Chapter 10: Multiple linear regression models

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about:

- Models with several explanatory variables
- Exploring pairwise relationships between all variables
- Multiple linear regression and the problem of multi-collinearity
- Fixing multi-collinearity
- Relevant `R`-code.

Section 10.1
Example: Modelling birth weights using several explanatory variables

Multiple explanatory variables

We have learned how to model the effects of numeric and/or factor explanatory variables using linear models.

More generally, we can (in principle) fit as many explanatory variables as we like. However, we shall see that this is not always a good idea.

Caution needs to be applied.

By way of example, let us examine which variables might explain the birth weight of babies.

Example: Birth weight of babies

`bwt`	birth weight in ounces (=28.35gm)
`gestation`	length of pregnancy in days
`not.first.born`	0=first born, 1=not first-born
`age`	mother's age in years
`height`	mother's height in inches
`weight`	mother's pre-pregnancy weight in pounds
`smoke`	smoking status of mother 0=not, 1=smoker.

The response variable is the baby's birth weight (`bwt`).

This dataset[1] was obtained from
http://www.stat.berkeley.edu/users/statlabs/labs.html.

The dataset has 1174 observations.

[1] It accompanies the excellent text Stat Labs: Mathematical Statistics through Applications Springer-Verlag (2001) by Deborah Nolan and Terry Speed.

Section 10.2
Exploring relationships between the variables

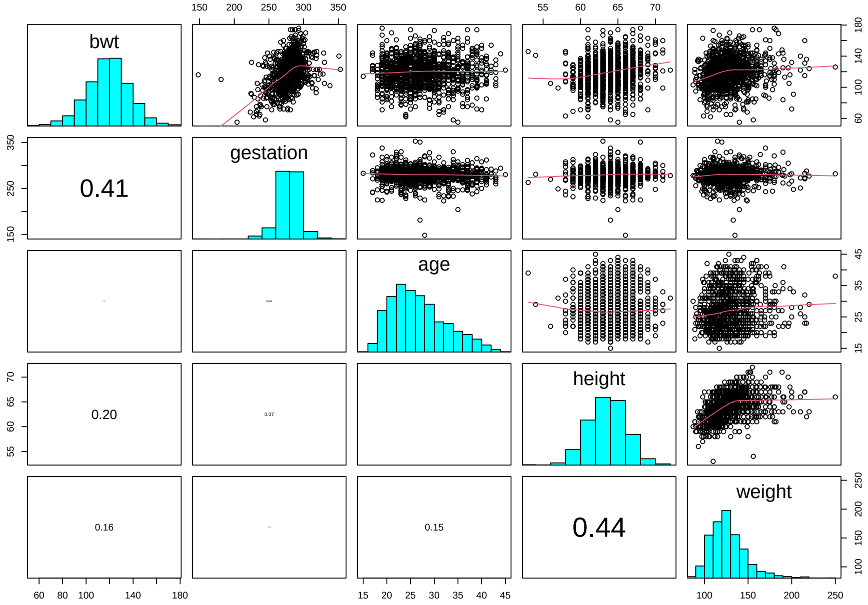
Birth weight of babies

Let us first inspect the relationships between the numerical explanatory variables and the response variable. The numeric explanatories are `gestation`, `age`, `height` and `weight`.

The five variables are in columns 1,2,4,5 and 6 in the data frame `Babies.df`.

```
> ## Invoke the s20x library
> library(s20x)
> ## Importing data into R
> Babies.df = read.table("Data/babies_data.txt", header=T)
> ## Create the pairs plot of the five numeric variables
> pairs20x(Babies.df[,c(1,2,4,5,6)])
```

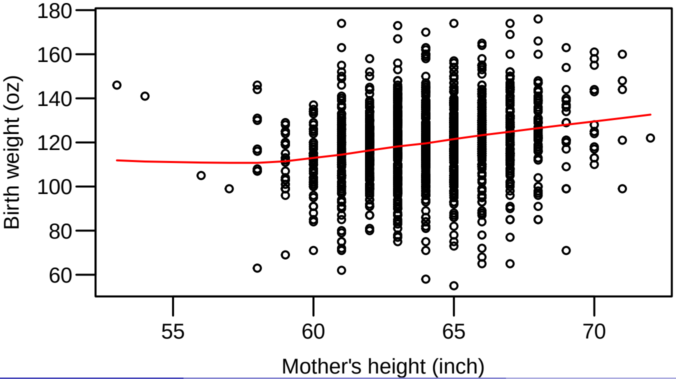
Birth weight of babies...



Birth weight of babies. . .

pairs20x gives a histogram of each variable in the diagonal cells. Above the diagonal, in the (i, j) cell $(i < j)$ it gives scatter plots of variable i (y-axis) against variable j (x-axis). To illustrate, variable 1 is bwt and variable 4 is the mother's height (height). The scatter plot in cell (1,4) is

```
> plot(bwt ~ height,data = Babies.df,
+      xlab="Mother's height (inch)",ylab="Birth weight (oz)")
> lines(lowess(Babies.df$height,Babies.df$bwt))
```



Birth weight of babies...

The correlation coefficient between height and bwt is in cell (4,1). It is 0.20, indicating that a straight-line relationship is, at best, weak.

This correlation coefficient can only measure the strength of a straight line relationship between x (height) and a y (bwt). It can be useful but can, on occasion, be misleading. In other words, look at the scatter plot and use it only if the relationship looks like a straight line.

