Lecture notes to Stock and Watson chapter 4

Introductory linear regression

Tore Schweder

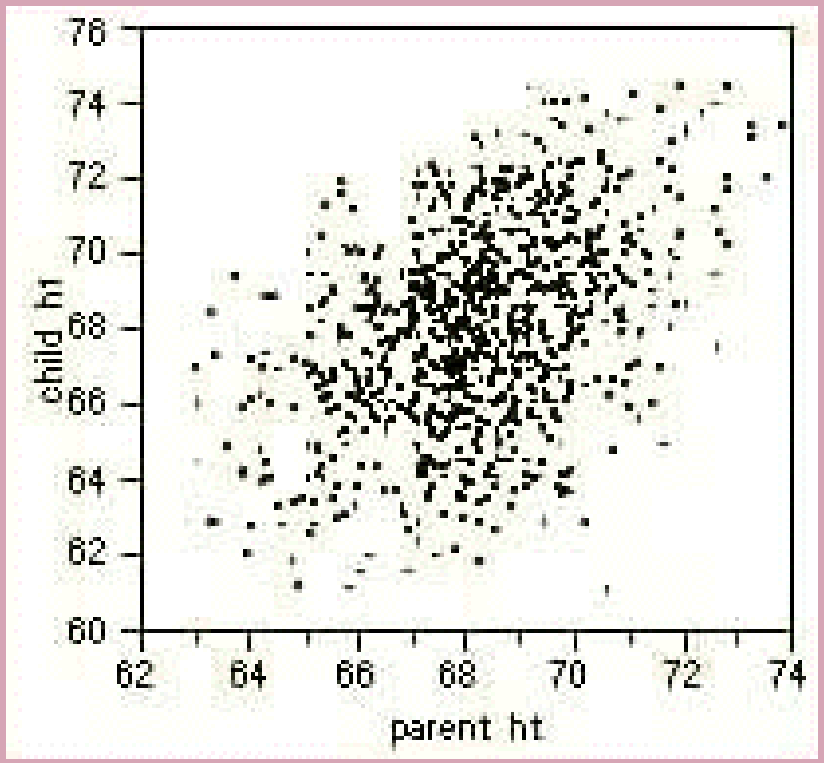
Sept 2009

Regression

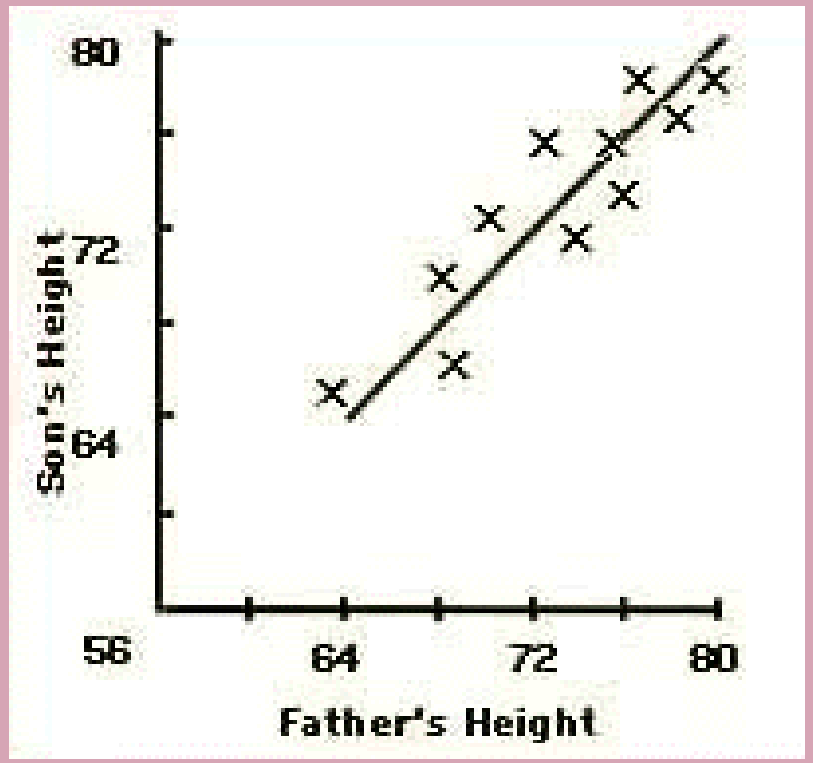
"Regression" is due to Francis Galton (1822-1911): how is a son's height related to his father's height?

