.11  -0.207  0.83591
racewhite          574.25     199.50   2.878  0.00447 **
smoke1            -350.50     275.47  -1.272  0.20486
raceother:smoke1   293.43     351.13   0.836  0.40443
racewhite:smoke1  -250.87     309.00  -0.812  0.41792
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$\dim(M) = 6$: 6 parameters

Example: Low Birth Weight

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2854.50     170.84  16.708  < 2e-16 ***
raceother          -40.26     194.11  -0.207  0.83591
racewhite          574.25     199.50   2.878  0.00447 **
smoke1            -350.50     275.47  -1.272  0.20486
raceother:smoke1   293.43     351.13   0.836  0.40443
racewhite:smoke1  -250.87     309.00  -0.812  0.41792
```

Estimated birth weight in the groups:

	Race		
Smoke	Black	Other	White
0	2850.50	2850.50 -40.26 = 2814.24	2850.50 +574.25 =3428.75
1	2850.50 -350.50 =2504.00	2850 -40.26 +293.43 =2757.17	2850 +574.25 -250.87 =2827.38

Example: Low Birth Weight

- Different dummy coding for description:

```
> model1a <- lm(bwt ~ race:smoke-1, data = lbw)
> summary(model1a)

Call:
lm(formula = bwt ~ race:smoke - 1, data = lbw)

Residuals:
     Min       1Q   Median       3Q      Max
-2407.75  -417.38    31.25   462.50  1561.25

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
raceblack:smoke0  2854.50     170.84   16.71   <2e-16 ***
raceother:smoke0  2814.24      92.15   30.54   <2e-16 ***
racewhite:smoke0  3428.75     103.02   33.28   <2e-16 ***
raceblack:smoke1  2504.00     216.10   11.59   <2e-16 ***
raceother:smoke1  2757.17     197.27   13.98   <2e-16 ***
racewhite:smoke1  2827.38      94.77   29.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: Low Birth Weight

- Model 1:

$$M_1: Y_{rsi} = \alpha + \beta_r + \gamma_s + \delta_{rs} + \varepsilon_{rsi}, i = 1, \dots, n_{rs}$$

- We want to compare to the simpler additive model :

$$M_2: Y_{rsi} = \alpha + \beta_r + \gamma_s + \varepsilon_{rsi}, i = 1, \dots, n_{rs}$$

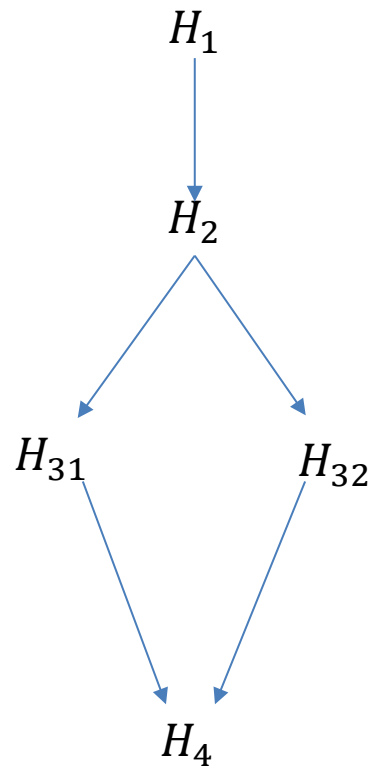
- Ie. we wish to test if the two interaction parameter δ_{rs} can be set to 0.

Example: Low Birth Weight

- **Hypotheses in two-way ANOVA:**

H_2: Additive hypothesis ($\delta_{rs} = 0$) :

$$M_2: \mu_{rsi} = \alpha + \beta_r + \gamma_s$$



H_{31}: No effect of smoking ($\gamma_s = 0$):

$$M_{31}: \mu_{rsi} = \alpha + \beta_r$$

H_{32}: No effect of race ($\beta_r = 0$):

$$M_{32}: \mu_{rsi} = \alpha + \gamma_s$$

H_4: No effect of neither smoking nor race ($\beta_r = 0$ or $\gamma_s = 0$):

$$M_4: \mu_{rsi} = \alpha$$

Sequential Testing

- **Backwards Selection.**

Start with a saturated model, evaluate it and simplify it, moving downwards in the test graph.

- (**Forwards selection.**

  Start with the simplest model, and complicate it, moving upwards in the test tree.)

- Problem with forward selection: Simpler but **WRONG**. Until the test for the final model, ALL models in tests are wrong, and test values more or less arbitrary!

- **Always use backwards selection**. In complicated test graphs, use forward selection from the final model for completion (sensitivity analysis), **after** backwards selection.

Example: Low Birth Weight

```
> drop1(model1,test="F")
Single term deletions

Model:
bwt ~ race + smoke + race:smoke
           Df Sum of Sq      RSS    AIC F value Pr(>F)
<none>                   85459758 2473.1
race:smoke  2   2108643 87568401 2473.7  2.2577 0.1075
```

We accept H_2 (p=0.11).

```
> model2<-update(model1,~.-race:smoke)
> drop1(model2,test="F")
Single term deletions

Model:
bwt ~ race + smoke
       Df Sum of Sq      RSS    AIC F value    Pr(>F)
<none>              87568401 2473.7
race    2   8749453 96317854 2487.7  9.2422 0.0001494 ***
smoke   1   7298537 94866938 2486.9 15.4191 0.0001214 ***
```

No terms are significant at the 5% test level, and the analysis stops here.

Example: Low Birth Weight

Summary of analysis:

```
> data.frame(round(tab, 2),
+            "p-value" = format.pval(coef(summary(model2))[ , 4],
+            digits = 3, eps = 1e-3))
            Estimate Std..Error   Lower   Upper p.value
(Intercept)  2884.32     141.29 2517.43 3163.07 < 0.001
raceother      -3.64     160.54 -423.24  313.08 0.98193
racewhite     450.54     153.07  180.63  752.52 0.00366
smoke1       -428.03     109.00 -894.02 -212.98 < 0.001
```

- For two mothers of the **same race** where one is a smoker and the other nonsmoker, the birth weight of the smoker's baby will on average be 428g less than for the nonsmoker. The difference could be as much as 894g or as little as 213g, based on a 95% CI.

- For two mothers with the **same smoking status** where one belongs the black group and the other belongs to the white group, the birth weight of the baby for the mother that belongs to the black group will on average be 451g less than the baby if the mother belonging to the white group. The difference could be as much as 753g or as little as 181g, based on a 95% CI.

