Project 1: Analytics and Data Science

DA 420

Marjorie Blanco

Anscombe Quartet

```
# Read Anscombe data
anscombe <- read.csv("data/Anscombe.csv")
kable(anscombe)
```

Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Descriptive statistics

```
# calculate the mean and standard deviation
kable(anscombe %>% select(-Observation) %>%
    summarise_all(funs( mean(.), sd(.))) %>%
  gather())
```

key	value
x1_mean	9.000000
y1_mean	7.500909
x2_mean	9.000000
y2_mean	7.500909
x3_mean	9.000000
y3_mean	7.500000
x4_mean	9.000000
y4_mean	7.500909
x1_sd	3.316625
y1_sd	2.031568
x2_sd	3.316625
y2_sd	2.031657

key	value
x3_sd	3.316625
y3_sd	2.030424
x4_sd	3.316625
y4_sd	2.030578

Linear Model

```
fit.model <- lm(y1~x1, anscombe)
summary(fit.model)
```

```
##
## 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
```

```
anscombe$x1_1 <- anscombe$x1
anscombe$y1_1 <- fit.model$fitted.values -  fit.model$residuals

fit.model <- lm(y2~x2, anscombe)
summary(fit.model)
```

```
##
## 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
```

```
anscombe$x2_1 <- anscombe$x2
anscombe$y2_1 <- fit.model$fitted.values -  fit.model$residuals

fit.model <- lm(y3~x3, anscombe)
summary(fit.model)
```

```
##
## Call:
## lm(formula = y3 ~ x3, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## x3            0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```
anscombe$x3_1 <- anscombe$x3
anscombe$y3_1 <- fit.model$fitted.values -  fit.model$residuals

fit.model <- lm(y4~x4, anscombe)
summary(fit.model)
```

```
##
## Call:
## lm(formula = y4 ~ x4, data = anscombe)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0017     1.1239   2.671  0.02559 *
## x4            0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

```r
anscombe$x4_1 <- anscombe$x4
anscombe$y4_1 <- fit.model$fitted.values -  fit.model$residuals
```

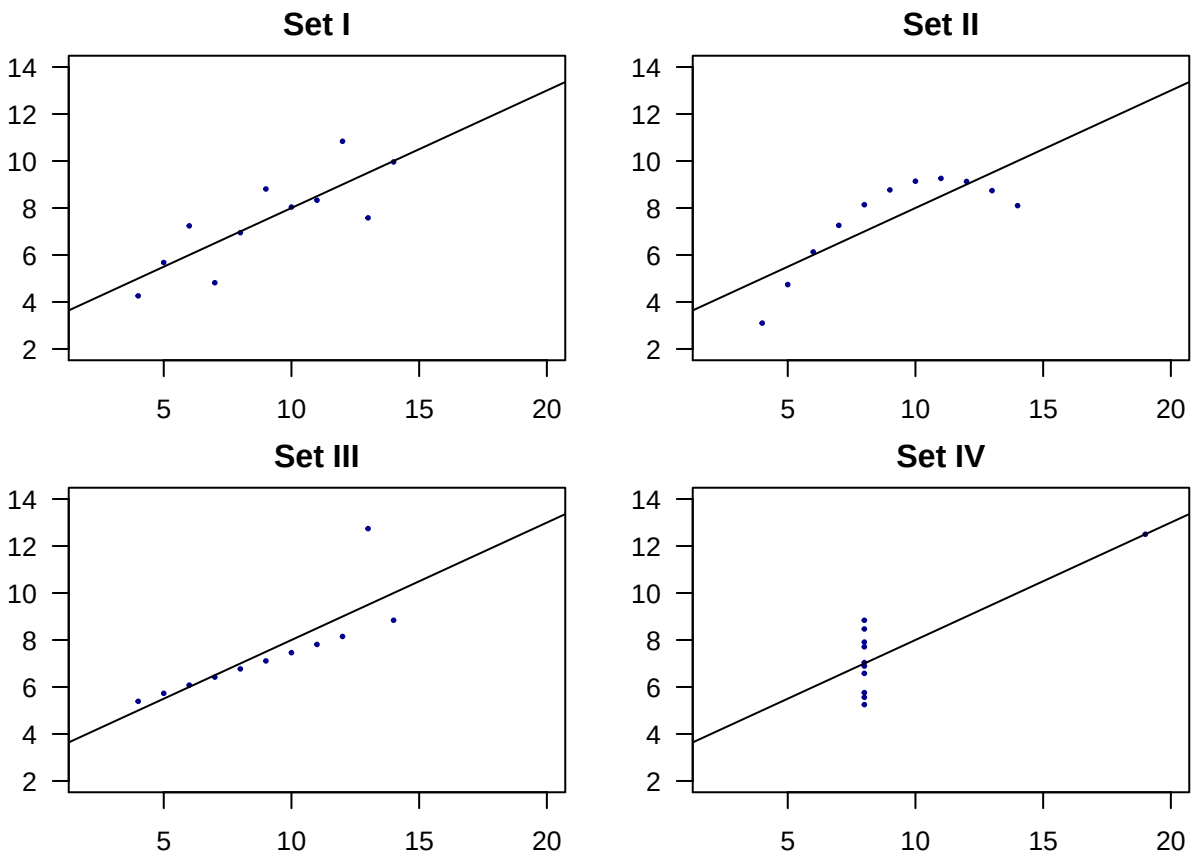
Original Plots