Figure A

Galton's Original Data



Hand-Drawn Data With Regression Line



Regression towards the mean

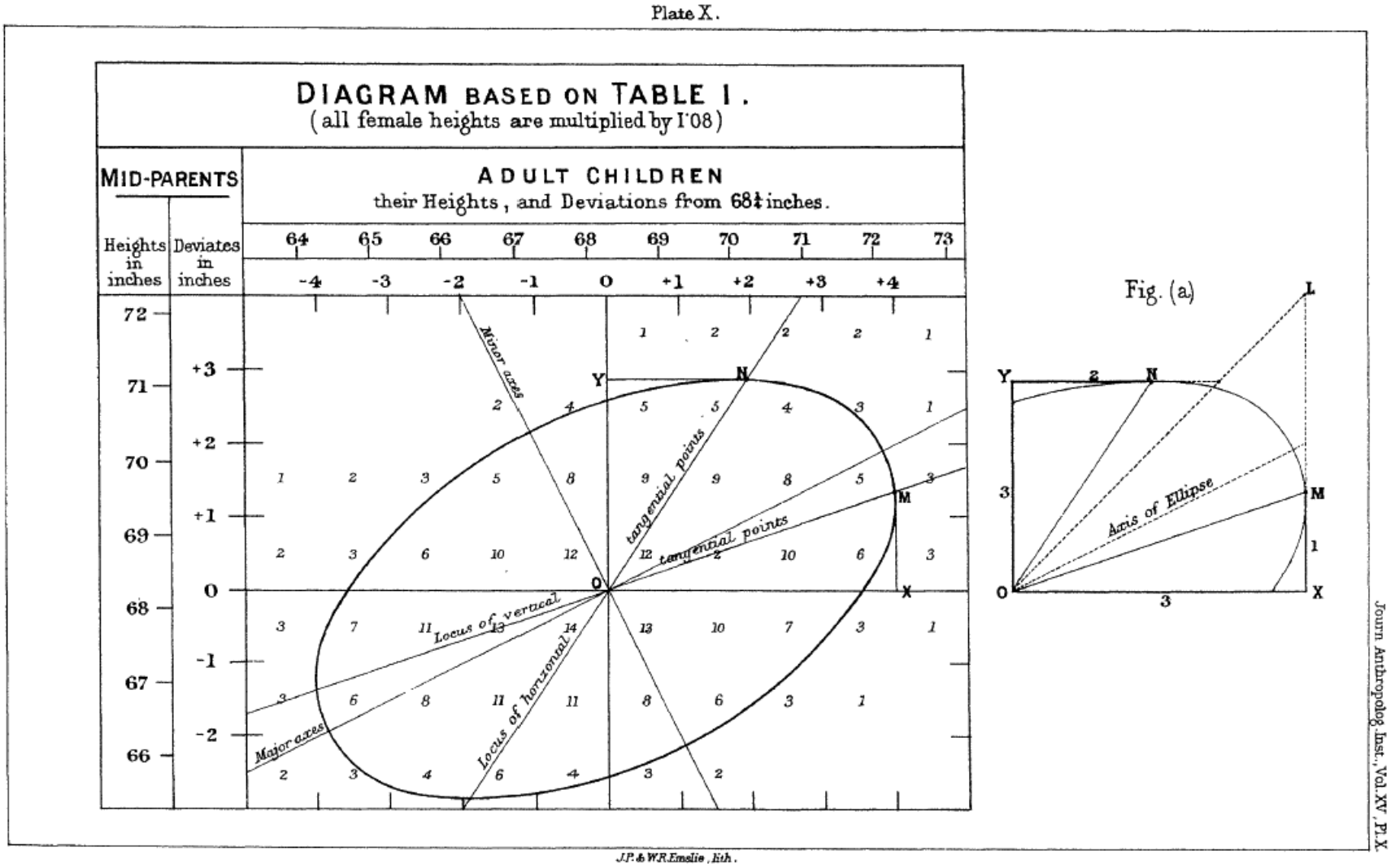


Figure: Galton's original diagram. Parents-child pairs by heigth.

