

# V. Statistical analysis in R (solution)

Data Science Lab, University of Copenhagen

2026-05-06

## Table of contents

Load libraries . . . . .	1
Import data . . . . .	2
Regression . . . . .	7
ANOVA . . . . .	10
Models with numerical as well as categorical predictors . . . . .	13
Logistic regression . . . . .	16
Linear mixed models . . . . .	17

## Load libraries

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.2.0      v readr      2.1.6
v forcats    1.0.1      v stringr    1.6.0
v ggplot2     4.0.2      v tibble     3.3.1
v lubridate  1.9.5      v tidyr      1.3.2
v purrr       1.2.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(emmeans)
```

Welcome to emmeans.

Caution: You lose important information if you filter this package's results.  
See '? tidy'

## Import data

1. We load the `birthwt` dataset from the `MASS` package and make the relevant wrangling of the dataset. In the end, notice how levels 3-6 of `ftvFac` are collapsed into one level denoted "MoreThanOnce".

```
# Load "birthwt" dataset from the MASS-package
data(birthwt, package = "MASS")
```

2. For nice printing etc we reformulate `birthwt` as a tibble.

```
### Make a tibble with the data
birthData <- as_tibble(birthwt)
birthData
```

```
# A tibble: 189 x 10
```

	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	0	19	182	2	0	0	0	1	0	2523
2	0	33	155	3	0	0	0	0	3	2551
3	0	20	105	1	1	0	0	0	1	2557
4	0	21	108	1	1	0	0	1	2	2594
5	0	18	107	1	1	0	0	1	0	2600
6	0	21	124	3	0	0	0	0	0	2622
7	0	22	118	1	0	0	0	0	1	2637
8	0	17	103	3	0	0	0	0	1	2637
9	0	29	123	1	1	0	0	0	1	2663
10	0	26	113	1	1	0	0	0	0	2665

```
# i 179 more rows
```

3. We recode the `smoke` variable from a numerical variable to a categorical one.

```
### Make smoke into a factor
birthData <- mutate(birthData, smoke = factor(smoke))
```

4. We examine the `ftv` variable, then construct a categorical version of the `ftv` variable, then make a version where levels 2 and above are collapsed into just one level.

```
# Check the ftv variable, make it a factor and collapse some of the levels
table(birthData$ftv)
```

	0	1	2	3	4	6
100	47	30	7	4	1	

```
birthData <- birthData %>%
  mutate(
    ftvFac = factor(ftv),
    visits = fct_collapse(ftvFac, Never="0", Once="1", other_level="MoreThanOnce")
  )

table(birthData$visits, birthData$ftvFac)
```

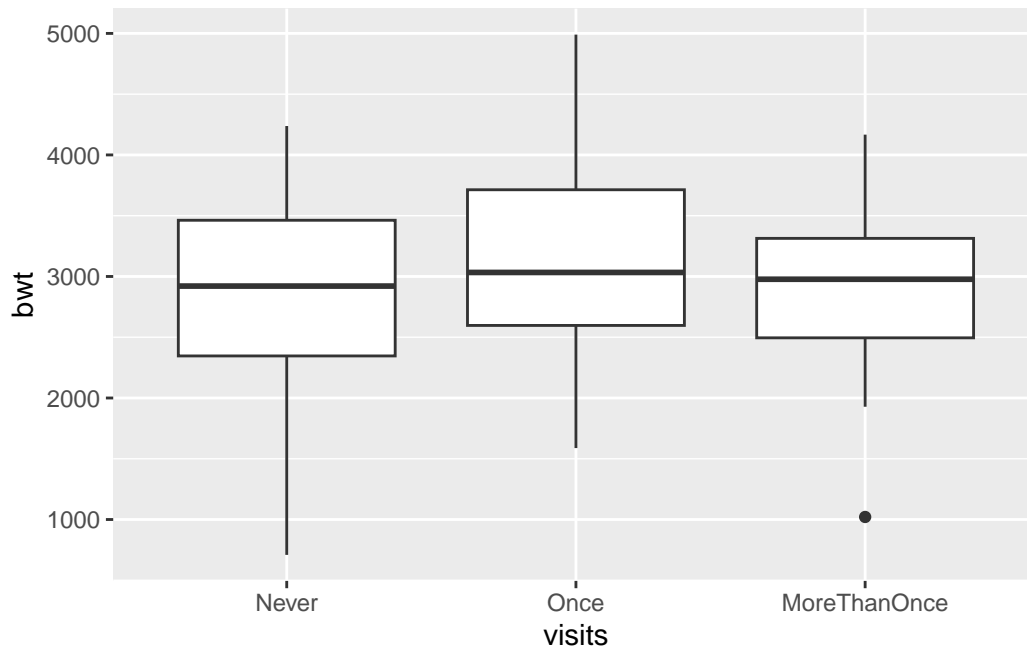
		0	1	2	3	4	6
Never	100	0	0	0	0	0	0
Once	0	47	0	0	0	0	0
MoreThanOnce	0	0	30	7	4	1	