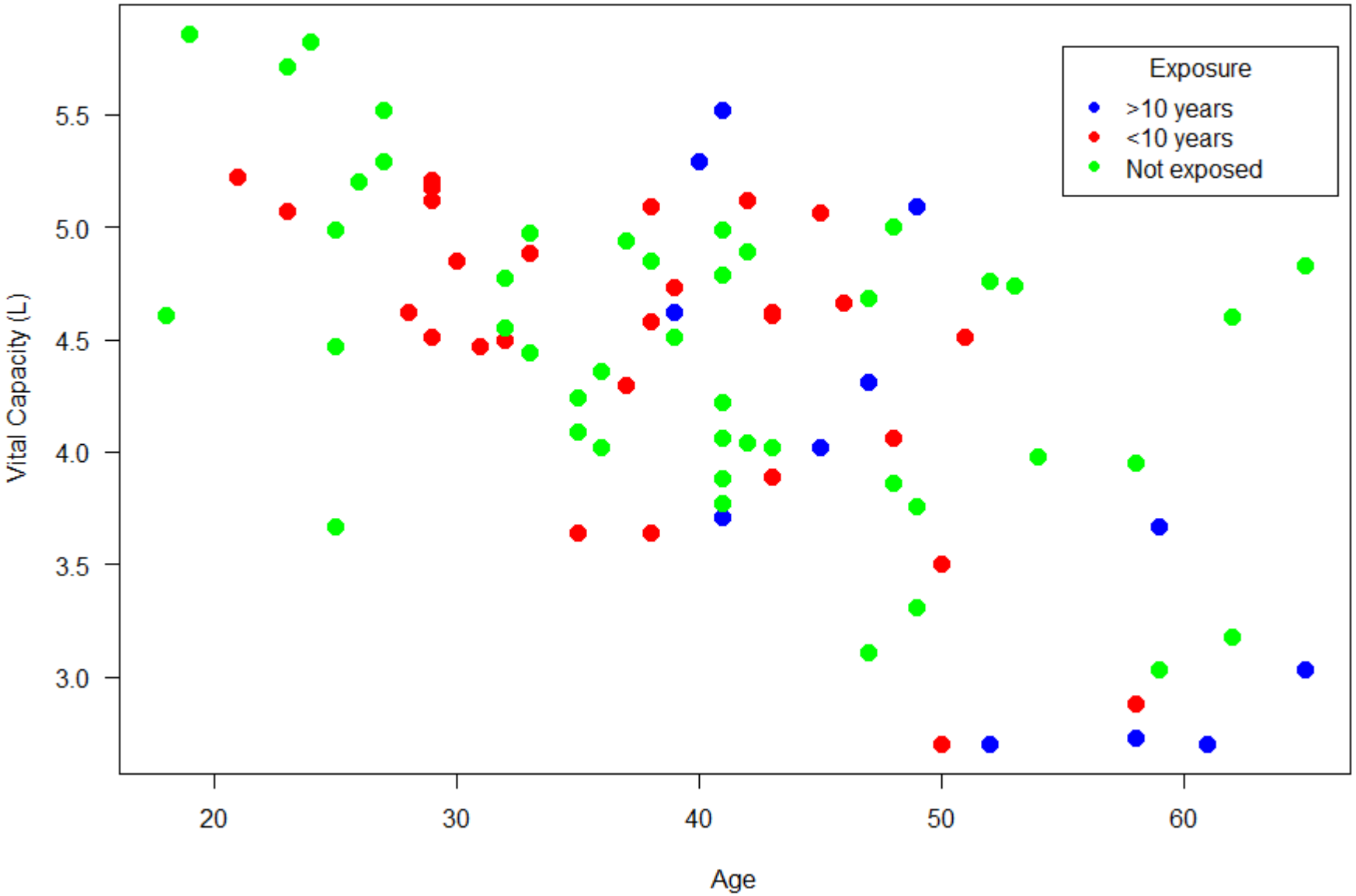
Example: Vital Capacity and Cadmium

- Data from a study of the effect of exposure to cadmium on the vital capacity. *Armitage & G Berry: Statistical methods in medical research. 2nd ed. Blackwell 1987.*

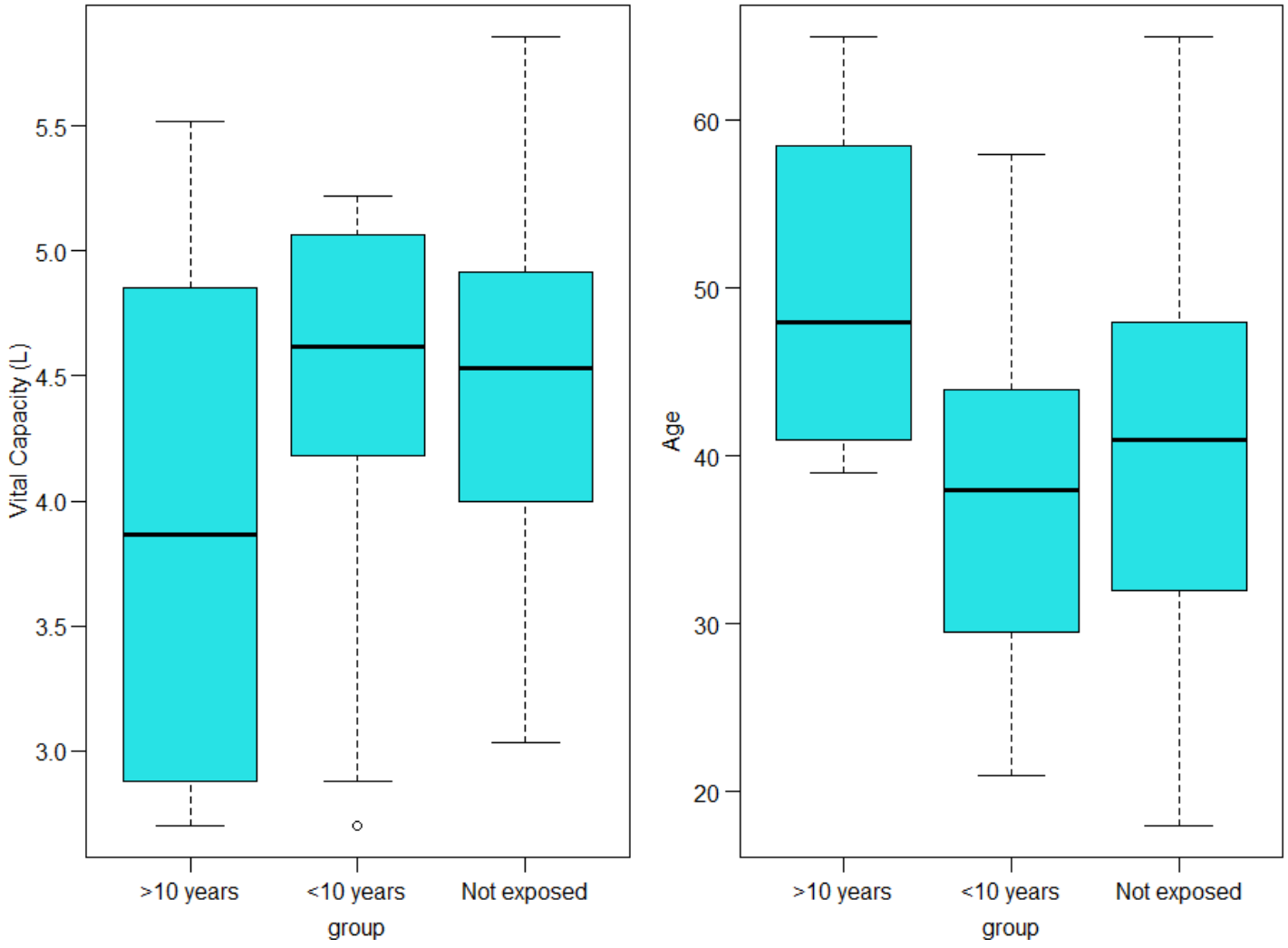
- *Vital capacity* is the maximum amount of air a person can expel from the lungs after a maximum inhalation.

- Measurements of vital capacity (L), age and exposure to cadmium (> 10 years, < 10 years, not exposed).

