Anscombe Regression Example

Data

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

Basic S/R Code and Output

```
R : Copyright 2003, The R Development Core Team
Version 1.7.1  (2003-06-16)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type ‘license()’ or ‘licence()’ for distribution details.

R is a collaborative project with many contributors.
Type ‘contributors()’ for more information.

Type ‘demo()’ for some demos, ‘help()’ for on-line help, or
‘help.start()’ for a HTML browser interface to help.
Type ‘q()’ to quit R.

[Previously saved workspace restored]

> options(STERM=’iESS’, editor=’gnuclient.exe’)

> attach(anscombe)
> anscombe <- read.table("anscombe.dat", header=T)
> mean(anscombe)
       x       y1       y2       y3       x4       y4
9.000000 7.500909 7.500909 7.500000 9.000000 7.500909
```

Plot Data and Ordinary Least Squares Line

```
> postscript(file="ansc1.ps")
> plot(x,y1)
> lm1 <- lm(y1 ~ x, data=anscombe)
> summary(lm1)

Call:
lm(formula = y1 ~ x, 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 *
x             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: 0.002170

> abline(coefficients(lm1))
> dev.off()
```

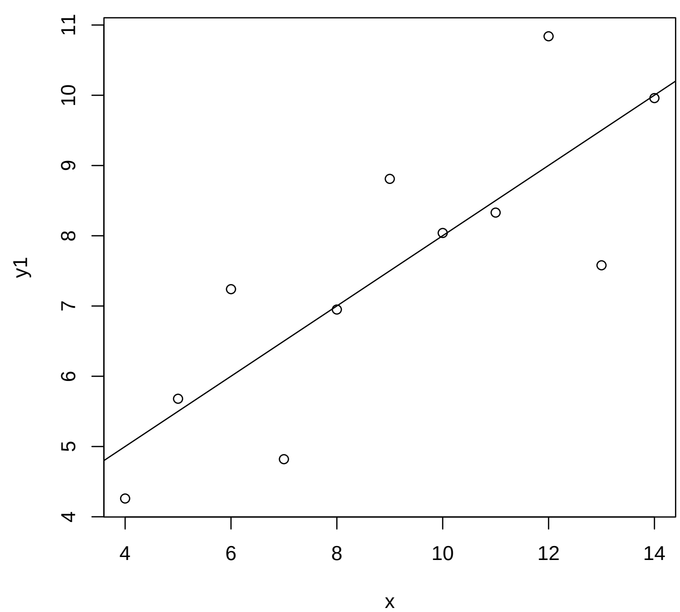


Figure 1: Scatter Plot of pairs (x, y1) from Anscombe example

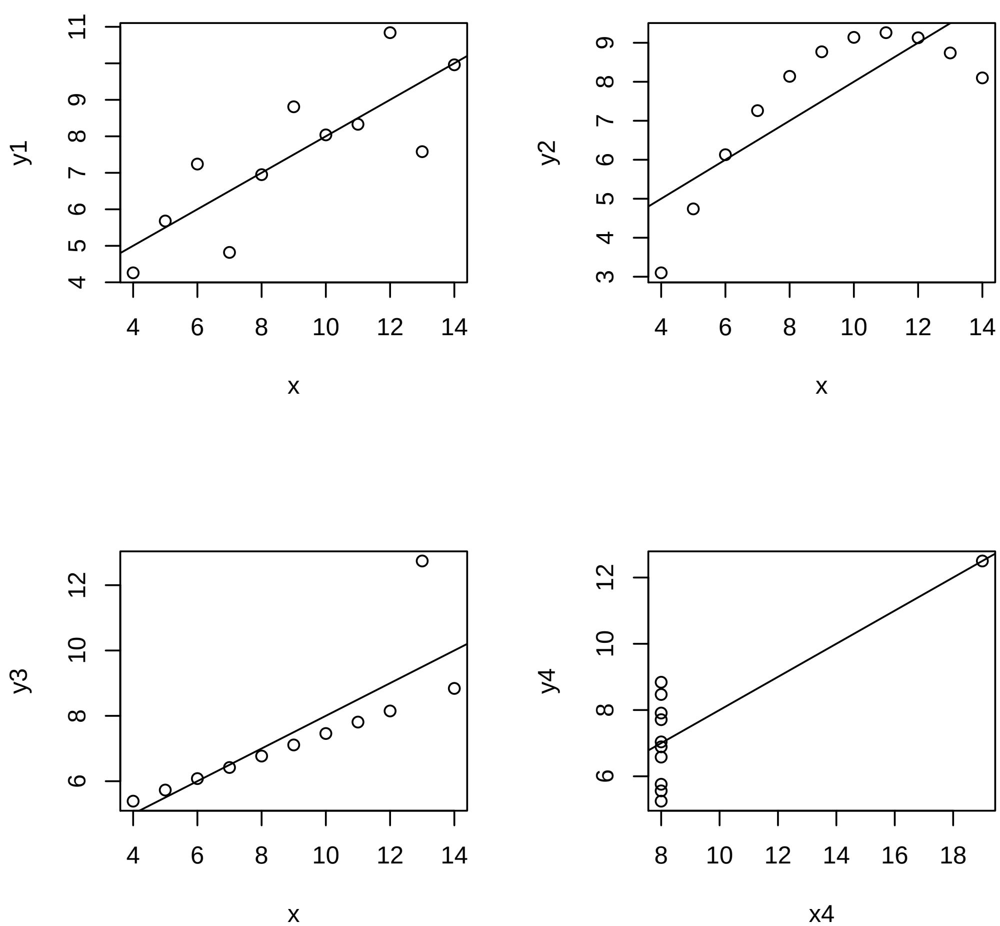
Summaries from Other Regressions

```
> summary(lm(y2 ~ x, 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 *
x              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

> summary(lm(y3 ~ x, 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 *
x             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

> summary(lm(y4 ~ x4, data=anscombe))

Residuals:
       Min         1Q     Median         3Q        Max
-1.751e+00 -8.310e-01  1.258e-16  8.090e-01  1.839e+00

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
```

Which Summaries are the Same?

Create Plots of Other Data Sets with OLS line

```
> lm2 <- lm(y2 ~ x, data=anscombe) # CREATE A LINEAR MODEL OBJECT
> lm3 <- lm(y3 ~ x, data=anscombe)
> lm4 <- lm(y4 ~ x4, data=anscombe)

> pdf(file="anscombe-plots.pdf", horizontal=T) # Create a PDF output file
> par(mfrow=c(2,2))
> plot(x,y1)
> abline(coefficients(lm1))
> plot(x,y2)
> abline(coefficients(lm2))
> plot(x,y3)
> abline(coefficients(lm2))
> plot(x4,y4)
> abline(coefficients(lm2))
> dev.off()
```



Summary

- data set 1 is clearly linear with some scatter
- data set 2 is clearly quadratic
- data set 3 has an “outlier”
- data set 4 – poor experimental design! the extreme x value far removed from the bulk of the data has undue influence on the estimate “wagging the dog”

One can always find the best fitting least squares line: the value of β_0 and β_1 that minimize

$$\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i)^2$$

While all data sets had the same OLS estimates, correlations, etc, the estimated slopes and intercepts are misleading because the estimation is done in the context of an assumed linear model.