Linear Models Part II

Mandy Vogel

University Leipzig

July 19, 2015

Overview

Recap

- Simple Linear Model
- R-squared
- Visualization

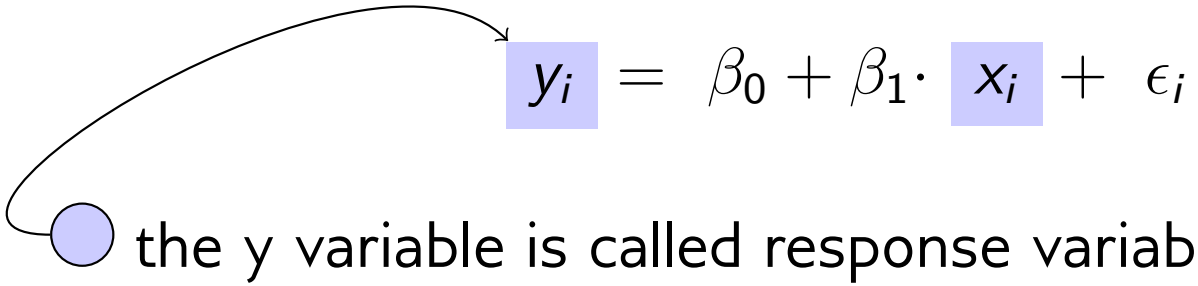
Exercises

Table of Contents I

Simple Linear Model

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Simple Linear Model

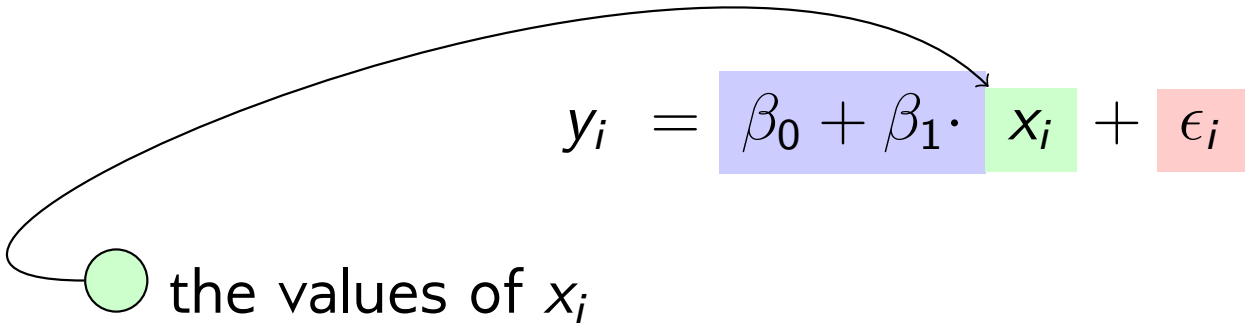


Simple Linear Model

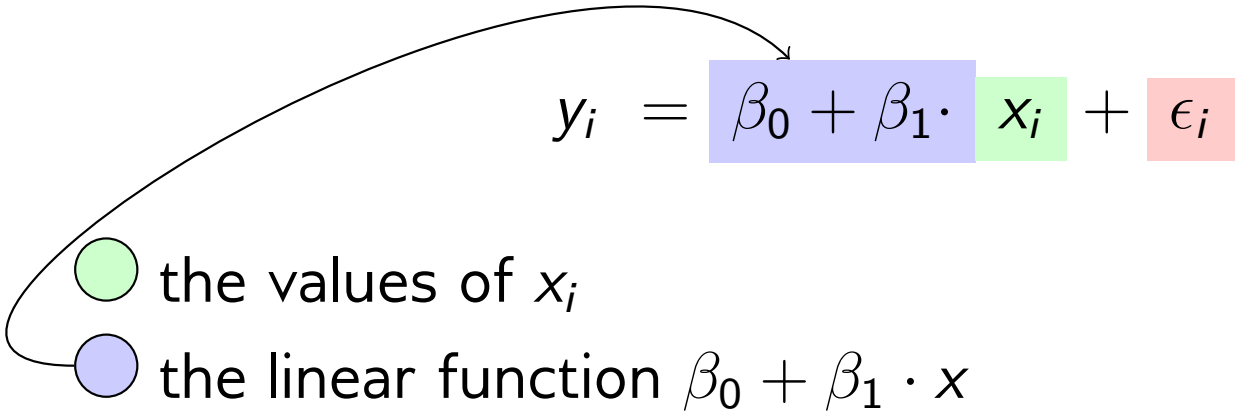
$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