Example: Vital Capacity and Cadmium

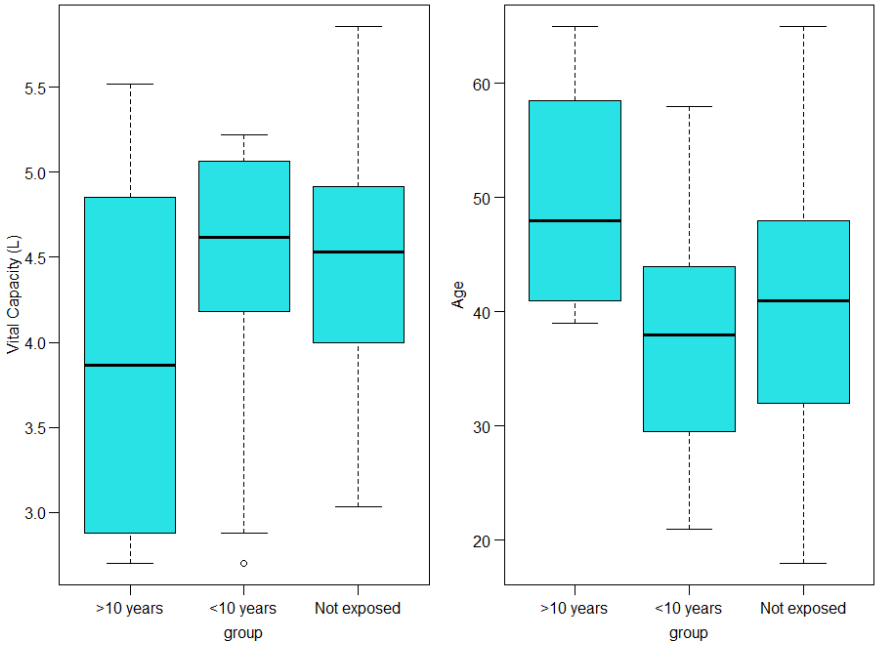


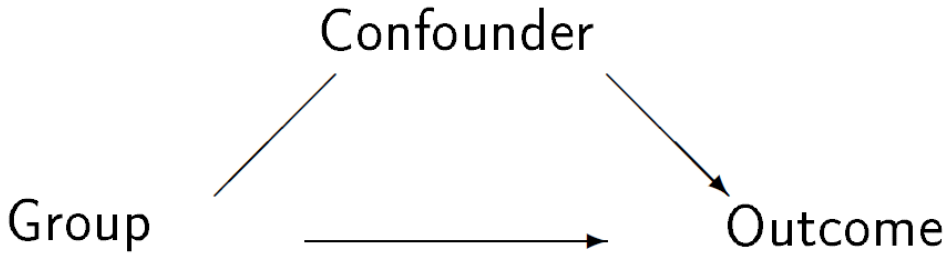
Example: Vital Capacity and Cadmium



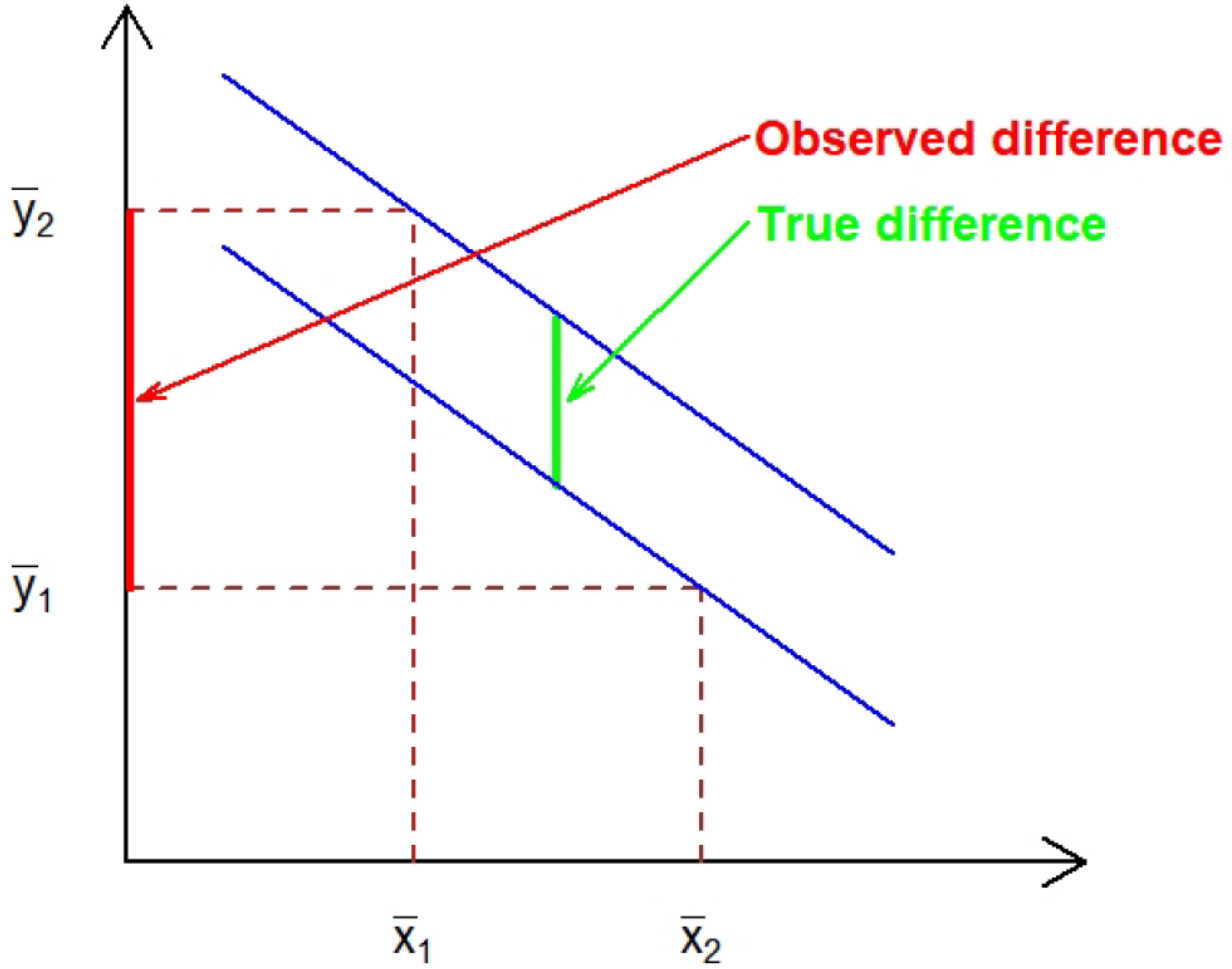
Confounding

- Comparing groups that are not quite comparable (e.g. cadmium exposure).
- Confounder: A variable that
  - Has an effect on the outcome.
  - Is associated to group (different ages in groups)
- This can cause **bias**.





Confounding



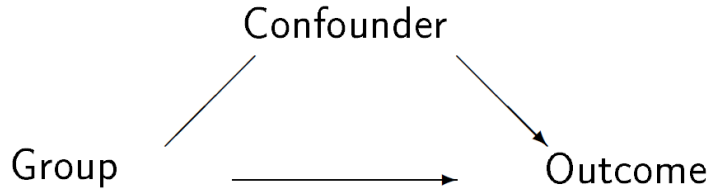
Confounding

- Handling confounders:

  Include the skew covariate in the model.
- We will answer a different question:

  Instead of “Is the vital capacity associated with exposure to cadmium?”, we will investigate

  “Is the vital capacity, corrected for age, associated with exposure to cadmium?”