The geometry of linear regression

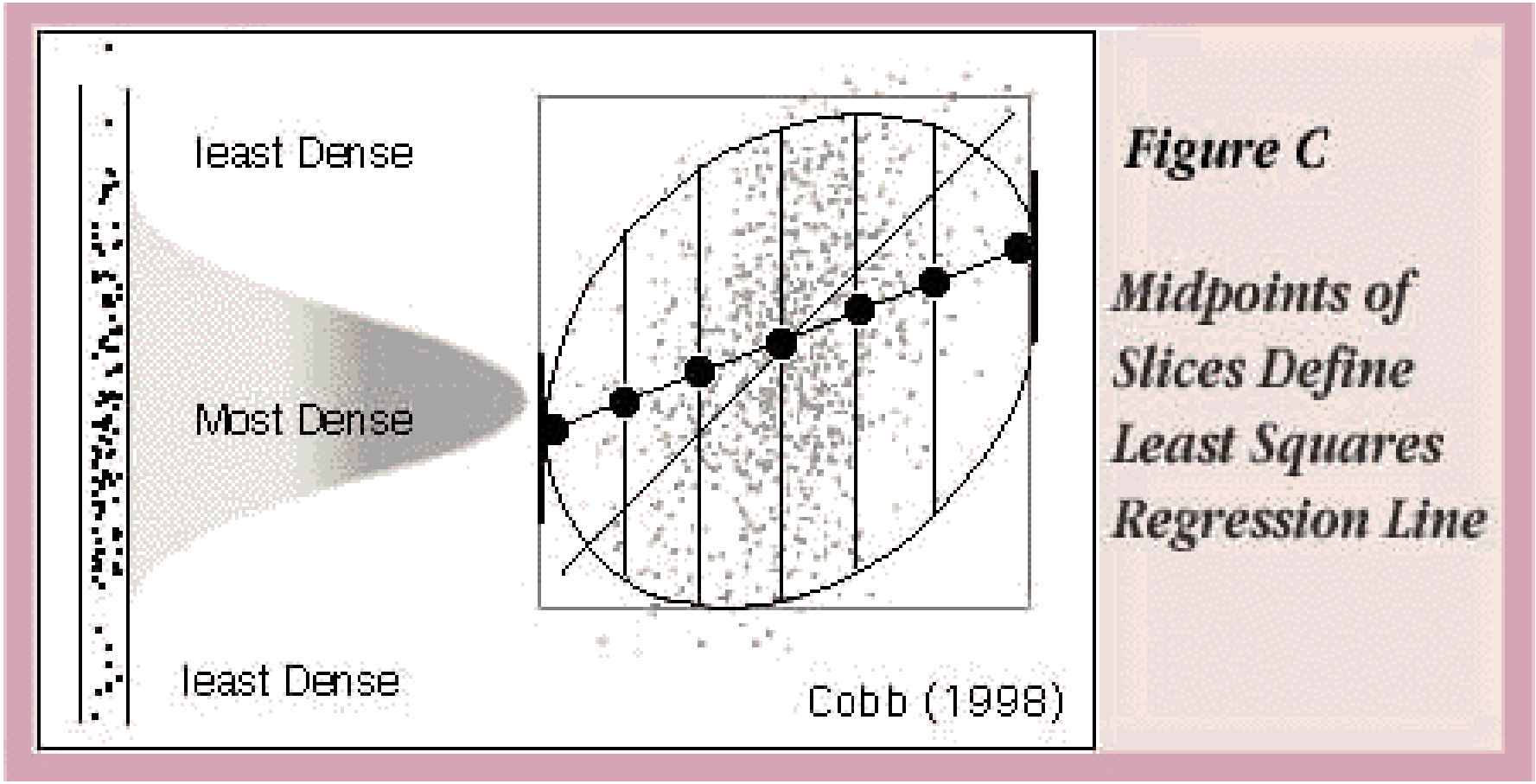


Figure: The regression curve is $y = f(x) = E[Y|X = x]$.

The linear regression model

Question: how is a response variable Y related to a stimulus variable X (explanatory/control/explanatory)? Assuming a linear relation,

1. what is the slope?

The linear regression model

Question: how is a response variable Y related to a stimulus variable X (explanatory/control/explanatory)? Assuming a linear relation,

1. what is the slope?
2. is the slope positive?

