Serik Sagitov, Chalmers and GU, March 2, 2018

Solutions chapter 14

Problem 14.2

Ten pairs

x	0.34	1.38	-0.65	0.68	1.40	-0.88	-0.30	-1.18	0.50	-1.75
y	0.27	1.34	-0.53	0.35	1.28	-0.98	-0.72	-0.81	0.64	-1.59

with

$$\bar{x} = -0.046, \quad \bar{y} = -0.075, \quad s_x = 1.076, \quad s_y = 0.996, \quad r = 0.98.$$

Draw a scatter plot using

x	-1.75	-1.18	-0.88	-0.65	-0.30	0.34	0.50	0.68	1.38	1.40
y	-1.59	-0.81	-0.98	-0.53	-0.72	0.27	0.64	0.35	1.34	1.28

(a) Simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Fitting a straight line using

$$y - \bar{y} = r \cdot \tfrac{s_y}{s_x}(x - \bar{x})$$

we get the predicted response

$$\hat{y} = -0.033 + 0.904 \cdot x.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_y^2 (1 - r^2) = 0.05.$$

(b) Simple linear regression model

$$X = \beta_0 + \beta_1 y + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Fitting a straight line using

$$x - \bar{x} = r \cdot \tfrac{s_x}{s_y}(y - \bar{y})$$

we get the predicted response

$$\hat{x} = 0.033 + 1.055 \cdot y.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_x^2 (1 - r^2) = 0.06.$$

(c) First fitted line

$$y = -0.033 + 0.904 \cdot x$$

is different from the second

$$y = -0.031 + 0.948 \cdot x.$$

Problem 14.4

Two consecutive grades

X = the high school GPA (grade point average),
Y = the freshman GPA.

Allow two different intercepts for females and males

$$Y_F = \beta_F + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2),$$
$$Y_M = \beta_M + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Using an extra explanatory variable f which equal 1 for females and 0 for males, we rewrite this model in the form of a multiple regression

$$Y = f\beta_F + (1 - f)\beta_F + \beta_1 x + \epsilon = \beta_0 + \beta_1 x + \beta_2 f + \epsilon,$$

where

$$\beta_0 = \beta_M, \quad \beta_2 = \beta_F - \beta_M.$$

Here $p = 3$ and the design matrix is

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & f_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & f_n \end{pmatrix}.$$

After $\beta_0, \beta_1, \beta_2$ are estimated, we compute

$$\beta_M = \beta_0, \quad \beta_F = \beta_0 + \beta_2.$$

A null hypothesis of interest $\beta_2 = 0$.

Problem 14.14

Simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Using n pairs of (x_i, y_i) we fit a regression line by

$$y = b_0 + b_1 x, \quad \mathrm{Var}(b_0) = \tfrac{\sigma^2 \overline{x^2}}{(n-1)s_x^2}, \quad \mathrm{Var}(b_1) = \tfrac{\sigma^2}{(n-1)s_x^2}, \quad \mathrm{Cov}(b_0, b_1) = -\tfrac{\sigma^2 \bar{x}}{(n-1)s_x^2}.$$

For a given $x = x_0$, we wish to predict the value of a new observation

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon$$

by

$$\hat{y}_0 = b_0 + b_1 x_0.$$

(a) The predicted value $\hat{y}_0$ and actual observation Y_0 are independent random variables, therefore

$$\mathrm{Var}(Y_0 - \hat{y}_0) = \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \mathrm{Var}(b_0 + b_1 x_0) = \sigma^2 C_n^2,$$

where

$$C_n^2 = 1 + \frac{\mathrm{Var}(b_0)+\mathrm{Var}(b_1)x_0^2-2x_0\mathrm{Cov}(b_0,b_1)}{\sigma^2} = 1 + \frac{\overline{x^2}+x_0^2-2\bar{x}x_0}{(n-1)s_x^2} = 1 + \frac{\overline{x^2}-\bar{x}^2+(x_0-\bar{x})^2}{(n-1)s_x^2} = 1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}.$$

(b) 95% prediction interval for the new observation Y_0 is obtained from

$$\frac{Y_0-\hat{y}_0}{sC_n} \sim \mathrm{t}_{n-2}.$$

Since

$$0.95 = \mathrm{P}(|Y_0 - \hat{y}_0| \le t_{n-2}(0.025) \cdot sC_n) = \mathrm{P}(Y_0 \in \hat{y}_0 \pm t_{n-2}(0.025) \cdot sC_n),$$

we conclude that a 95% prediction interval for the new observation Y_0 is given by

$$b_0 + b_1x_0 \pm t_{n-2}(0.025) \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}.$$

The further x_0 is from $\bar{x}$, the more uncertain becomes the prediction.