Example: Vital Capacity and Cadmium

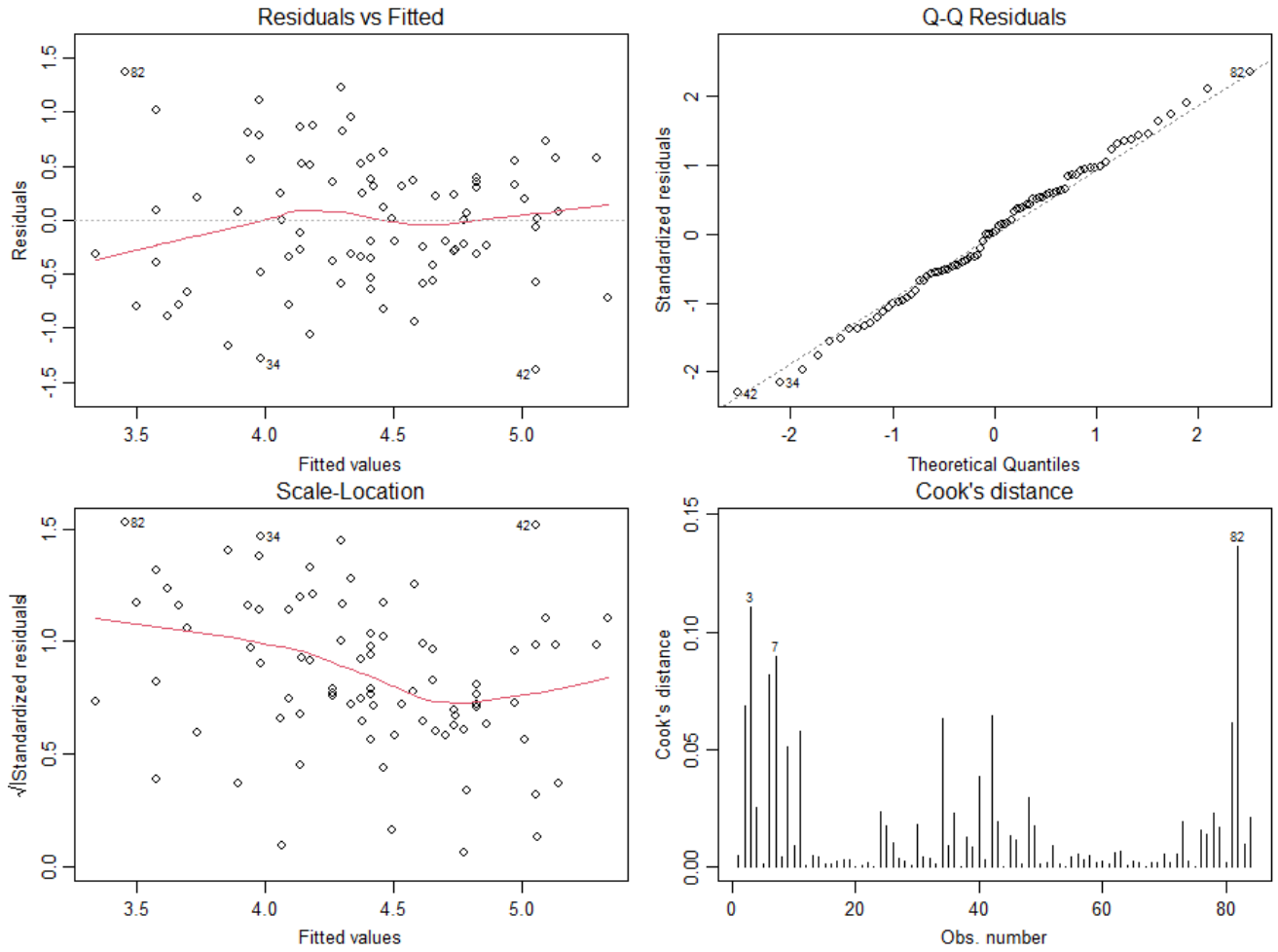
- Model for Vital capacity, dummy-coding the exposure factor:

$$Y_i = \beta_0 + \beta_{<10} X_{<10,i} + \beta_{\geq 10} X_{\geq 10,i} + \beta_{age} X_{age,i} + \varepsilon_i, \qquad \mathrm{i} = 1, \ldots, 84$$

```
CADdata$expo[CADdata$group==3] <- 1
CADdata$expo[CADdata$group==2] <- 2
CADdata$expo[CADdata$group==1] <- 3

CADdata$expo<-as.factor(CADdata$expo)

Model1<-lm(vitcap ~ expo + age, data = CADdata)
```

