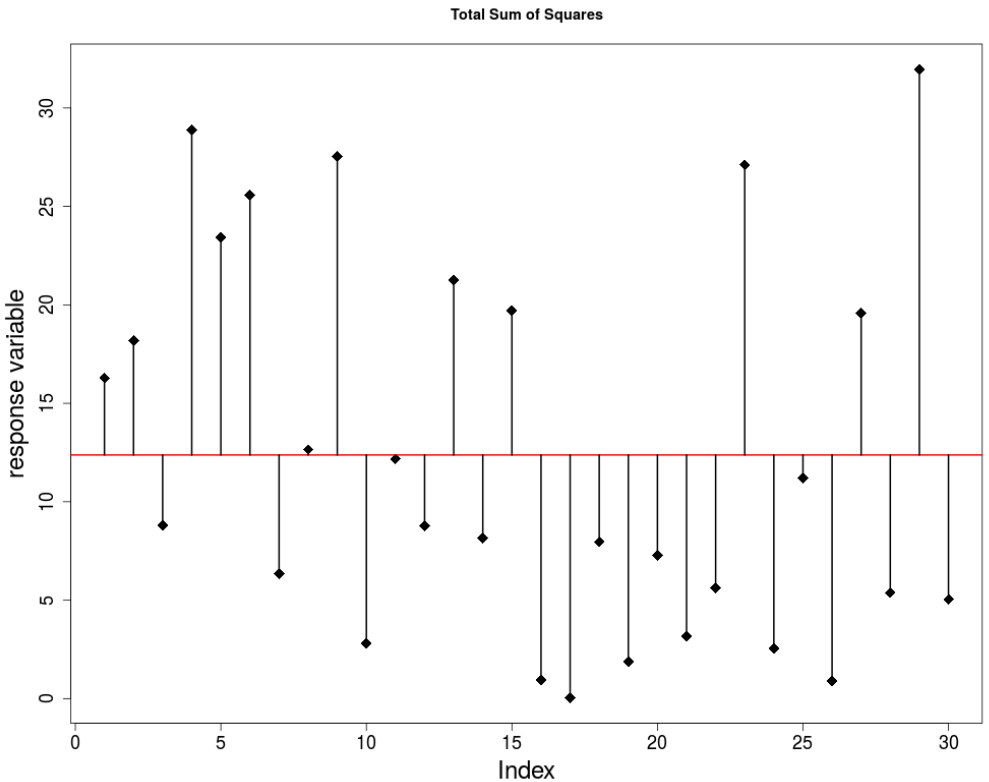
Linear Models

Mandy Vogel

July 18, 2016

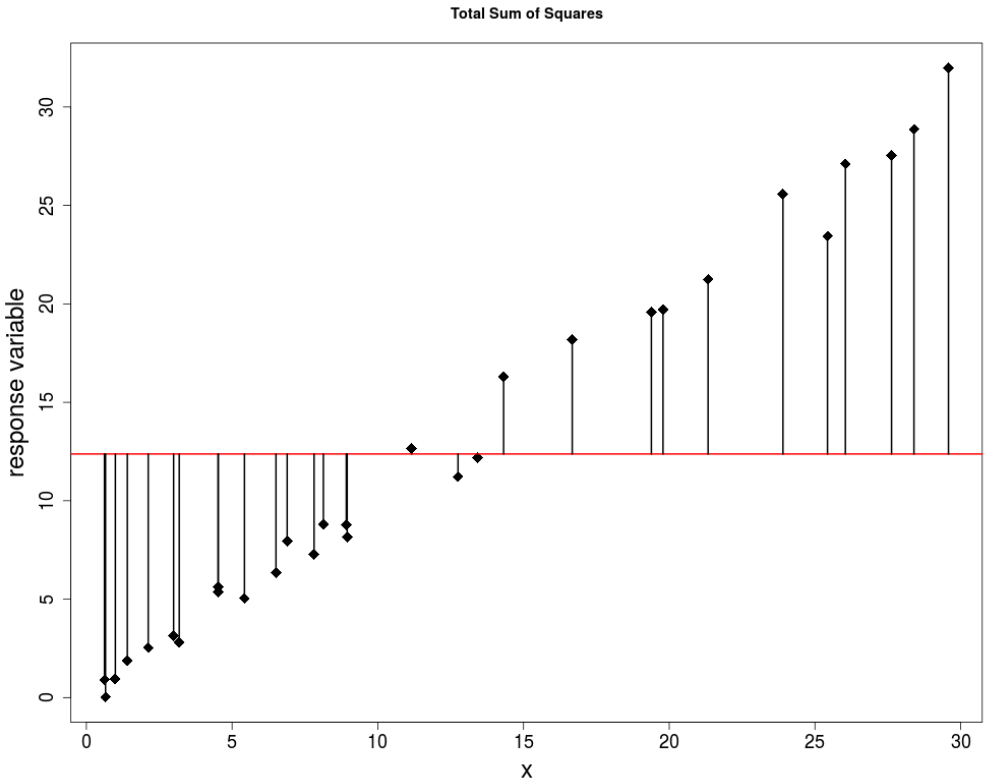
Table of Contents

The Null model



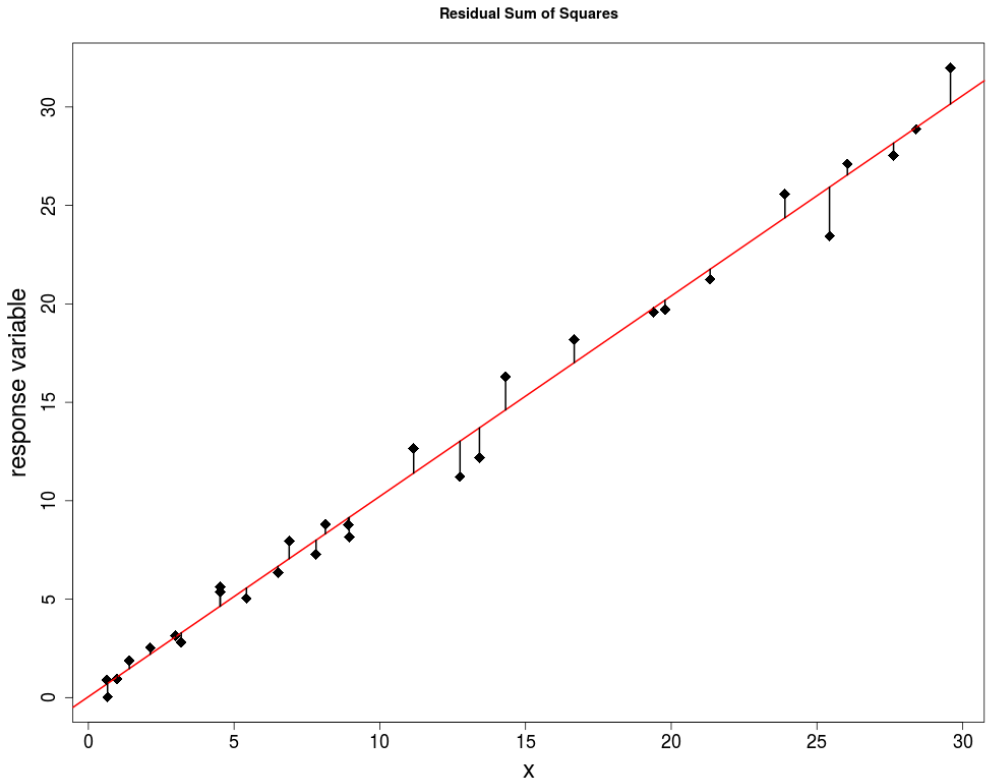
- Just one parameter, the overall mean $\bar{y}$
- Fit: none; $SSE = SSY$
- Degrees of freedom: $n - 1$
- Explanatory power of the model: none

Adding Information



- model with $0 \leq p' \leq p$ parameters
- Fit: less than the maximal model, but not significantly so
- Degrees of freedom: $n - p' - 1$
- Explanatory power of the model: $r^2 = \frac{SSR}{SSY}$

Adding Information



- model with $0 \leq p' \leq p$ parameters
- Fit: less than the maximal model, but not significantly so
- Degrees of freedom: $n - p' - 1$
- Explanatory power of the model: $r^2 = \frac{SSR}{SSY}$

The births data

A data frame with 500 observations on the following 8 variables.