```
birthData %>% group_by(ftvFac, visits) %>% count()
```

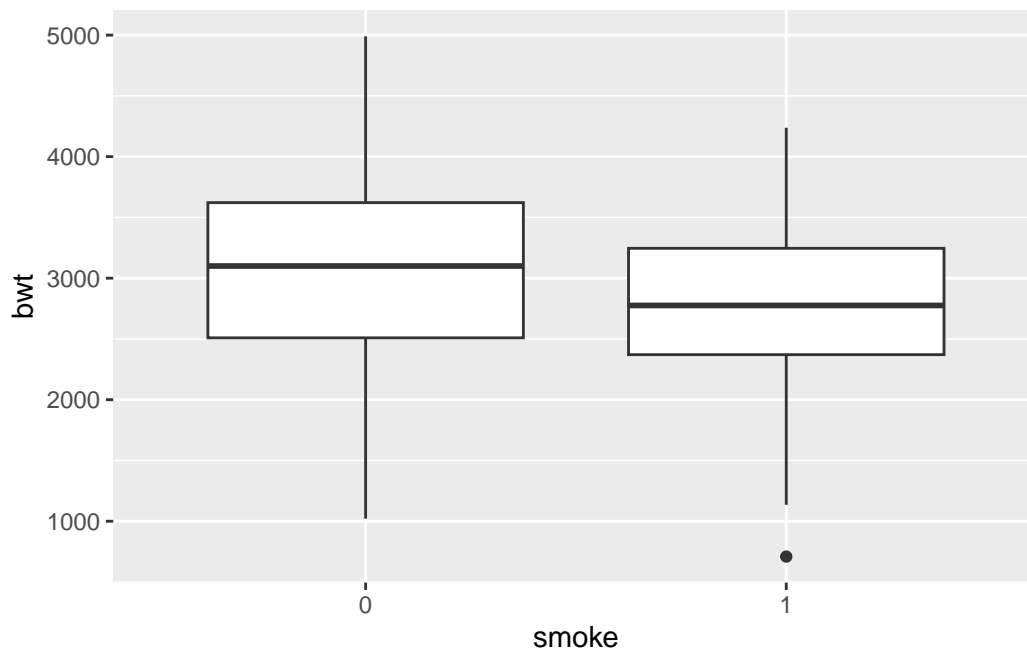
```
# A tibble: 6 x 3
# Groups:   ftvFac, visits [6]
  ftvFac visits      n
  <fct>   <fct>   <int>
1 0      Never    100
2 1      Once     47
3 2      MoreThanOnce 30
4 3      MoreThanOnce 7
5 4      MoreThanOnce 4
6 6      MoreThanOnce 1
```

5. We make boxplots for `bwt` for different levels of one or more variables. Birth weight appears to be smaller for infants born by smokers compared to non-smokers, while birth weight does not appear to depend on the frequency of visits to the doctor. This is just preliminary considerations; formal analyses are carried out later.

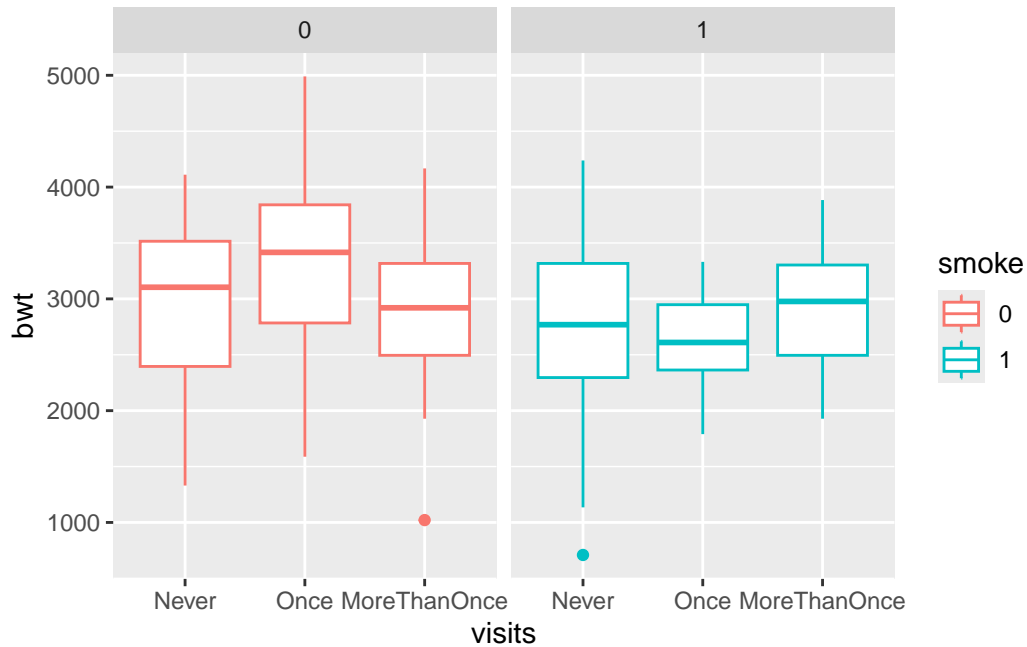
```
### Parallel boxplots
ggplot(birthData, aes(x=visits, y=bwt)) + geom_boxplot()
```



