Problem Set 10

You are strongly encouraged to solve the following exercise before the final exam:

To explore the potential effect of age on systolic blood pressure (SBP), data of 33 women, aged 22-81 was collected. It is presented here in Table 1.

(a) Figure 1 displays a scatter plot of the data in Table 1. What is your early assessment of the idea to fit a simple linear regression model to the data?

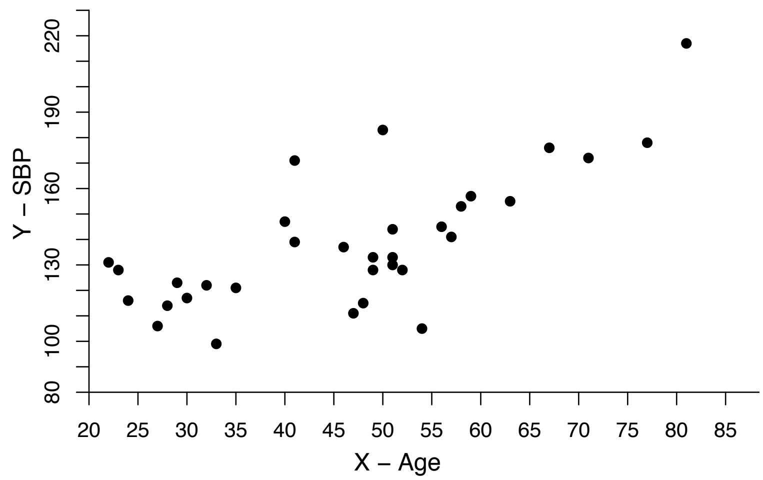

Figure 1: A scatter plot of the data displayed in Table 1.

a linear trend but maybe too much noise.

Use Least Square Estimators formula

(b) Calculate the straight line equation for the simple linear regression model and interpret the estimated slope. What proportion of the variability in SBP can be explained by age?

R^2 = r^2, so find sample correlation coefficient

(c) Use Table 2 to estimate the random noise variance. Test the hypothesis on the existence of a linear trend (at the 5% level) and provide a 95% confidence interval for the slope.

slope estimate follows t_{n-2} distribution

(d) Provide a point estimate and a 95% confidence interval for the mean SBP of 37 year old women. (use $t_{31,0.975} = 2.04$)

(e) Express your opinion on the validity of the model assumptions, based on the residual plots in Figure 2.

	Age (X)	SBP (Y)	
1	22	131	
2	23	128	
3	24	116	
4	27	106	
5	28	114	
6	29	123	
7	30	117	
8	32	122	
9	33	99	
10	35	121	
11	40	147	
12	41	139	
13	41	171	
14	46	137	
15	47	111	
16	48	115	
17	49	133	
18	49	128	
19	50	183	
20	51	130	
21	51	133	
22	51	144	
23	52	128	
24	54	105	
25	56	145	
26	57	141	
27	58	153	
28	59	157	
29	63	155	
30	67	176	
31	71	172	
32	77	178	
33	81	217	
	$\sum x_i = 1542$	$\sum x_i^2 = 79{,}716$	
	$\sum y_i = 4{,}575$	$\sum y_i^2 = 656{,}481$	$\sum x_i y_i = 223{,}144$

Table 1: Raw data for the SBP vs. Age example.

	Age (X)	SBP (Y)	$\widehat{Y}$	e
1	22	131	108.4	22.6
2	23	128	109.6	18.4
3	24	116	110.9	5.1
4	27	106	114.5	-8.5
5	28	114	115.7	-1.7
6	29	123	117.0	6.0
7	30	117	118.2	-1.2
8	32	122	120.6	1.4
9	33	99	121.9	-22.9
10	35	121	124.3	-3.3
11	40	147	130.4	16.6
12	41	139	131.6	7.4
13	41	171	131.6	39.4
14	46	137	137.7	-0.7
15	47	111	139.0	-28.0
16	48	115	140.2	-25.2
17	49	133	141.4	-8.4
18	49	128	141.4	-13.4
19	50	183	142.6	40.4
20	51	130	143.9	-13.9
21	51	133	143.9	-10.9
22	51	144	143.9	0.1
23	52	128	145.1	-13.1
24	54	105	147.5	-42.5
25	56	145	150.0	-5.0
26	57	141	151.2	-10.2
27	58	153	152.4	0.6
28	59	157	153.6	3.4
29	63	155	158.5	-3.5
30	67	176	163.4	12.6
31	71	172	168.3	3.7
32	77	178	175.6	2.4
33	81	217	180.5	36.5
				$\sum e_i^2 = 10769.7$