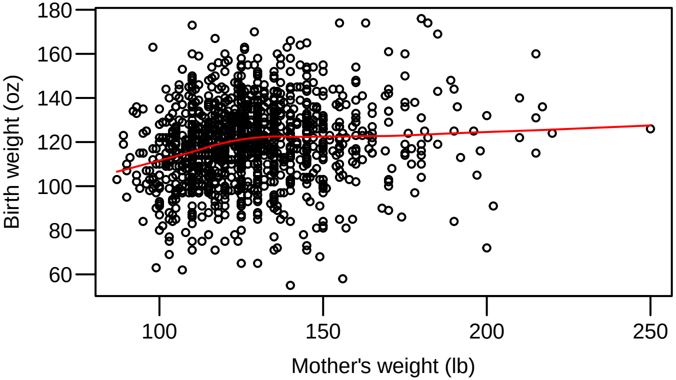
Note: In a simple linear regression of y on x, the resulting R^2 value is the square of the sample correlation coefficient. To illustrate:

```
> summary(lm(bwt ~ height,data = Babies.df))$r.squared
[1] 0.04149539
> cor(Babies.df$bwt,Babies.df$height)^2
[1] 0.04149539
```

Birth weight of babies...

Looking at the pairs plot again, we also see a somewhat weak relationship between bwt and mother's weight.

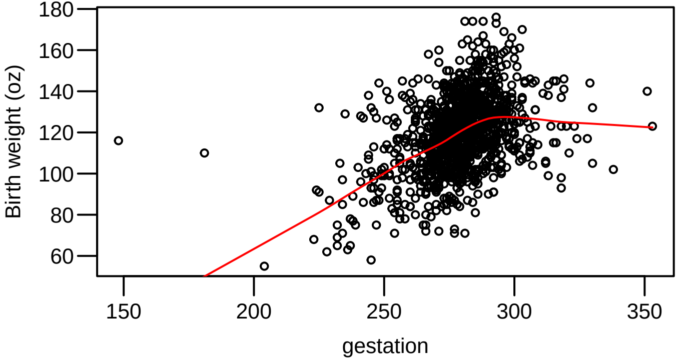
```
> plot(bwt ~ weight,data = Babies.df,
+          xlab="Mother's weight (lb)",ylab="Birth weight (oz)")
> lines(lowess(Babies.df$weight,Babies.df$bwt))
```



Birth weight of babies...

There is a stronger relationship between the `gestation` time for the babies and its `bwt` which is not surprising, as the longer the child is in the mother's womb the longer the child has had time to have nutrition and grow. But, this relationship distinctly flattens out beyond a certain gestational age – some people call this a "hockey stick" curve.

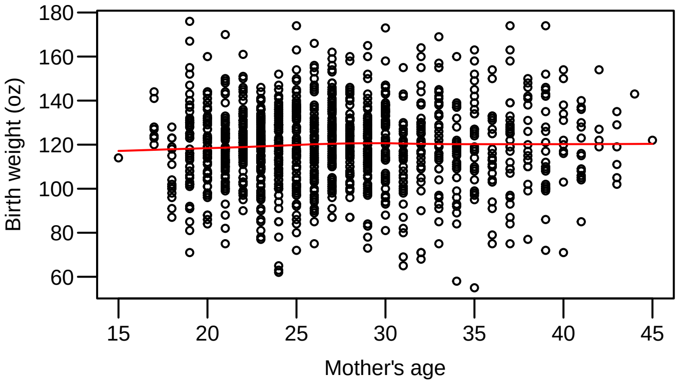
```
> plot(bwt ~ gestation,data = Babies.df,ylab="Birth weight (oz)")
> lines(lowess(Babies.df$gestation,Babies.df$bwt))
```



Birth weight of babies...

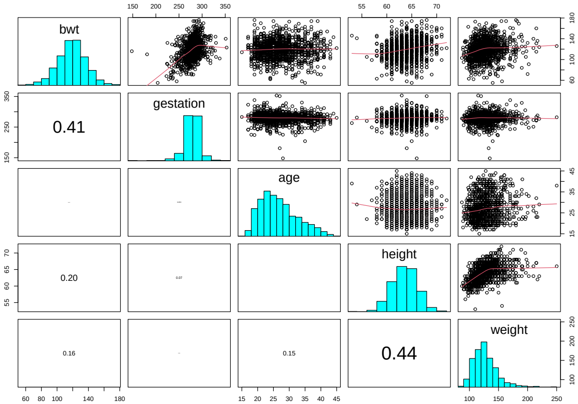
There does not seem to be any relationship between a mother's age and her child's `bwt`.

```
> plot(bwt ~ age,data = Babies.df,xlab="Mother's age",ylab="Birth weight (oz)")
> lines(lowess(Babies.df$age,Babies.df$bwt))
```



Birth weight of babies. . .

Note: There seem to be some outlying data points in these plots. There does not appear to be much of a relationship between the x variables, except between `height` and `weight`.

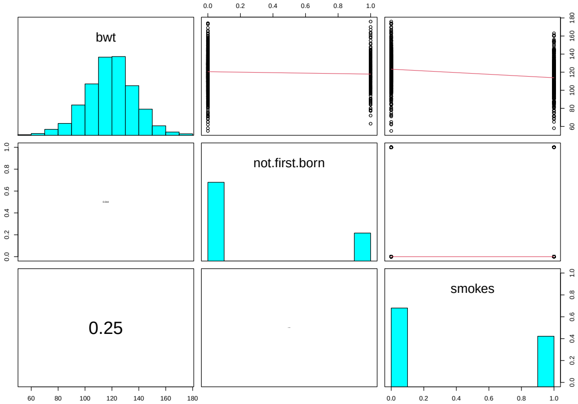


Birth weight of babies. . .