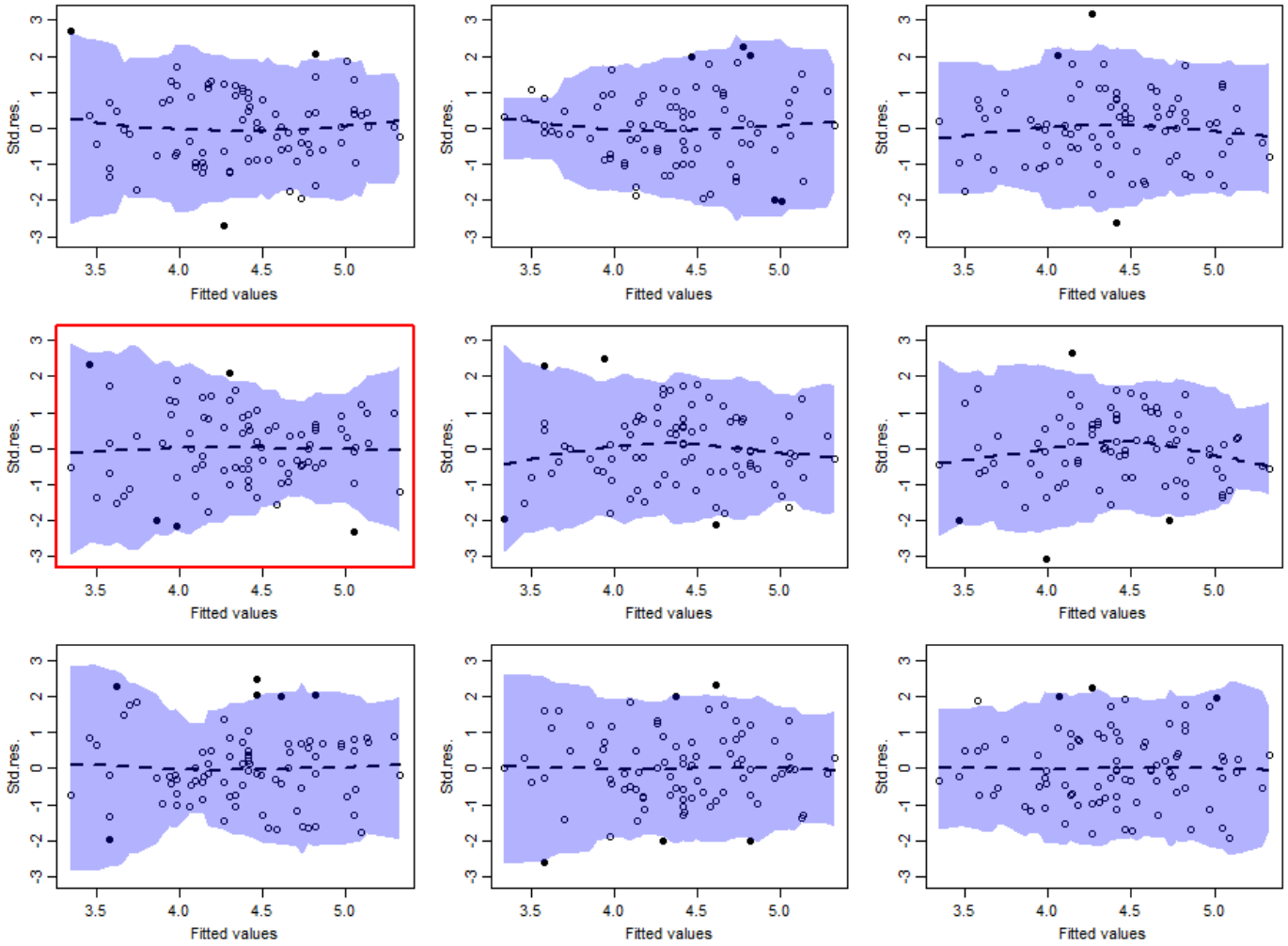
Example: Vital Capacity and Cadmium

- Model control:



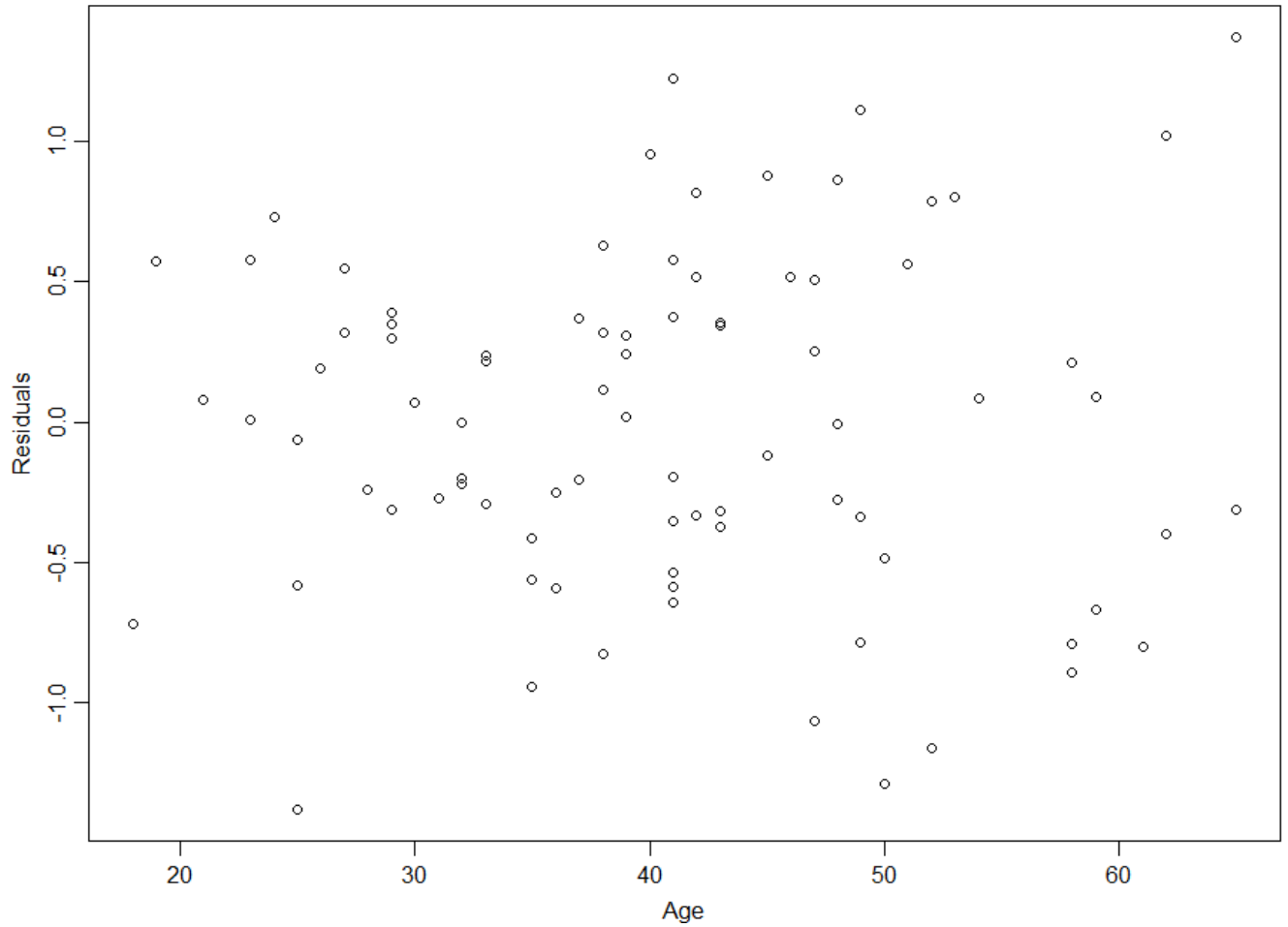
Example: Vital Capacity and Cadmium

```
wallyplot(model1)
```



Example: Vital Capacity and Cadmium

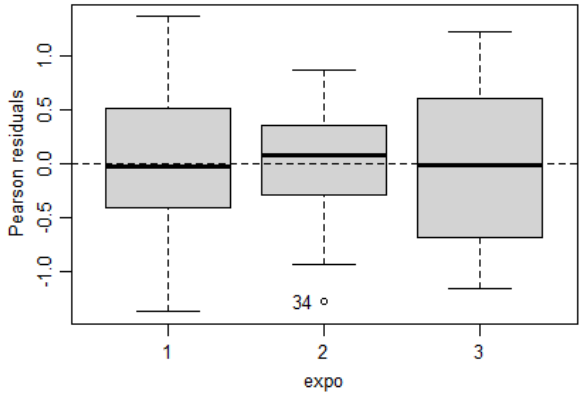
- Residuals vs. Age:

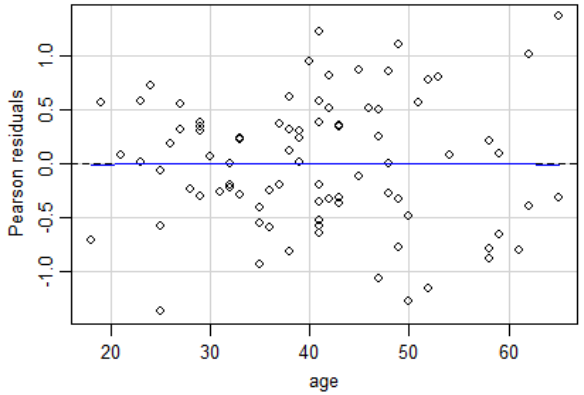


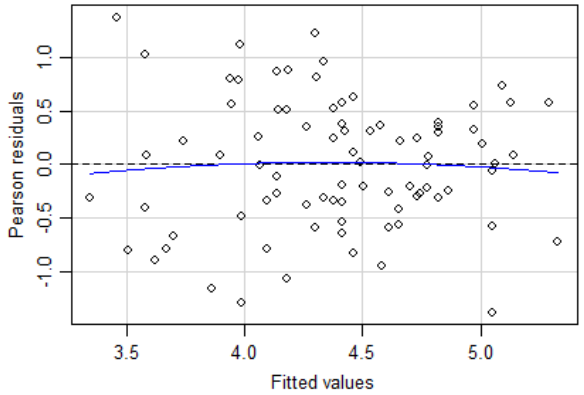
Example: Vital Capacity and Cadmium

- Further investigating curvature

```
> residualPlots(model1)
           Test stat Pr(>|Test stat|)
expo
age          -0.0535           0.9575
Tukey test   -0.3918           0.6952
```