Table 2: The original data table along with the fitted values and the model residuals.

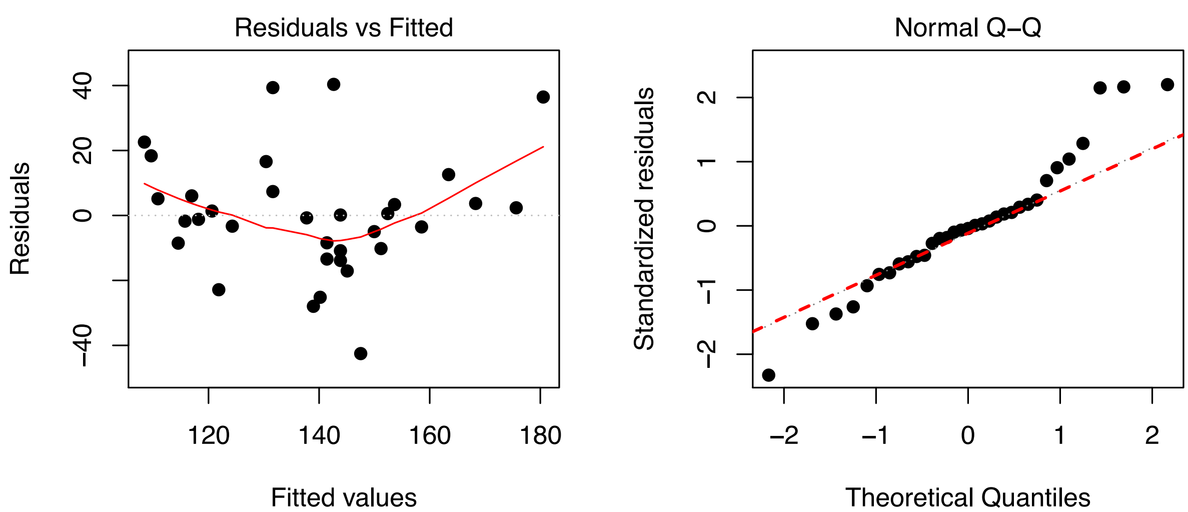

Figure 2: Residual plots for the linear fit.

Solution:

(a) There definitely appears to be an upward trend, but with the abundance of noise it is hard to tell whether or not it is linear. In any case, the ability of age alone to explain differences in SBP seems to be limited.

(b) $$\overline{x} = \frac{1542}{33} = 46.73, \quad S_X^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1} = \frac{79716 - 33 \times (46.73)^2}{32} = 239.45,$$

$$\overline{y} = \frac{4575}{33} = 138.64, \quad S_{XY} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{n-1} = \frac{223144 - 33 \times 46.73 \times 138.64}{32} = 292.71,$$

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{292.71}{239.45} = 1.222 \quad \text{and} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 138.64 - 1.222 \times 46.73 = 81.52,$$

and the linear fit equation is thus

$$\widehat{y}(x) = 81.52 + 1.222x.$$

We estimate the increase in average systolic blood pressure at 1.222 per one year of aging. In addition,

$$S_Y^2 = \frac{\sum y_i^2 - n\overline{y}^2}{n-1} = \frac{656481 - 33 \times (138.64)^2}{32} = 694.36$$

$$\Longrightarrow R^2 = r_{XY}^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2} = \frac{(292.71)^2}{239.45 \times 694.36} = 0.5153,$$

hence only 51.53% of the variation in the SBP values can be explained by age, as expected perhaps.