Let us look at the categorical (factor) explanatory variables against the baby's birth weight `bwt`.

The categorical variables are `not.first.born` and `smoke`, in columns 3 and 7 of the data frame `Babies.df`.

```
> pairs20x(Babies.df[,c(1,3,7)])
```



Birth weight of babies...

We see a slight decrease in babies `bwt` if the mother smokes. This increases the chance of a mother having a low birth weight baby if she smokes – perhaps another reason to avoid tobacco!

The variable `not.first.born` does not appear to have too much of an effect. This is perhaps not a surprise given that this variable may not be as important as it once was as family size has deceased markedly in the developed world (this is US data) and prenatal care has improved.

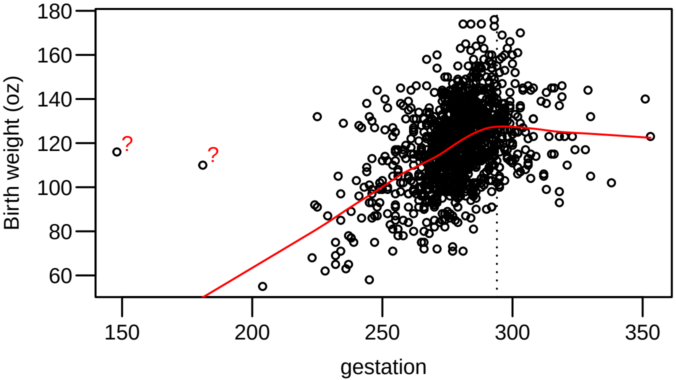
We will know begin our linear modelling of these data...

Birth weight of babies. . .

Relationship between birth weight and gestational age

Let us start with an understanding of `gestation` to explain `bwt` since it is the strongest relationship. The atypical data points have been marked with question marks. We will add other explanatory variables later.

```
> plot(bwt ~ gestation, data = Babies.df, ylab="Birth weight (oz)" )
> lines(lowess(Babies.df$gestation, Babies.df$bwt), col = "red")
> text(c(152, 185),c(120, 115), "?", col = "red")
> abline(v = 294,lty = 3)
```



Birth weight of babies. . .

Relationship between birth weight and gestational age. . .

Let us identify the two points denoted by the '?' symbol.

We can easily identify them in the plot as they have `gestation` < 200.

They look extremely implausible as they have typical birth-weight but have a gestational age that is extremely low for these data.

```
> id=(Babies.df$gestation<200)
> Babies.df[id,]
    bwt gestation not.first.born age height weight smokes
239 116       148              0  28     66    135      0
820 110       181              0  27     64    133      0
```

These points (observations 239 and 820) may be be unduly influential.

Birth weight of babies...

Relationship between birth weight and gestational age...

The above plot has a vertical line at 294 days. The relevance of 294 days is explained in the article "How Your Baby Grows During Pregnancy".

Most babies are born before 42 weeks $= 42 \times 7 = 294$ days. It seems that beyond this point babies cease to grow and hence the 'flattening out' and/or decrease. In other words, it looks like the effect of gestational age depends on whether the baby is overdue or not. That is, the effect of gestational age appears to change with overdue status.

We want to fit a model that fits a straight line for `gestation` ≤ 294 and then changes the slope of that line when `gestation` > 294.

We'll need to put our statistical thinking caps on, and devise a way to fit such a model.

Birth weight of babies...

Relationship between birth weight and gestational age...

For $\texttt{gestation} \leq 294$ days we'll use the familiar simple linear regression model

$$\mathsf{E}[\texttt{bwt}] = \beta_0 + \texttt{gestation} \times \beta_1$$

We'd like to extend this model by adding an extra term so that the slope changes when $\texttt{gestation} > 294$. That is,

$$\mathsf{E}[\texttt{bwt}] = \beta_0 + \texttt{gestation} \times \beta_1 + v \times \beta_2$$

where v is some suitable explanatory variable. What should v be?

- For $\texttt{gestation} \leq 294$ the extended model is just the simple linear regression model, so that means $v = 0$ when $\texttt{gestation} \leq 294$.
- For $\texttt{gestation} > 294$ we need another slope effect for gestational age. In fact, we need $v = \texttt{gestation} - 294$.[2]

[2] We substract the 294 so that the simple linear regression model and extended model have the same value when $\texttt{gestation} = 294$, because then $v = 0$.

Birth weight of babies...

Relationship between birth weight and gestational age...

Let's create the new explanatory $v = \texttt{gestation} - 294$ that is described above. We'll give it the name `ODdays` because it is the number of days that the baby is overdue.