The linear regression model

Question: how is a response variable Y related to a stimulus variable X (explanatory/control/explanatory)? Assuming a linear relation,

1. what is the slope?
2. is the slope positive?
3. how good does the estimated line fit the observed data?

The linear regression model

Question: how is a response variable Y related to a stimulus variable X (explanatory/control/explanatory)? Assuming a linear relation,

1. what is the slope?
2. is the slope positive?
3. how good does the estimated line fit the observed data?

- Example: $Y =$ growth in BNP, $X =$ inflation previous year.

The linear regression model

Question: how is a response variable Y related to a stimulus variable X (explanatory/control/explanatory)? Assuming a linear relation,

1. what is the slope?
2. is the slope positive?
3. how good does the estimated line fit the observed data?

- Example: $Y =$ growth in BNP, $X =$ inflation previous year.
- D : A sample of size n of pairs explanatory variables X and response variable Y.

The linear regression model

Question: how is a response variable Y related to a stimulus variable X (explanatory/control/explanatory)? Assuming a linear relation,

1. what is the slope?
2. is the slope positive?
3. how good does the estimated line fit the observed data?

- Example: $Y =$ growth in BNP, $X =$ inflation previous year.
- D : A sample of size n of pairs explanatory variables X and response variable Y.
- **M :** $(X_1, Y_1), \cdots, (X_n, Y_n)$ is an iid random sample from an infinite population. $Y = \beta_0 + \beta_1 X + u$;
  $E[Y|X = x] = \beta_0 + \beta_1 x \Leftrightarrow Eu = 0, \quad cov(u, X) = 0$. β_0, β_1 are parameters.

OLS

Least squares and regression was known by Laplace and Gauss (80 years before Galton).

$$\begin{aligned}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) &= \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_i)\right)^2 \\ \widehat{\beta}_1 &= \frac{s_{XY}}{s_X^2} = r\frac{s_Y}{s_X}, \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}\end{aligned}$$

- Predicted response given the stimuli: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ is the fitted value.

OLS

Least squares and regression was known by Laplace and Gauss (80 years before Galton).

$$\begin{aligned}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) &= \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_i)\right)^2 \\ \widehat{\beta}_1 &= \frac{s_{XY}}{s_X^2} = r\frac{s_Y}{s_X}, \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}\end{aligned}$$

- Predicted response given the stimuli: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ is the fitted value.
- (Empirical) residual: $\widehat{u}_i = Y_i - \widehat{Y}_i$ is the vertical distance from the fitted value to the observed value of Y. $\overline{\widehat{u}} = 0, \quad r_{\widehat{u}X} = 0$

Measures of fit

- Homoscedasticity when $var(Y|X = x) = \sigma^2$ is independent of x.

Measures of fit

- Homoscedasticity when $var(Y|X = x) = \sigma^2$ is independent of x.
- Then $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n-2}RSS$ is unbiased.

Measures of fit

- Homoscedasticity when $var(Y|X = x) = \sigma^2$ is independent of x.
- Then $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n-2}RSS$ is unbiased.
- $TSS = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 = RSS + ESS$

Measures of fit

- Homoscedasticity when $var(Y|X = x) = \sigma^2$ is independent of x.
- Then $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n-2}RSS$ is unbiased.
- $TSS = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 = RSS + ESS$
- $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = r_{XY}^2$

Be aware of extreme observations!

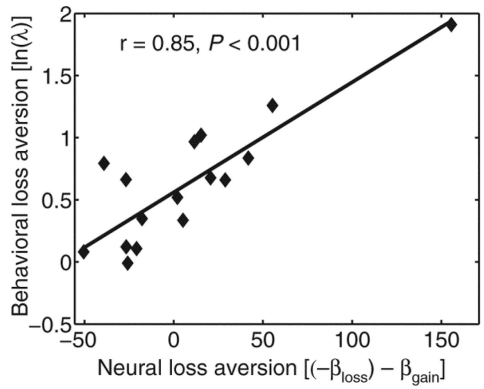


Figure: The OLS line hinges on one point.