(c) As argued in class, the noise variance can be estimated by

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{10769.71}{31} = 347.41.$$

Testing for a linear trend is based on evaluating $\mathcal{T} = \frac{\widehat{\beta}_1}{\frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}}}$ with respect to the t_{n-2} distribution. Here

$$\mathcal{T} = \frac{1.222}{\frac{\sqrt{347.41}}{\sqrt{(33-1) \cdot 239.45}}} = 5.73 >> 2.04 = t_{31,0.975}$$

(no row for 31 degrees of freedom in the table, but clearly $t_{31,0.975} < t_{30,0.975} = 2.042$), and it can be verified that the p-value is 2.57×10^{-6}, so if we trust the model, the existence of a linear trend is undeniable. Similarly,

$$\widehat{\beta}_1 \pm \frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}} t_{n-2,1-\alpha/2} = 1.222 \pm \frac{\sqrt{347.41}}{\sqrt{(33-1)\cdot 239.45}} \underbrace{t_{31,0.975}}_{2.04}$$

$$= 1.222 \pm 0.434 = [0.788, 1.656]$$

is a 95% confidence interval for β_1.

(d) A point estimate for the mean SBP of 37 year old women would be

$$\widehat{y}(37) = 81.52 + 1.222 \times 37 = 126.73,$$

whereas a 95% confidence interval for the mean is given by

$$\widehat{y}(x_0) \pm t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_j (x_j - \overline{x})^2}}$$

$$= 126.73 \pm \underbrace{t_{31,0.975}}_{2.04} \sqrt{347.41} \sqrt{\frac{1}{33} + \frac{(37-46.73)^2}{(33-1)\cdot 239.45}} = 126.73 \pm 7.85$$

$$= [118.88, 134.58].$$

The fitted line and the 95% confidence bands are displayed in Figure 3.

(e) The "belly" in the Residuals vs. Fitted plot implies model misspecification. This is perhaps the reason why the standardized residuals in the Q-Q plot appear to have heavier tails than expected. It is possible that a quadratic model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

could prove a better fit for this particular dataset.

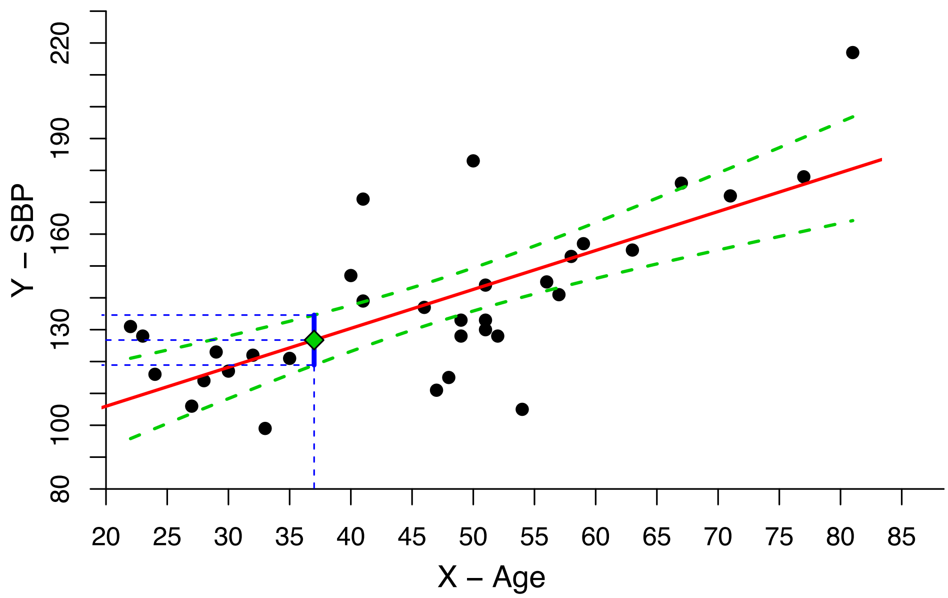

Figure 3: Point estimate and 95% confidence bands for the mean SBP of women, based on the linear regression model.