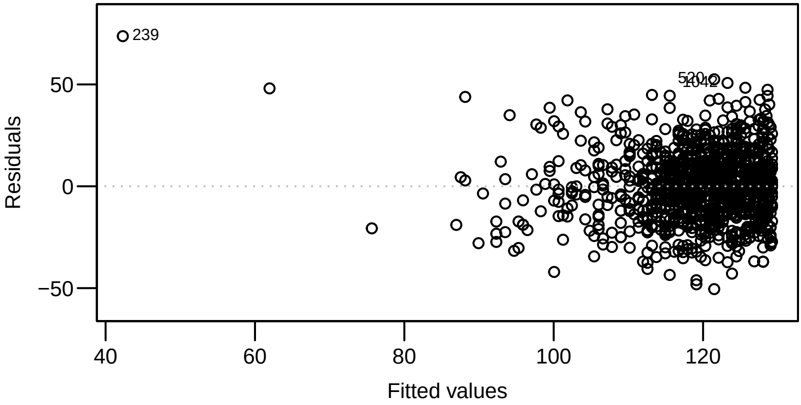
```
> head(Babies.df,12) #Print first 12 lines of dataframe
   bwt gestation not.first.born age height weight smokes ODdays
1  120       284              0  27     62    100      0      0
2  113       282              0  33     64    135      0      0
3  128       279              0  28     64    115      1      0
4  108       282              0  23     67    125      1      0
5  136       286              0  25     62     93      0      0
6  138       244              0  33     62    178      0      0
7  132       245              0  23     65    140      0      0
8  120       289              0  25     62    125      0      0
9  143       299              0  30     66    136      1      5
10 140       351              0  27     68    120      0     57
11 144       282              0  32     64    124      1      0
12 141       279              0  23     63    128      1      0
```

Section 10.3
Fitting the initial model

Birth weight of babies. . .

Our initial fitted model is the hockey stick model for the effect of gestational age.

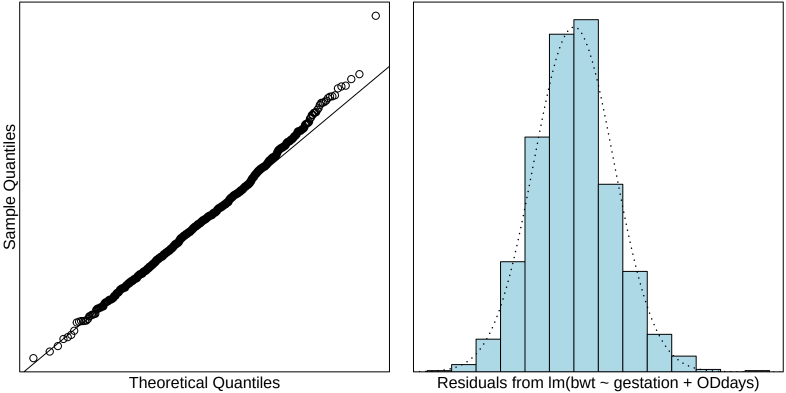
```
> bwt.fit=lm(bwt~ gestation+ODdays, data = Babies.df)
> plot(bwt.fit, which = 1, add.smooth = FALSE)
```



Observation 239 is a problem.

Birth weight of babies...

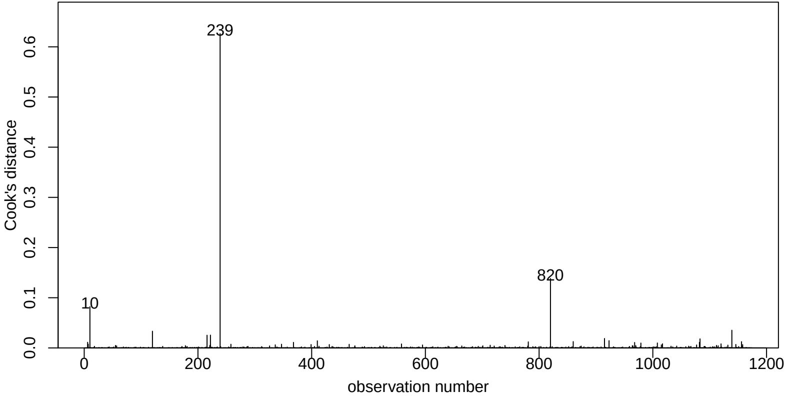
```
> normcheck(bwt.fit)
```



Other than observation 239, things look pretty good.

Birth weight of babies...

```
> cooks20x(bwt.fit)
```

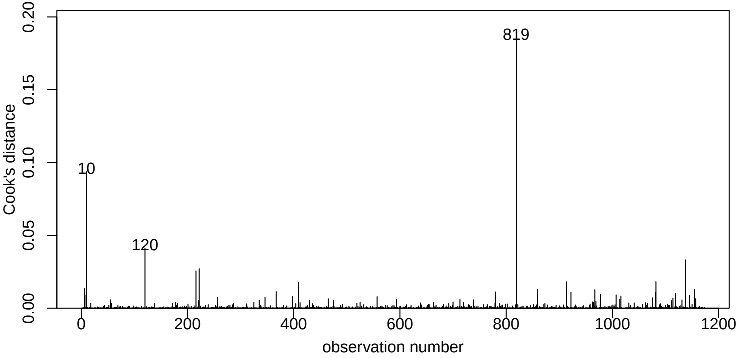


Point 239 is unduly influential. This baby has a gestational age of just 148 days, and yet has a weight typical of a full term baby. It is clearly a data-entry mistake and we will remove this data point.

Birth weight of babies...