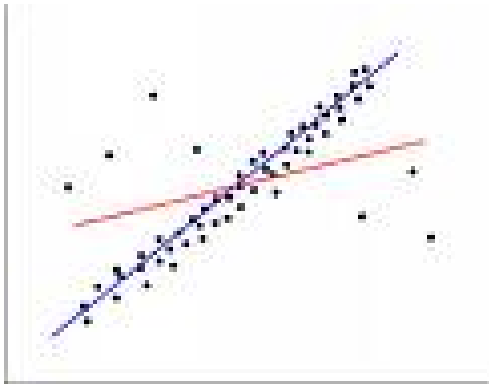
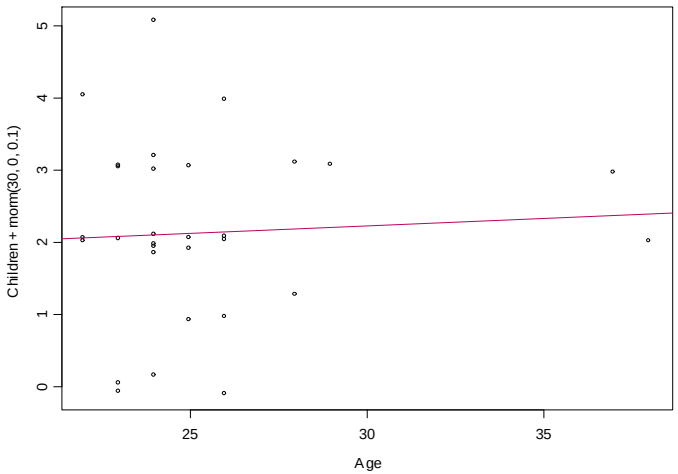


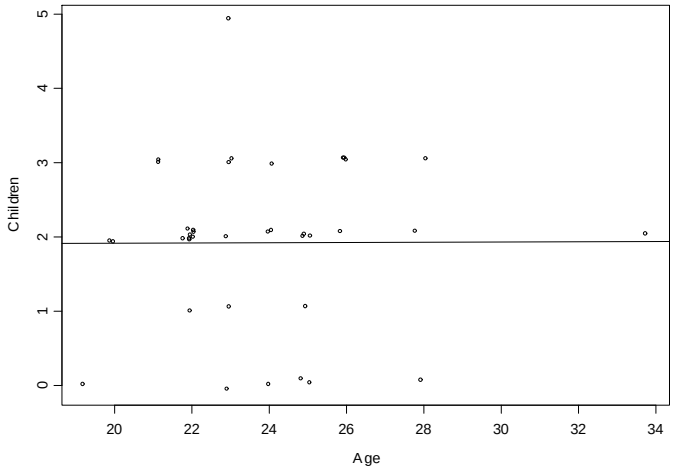
Figure: Heavy tails in the conditional distribution of $Y|X$ distorts the OLS line.

Desired number of children and age for beginning master students

	n	$\overline{Children}$	$\overline{Age}$	OLS	ρ	R^2
2008	38	1.92	23.8	$C = 1.88 + 0.002A$	0.004	0.00002
2009	29	2.13	25.4	$C = 1.61 + 0.021A$	0.06	0.004



2009



2008

Desired number of children and number of siblings, 2009

```
. tabulate Children Siblings
```

Children	Siblings 0	1	2	3	4	6	Total
0	2	1	0	1	0	0	4
1	1	1	1	0	0	0	3
2	2	3	7	0	0	0	12
3	1	1	3	2	1	0	8
4	0	0	1	0	0	1	2
5	0	0	1	0	0	0	1
Total	6	6	13	3	1	1	30

```
. reg children siblings
```

Source	SS	df	MS
Model	8.93518887	1	8.93518887
Residual	34.5314778	28	1.23326706
Total	43.4666667	29	1.49885057

Number of obs	= 30
F(1, 28)	= 7.25
Prob > F	= 0.0119
R-squared	= 0.2056
Adj R-squared	= 0.1772
Root MSE	= 1.1105

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
siblings	.4214712	.156583	2.69	0.012	.1007255	.7422169
_cons	1.416832	.3346141	4.23	0.000	.7314064	2.102258