`id:`	Identity number for mother and baby.
`bweight:`	Birth weight of baby.
`lowbw:`	Indicator for birth weight less than 2500 g.
`gestwks:`	Gestation period.
`preterm:`	Indicator for gestation period less than 37 weeks.
`matage:`	Maternal age.
`hyp:`	Indicator for maternal hypertension.
`sex:`	Sex of baby: 1:Male, 2:Female.

From: Michael Hills and Bianca De Stavola (2002). A Short Introduction to Stata 8 for Biostatistics, Timberlake Consultants Ltd URL: http://www.timberlake.co.uk

Transform Data

```
> births <- transform(births,
+        lowbw = factor(lowbw, labels=c("normal","low")),
+        preterm = factor(preterm, labels=c("normal","preterm")),
+        hyp = factor(hyp, labels=c("normal","hyper")),
+        sex = factor(sex, labels=c("M","F")),
+        gest4 = cut(gestwks, breaks=c(20,35,37,39,45),
+        right=F))
```

(The original and the corrected version are contained in the data folder.)

Variables in Models

The response variable must be numeric. Main types are

- Metric (a measurement with units); the easiest case, we will begin with this
- Binary (two values code 0/1)
- Count (aggregated data)
- Failure (does the subject fail at end of follow up)

Explanatory variables can be

- Numeric
- Factor

Metric Response, Numeric explanatory variable

Assuming that the relationship of `bweight` with `gestwks` is roughly linear we can find the linear effect on `bweight` of a unit increase in `gestwks` with

```
> m <- lm(bweight ~ gestwks, data=births)
```

- `lm()` is the linear model function
- `bweight ~ gestwks` is the model formula
- `m` is a model object (containing all information about our model), there are certain functions to extract these information, e.g.:

```
> coef(m)
(Intercept)     gestwks
 -4489.1398    196.9726
```

One extra week of gestation produces an extra 197g of baby.

Extractor functions

```
> summary(m)

Call:
lm(formula = bweight ~ gestwks, data = births)

Residuals:
     Min       1Q   Median       3Q      Max
-1698.40  -280.14    -3.64   287.61  1382.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4489.140    340.899  -13.17   <2e-16 ***
gestwks       196.973      8.788   22.41   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 449.7 on 488 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.5073, Adjusted R-squared: 0.5062
F-statistic: 502.4 on 1 and 488 DF,  p-value: < 2.2e-16
```

Extractor functions

```
> coef(m)
(Intercept)     gestwks
 -4489.1398    196.9726
> confint(m)
                  2.5 %     97.5 %
(Intercept) -5158.9503 -3819.3293
gestwks       179.7054   214.2399
```

Other Useful Functions

The model object is a list of different elements each of which can be accessed separately (see `str(m)` for the full list).
Other useful functions:

- `print(m)` simple display
- `plot(m)` produces various diagnostic plots based on residuals
- `fitted(m)` returns a vector of fitted values
- `resid(m)` returns a vector of residuals
- `predict(m, newdata)` predicts the response for new values of the explanatory variables
- `deviance(m)` residual sum of squares
- `df.residual(m)` for the residual degrees of freedom
- `vcov(m)` variance-covariance matrix

Explanatory Variable is a Factor

The effect of `hyp` (2-level factor) on `bweight` is obtained with

```
> m <- lm(bweight ~ hyp, data=births)
> coef(m)
(Intercept)     hyphyper
  3198.9042   -430.6959
```

Omitting the intercept gives the mean `bweight` at the two levels of `hyp`

```
> m <- lm(bweight ~ -1 + hyp, data=births)
> coef(m)
hypnormal  hyphyper
 3198.904  2768.208
```

