Solution for exercise 2

Stephan Peischl

2/23/2020

Contents

Exercise 2

In this exercise, we consider four data sets constructed by the British statistician Frank Anscombe. Each data set consists of a response variable Y and an explanatory variable X. The data sets can be made available in R using the command

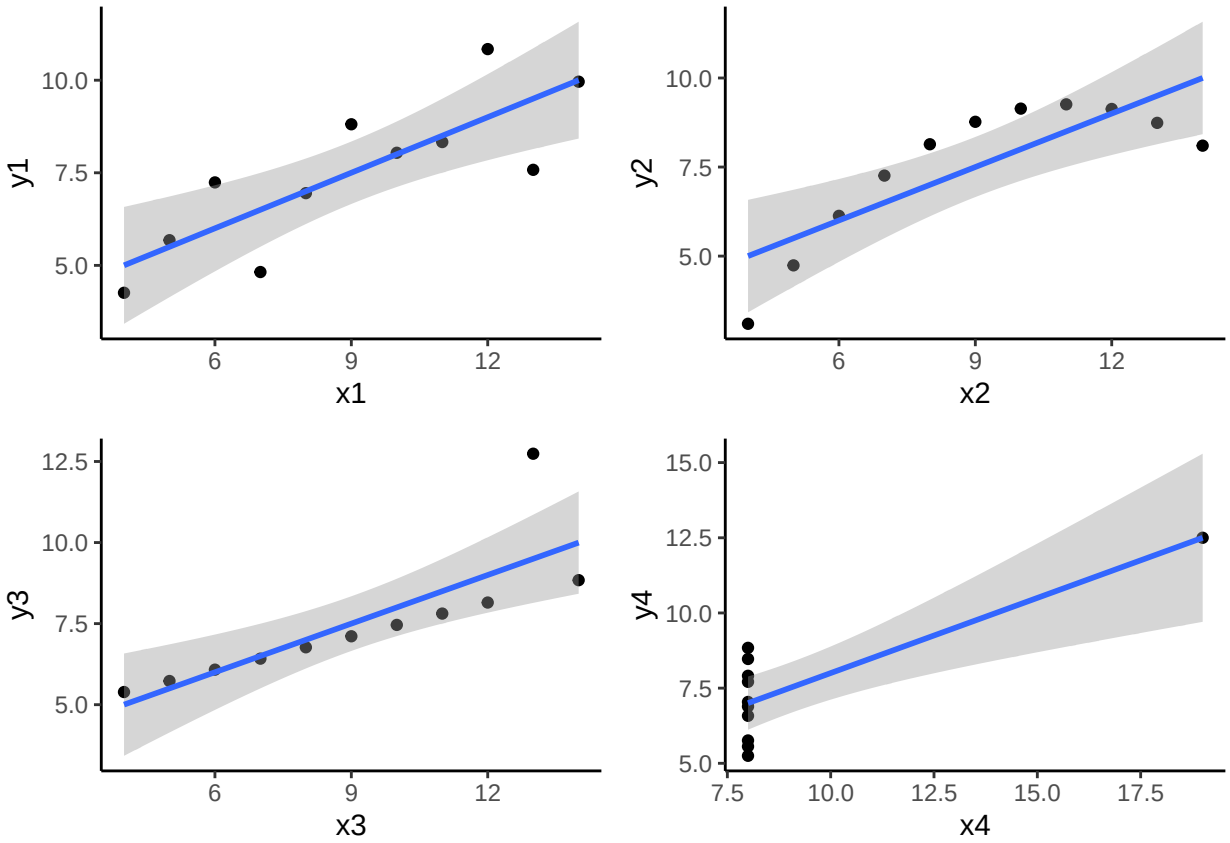
```
data(anscombe)
```

After this, a data frame anscombe is available with the four variable pairs (x1, y1), (x2, y2), (x3, y3), and (x4, y4).