- the y variable is called response variable
- the x variable is called the predictor variable, covariate, or regressor

y_i depends on



y_i depends on



y_i depends on

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

- the values of x_i
- the linear function $\beta_0 + \beta_1 \cdot x$
- the value of the random variable $\epsilon_i \sim \mathcal{N}(0, \sigma)$

In R this becomes

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

`lm( y ~ x )`

Remember

- we had birth weight as response variable (`bweight`)
- gestational age as a continous explanatory variable (`gestwks`)
- hypertension status as discrete explanatory variable with two levels (yes and no) (`hyp`)

Our Example I

- one continous explanatory variable
- so we estimate one intercept (β_0 from above) and one slope (β_1 from above)

Our Example I

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

R-squared

- the R^2 is the cœfficient of determination
- in case of a simple linear regression $R^2 = r^2$ where r is the correlation cœfficient
- more general:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \overline{y})^2} = \frac{\sum (\widehat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}$$

- that means $R^2 \cdot 100\%$ of the variation is explained by the model
- the adjusted R^2 is adjusted for the number of cœfficients (penalize models for having many covariates)

Our Example I - visualization

- we build first a scatterplot (`geom_abline()`)
- extract the cœfficients from the model (`coef(m)`)

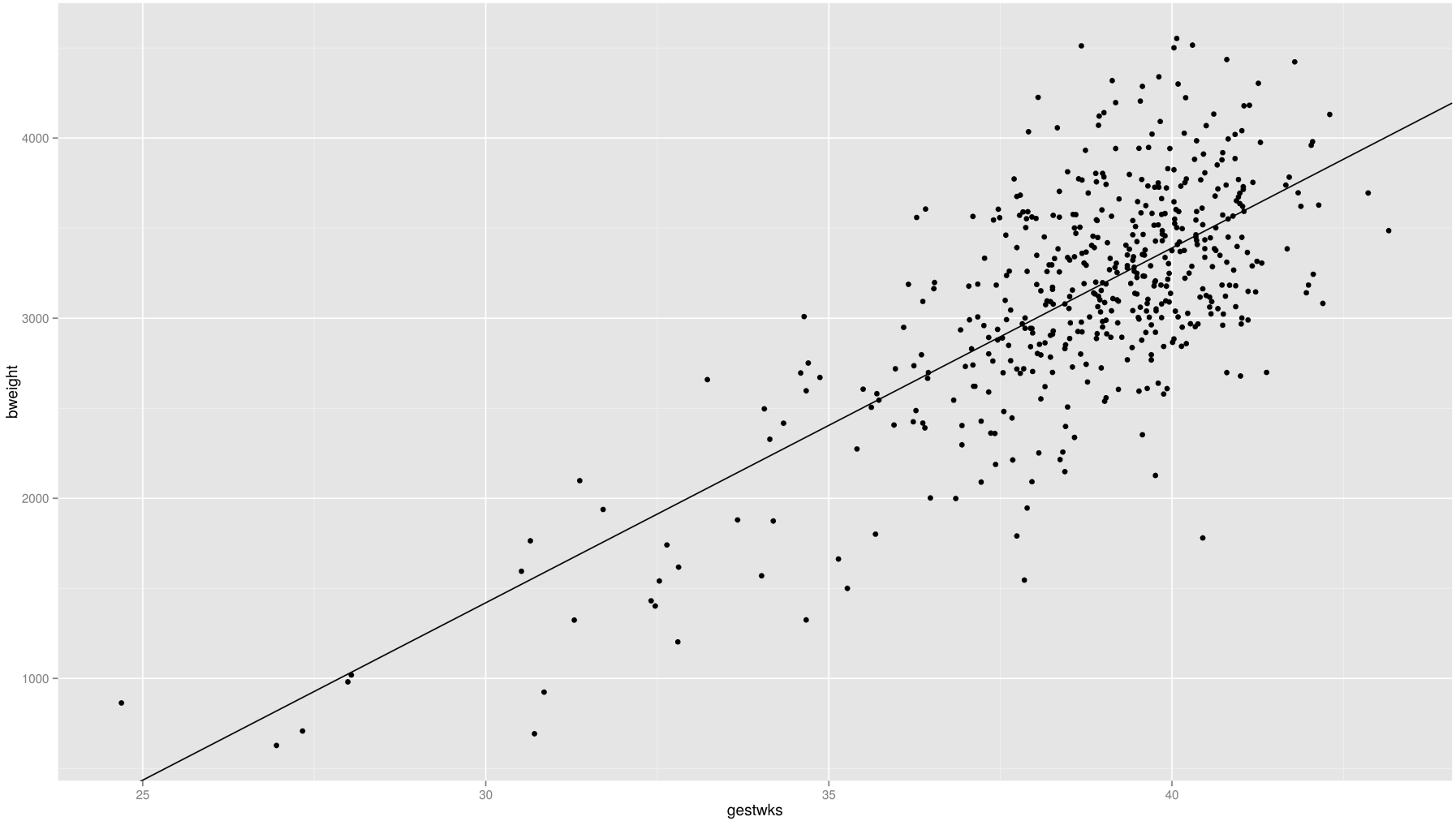
```
> coef(m)
(Intercept)     gestwks
 -4489.1398    196.9726
```