A Multivariable Model

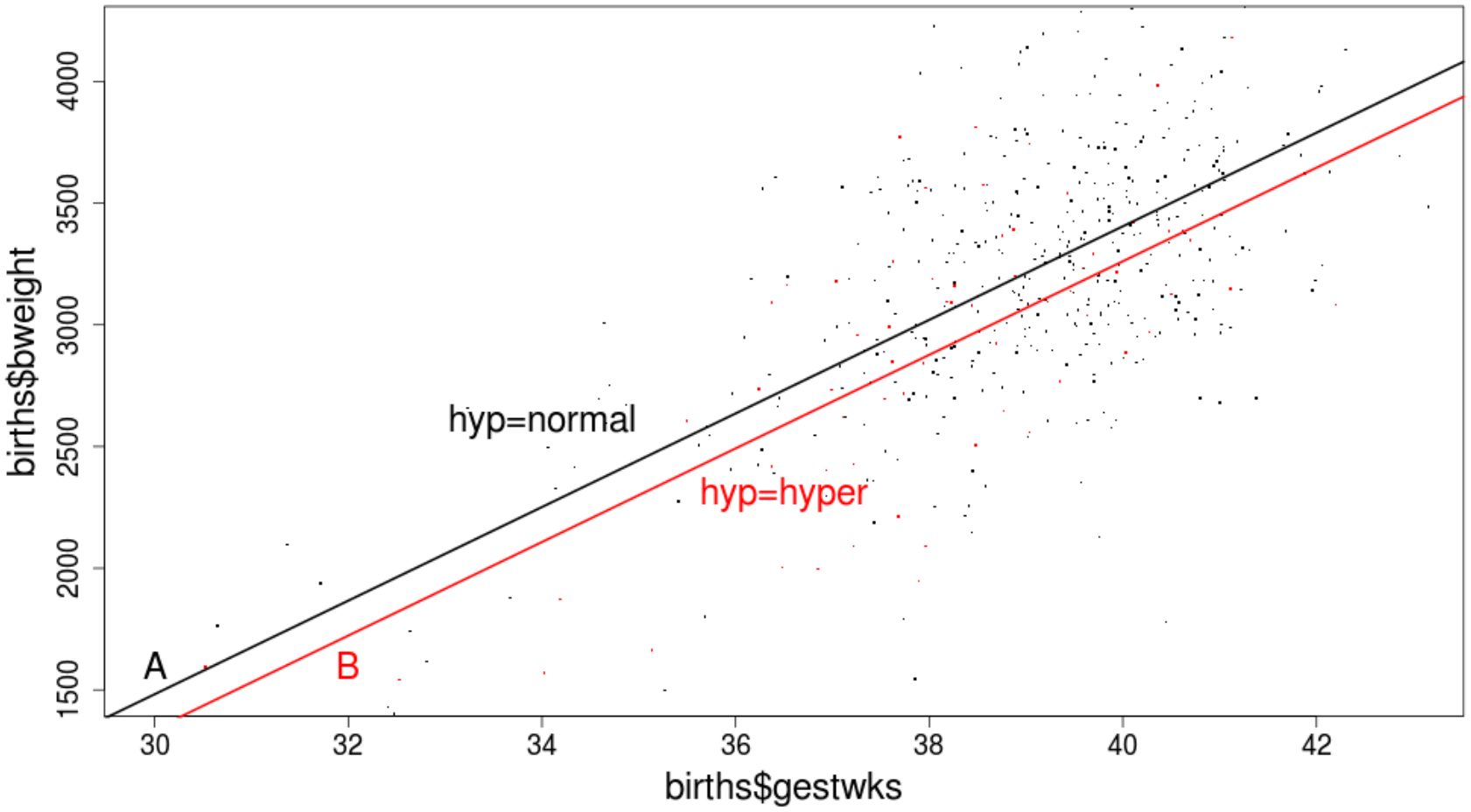
The joint effect of `hyp` and `gestwks` on `bweight` is obtained with

```
> m <- lm(bweight ~ hyp + gestwks, data=births)
```

```
             Estimate
(Intercept) -4285.002
hyphyper     -143.675 (level 2 vs. level 1)
gestwks       192.238 (increase per week)
```

The effect of `hyp` is attenuated (from -430.7 to -143.7). This suggests that much of the effect of hypertension on birth weight is mediated through a shorter gestation period.

A Model With Both `gestwks` and `hyp`



The effect of `gestwks` is the slope of the lines A and B (assumed to be the same). The effect of `hyp` ist the vertical distance between them.