Let us refit with observation 239 removed.

```
> bwt.fit2=lm(bwt~ gestation+ODdays,data = Babies.df[-239,])
> cooks20x(bwt.fit2)
```



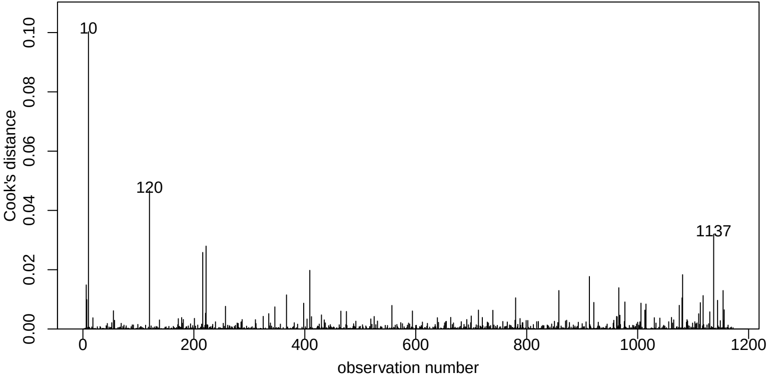
Although observation 820 [3] is not unduly influential, we shall make a judgement call, and remove it.

[3] Note that it is now identified as point 819 in this plot, but it was point 820 before we dropped point 239.

Birth weight of babies...

We refit the model using the reduced data.

```
> #This time we demonstrate using the subset argument to remove points
> bwt.fit3=lm(bwt~ gestation+ODdays,data = Babies.df, subset = -c(239, 820))
> cooks20x(bwt.fit3)
```

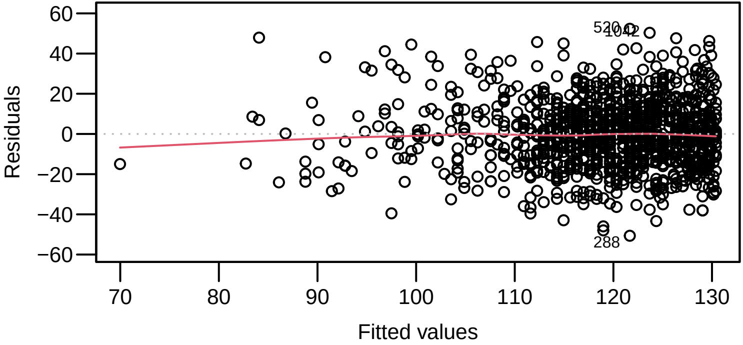


Now we have no unduly influential data points.

Birth weight of babies...

Let us recheck the residuals now that we have removed these two points.

```
> plot(bwt.fit3,which=1)
```

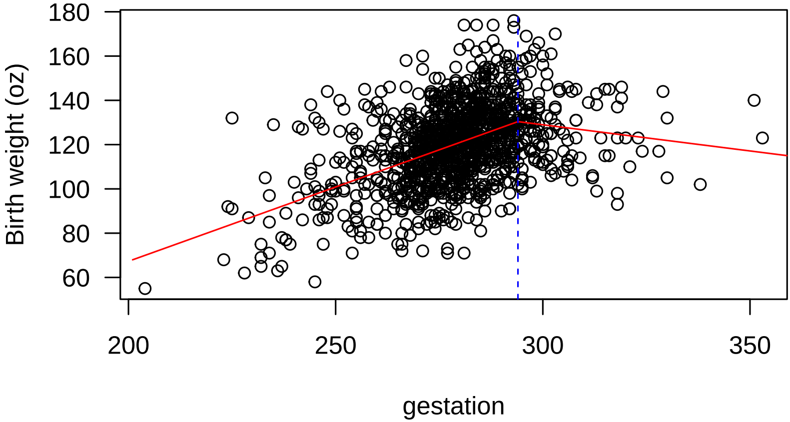


EOV seems fine now, and the residuals seem to be centred around zero.

Birth weight of babies. . .

Let's take a look at our fitted hockey stick model.

```
> gestation.seq=201:360 #Explanatory values at which to get predictions
> ODdays.seq=ifelse(gestation.seq<=294,0,gestation.seq-294)
> fit.seq=predict(bwt.fit3,new=data.frame(gestation=gestation.seq,
+                                          ODdays=ODdays.seq))
> plot(bwt~gestation,data=Babies.df[-c(239, 820),],ylab="Birth weight (oz)")
> lines(gestation.seq,fit.seq,col="red"); abline(v=294,lty=2,col="blue")
```



Model checks are good and no influential points remain, so we can trust this fit. Let's interpret the output.

```
> summary(bwt.fit3)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -66.95336   10.42810   -6.42 1.97e-10 ***
gestation     0.67124    0.03757   17.87  < 2e-16 ***
ODdays       -0.90783    0.11745   -7.73 2.31e-14 ***
---
Residual standard error: 16.23 on 1169 degrees of freedom
Multiple R-squared:  0.2188,Adjusted R-squared:  0.2174
F-statistic: 163.7 on 2 and 1169 DF,  p-value: < 2.2e-16
```

The fitted model is:

$$E[\texttt{bwt}] = -66.95 + 0.67 \times \texttt{gestation} - 0.91 \times \texttt{ODdays}$$

Birth weight of babies...

So, for $\texttt{gestation} \leq 294$ days (i.e., $\texttt{ODdays = 0}$)

$$E[\texttt{bwt}] = -66.95 + 0.67 \times \texttt{gestation}$$

That is, on average, babies initially grow at 0.67 oz per day until about 130 oz[4] at week 42 (i.e., day 294).

For $\texttt{gestation} > 294$ days (i.e., $\texttt{ODdays = gestation-294}$)

$$\begin{aligned} E[\texttt{bwt}] &= -66.95 + 0.67 \times \texttt{gestation} - 0.91 \times (\texttt{gestation-294}) \\ &= 199.95 - 0.24 \times \texttt{gestation} \end{aligned}$$

So, on average, it is estimated that overdue babies lose about 0.24 oz per day after week 42.[5]

---

[4] $130 \approx -66.95 + 0.67 \times 294$

[5] **Question:** How could we test whether this is significantly different from zero?