- using `geom_abline()` which take the intercept and the slope as arguments

Our Example I - visualization

```
> ggplot(births, aes(x = gestwks,y = bweight)) +
+     geom_point() +
+     geom_abline(intercept = coef(m)[1],
+                 slope = coef(m)[2])
Warnmeldung:
Removed 10 rows containing missing values (geom_point).
```

Our Example I - visualization



The effect of `gestwks` is the slope of the line

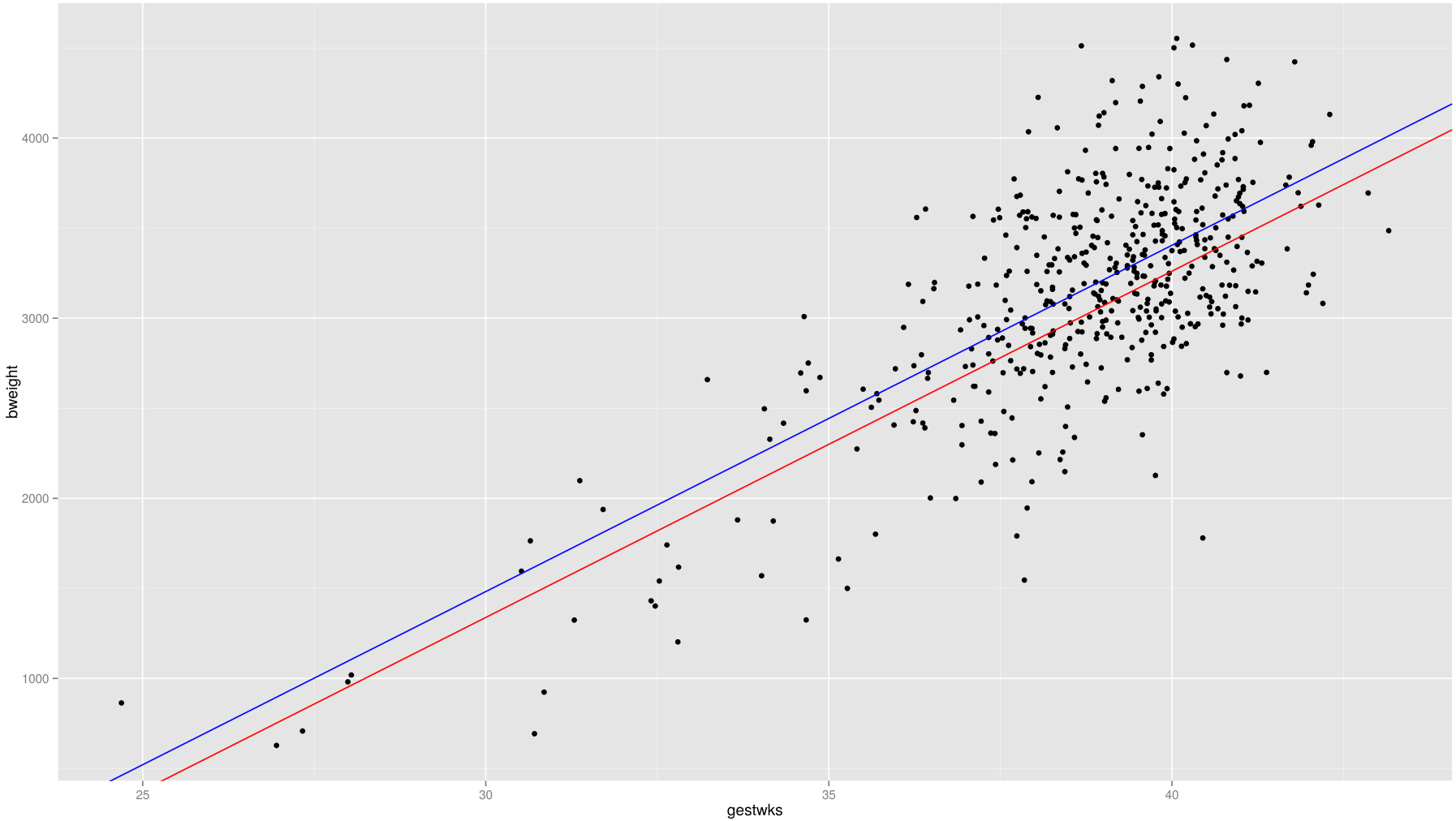
Our Example II

- we add one discrete explanatory variable
- so we estimate two intercepts, one for each level of `hyp`
- and one slope (β_1 from above)

Our Example II

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-1711.04  -283.13    -9.86   283.92  1361.22

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4285.002    349.322 -12.267   <2e-16 ***
hyphyper     -143.675     58.820  -2.443   0.0149 *
gestwks       192.238      8.956  21.465   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 447.5 on 487 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.5132, Adjusted R-squared:  0.5112