Interaction Models in `lm`

To specify an interaction term in `lm`, change the model formula from

Input

```
> m <- lm(bweight ~ hyp + gestwks, data=births)
```

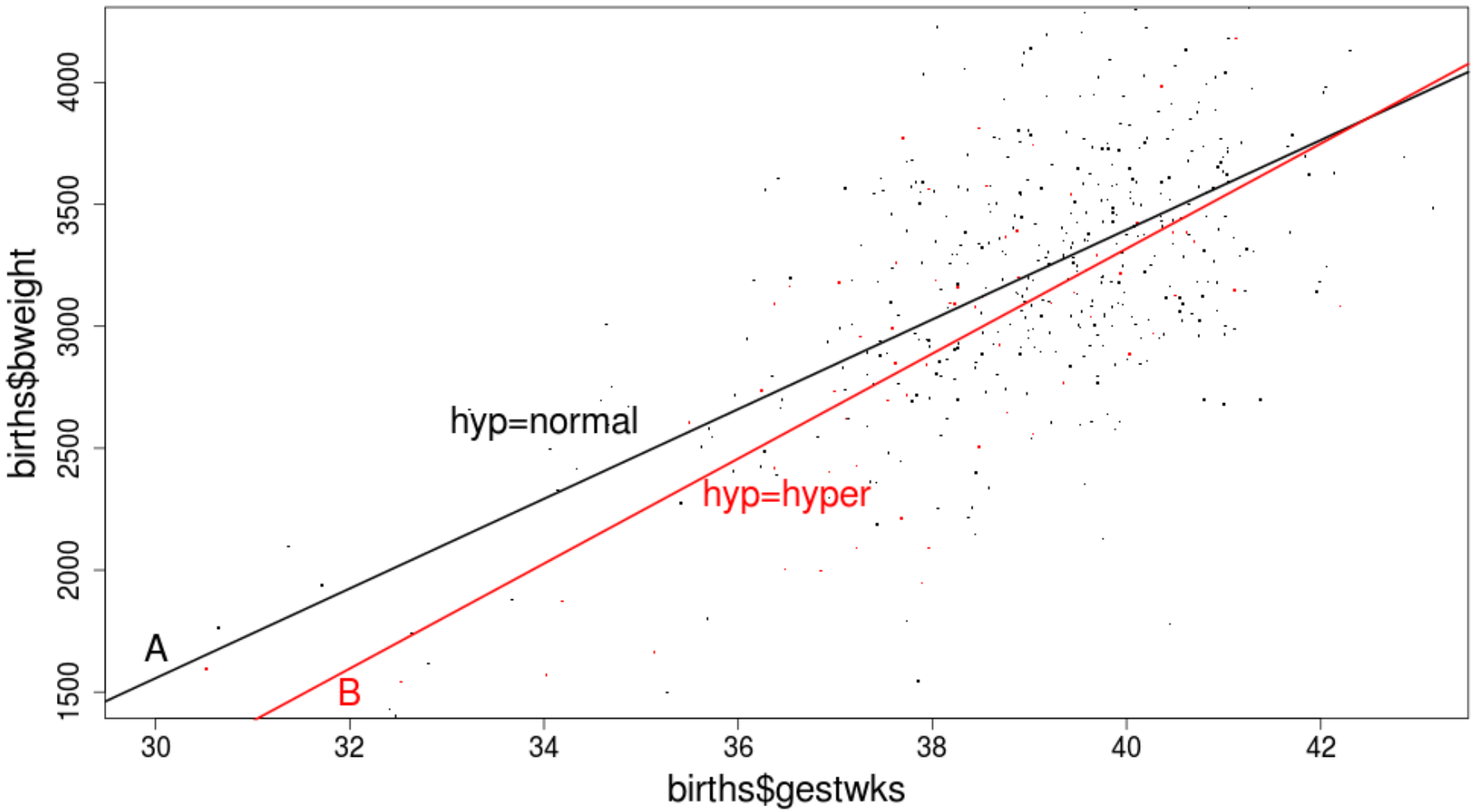
to

```
> m <- lm(bweight ~ hyp + gestwks + hyp:gestwks, data=births)
```

or shorter

```
> m <- lm(bweight ~ hyp * gestwks, data=births)
```

Interaction Between `gestwks` and `hyp`



Interactions Models in `lm`

Output

```
                  Estimate
(Intercept)       -3960.82
hyphyper          -1332.66 (level 2 vs level 1 - intercept)
gestwks             183.91
hyphyper:gestwks     31.39 (level 2 vs level 1 - slope)
```

Now the effect of `hyp` more difficult to explain, because it is not constant. The effect of -1332 is valid on a hypothetical gestational age of 0. Which doesn't make sense. You could scale the `gestwks` variable.

```
> births$gwsc <- births$gestwks-40
> m <- lm(bweight ~ hyp * gwsc, data=births)
```