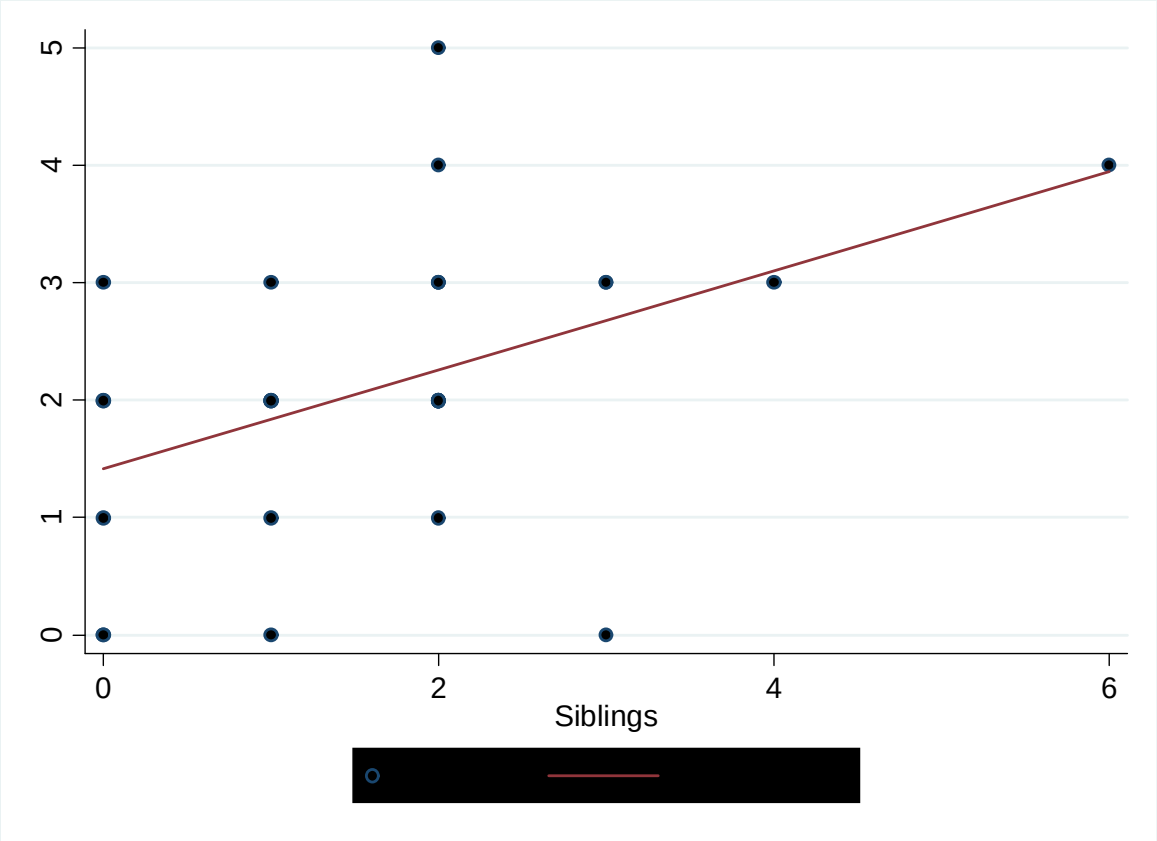


Figure: STATA scatter diagram of Children vs. Siblings, with OLS line. Some points represent several observations.

Is the (slightly) positive relation between Age and Children due to Siblings being an intermediate variable? Age$\longrightarrow$Siblings$\longrightarrow$Children, but controling for siblings, no effect of Age on Siblings?