F-statistic: 256.7 on 2 and 487 DF,  p-value: < 2.2e-16
```

Our Example II - visualization



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.

Our Example III

- we add the interaction between `hyp` and `gestwks`
- so we estimate two intercepts and two slopes, one of each for each level of `hyp`

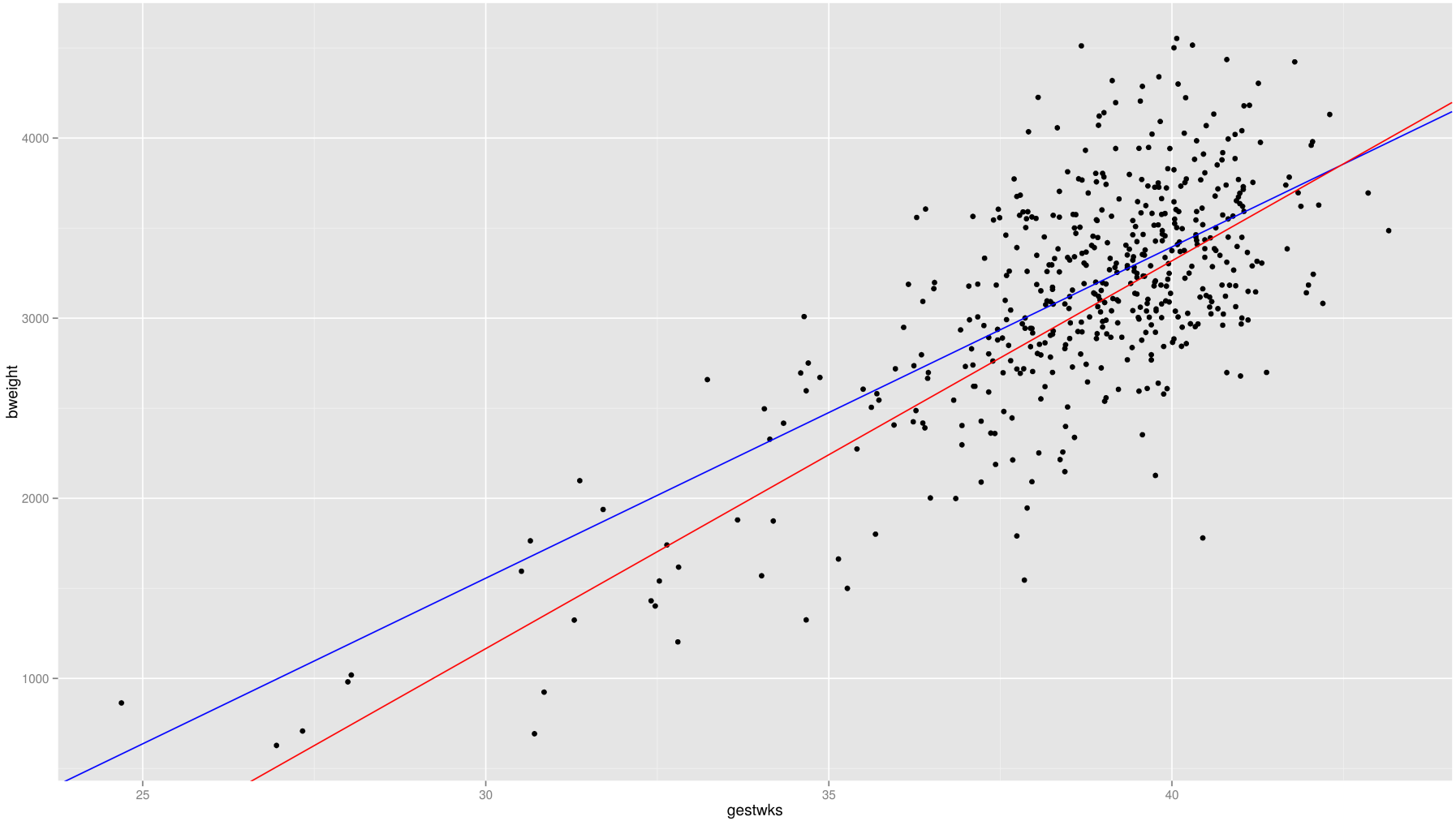
Our Example III

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-1698.36  -291.13    -5.14   284.48  1359.16

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3960.82     406.76  -9.738   <2e-16 ***
hyphyper         -1332.66     769.60  -1.732    0.084 .
gestwks            183.91      10.43  17.626   <2e-16 ***
hyphyper:gestwks    31.39      20.26   1.549    0.122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 446.8 on 486 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.5156, Adjusted R-squared:  0.5126
F-statistic: 172.4 on 3 and 486 DF,  p-value: < 2.2e-16
```