Interactions Models in `lm`

Input/Output

```
                Estimate
(Intercept)   3395.60329
hyphyper       -77.25215 (level 2 vs level 1 - intercept)
gwsc           183.91048
hyphyper:gwsc   31.38510 (level 2 vs level 1 - slope)
```

How much is explained? - aov

In the Null-Model we have seen that $SSE = SSY$ (the error sum of squares is equal to the total sum of squares in y) and therefore the Null-Model explaines nothing of the overall variance. So the fraction how much of the overall variance is explained by our model regarding to the overall variance is a first measure for the fit of the model...

- the simple model with one explanatory variable

```
> m <- lm(bweight ~ gestwks, data=births)
> anova(m)
Analysis of Variance Table

Response: bweight
            Df    Sum Sq   Mean Sq F value    Pr(>F)
gestwks      1 101603845 101603845  502.36 < 2.2e-16 ***
Residuals  488  98698698    202251
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

How much is explained? - aov

- in the second column of the summary we see the regression sum of squares (SSR) in the first line and in the second line the error sum of squares (SSE). So the total sum of squares (SSY - a measure for the overall variation) is the sum of both:

```
> sum(anova(m)$Sum)
[1] 200302543
```

- and the fraction is

```
> anova(m)$Sum[1]/sum(anova(m)$Sum)
[1] 0.5072519
```

How much is explained? - aov

- this is r-squared

```
> summary(m)$r.squared
[1] 0.5072519
```

- which you can extract from the summary of the model

```
> summary(m)
Call:
lm(formula = bweight ~ gestwks, data = births)
Residuals:
     Min       1Q   Median       3Q      Max
-1698.40  -280.14    -3.64   287.61  1382.24
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4489.140    340.899  -13.17   <2e-16 ***
gestwks       196.973      8.788   22.41   <2e-16 ***

Residual standard error: 449.7 on 488 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.5073,Adjusted R-squared: 0.5062
F-statistic: 502.4 on 1 and 488 DF,  p-value: < 2.2e-16
```

Exercises I

1. load the nhanes data
2. how many observations, how many variables?
3. how old are the participants (summary statistics, mean, sd)
4. plot waist circumference vs age
5. model the respective data in a linear model, extract and interpret the coefficients. Extract also the confidence intervals.
6. add sex as a covariate. interpret.

Generalized Linear Models

Input

```
> m <- lm(bweight ~ hyp, data=births)
> m <- glm(bweight ~ hyp, family=gaussian, data=births)
```

give the same answer. The model formula is the same for both, but for `glm` it is necessary to specify the family of likelihoods which will be used to fit the model.
The `glm` function allows us to fit other models including logistic regression and Poisson regression.

Predicting Low Birth Weight

We are more interested in predicting birth weight under 2500g (`lowbw`). This requires a model where the outcome is not metric, but binary. For a binary response we use a `glm` with a *binomial* family.

Input/Output

```
> m <- glm(lowbw ~ hyp, family=binomial, data=births)
> ci.lin(m, Exp=T)[,5:7]
             exp(Est.)       2.5%     97.5%
(Intercept) 0.1030928 0.07445162 0.1427521
hyphyper    3.7307692 2.02747522 6.8650107
```

This returns estimates of the log odds (Intercept) or log odds ratios (for the parameters). To present the results in terms of odds ratios we use the `Exp=TRUE` option to `ci.lin`.

Controlling

Controlling the effect of `hyp` on `lowbw` for `sex`

Input/Output

```
> m <- glm(lowbw ~ hyp+sex, family=binomial, data=births)
> ci.lin(m,Exp=T)
               exp(Est.)
(Intercept)    0.0813691
hyphyper       3.9060041 (hyp controlled for sex)
sexF           1.5641095 (sex controlled for hyp)
```

When you control for a variable you are assuming that any interaction can be ignored.

Interaction (effect modification)

Input/Output

```
> m <- glm(lowbw ~ hyp + sex + hyp:sex,
+            family=binomial, data=births)
> ci.lin(m,Exp=T)[,5:7]
                exp(Est.)
(Intercept)   0.07281553
hyphyper      5.31612903
sexF          1.88644689
hyphyper:sexF 0.52168285
```

Alternatively, use

Input/Output