Solution

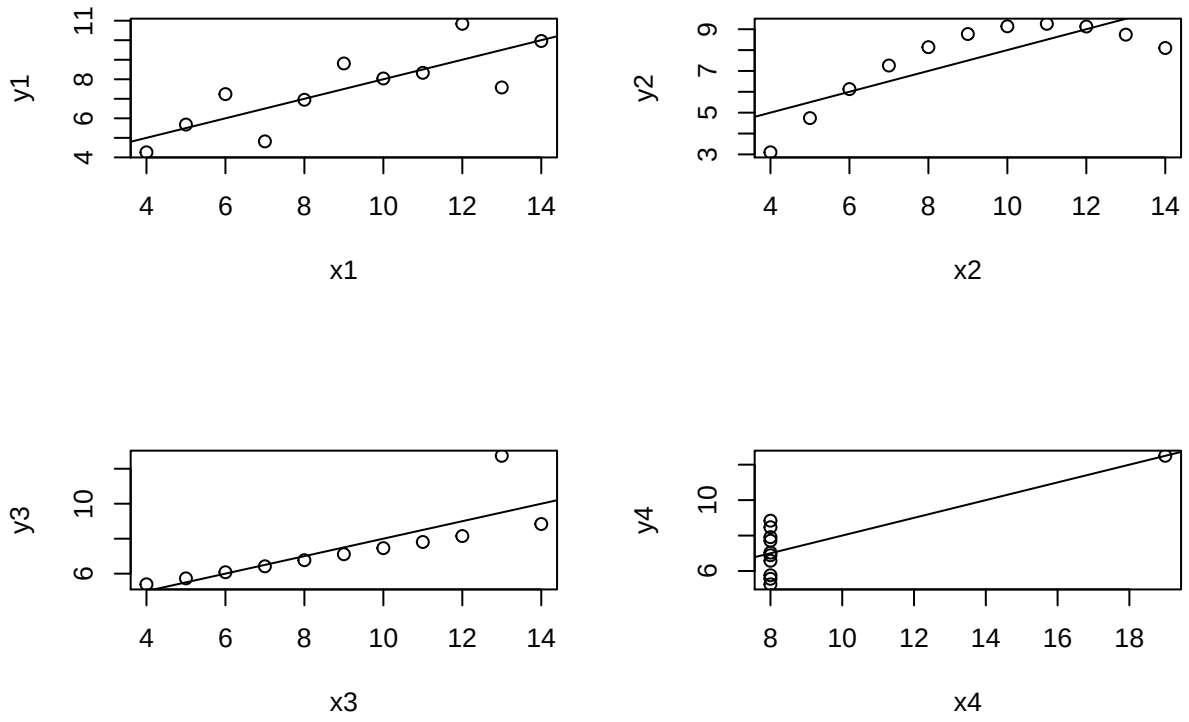
a) We first plot the four data sets as scatterplots and add a linear regression line. I use ggplot2 here:

```
data(anscombe)
p1 = ggplot(data = anscombe,aes(x=x1,y=y1)) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_classic()

p2 = ggplot(data = anscombe,aes(x=x2,y=y2)) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_classic()

p3 = ggplot(data = anscombe,aes(x=x3,y=y3)) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_classic()

p4 = ggplot(data = anscombe,aes(x=x4,y=y4)) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_classic()


grid.arrange(p1,p2,p3,p4,nrow = 2)
```



In base R the plots can be done like this:

```
model1 = lm(y1~x1,data=anscombe)
model2 = lm(y2~x2,data=anscombe)
model3 = lm(y3~x3,data=anscombe)
model4 = lm(y4~x4,data=anscombe)

par(mfrow=c(2,2))

plot(anscombe$x1,anscombe$y1,xlab="x1",ylab="y1")
abline(model1)

plot(anscombe$x2,anscombe$y2,xlab="x2",ylab="y2")
abline(model2)

plot(anscombe$x3,anscombe$y3,xlab="x3",ylab="y3")
abline(model3)

plot(anscombe$x4,anscombe$y4,xlab="x4",ylab="y4")
abline(model4)
```



These models look pretty similar but the underlying data looks very different. You can see that only in the first case a linear regression is reasonable. In the second case the relationship between X and Y is not linear but quadratic. In the third case an outlier has a huge influence on the estimated parameters. In the forth case the regression line is only dependent on one point.

b) We next explore the fitted models in more detail. To do so, we can look at the summary of each model:

```
summary(model1)
```

```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = y2 ~ x2, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## x2             0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

To make the comparsion easier, I extract the fitted parameters and store them in a data frame. I then use the knitr function kable to display it in a nice table format.

```
model.params = data.frame(
  Model1 = c(model1$coefficients,summary(model1)$sigma,summary(model1)$r.squared),
  Model2 = c(model2$coefficients,summary(model2)$sigma,summary(model2)$r.squared),
  Model3 = c( model3$coefficients,summary(model3)$sigma,summary(model3)$r.squared),
  Model4 =c(model4$coefficients,summary(model4)$sigma,summary(model4)$r.squared))

rownames(model.params)= c("intercept","slope","sigma","R2")

knitr::kable(model.params)
```

	Model1	Model2	Model3	Model4
intercept	3.0000909	3.000909	3.0024545	3.0017273
slope	0.5000909	0.500000	0.4997273	0.4999091
sigma	1.2366033	1.237214	1.2363114	1.2356955
R2	0.6665425	0.666242	0.6663240	0.6667073

For all four models, the estimated parameters as well as R^2 are almost identical. It is absolutely necessary to always make a visual check of the data and the fitted model. If the linearity assumption does not hold, quality measures such as R^2 have little meaning.