Our Example III

```
> coef(m)
     (Intercept)          hyphyper           gestwks hyphyper:gestwks
      -3960.8157       -1332.6563          183.9105           31.3851
> ggplot(births, aes(x = gestwks,y = bweight)) +
+     geom_point() +
+     geom_abline(intercept = coef(m)[1],
+                 slope = coef(m)[3],colour="blue") +
+     geom_abline(intercept = coef(m)[1] + coef(m)[2],
+                 slope = coef(m)[3] + coef(m)[4],colour="red")
```

Our Example III - visualization



The effect of `gestwks` differs between groups.

Our Example III

- because of the difficulty to explain the cœfficients we first center the variable `gestwks`

```
> births$gwsc <- births$gestwks-40
> m <- lm(bweight ~ hyp * gwsc, data=births)
> coef(m)
  (Intercept)      hyphyper          gwsc hyphyper:gwsc
   3395.60329     -77.25215     183.91048      31.38510
```

Table of Contents I

Exercises I

1. For the Cars93 (MASS) data set, answer the following:
   1.1 For MPG. highway modeled by Horsepower, find the simple regression cœfficients. What is the predicted mileage for a car with 225 horsepower?
   1.2 Fit the linear model with MPG. highway modeled by Weight. Find the predicted highway mileage of a 6,400 pound HUMMER H2 and a 2,524 pound MINI Cooper.
   1.3 Fit the linear model Max .Price modeled by Min .Price. Why might you expect the slope to be around 1 ?

   Can you think of any other linear relationships among the variables?
2. For the data set MLBattend (UsingR) concerning major league baseball attendance, fit a linear model of attendance modeled by wins. What is the predicted increase in attendance if a team that won 80 games last year wins 90 this year?

Exercises II

3. People often predict children's future height by using their 2-year-old height. A common rule is to double the height. Table 10.2 contains data for eight people's heights as 2-year-olds and as adults. Using the data, what is the predicted adult height for a 2-year-old who is 33 inches tall?

Age 2 (inch)	39	30	32	34	35	36	36	30
Adult (in.)	71	63	63	67	68	68	70	64