Constructed model and simulated data

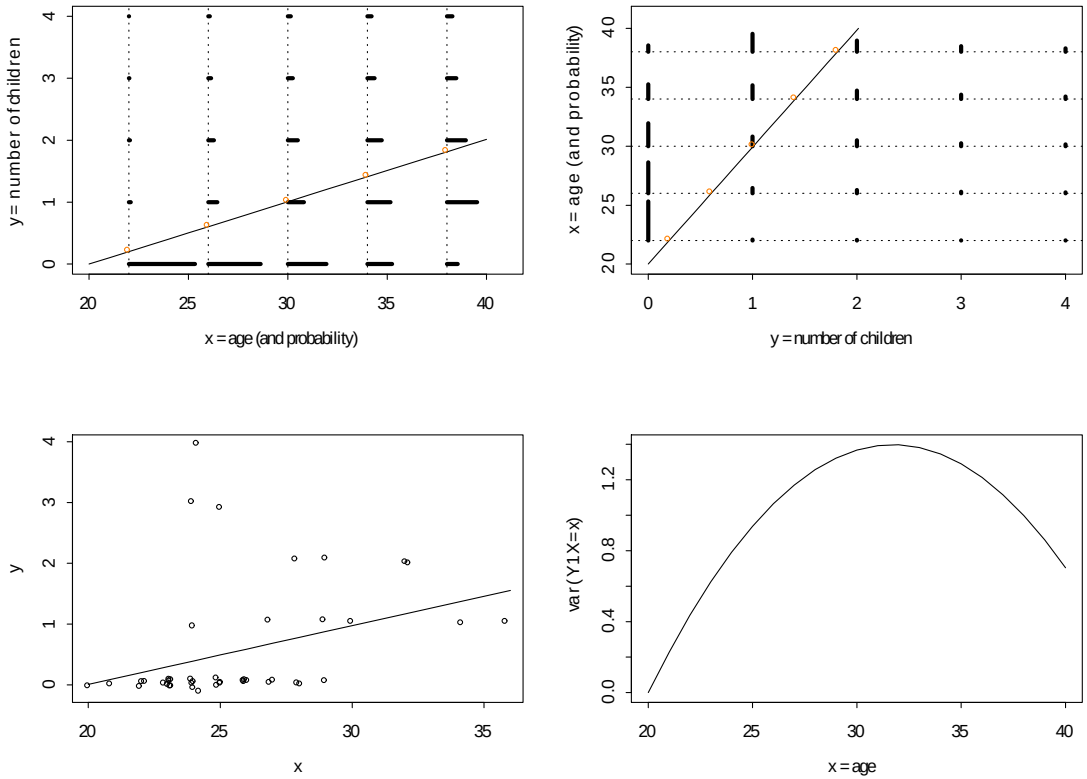


Figure: UL: $E[Y = y|X = x] = -2.02 + 0.101x$, and $P[Y = y|X = x]$ shown by horizontal thick lines for some values of x; UR: the same, but axes interchanged; LL the scatter plot of a simulated sample of size $n = 43$, the points are slightly jittered, with the OLS line $y = -1.9 + 0.097x$; LR: $var[Y|X = x]$.