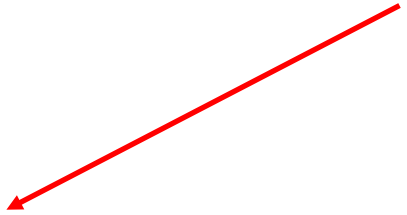




Example: Vital Capacity and Cadmium

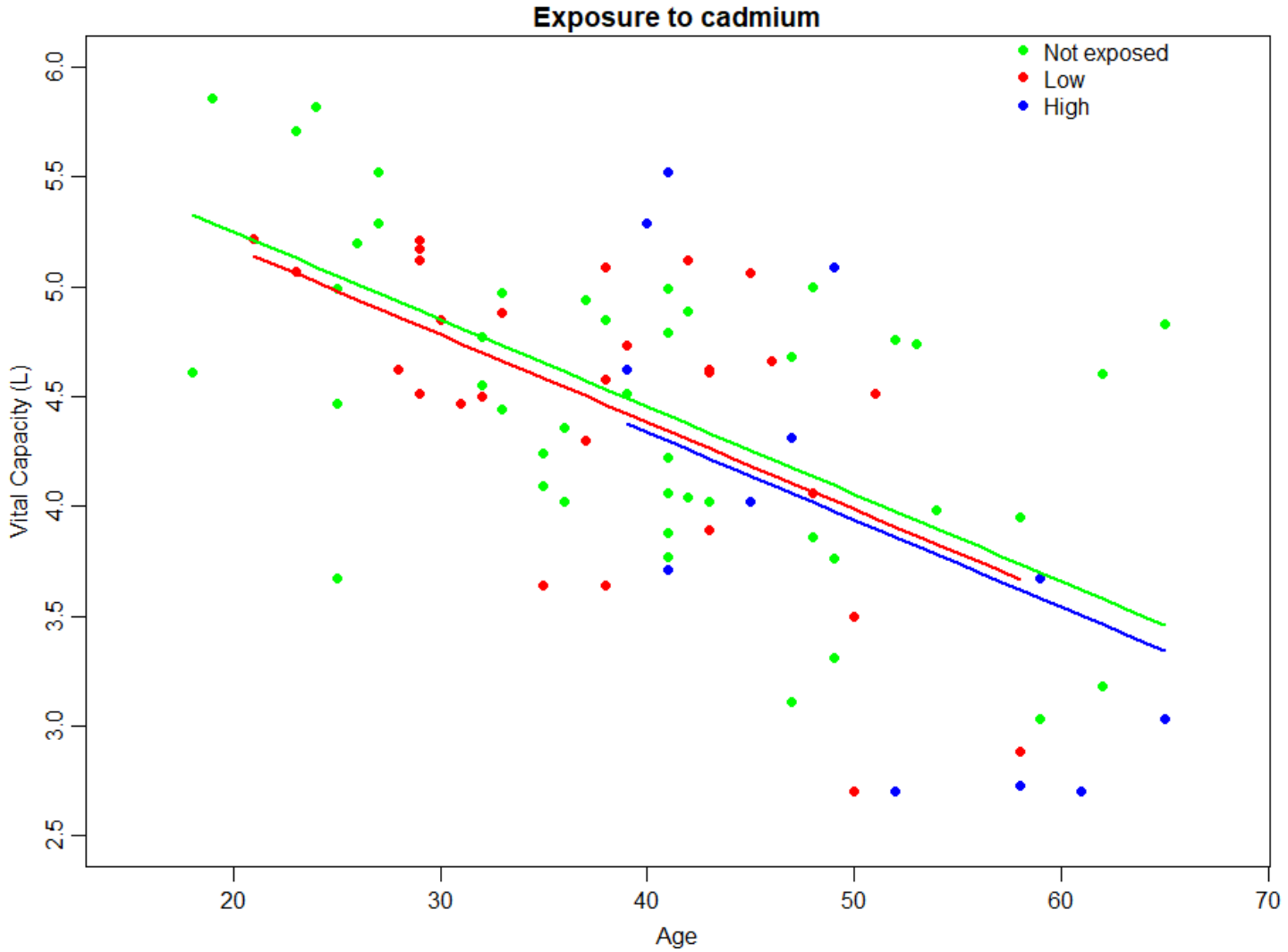
- Model control went well.

```
> summary(model1)

Call:
lm(formula = vitcap ~ expo + age, data = CADdata)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.044917   0.268025  22.554  < 2e-16 ***
expo2       -0.070198   0.148669  -0.472    0.638
expo3       -0.116935   0.209236  -0.559    0.578
age         -0.039775   0.006322  -6.291 1.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6127 on 80 degrees of freedom
Multiple R-squared:  0.3696,    Adjusted R-squared:  0.3459
F-statistic: 15.63 on 3 and 80 DF,  p-value: 4.323e-08

> drop1(model1,test="F")
Single term deletions

Model:
vitcap ~ expo + age
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              30.035 -78.391
expo    2    0.1617 30.196 -81.940  0.2153    0.8067
age     1   14.8589 44.894 -46.628 39.5781 1.572e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



When correcting for age, exposure to cadmium does NOT seem to be significant in model 1

Example: Vital Capacity and Cadmium



Example: Vital Capacity and Cadmium

- The vital capacity decreases with -0.04 L per year.
- Is it reasonable that the vital capacity decreases with the same rate in all three exposure groups?
- Allow different slopes in the three groups → Include an interaction between age and group.

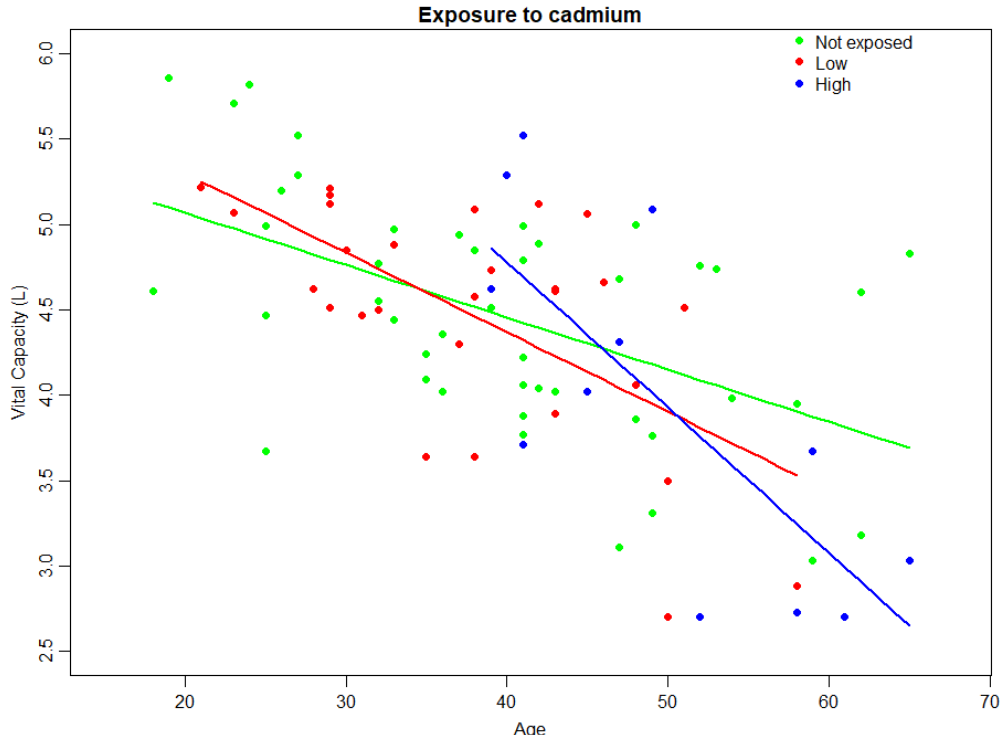
```
> model2 <- lm(vitcap ~ age + expo + age:expo, data = CADdata)
> summary(model2)