```
ggplot(birthData, aes(x=smoke, y=bwt)) + geom_boxplot()
```

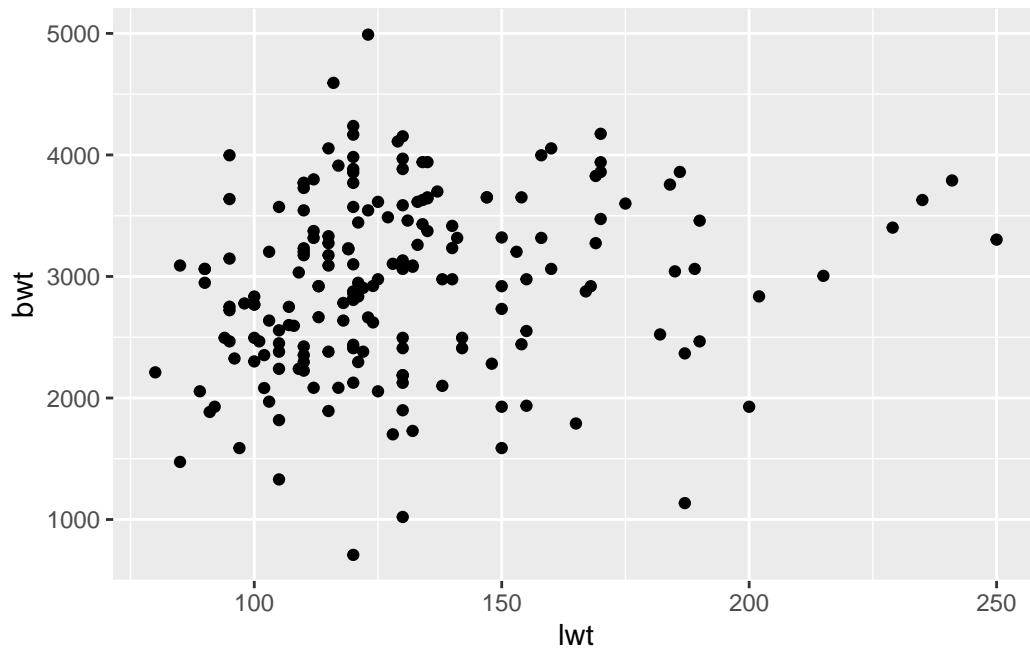