```r
# 2x2 row with 2 margin
par(mfrow=c(2,2), mar=c(2, 2, 2, 2))
plot(anscombe$x1, anscombe$y1, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y1~anscombe$x1))
title("Set I")

plot(anscombe$x2, anscombe$y2, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y2~anscombe$x2))
title("Set II")

plot(anscombe$x3, anscombe$y3, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y3~anscombe$x3))
title("Set III")

plot(anscombe$x4, anscombe$y4, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y4~anscombe$x4))
title("Set IV")
```



Modified Anscombe Quartet

To modify the original data while keeping the mean and standard deviation unchaged I fliped the dependent variable on the linear regression line. The expected impact is that the $\beta 0$ (intercept) and $\beta 1$ will not change. The mean and standard deviation for the explantory variables will not change.

```
kable(anscombe %>% select(-(x1:y4)))
```

Observation	x1_1	y1_1	x2_1	y2_1	x3_1	y3_1	x4_1	y4_1
1	10	7.962000	10	6.861818	10	8.539454	8	7.422
2	8	7.051636	8	5.861818	8	7.230545	8	8.242
3	13	11.422546	13	10.261818	13	6.257818	8	6.292
4	9	6.191818	9	6.231818	9	7.890000	8	5.162
5	11	8.672182	11	7.741818	11	9.188909	8	5.532
6	14	10.042727	14	11.901818	14	11.157273	8	6.962
7	6	4.761273	6	5.871818	6	5.921636	8	8.752
8	4	5.740909	4	6.901818	4	4.612727	19	12.500
9	12	7.162364	12	8.871818	12	9.848364	8	8.442
10	7	8.181454	7	5.741818	7	6.581091	8	6.092
11	5	5.321091	5	6.261818	5	5.272182	8	7.112

Descriptive statistics

```
# calculate the mean and standard deviation
kable(anscombe %>% select(-(Observation:y4)) %>%
    summarise_all(funs( mean(.), sd(.))) %>%
  gather())
```

key	value
x1_1_mean	9.000000
y1_1_mean	7.500909
x2_1_mean	9.000000
y2_1_mean	7.500909
x3_1_mean	9.000000
y3_1_mean	7.500000
x4_1_mean	9.000000
y4_1_mean	7.500909
x1_1_sd	3.316625
y1_1_sd	2.031568
x2_1_sd	3.316625
y2_1_sd	2.031657
x3_1_sd	3.316625
y3_1_sd	2.030424
x4_1_sd	3.316625
y4_1_sd	2.030578

Modified Linear Model

```
fit.model <- lm(y1_1~x1_1, anscombe)
summary(fit.model)
```

```
##
## Call:
## lm(formula = y1_1 ~ x1_1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83882 -0.70941  0.04136  0.45577  1.92127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1_1          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
```

```r
fit.model <- lm(y2_1~x2_1, anscombe)
summary(fit.model)
```

```
##
## Call:
## lm(formula = y2_1 ~ x2_1, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2691 -0.9491 -0.1291  0.7609  1.9009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## x2_1           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
```

```r
fit.model <- lm(y3_1~x3_1, anscombe)
summary(fit.model)
```

```
##
## Call:
## lm(formula = y3_1 ~ x3_1, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2411 -0.1540  0.2303  0.6146  1.1586
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## x3_1          0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```r
fit.model <- lm(y4_1~x4_1, anscombe)
summary(fit.model)
```