Problem 14.23

Data collected for

x = midterm grade,
y = final grade,

gave

$$r = 0.5, \quad \bar{x} = \bar{y} = 75, \quad s_x = s_y = 10.$$

(a) Given $x = 95$, we predict the final score by

$$\hat{y} = 75 + 0.5(95 - 75) = 85.$$

Regression to mediocracy.

(b) Given $y = 85$ and we do not know the midterm score, we predict the midterm score by

$$\hat{x} = 75 + 0.5(85 - 75) = 80.$$

Problem 14.33

Let

$$Y = X + \beta Z,$$

where $X \in \mathrm{N}(0,1)$ and $Z \in \mathrm{N}(0,1)$ are independent.

(a) Find the correlation coefficient ρ for (X,Y). Since $\mathrm{E}X = 0$, we have

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) = \mathrm{E}(X^2 + \beta XZ) = 1, \quad \mathrm{Var}Y = \mathrm{Var}X + \mathrm{Var}Z = 1 + \beta^2,$$

and we see that the correlation coefficient is always positive

$$\rho = \frac{1}{\sqrt{1+\beta^2}}.$$

(b) Use (a) to generate five samples

$$(x_1, y_1), \ldots, (x_{20}, y_{20})$$

with different

$$\rho = -0.9, \quad -0.5, \quad 0, \quad 0.5, \quad 0.9,$$

and compute the sample correlation coefficients.

From $\rho = \frac{1}{\sqrt{1+\beta^2}}$, we get $\beta = \sqrt{\rho^{-2} - 1}$ so that

$$\rho = 0.5 \Rightarrow \beta = 1.73, \qquad \rho = 0.9 \Rightarrow \beta = 0.48.$$

How to generate a sample with $\rho = -0.9$ using Matlab:

```
X=randn(20,1);
Z=randn(20,1);
Y=-X+0.48*Z;
r=corrcoeff(X,Y)
```

How to generate a sample with $\rho = 0$ using Matlab:

```
X=randn(20,1);
Y=randn(20,1);
r=corrcoeff(X,Y)
```

Simulation results

ρ	-0.9	-0.5	0	0.5	0.9
r	-0.92	-0.45	-0.20	0.32	0.92

Problem 14.42

Data

velocity of a car x	20.5	20.5	30.5	40.5	48.8	57.8
stopping distance y	15.4	13.3	33.9	73.1	113.0	142.6

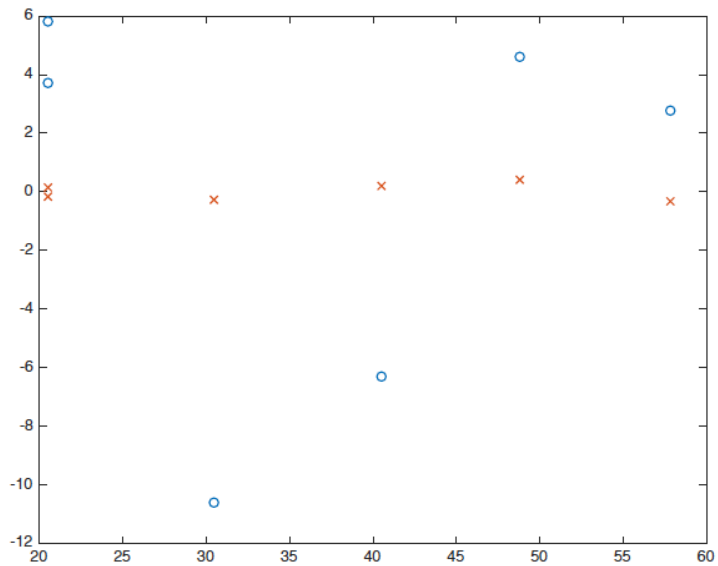
Matlab commands (x and y are columns)

```
[b,bint,res,rint,stats]=regress(y,[ones(6,1),x])
[b,bint,res,rint,stats]=regress(sqrt(y),[ones(6,1),x])
```

give two sets of residuals - see the plot. Two simple linear regression models

$$y = -62.05 + 3.49 \cdot x, \quad r^2 = 0.984,$$
$$\sqrt{y} = -0.88 + 0.2 \cdot x, \quad r^2 = 0.993.$$

Can you suggest any physical reason that explains why the second model is better?


6
4
2
0
-2
-4
-6
-8
-10
-12
20
25
30
35
40
45
50
55
60