```
ggplot(birthData, aes(x=visits, y=bwt, color=smoke)) +  
  geom_boxplot() + facet_wrap(~smoke)
```

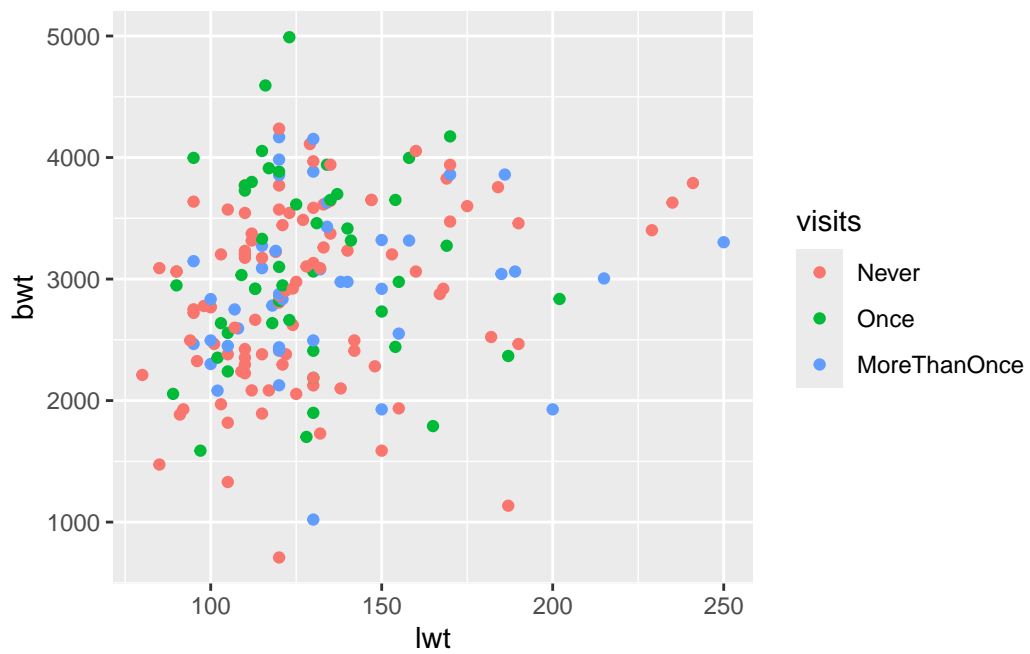


6. Scatterplot with `lwt` on the  $x$ -axis and `bwt` on the  $y$ -axis. The color and shape of the points can be controlled by the options `color` and `shape`. There is no clear association between the two variables (mother's weight before pregnancy and infant weight), but let us see in the analysis.

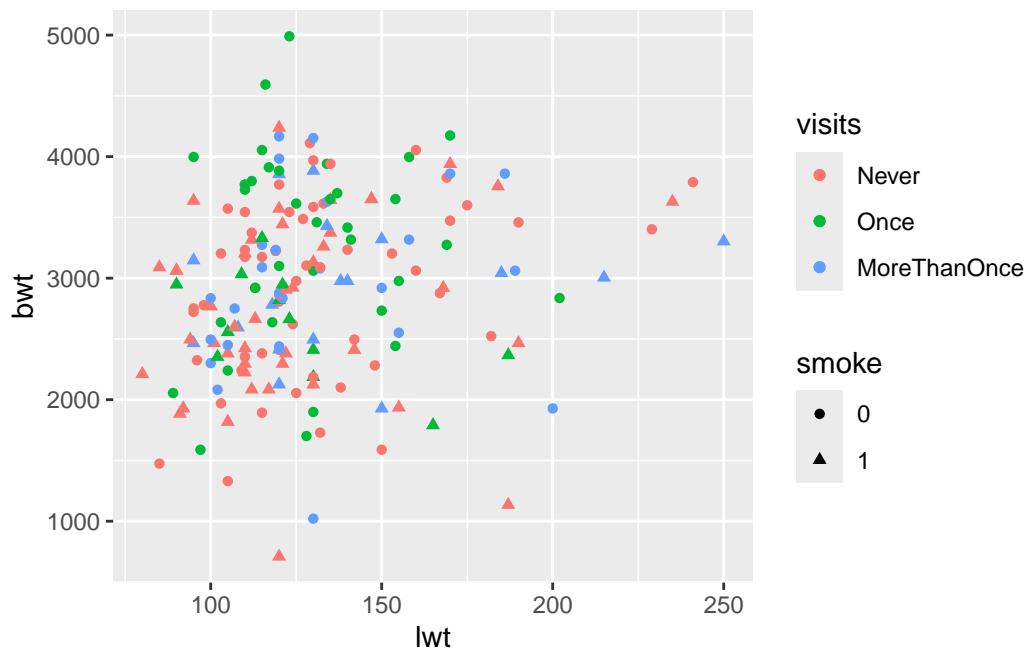
```
### Scatter plot
ggplot(birthData, aes(x=lwt, y=bwt)) + geom_point()
```



```
### Points coloured after visits
ggplot(birthData, aes(x=lwt, y=bwt, color=visits)) + geom_point()
```



```
### Points coloured after visits, point types after smoke status
ggplot(birthData, aes(x=lwt, y=bwt, color=visits, shape=smoke)) + geom_point()
```



## Regression

- We use the `lm` function to fit the simple linear regression model with infant's birth weight as response and mothers weight before pregnancy as predictor. Then, we extract estimates and confidence intervals (CIs). Specifically we get a 95% CI 1.05-7.81 for the slope parameter. Notice that zero is not included in the CI, in line with the p-value of 0.01 which appears in the summary. So, despite the impression from the scatter plot above, there is a significant association between the two variables.

```
### Simple linear regression
reg1 <- lm(bwt ~ lwt, data=birthData)
summary(reg1)
```

Call:

```
lm(formula = bwt ~ lwt, data = birthData)
```

Residuals:

Min	1Q	Median	3Q	Max
-2192.12	-497.97	-3.84	508.32	2075.60

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2369.624	228.493	10.371	<2e-16 ***
lwt	4.429	1.713	2.585	0.0105 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 718.4 on 187 degrees of freedom

Multiple R-squared: 0.0345, Adjusted R-squared: 0.02933

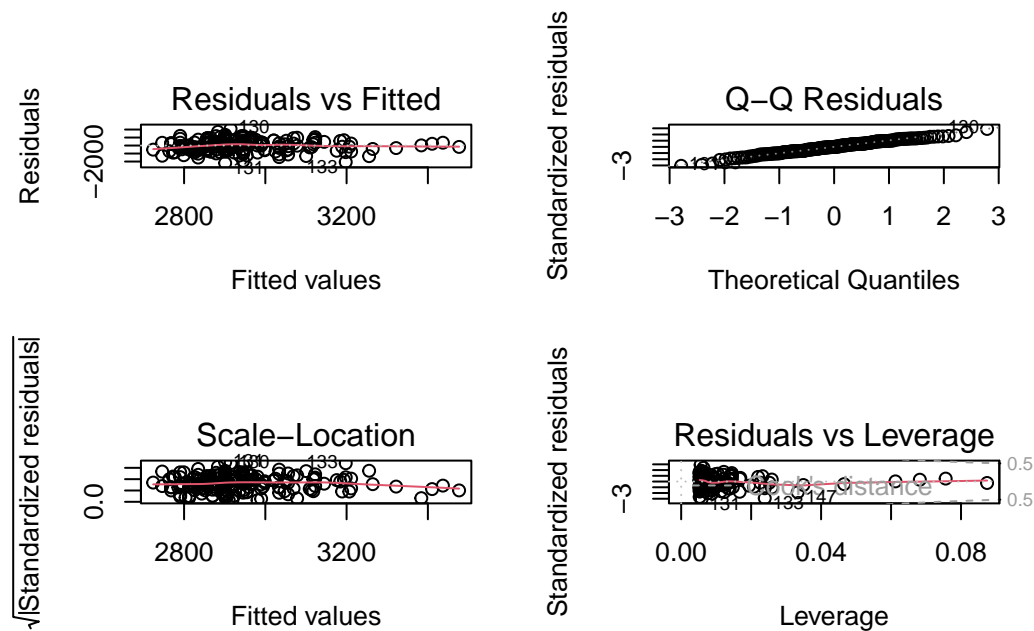
F-statistic: 6.681 on 1 and 187 DF, p-value: 0.0105

```
confint(reg1)
```

	2.5 %	97.5 %
(Intercept)	1918.867879	2820.37916
lwt	1.048845	7.80937

8. We make the usual model validation plots. None of them gives us any reasons to worry about the validity of the model assumptions.

```
### Model validation
par(mfrow=c(2,2))
plot(reg1)
```





9. We include extra predictors in the model, so we now get a multiple linear regression. The regression coefficients are estimated to 4.12 for mother's weight (`lwt`), 7.49 for age and 15.4 for number of visits to the doctor (`ftv`). The interpretation of the last one, say, is that the birth weight is expected to increase by 15.4 grams per extra visit by the doctor, all other predictors kept unchanged. Notice that the estimate is not significantly different from zero ( $p=0.76$ ) though.

```
### Include age and ftv as covariate
reg2 <- lm(bwt ~ lwt + age + ftv, data=birthData)
summary(reg2)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2223.505618	301.567140	7.3731694	5.402210e-12
<code>lwt</code>	4.120744	1.757787	2.3442798	2.012498e-02
<code>age</code>	7.485748	10.285178	0.7278189	4.676446e-01
<code>ftv</code>	15.363944	51.113921	0.3005824	7.640705e-01

10. We make a new (boring dataset) with just one data line and values of each of the three predictors from the multiple linear regression from the previous question. The `predict` function computes the corresponding predicted value for the birth weight. It needs to know the predictor values as well as which model to use for the prediction (you could try using `reg1` instead of `reg2`; then you would get a slightly different prediction). The `interval` option makes it possible to supplement the point prediction with a 95% prediction interval, i.e., an interval that contains the birth weight for 95% of infants with the given predictor characteristics.

```
### Prediction
newData <- data.frame(lwt=100, age=25, ftv=0)
newData
```

```
  lwt age ftv
1 100  25   0
```

```
predict(reg2, newData)
```

```
1
2822.724
```

```
predict(reg2, newData, interval="prediction")
```

```
      fit      lwr      upr
1 2822.724 1390.148 4255.299
```

11. We use a different dataset for the regression, namely a subset constructed with the `filter` function. There is just for the sake of the example, so we refrain from looking at the results in any detail.

```
### Use only data from mothers with weight below 160
reg3 <- lm(bwt ~ lwt + age + ftv, data=filter(birthData, lwt<160))
summary(reg3)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1824.063262	444.508301	4.1035528	0.0000651941
lwt	7.182184	3.341787	2.1492047	0.0331503757
age	9.178577	11.624301	0.7896024	0.4309508695
ftv	16.986307	60.011783	0.2830495	0.7775117320

## ANOVA

12. We use `lm` to fit the oneway ANOVA. The non-smokers (`smoke=0`) are used as reference, so the expected birth weight for that group is given by the intercept, 3056 grams. For the estimate for smokers, we must add the contrast/difference caused by the smoke variable (which is negative). We get  $3056 - 284 = 2772$ . With `emmeans` we get both numbers automatically (but then not the difference). The p-value for the effect of smoking is 0.0087, which can be found both from the `summary` and the `pairs` output, so there is a statistically significant effect of smoking.