```
##
## Call:
## lm(formula = y4_1 ~ x4_1, data = anscombe)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -1.839 -0.809  0.000  0.831  1.751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0017     1.1239   2.671  0.02559 *
## x4_1          0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

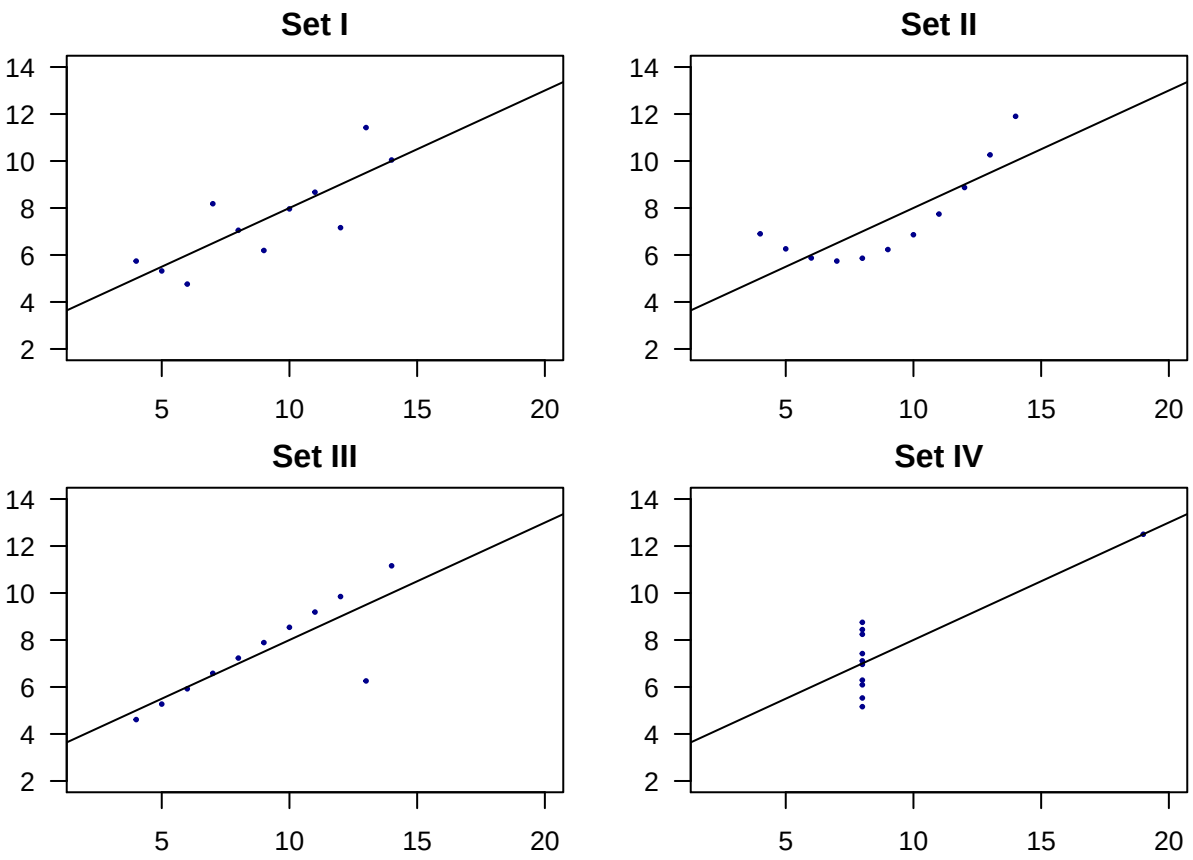
Modified Plots

```r
# 2x2 row with 2 margin
par(mfrow=c(2,2), mar=c(2, 2, 2, 2))
plot(anscombe$x1_1, anscombe$y1_1, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y1_1~anscombe$x1_1))
title("Set I")

plot(anscombe$x2_1, anscombe$y2_1, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y2_1~anscombe$x2_1))
title("Set II")

plot(anscombe$x3_1, anscombe$y3_1, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y3_1~anscombe$x3_1))
title("Set III")

plot(anscombe$x4_1, anscombe$y4_1, xlim=c(2, 20), ylim=c(2, 14), pch = 19,
    col = "darkblue", cex = 0.3, las = 1, xlab = "", ylab = "")
abline(lm(anscombe$y4_1~anscombe$x4_1))
title("Set IV")
```


Set I
Set II
Set III
Set IV
14
12
10
8
6
4
2
5
10
15
20