Call:
lm(formula = vitcap ~ age + expo + age:expo, data = CADdata)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.680291   0.313426  18.123  < 2e-16 ***
age         -0.030613   0.007547  -4.056 0.000117 ***
expo2        0.549740   0.575884   0.955 0.342728
expo3        2.503148   1.041842   2.403 0.018655 *
age:expo2   -0.015919   0.014547  -1.094 0.277170
age:expo3   -0.054498   0.021070  -2.587 0.011554 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> drop1(model2, test = "F")
Single term deletions

Model:
vitcap ~ age + expo + age:expo
         Df Sum of Sq    RSS     AIC F value  Pr(>F)
<none>                 27.535 -81.689
age:expo  2    2.4995 30.035 -78.391  3.5402 0.03376 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

When correcting for age, exposure to cadmium IS significant in model 2

Example: Vital Capacity and Cadmium

- Redoing dummy coding for easy plotting:

```
> model2B<-lm(vitcap ~ 0 + expo + age:expo,
              data = CADdata)
> my.coef <- coef(model2B)
> my.coef
      expo1       expo2       expo3   expo1:age   expo2:age   expo3:age
 5.68029060  6.23003098  8.18343838 -0.03061267 -0.04653201 -0.08511099
```



- The model control of model 1 did not reveal anything problematic.
- Yet, the conclusion was turned upside down by the addition of an interaction.
- You may not see the need for an interaction unless you investigate it.

Model Diagnostics Revisited

- With h_{ii} the *leverage* of observation i, the diagonal element of $P_M = X(X^TX)^{-1}X^T$:

$h_{ii} = X_i(X^TX)^{-1}X_i^T$;
$e_i = Y_i - \hat{Y}_i$,

a representation of Cooks Distance is

$$D_i = \frac{e_i^2}{p\hat{\sigma}^2}\frac{h_{ii}}{(1-h_{ii})^2}$$

Note that Cook's distance is high when both the leverage h_{ii} and the (squared) residual e_i^2 is high.

Model Diagnostics Revisited

- Assume no constant term, and that all explantory variables are mean-centered

$$h_{ii} = X_i(X^T X)^{-1} X_i^T = \|X_i\|^2_{(X^T X)^{-1}}$$

is a distance measure in the feature space wrt. The variation of the features. High leverage means far away from the average in the feature space.

with

$$D_i = \frac{e_i^2}{p\hat{\sigma}^2} \frac{h_{ii}}{(1-h_{ii})^2}$$

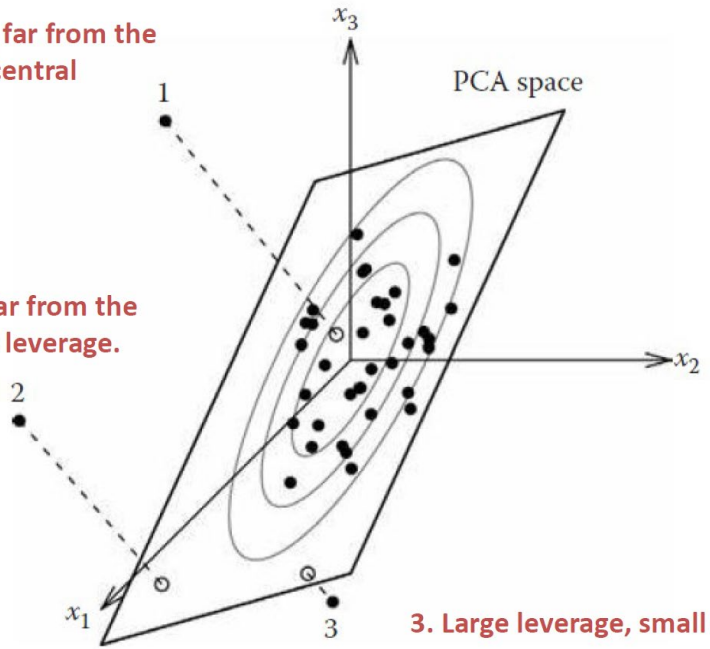
Cook's distance is high when both the leverage h_{ii} and the (squared) residual e_i^2 is high. Conforms with the criteria for problematic data in PCA analysis:

Model Diagnostics Revisited

Diagnostic Plots – Residuals and Leverage

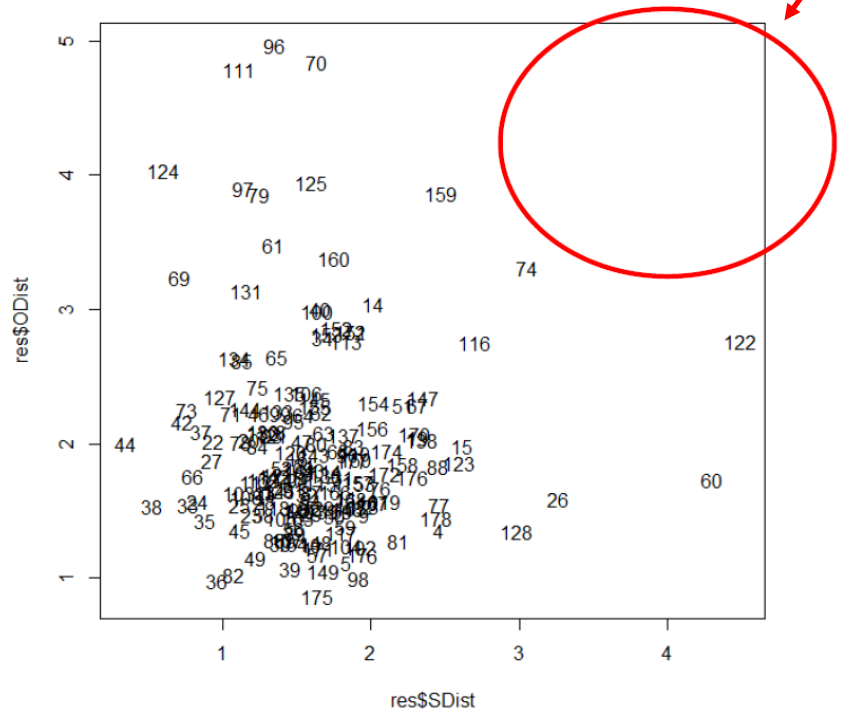
1. Large residual, far from the PCA space but central

2. Large residual, far from the PCA space, large leverage.



3. Large leverage, small residual

Points corresponding to large values of Cooks distance



Model Diagnostics Revisited

- DFFITS: Studentized DFFIT (Difference in Fit).

$$DFFIT_i = \hat{y}_i - \hat{y}(i)_i$$
$$DFFITS_i = \frac{\hat{y}_i - \hat{y}(i)_i}{\sigma_{(i)}\sqrt{h_{ii}}}$$

DFFITS is a measure similar to Cooks distance and may be converted to it.