```
### Oneway ANOVA against smoke
oneway1 <- lm(bwt ~ smoke, data=birthData)
summary(oneway1)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3055.6957	66.93326	45.652875	2.463035e-103
smoke1	-283.7767	106.96877	-2.652893	8.666726e-03

```
emmeans(oneway1, ~smoke)
```

smoke	emmean	SE	df	lower.CL	upper.CL
0	3056	66.9	187	2924	3188
1	2772	83.4	187	2607	2937

Confidence level used: 0.95

```
pairs(emmeans(oneway1, ~smoke))
```

contrast	estimate	SE	df	t.ratio	p.value
smoke0 - smoke1	284	107	187	2.653	0.0087

13. We now fit a another oneway ANOVA where `visits` from question 4 is used as the explanatory variable. The groupwise estimated birth weights are most easily found with

emmeans and are 2865 g, 3108 g, and 2951 g, respectively. The output from `pairs` shows that none of the groups differ statistically significantly from each other (p-values are 0.14 or larger). Finally, the overall test gives a p-value of 0.17 (`drop1`), confirming that groups do not differ significantly.

```
### Oneway ANOVA against visits
oneway2 <- lm(bwt ~ visits, data=birthData)
emmeans(oneway2,~visits)
```

visits	emmean	SE	df	lower.CL	upper.CL
Never	2865	72.6	186	2722	3008
Once	3108	106.0	186	2899	3317
MoreThanOnce	2951	112.0	186	2730	3172

Confidence level used: 0.95

```
pairs(emmeans(oneway2,~visits))
```

contrast	estimate	SE	df	t.ratio	p.value
Never - Once	-242.9	128	186	-1.891	0.1441
Never - MoreThanOnce	-85.7	134	186	-0.642	0.7970
Once - MoreThanOnce	157.1	154	186	1.019	0.5658

P value adjustment: tukey method for comparing a family of 3 estimates

```
drop1(oneway2,test="F")
```

Single term deletions

Model:

bwt ~ visits

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			98081730	2493.2		
visits	2	1887925	99969656	2492.8	1.7901	0.1698

14. A twoway ANOVA with both `smoke` and `visits` is fitted. It is the model *without* interaction because there is a plus (+) between the two variables. Reference groups are chosen for each variable (`smoke=0` corresponding to non-smokers and `visits=Never`), and the intercept 2981 g is the estimated birth weight for the reference combination. For the contrast, the birth weight is estimated to decrease by 257 g if the mother is a smoker, and to increase by 193 g and 74 g if the mother visits the doctor once or more than once, respectively. We also get p-values for each of these contrasts, and conclusions are unchanged from the oneway ANOVAs: There is a significant effect of smoking but not a significant effect of visiting the doctor.

```
### Twoway ANOVA without interaction
twoway1 <- lm(bwt ~ visits + smoke, data=birthData)
summary(twoway1)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2980.91672	86.73734	34.3671676	3.021284e-82
visitsOnce	192.77220	128.59583	1.4990548	1.355639e-01
visitsMoreThanOnce	74.10202	131.98092	0.5614601	5.751634e-01
smoke1	-257.28159	108.37493	-2.3739954	1.862096e-02

15. The twoway ANOVA *with* interaction is fitted by using a multiplication sign or star (\*) between the two variables. This allows for the effect of smoking to differ between visits groups, or vice versa. The hypothesis of no interaction can be carried out with the `anova` function, where the models to be compared (the models with and without interaction) are specified explicitly, or with `drop1`. The same test is carried out, so we get the same p-value, 0.14. So, the interaction effect is not significant. By the way, the function name `anova` is a bit annoying since it can be used for many model types, not only ANOVA models.

```
### Twoway ANOVA with interaction, test for interaction in two ways
twoway2 <- lm(bwt ~ visits * smoke, data=birthData)
anova(twoway2, twoway1)
```

#### Analysis of Variance Table

```
Model 1: bwt ~ visits * smoke
Model 2: bwt ~ visits + smoke
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	183	93189162				
2	185	95182096	-2	-1992934	1.9568	0.1443

```
drop1(twoway2, test="F")
```

#### Single term deletions

```
Model:
bwt ~ visits * smoke
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			93189162	2489.5		
visits:smoke	2	1992934	95182096	2489.5	1.9568	0.1443

16. The `emmeans` function can be used to compute group-specific estimates, on average over the other variables in the model. We get the estimated birth weights of 3066 g and 2772 g, respectively. This is not exactly the same as in question 12: The estimates in question 12

were simple averages in the two groups, while the new estimates are averages of estimates. This is not the same when the dataset is unbalanced wrt. to the two variables.

```
emmeans(twoway2, ~smoke)
```

NOTE: Results may be misleading due to involvement in interactions

smoke	emmean	SE	df	lower.CL	upper.CL
0	3066	70.1	183	2928	3205
1	2766	96.4	183	2576	2956

Results are averaged over the levels of: visits

Confidence level used: 0.95

## Models with numerical as well as categorical predictors

17. `model1` includes both a numerical and a categorical explanatory variable. Birth weight is estimated to increase 4.24 g for each extra pound of mother's weight (no matter the smoking status) and to differ 272 g for smoking and non-smoking mothers. The intercept is the estimated birth weight for a non-smoking mother with a weight of zero pounds! This is of course nonsense - as is often the case with regression type models when zero values of a numerical predictor cannot occur in practice.