Constructed model - repeated simulations

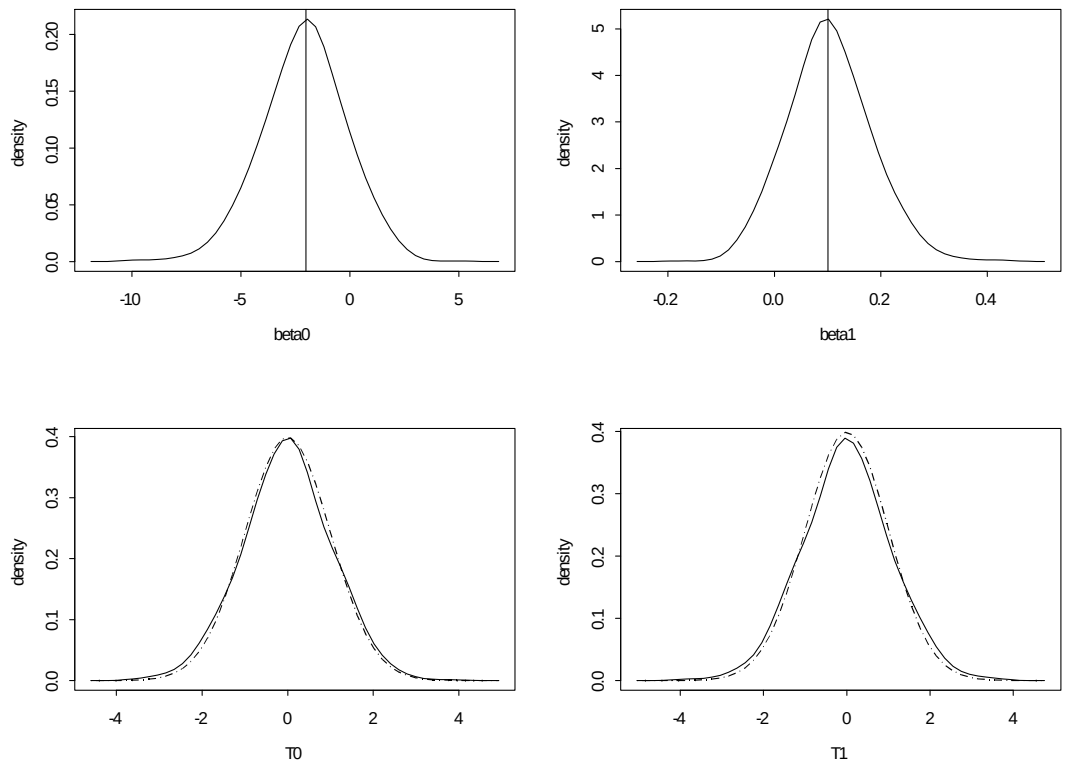


Figure: Approximate densities of $\widehat{\beta}_0$ (UL), $\widehat{\beta}_1$ (UR), $T_0 = \left(\widehat{\beta}_0 - \beta_0\right) / SE\left(\widehat{\beta}_0\right)$ (LL), and $T_1 = \left(\widehat{\beta}_1 - \beta_1\right) / SE\left(\widehat{\beta}_1\right)$ (LL), based on 1000 replicates of simulated data of size $n = 43$ in constructed model for (X, Y) .Dotted curve is the normal density.

Constructed model - repeated simulations (cont)

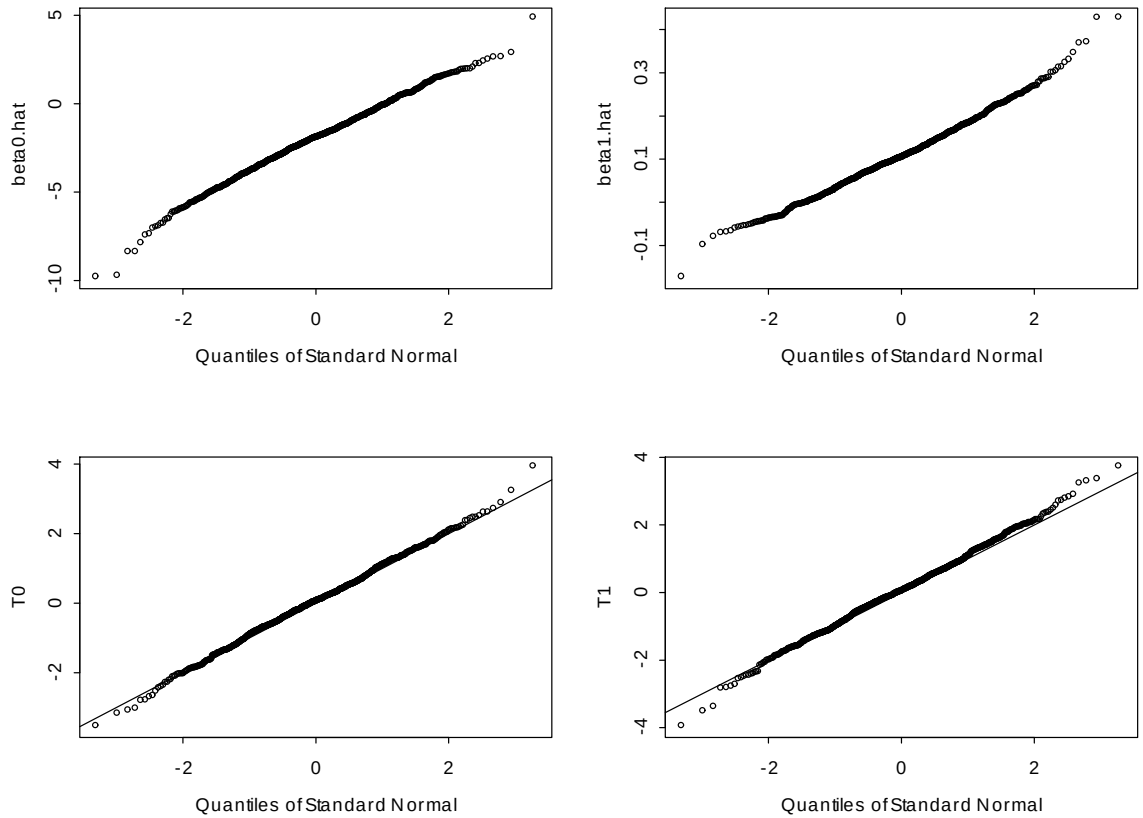


Figure: Normal probability plots (QQ-plots against N(0,1)) for the same simulated material as in previous figure.

Problems to be done in class

- SW: 4.3