- DFBETAS – Studentizised DFBETA.

$$DFBETAS_j = \frac{\hat{\theta}_j - \hat{\theta}(i)_j}{\sigma_{(i)}\sqrt{(X^T X)^{-1}_{jj}}}$$

For each observation i, the set of DFBETAS measures the effect on parameter estimation of removing that observations in standard errors.

Model Diagnostics Revisited

- **Multicolinearity:** Can seriously inflate the variance of parameter estimates due to $X^T X$ being close to singular.

Definition 3.15

We define the *tolerance (TOL)* and the *variance inflation (VIF)* as

$$\text{TOL}_i = 1 - R^2\,(x_i | \text{all other } x\text{-variables})$$
$$\text{VIF}_i = \frac{1}{\text{TOL}_i}$$

As a rule of thumb, TOL < 0.1 or equivalently VIF > 10 indicates a multi-collinearity problem.

Definition 3.16

We define the *condition number* as the square root of the largest eigenvalue of $\mathbf{x}\mathbf{x}^T$ divided by the smallest. The condition number should be below 15. If it is above 30, it is a matter of serious concern.

- Best method to handle it: Be careful when modeling.

INTRODUCTION TO MULTIVARIATE ANALYSIS OF VARIANCE

Hotelling's T-Square in the One-Sample Case

We consider independent random variables $X_1, \ldots, X_n$, with $X_i \sim N_p(\mu, \Sigma)$, $i = 1, \ldots, n$.

Parameter estimates:

$$\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X_i \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$$

$$\hat{\Sigma} = S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \sim W\left(n-1, \frac{1}{n-1}\Sigma\right)$$

Suppose that we whish to test the hypothesis

$$H_o: \mu = \mu_o \quad against \quad H_1: \mu \neq \mu_0$$

Hotelling's T-Square in the One-Sample Case

Theorem 4.2

We will use the notation

$$T^2 = n(\bar{X} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\bar{X} - \boldsymbol{\mu}_0),$$

where $\bar{X}$, $\boldsymbol{\mu}_0$ and $\mathbf{S}$ are as stated in the introduction to this section. Then the critical area for a ratio test of H_0 against H_1 at level α is

$$C = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \frac{n-p}{(n-1)p} t^2 > \mathrm{F}(p, n-p)_{1-\alpha}\},$$

where t^2 is the observed value of T^2.

Hotelling's T-Square in the One-Sample Case

EXAMPLE . At the Laboratory of Heating- and Climate-technique, DTU, one has measured the following in an experiment

i) the height in cm.

ii) evaporation loss in g/m^2 skin during a 3 hour periode

iii) mean temperature in °C. This temperature is found by measuring the skin temperature at 14 different locations every minute for 5 minutes (same locations every time). The mean temperature is then an average of all $14 \times 5 = 70$ measurements,

on 16 men and 16 women.

Hotelling's T-Square in the One-Sample Case

We use the notation

$$
\begin{aligned}
\bar{\boldsymbol{X}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i \quad \in \quad \mathrm{N}_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right) \\
\bar{\boldsymbol{Y}} &= \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{Y}_i \quad \in \quad \mathrm{N}_p\left(\boldsymbol{\nu}, \frac{1}{m}\boldsymbol{\Sigma}\right) \\
\mathbf{S}_1 &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T \\
\mathbf{S}_2 &= \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^T \\
\mathbf{S} &= \frac{(n-1)\mathbf{S}_1 + (m-1)\mathbf{S}_2}{n+m-2} \quad \in \quad \mathrm{W}\left(n+m-2, \frac{1}{n+m-2}\boldsymbol{\Sigma}\right)
\end{aligned}
$$

Hotelling's T-Square in the One-Sample Case

Theorem 4.9

We use the same notation as given above. Now, let

$$T^2 = \frac{nm}{n+m}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}).$$

Then the critical region for a test of H_0 against H_1 at level α is equal to

$$C = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \frac{n+m-p-1}{(n+m-2)p} t^2 > \mathrm{F}(p, n+m-p-1)_{1-\alpha}\}$$

Here t^2 is the observed value of T^2.

Hotelling's T-Square in the One Sample Case

```
> heatclimate<-read.table("Data/heatclimate.txt",header=T, sep=" ")

> heatclimate$sex<-as.factor(heatclimate$sex)
> summary(heatclimate)
 sex        height           evap            temp
 f:16   Min.   :157.0   Min.   :12.60   Min.   :31.90
 m:16   1st Qu.:166.2   1st Qu.:18.48   1st Qu.:33.20
        Median :173.0   Median :20.75   Median :33.55
        Mean   :172.9   Mean   :22.49   Mean   :33.52
        3rd Qu.:180.0   3rd Qu.:25.65   3rd Qu.:33.90
        Max.   :190.0   Max.   :45.40   Max.   :34.80
```

- We want to compare the multivariate data for males and females. We will disregard the height.

Hotelling's T-Square in the One Sample Case

```
> X<-as.matrix(heatclimate[heatclimate$sex=="f",3:4])
> Y<-as.matrix(heatclimate[heatclimate$sex=="m",3:4])
> n<-dim(X)[1]
> m<-dim(Y)[1]
> p<-dim(X)[2]
> Xbar<-colMeans(X)
> Ybar<-colMeans(Y)
>
> Xbar2<-matrix(rep(Xbar,16),ncol=p,byrow=T)
> Ybar2<-matrix(rep(Ybar,16),ncol=p,byrow=T)
>
> S1<-t(X-Xbar2)%*%(X-Xbar2)/(n-1)
> S2<-t(Y-Ybar2)%*%(Y-Ybar2)/(m-1)
>
> S<-((n-1)*S1+(m-1)*S2)/(n+m-2)
> S
           evap       temp
evap 45.5081250 -0.3004167
temp -0.3004167  0.2985000
```

Hotelling's T-Square in the One Sample Case

```
> T2<-(n*m/(n+m))*t(Xbar-Ybar)%*%solve(S)%*%(Xbar-Ybar)
> T2
          [,1]
[1,] 4.612128
> 1-pf(T2,p,n+n-p+1)
            [,1]
[1,] 0.01764176
```

The data do not support that males and females have the same measurements ($p<0.02$).

Exercises

- Exam 2007 problem 8,
- Exam 2008 problem 3 (disregard pin 3.7)
- Exercise 4.3 (disregard task 5)
- Exercise 5.1