```
m <- glm(lowbw ~ hyp*sex, family=binomial, data=births)
```

Testing for Interaction

Input/Output

```
> m1 <- glm(lowbw ~ hyp+sex, family=binomial, data=births)
> m2 <- glm(lowbw ~ hyp*sex, family=binomial, data=births)
> anova(m1,m2,test="Chisq")

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       497     348.34
2       496     347.29  1   1.0561   0.3041
```

The anova function conducts an *analysis of variance* – an old-fashioned name for a test of significance between two nested models.

Stratified Effects

When there is a strong interaction it may be best to report stratified effects. Omitting the main effect of `hyp` in an interaction model gives us the effect of `hyp` within strata of `sex`.

Input/Output

```
m <- glm(lowbw ~ sex + sex:hyp, family=binomial,
+                                data=births)
> ci.lin(m,Exp=T)[,5:7]
                exp(Est.)
(Intercept)   0.07281553 % 15/206 nur normale Jungen
sexF          1.88644689
sexM:hyphyper 5.31612903
sexF:hyphyper 2.77333333
```

Note that $2.77/5.32 = 0.52$ is the interaction term.

Looking Inside the Black Box

The paradigm is the model

$$\mu = \alpha + \beta X + \gamma Z + \cdots$$

where X,Z,$\cdots$ are numeric explanatory variables. In a glm μ is replaced by some function of *mu* such as $\log(\mu)$ (link function). When X is a factor, on (say) 3 levels, it is replaced by X_1, X_2, X_3, die indicator variables for the levels of X.
Predicted values for $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ are

level	X_1	X_2	X_3	$\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
1	1	0	0	$\alpha + \beta_1$
2	0	1	0	$\alpha + \beta_2$
3	0	0	1	$\alpha + \beta_3$

Too Many Parameters

Drop α

level	X_1	X_2	X_3	$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
1	1	0	0	β_1
2	0	1	0	β_2
3	0	0	1	β_3

β_1 is the mean response at level 1, β_2 at level 2, β_3 at level 3.
Drop X_1

level	X_2	X_3	$\alpha + \beta_2 X_2 + \beta_3 X_3$
1	0	0	α
2	1	0	$\alpha + \beta_2$
3	0	1	$\alpha + \beta_3$

α is the mean response at level 1
β_2 und β_3 are the effects of levels 2 and 3 vs level 1. These are called *treatment contrasts*.

Two Factors

X on 3 levels, Z on 2 levels

$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 Z_1 + \gamma_2 Z_2$$

X_1, X_2, X_3 are the indicators for X and Z_1, Z_2 are the indicators for Z. Omitting X_1 and Z_1 the model becomes

$$\mu = \alpha + \beta_2 X_2 + \beta_3 X_3 + \gamma_2 Z_2$$

with predicted means

		Z	
		1	2
X	1	α	$\alpha + \gamma_2$
	2	$\alpha + \beta_2$	$\alpha + \beta_2 + \gamma_2$
	3	$\alpha + \beta_3$	$\alpha + \beta_3 + \gamma_2$

Interaction

Effect of Z the same at each level of X:

		Z	
		1	2
	1	α	$\alpha + \gamma_2$
X	2	$\alpha + \beta_2$	$\alpha + \beta_2 + \gamma_2$
	3	$\alpha + \beta_3$	$\alpha + \beta_3 + \gamma_2$

Effect of Z differs at different levels of X:

		Z	
		1	2
	1	α	$\alpha + \gamma_2$
X	2	$\alpha + \beta_2$	$\alpha + \beta_2 + \gamma_2 + \delta_{22}$
	3	$\alpha + \beta_3$	$\alpha + \beta_3 + \gamma_2 + \delta_{32}$

The δ parameters measure how much the effect of Z changes.

Nested or Stratified Effects

A slightly different way of parameterizing the model gives stratified effects:

		Z	
		1	2
X	1	β_1	$\beta_1\delta_{12}$
	2	β_2	$\beta_2 + \delta_{22}$
	3	β_3	$\beta_3 + \delta_{32}$

Same number of parameters as for interaction, but the δ's now measure the effects of Z at each level of X. In `R` this would be produced by the model formula `Y ~ -1 + X + X:Z`