```
### Model with linear (lwt,bwt) association.
### Intercept differ between smokers and non-smokers, one common slope.
model1 <- lm(bwt ~ lwt + smoke, data=birthData)
summary(model1)
```

Call:

```
lm(formula = bwt ~ lwt + smoke, data = birthData)
```

Residuals:

Min	1Q	Median	3Q	Max
-2030.90	-445.69	29.16	521.76	1967.76

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2501.125	230.836	10.835	<2e-16 ***
lwt	4.237	1.690	2.507	0.0130 *
smoke1	-272.081	105.591	-2.577	0.0107 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 707.8 on 186 degrees of freedom  
 Multiple R-squared: 0.06777, Adjusted R-squared: 0.05775  
 F-statistic: 6.761 on 2 and 186 DF, p-value: 0.001464

18. Interaction between a numerical and categorical variable is interpreted as different slopes (wrt. the numerical variable) for the different groups. Here, this means two straight lines ( $x=lwt$  and  $y=bwt$ ), one for smokers and one for non-smokers, with different intercepts and different slopes. The model from question 17 had different intercepts, but the same slope. Both the intercept and the slope are reported in terms of a reference groups (non-smokers) and a difference between smokers and non-smokers, and the two estimated regression lines are therefore given by:  $bwt = 2351 + 5.39 * lwt$  for non-smokers and  $bwt = (2351+41) + (5.39-2.42) * lwt = 2392 - 2.97 * lwt$  for smokers. The interaction term, as reported the last line in the coefficients table, estimates the difference in the weight effect between non-smokers and smokers, so this is where we should look to answer the last question. The p-value is 0.48, so there is no evidence that the effect of mother's weight affects the birth weight is different for smokers and non-smokers. The same test can be carried out by comparing to the additive model from question 17, using the `anova` function.

```
### Model with linear (lwt,bwt) association. Intercept and slope both differ between vi
model2 <- lm(bwt ~ lwt * smoke, data=birthData)
summary(model2)
```

Call:

```
lm(formula = bwt ~ lwt * smoke, data = birthData)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2038.80	-454.76	28.36	530.84	1976.84

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2350.578	312.733	7.516	2.35e-12 ***
lwt	5.387	2.335	2.307	0.0222 *
smoke1	41.384	451.187	0.092	0.9270
lwt:smoke1	-2.422	3.388	-0.715	0.4757

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 708.8 on 185 degrees of freedom  
 Multiple R-squared: 0.07034, Adjusted R-squared: 0.05527

F-statistic: 4.666 on 3 and 185 DF, p-value: 0.003621

```
### Test if slopes differ between visit groups
anova(model2, model1)
```

#### Analysis of Variance Table

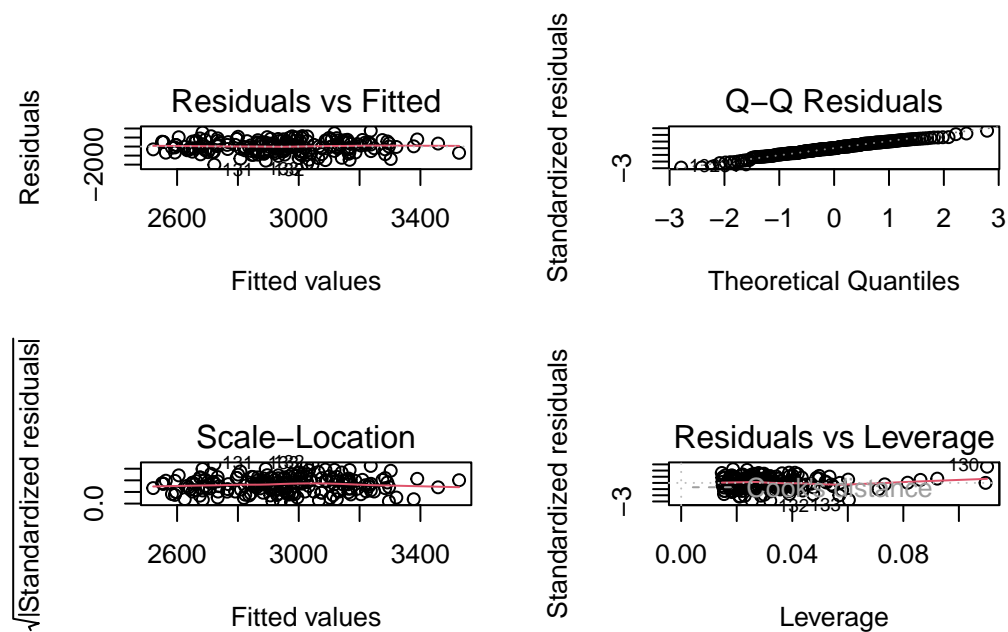
Model 1: bwt ~ lwt \* smoke

Model 2: bwt ~ lwt + smoke

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	185	92937722				
2	186	93194298	-1	-256576	0.5107	0.4757