Birth weight of babies...

Note that this model only explains about 22% of the variation in babies' birth weight, so it would be worth seeing if adding the other explanatory variables will help explain more.

In the `pairs20x` plot above we saw that `height` and `weight` had correlations of 0.20 and 0.16 with `bwt`.

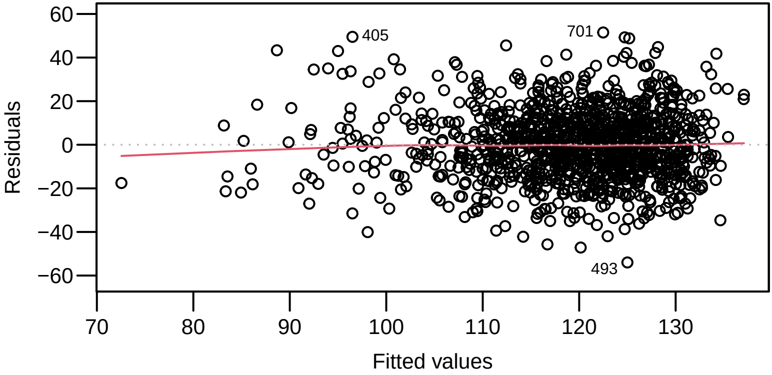
So let us see what we find when we introduce the `height` variable into the model. We will proceed with selecting variables one at a time (with reflection) – this is one of many multiple regression strategies!

Section 10.4
Multiple linear regression model:
Adding more terms to the model and the peril of multi-collinearity

Birth weight of babies. . .

Let us add the height variable and see how it works out.

```
> bwt.fit4 = lm(bwt ~ gestation + ODdays + height, data = Babies.df,
+      subset = -c(239,820))
> plot(bwt.fit4,which=1)
```



All seems okay. Let us make sure that this makes sense in terms of output.

Birth weight of babies...

```
> summary(bwt.fit4)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -139.20571   15.05961  -9.244  < 2e-16 ***
gestation      0.65219    0.03703  17.613  < 2e-16 ***
ODdays        -0.89039    0.11543  -7.714 2.61e-14 ***
height         1.21083    0.18495   6.547 8.79e-11 ***
---
Residual standard error: 15.94 on 1168 degrees of freedom
Multiple R-squared:  0.2464,Adjusted R-squared:  0.2445 
F-statistic: 127.3 on 3 and 1168 DF,  p-value: < 2.2e-16
```

This seems to make sense, whereby mother's height is positively related to a baby's birth weight (on average).

Note: We will drop the checking of fitted vs residuals plots as it has been okay to date and it is starting to get a little tedious. We will recheck this once we get to the final model.

Birth weight of babies...

Let us add `weight` to the model. We're going to save some typing and use the `update` function to update our model.[6]

```
> bwt.fit5 = update(bwt.fit4,~. + weight)
> summary(bwt.fit5)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -131.68169   15.14974  -8.692  < 2e-16 ***
gestation      0.65624    0.03688  17.795  < 2e-16 ***
ODdays        -0.90868    0.11502  -7.900 6.41e-15 ***
height         0.90486    0.20453   4.424 1.06e-05 ***
weight         0.08535    0.02485   3.434 0.000615 ***
---
Residual standard error: 15.87 on 1167 degrees of freedom
Multiple R-squared:  0.254,Adjusted R-squared:  0.2514 
F-statistic: 99.32 on 4 and 1167 DF,  p-value: < 2.2e-16
```

This makes sense. Heavier mothers can be expected to have heavier babies.

[6] In the above use of `update` the `~.` term is used to denote the model containing the explanatory variables in `bwt.fit4`.

Birth weight of babies...

The mother being very underweight or excessively overweight can have negative effects on their babies health, but neither `height` or `weight` directly measures this.

We will construct a new variable, body mass index `bmi`.

$$BMI = \frac{\texttt{mass in kg}}{\texttt{height in metres}^2} = \frac{\texttt{mass in lb}}{\texttt{height in inches}^2} \times 703$$

The World Health Organisation classifies BMIs in the range 18.5–25 as healthy, 25–30 as overweight, and 30+ as obese.

Birth weight of babies. . .

Let us add `bmi` to the current model.

```
> # Create the variable BMI and add it to the model
> Babies.df$bmi = (Babies.df$weight / (Babies.df$height^2) ) * 703
> bwt.fit6 = update(bwt.fit5,~. + bmi)
> summary(bwt.fit6)
```