19. We now include more variables in the model. The model validation plots cause no worries. Smoking is estimated to decrease infant birth weight by 244 g (SE 107 g), and the effect is statistically significant ( $p=0.017$ ).

```
### Model with many effects (no interactions)
model3 <- lm(bwt ~ lwt + smoke + age + visits, data=birthData)
par(mfrow=c(2,2))
plot(model3)
```



```
summary(model3)
```

```
Call:
lm(formula = bwt ~ lwt + smoke + age + visits, data = birthData)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2024.13	-491.65	6.56	507.56	1753.25

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2346.442	301.420	7.785	5e-13	***
lwt	4.167	1.727	2.413	0.0168	*
smoke1	-244.402	107.187	-2.280	0.0238	*
age	4.300	10.252	0.419	0.6754	
visitsOnce	184.314	130.194	1.416	0.1586	
visitsMoreThanOnce	32.318	133.410	0.242	0.8089	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 708.6 on 183 degrees of freedom

Multiple R-squared: 0.08076, Adjusted R-squared: 0.05564

F-statistic: 3.215 on 5 and 183 DF, p-value: 0.008306

## Logistic regression

20. We now fit a logistic regression, modelling the probability of a low birthweight defines as <2500 g. The effect of smoking is just not significant ( $p=0.06$ ), which most likely reflects that we loose information going from the actual birth weights to the binary version. Notice that the estimated coefficient for smoking is now positive (0.62). This means that smoking increases the probability of getting a child with low birth weight; hence the direction of a potential effect is still the same: Smoking increases the likelihood of a smaller birth weight.

```
### Logistic regression with many predictors (no interactions)
logreg1 <- glm(low ~ lwt + smoke + age + visits, data=birthData, family="binomial")
summary(logreg1)
```

Call:

```
glm(formula = low ~ lwt + smoke + age + visits, family = "binomial",
     data = birthData)
```

Coefficients:



	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.370311	1.016623	1.348	0.1777
lwt	-0.012276	0.006138	-2.000	0.0455 *
smoke1	0.619077	0.330428	1.874	0.0610 .
age	-0.031764	0.033933	-0.936	0.3492
visitsOnce	-0.413043	0.424527	-0.973	0.3306
visitsMoreThanOnce	-0.148285	0.420965	-0.352	0.7247

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom  
Residual deviance: 221.91 on 183 degrees of freedom  
AIC: 233.91

Number of Fisher Scoring iterations: 4

## Linear mixed models

21. We first make the artificial center variable as described in the exercise. You should of course never make such a construct unless it reflects the data collection.

```
### Make artificial center variable
set.seed(123)
center <- sample(rep(1:19, each=10)[1:189])
birthData <- mutate(birthData, center=factor(center))
```

22. We use the new variable as a random factor in a linear mixed model (LMM) and identify the relevant estimate for smoking to be 244 g (SE 107). This is exactly the same as in the model without random effects (question 19). That would typically not be the case, but happens here because the center-to-center standard deviation is essentially estimated to zero (3.109e-06), which is not too surprising since the center variable had nothing to do with actual data collection. Notice that the summary of an `lmer` object does not provide p-values. There are ways to get those p-values, for example by using `emmeans` or the `lmerTest` package instead of the the `lme4` package, but we will not go into details about that here.

```
### Remember to install lme4 before this can run
# install.packages("lme4")
library(lme4)
```

Indlæser krævet pakke: Matrix

Vedhæfter pakke: 'Matrix'

De følgende objekter er maskerede fra 'package:tidyr':

expand, pack, unpack

```
### Linear mixed model with random effect of center  
lmm1 <- lmer(bwt ~ lwt + smoke + age + visits + (1|center), data=birthData)
```

boundary (singular) fit: see help('isSingular')

```
summary(lmm1)
```

Linear mixed model fit by REML ['lmerMod']

Formula: bwt ~ lwt + smoke + age + visits + (1 | center)

Data: birthData

REML criterion at convergence: 2958

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-2.85637	-0.69380	0.00926	0.71625	2.47412

Random effects:

Groups	Name	Variance	Std.Dev.
center	(Intercept)	9.668e-12	3.109e-06
Residual		5.022e+05	7.086e+02

Number of obs: 189, groups: center, 19

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	2346.442	301.420	7.785
lwt	4.167	1.727	2.413
smoke1	-244.402	107.187	-2.280
age	4.300	10.252	0.419
visitsOnce	184.314	130.194	1.416
visitsMoreThanOnce	32.318	133.410	0.242

Correlation of Fixed Effects:

	(Intr)	lwt	smoke1	age	vstsOn
lwt		-0.613			
smoke1		-0.196	0.046		
age		-0.620	-0.173	0.003	

```
visitsOnce -0.026  0.048  0.160 -0.218  
vstsMrThnOn 0.055 -0.058  0.031 -0.190  0.335  
optimizer (nloptwrap) convergence code: 0 (OK)  
boundary (singular) fit: see help('isSingular')
```

**End of solution**