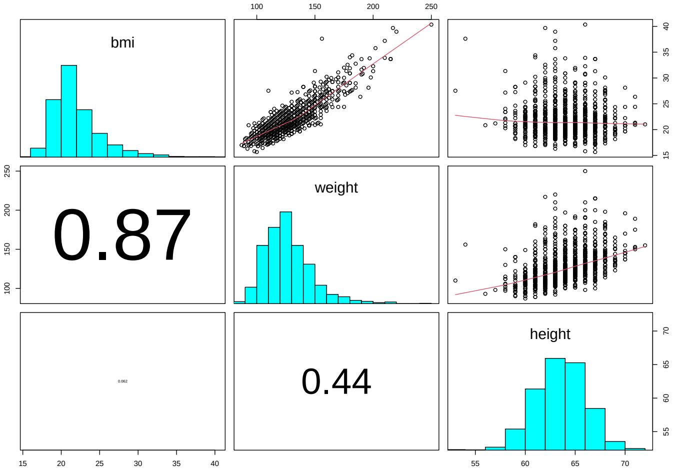
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -216.33575   79.63707  -2.717  0.00669 **
gestation      0.65629    0.03688  17.798  < 2e-16 ***
ODdays        -0.90980    0.11502  -7.910 5.94e-15 ***
height         2.22845    1.23940   1.798  0.07243 .
weight        -0.24252    0.30382  -0.798  0.42490
bmi            1.90870    1.76280   1.083  0.27914
---
Residual standard error: 15.87 on 1166 degrees of freedom
Multiple R-squared:  0.2547,Adjusted R-squared:  0.2515
F-statistic:  79.7 on 5 and 1166 DF,  p-value: < 2.2e-16
```

Hang on. Everything has gone weird!!! None of `weight`, `height` or `bmi` is statistically significant (at the 5% level). So what is going on?

Birth weight of babies...

Let's look at these three variables to see what is happening.

```
> pairs20x(Babies.df[-c(239,820), c(9,6,5)])
```



Not surprisingly, we see that `bmi` and `weight` seem to explain each other.

Birth weight of babies…

The problem is that we have a redundancy in our explanatory variables. Here, `bmi` is explained by `weight` and vice-versa. Note that adding `bmi` to the model barely changed R^2 and so is telling us that it did not increase our ability to explain variability in birth weight.

In essence the statistically significance (i.e., P-value) of an explanatory variable is measuring its contribution toward explaining variability in the response variable (in our case `bwt`) *having adjusted for any other explanatory variables in the model*.

So `bmi` explains little variability in `bwt` since `weight` has already explained most of that variability, and vice-versa.

This problem is given the name **multi-collinearity**[7].

In linear algebra, we say we have linear dependence (as opposed to linear independence) in these variables.

[7] The double 'l' is not a mistake.

Birth weight of babies. . .

Back to the drawing board. Let us refit this model with `bmi` and `height`, but without `weight`.

```
> bwt.fit7 = update(bwt.fit6,~. - weight)
> summary(bwt.fit7)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -153.99353   15.56725  -9.892  < 2e-16 ***
gestation      0.65633    0.03687  17.801  < 2e-16 ***
ODdays        -0.90933    0.11500  -7.907 6.06e-15 ***
height         1.25013    0.18440   6.779 1.91e-11 ***
bmi            0.50629    0.14415   3.512 0.000461 ***
---
Residual standard error: 15.87 on 1167 degrees of freedom
Multiple R-squared:  0.2543,Adjusted R-squared:  0.2518
F-statistic:  99.5 on 4 and 1167 DF,  p-value: < 2.2e-16
```

Let us next investigate whether the categorical variable (`smokes`) helps to explain further variability in `bwt`.

Birth weight of babies. . .

Let us add `smokes` to this analysis.

```
> bwt.fit8=update(bwt.fit7,~. + smokes)
> summary(bwt.fit8)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -144.07719   15.15079  -9.510  < 2e-16 ***
gestation      0.63198    0.03589  17.608  < 2e-16 ***
ODdays        -0.88104    0.11164  -7.892 6.84e-15 ***
height         1.28081    0.17898   7.156 1.46e-12 ***
bmi            0.41516    0.14029   2.959  0.00315 **
smokes        -7.93655    0.92711  -8.561  < 2e-16 ***
---
Residual standard error: 15.4 on 1166 degrees of freedom
Multiple R-squared:  0.2984,Adjusted R-squared:  0.2954
F-statistic: 99.18 on 5 and 1166 DF,  p-value: < 2.2e-16
```

As we might have suspected, a mother smoking is associated with decreased birth weight.

Birth weight of babies...

Let us see if `not.first.born` is useful:

```
> bwt.fit9=update(bwt.fit8,~. + not.first.born)
> summary(bwt.fit9)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -145.56797   15.08855  -9.648  < 2e-16 ***
gestation        0.64129    0.03583  17.897  < 2e-16 ***
ODdays          -0.89215    0.11119  -8.024 2.48e-15 ***
height           1.29912    0.17825   7.288 5.78e-13 ***
bmi              0.35469    0.14078   2.520 0.011882 *
smokes          -7.98201    0.92301  -8.648  < 2e-16 ***
not.first.born  -3.51137    1.02978  -3.410 0.000672 ***
---
Residual standard error: 15.33 on 1165 degrees of freedom
Multiple R-squared:  0.3053,Adjusted R-squared:  0.3018
F-statistic: 85.34 on 6 and 1165 DF,  p-value: < 2.2e-16
```

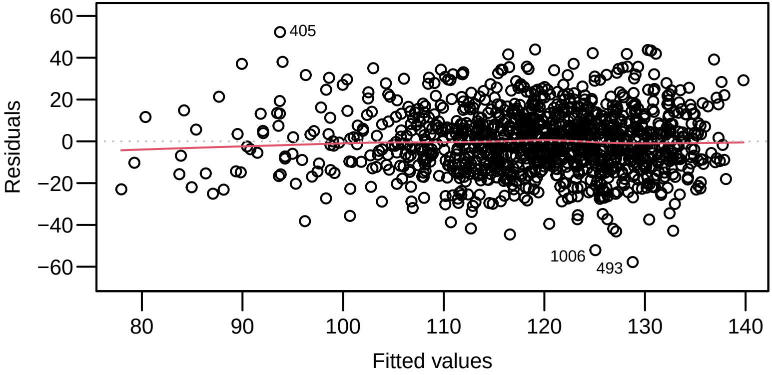
Hmmm, does the negative effect of `not.first.born` seem reasonable???

Birth weight of babies. . .

Let us check the assumptions on this final model:

Independence should be okay, as this is (hopefully) a random sample of data from a carefully designed study.

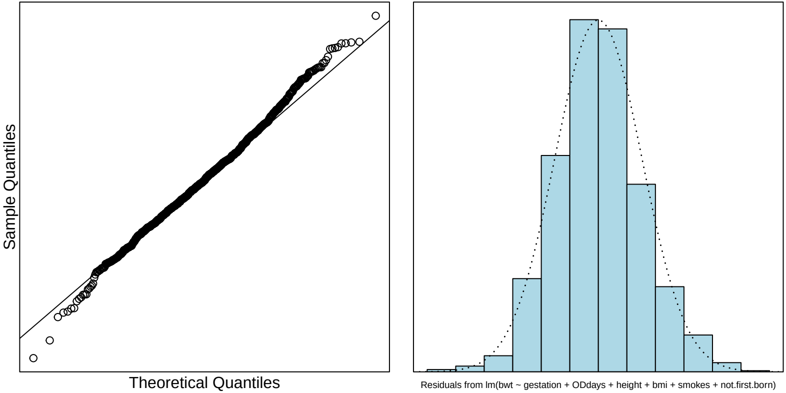
```
> plot(bwt.fit9,which=1)
```



No trend, and EOV assumption is fine.

Birth weight of babies...

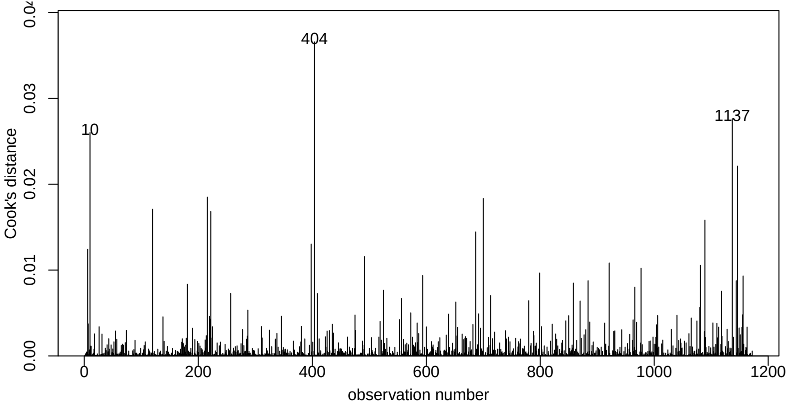
```
> normcheck(bwt.fit9)
```



Normality assumption looks fine.

Birth weight of babies...

```
> cooks20x(bwt.fit9)
```



No unduly influential points.

Birth weight of babies...

Let us get the CIs on this trusted output.

```
> confint(bwt.fit9)
                     2.5 %       97.5 %
(Intercept)    -175.17174865 -115.9641969
gestation         0.57098950    0.7115993
ODdays           -1.11029525   -0.6740001
height            0.94938608    1.6488470
bmi               0.07849275    0.6308947
smokes           -9.79295880   -6.1710576
not.first.born   -5.53179563   -1.4909493
```

See Case Study 10.1 for a detailed executive summary.

Birth weight of babies...

Closing remarks

Recall that we can fit as many explanatory variables as we like. So, did fitting all of these explanatory variables help us describe the variability of the birth weight of babies?

	What we did	Multiple R^2
`bwt.fit3`	Added `gestation+ODdays`	21.9%
`bwt.fit4`	Added `height`	24.6%
`bwt.fit5`	Added `weight`	25.4%
`bwt.fit6`	Added `bmi`	25.5%
`bwt.fit7`	Dropped `weight`	25.4%
`bwt.fit8`	Added `smokes`	29.8%
`bwt.fit9`	Added `not.first.born`	30.5%

Our final model, `bwt.fit9`, includes explanatory variables we deemed suitable and it has a Multiple R^2 of 30.5%.

Section 10.5
Closing remarks and relevant R-code

Closing remarks

In situations where there are many explanatory variables, some of which may be strongly correlated, selecting the best subset for the final model can be challenging.

Model selection is a crucial component of statistical modelling and machine learning, especially in the context of "big data" where there may be millions of observations and thousands of potential explanatory variables.

STATS 330 (Advanced Statistical Modelling) covers this topic in more detail, using techniques such as stepwise variable selection, AIC (Akaike's information criterion), and assessment of prediction error using cross validation.

Most of the R-code you need for this chapter

Note that this code comes with the usual code/checks discussed in chapters 1 and 2.

Useful tools for inspecting many relationships are:

```
> ## Create the pairs plot of the five numeric variables
> pairs20x(Babies.df[,c(1,2,4,5,6)])
```

and for the factor variables:

```
> pairs20x(Babies.df[,c(1,3,7)])
```

Then it is a process of repeatedly updating the model and using Occam's razor to determine a preferred model. E.g.,

```
> model2=update(model1, ~. + xvariable2)
```

This requires constant vigilance to avoid multi-collinearity

Also note that some times several different models may be selected that all make sense and are acceptable.