Exercise 2 : mother and daughter data

$(x_i; y_i)$ $i = 1, \dots, n$ $\quad n = 11$

a) the model is $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad i = 1, \dots, n$ with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$.
moreover, we need to have $S_x^2 \neq 0$, which is satisfied here.

b) We know that the mle are

$\hat\beta_1 = \bar y - \hat\beta_2 \bar x$

$\hat\beta_2 = \frac{S_{xy}}{S_x^2}$

$\Rightarrow \hat\beta_2 = \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^n (x_i - \bar x)^2} = \frac{\sum_{i=1}^n x_i y_i - \bar x \sum_{i=1}^n y_i - \bar y \sum_{i=1}^n x_i + n \bar x \bar y}{\sum_{i=1}^n x_i^2 + n\bar x^2 - 2\bar x \sum_{i=1}^n x_i + 2n\bar x^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar x \bar y}{\sum_{i=1}^n x_i^2 - n \bar x^2}$

$= \frac{284335.1 - 11 \cdot 159.76 \cdot 161.49}{291920.6 - 11 \cdot 159.76^2} = \frac{539.039}{1134.764} = 0.454$

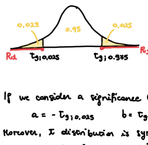
$\Rightarrow \hat\beta_1 = 161.49 - 0.454 \cdot 159.76 = 88.95$

c) compute $s^2 = \frac{1}{(n-2)} \sum_{i=1}^n (y_i - \hat y_i)^2 = \frac{1}{(n-2)} \sum_{i=1}^n e_i^2$

we need the residuals so we need the predicted values
$\hat y_i = \hat\beta_1 + \hat\beta_2 x_i = 88.95 + 0.454 \cdot x_i$
we have to compute $\hat y_i$ for all $i = 1, \dots, 11$
$i=1)$ $\hat y_1 = 88.95 + 0.454 \cdot 153.7 = 158.73$
$i=2)$ $\hat y_2 = 88.95 + 0.454 \cdot 156.7 = 160.09$
$i=3)$ ...
...
we do this computation for all $i = 1, \dots, 11$
Then, the residuals are
$i=1)$ $e_1 = y_1 - \hat y_1 = 163.1 - 158.73 = 4.35$
$i=2)$ $e_2 = y_2 - \hat y_2 = 159.5 - 160.09 = -0.60$
$i=3)$ ...
...
we obtain s^2 as $s^2 = \frac{1}{9}\left(4.35^2 + (-0.60)^2 + \dots + \dots\right) = 7.36$

d) $\begin{cases} H_0 : \beta_2 = 1 \\ H_1 : \beta_2 \neq 1 \end{cases}$
Under H_0 ("if H_0 true"), β_2 is assumed to be equal to 1
Hence, under H_0, $\hat\beta_2 \overset{H_0}{\sim} N(1, \underbrace{\text{var}(\hat\beta_2)}_{\text{contains } \sigma^2, \text{ unknown}})$
test statistic : $T = \frac{\hat\beta_2 - 1}{\sqrt{\widehat{\text{var}}(\hat\beta_2)}} \overset{H_0}{\sim} t_{n-2} = t_9$

We have a two-tailed test $\Rightarrow$ two-tailed reject region
$R_1 = (-\infty; a) \cup (b; +\infty) = R_{1,1} \cup R_{1,2}$



If we consider a significance level $\alpha = 0.05$,
$a = -t_{9;0.975}$ $\quad b = t_{9;0.975}$
Moreover, t distribution is symmetric so $a = -t_{9;0.975}$
$R_1 = (-\infty; -2.26) \cup (2.26; +\infty)$

With the data :
$t^{obs} = \frac{\hat\beta_2 - 1}{\sqrt{\widehat{\text{var}}(\hat\beta_2)}}$ we need the estimate $\widehat{\text{var}}(\hat\beta_2) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar x)^2} = \frac{7.36}{1134.764} = 0.0063$

$= \frac{0.454 - 1}{\sqrt{0.006}} = -7.04$

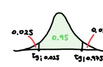
$t^{obs} \in R_1$ so we reject H_0 at a level $\alpha = 0.05$

If we want the p-value
$\alpha^{obs} = 2 \min\{ \mathbb{P}_{H_0}(T \geq t^{obs}) ; \mathbb{P}_{H_0}(T \leq t^{obs}) \}$
$= 2\, \mathbb{P}_{H_0}(T \leq t^{obs})$
$= 2\, \mathbb{P}_{H_0}(T \leq -7.04)$ where $T \overset{H_0}{\sim} t_9$
$= 2\, \mathbb{P}_{H_0}(T \geq 7.04)$
$= 2\left(1 - \mathbb{P}_{H_0}(T \leq 7.04)\right) < 0.002$
> 0.999
< 0.001



e) confidence interval with $1-\alpha = 0.95$
we want with β_2
we know that $\frac{\hat\beta_2 - \beta_2}{\sqrt{\widehat{\text{var}}(\hat\beta_2)}} \sim t_9$ (one value, unknown)

Hence $\mathbb{P}\left(t_{9;0.025} \leq \frac{\hat\beta_2 - \beta_2}{\sqrt{\widehat{\text{var}}(\hat\beta_2)}} \leq t_{9;0.975} \right)$



$\mathbb{P}\left(t_{9;0.025} \cdot \sqrt{\widehat{\text{var}}(\hat\beta_2)} \leq \hat\beta_2 - \beta_2 \leq t_{9;0.975} \sqrt{\widehat{\text{var}}(\hat\beta_2)} \right) = 0.95$

$\mathbb{P}\left(-\hat\beta_2 + t_{9;0.025} \cdot \sqrt{\widehat{\text{var}}(\hat\beta_2)} \leq -\beta_2 \leq -\hat\beta_2 + t_{9;0.975} \sqrt{\widehat{\text{var}}(\hat\beta_2)} \right) = 0.95$

$\mathbb{P}\left(\hat\beta_2 - t_{9;0.975} \sqrt{\widehat{\text{var}}(\hat\beta_2)} \leq \beta_2 \leq \hat\beta_2 - \underbrace{t_{9;0.025}}_{= -t_{9;0.975}} \cdot \sqrt{\widehat{\text{var}}(\hat\beta_2)} \right) = 0.95$

$\mathbb{P}\left(\hat\beta_2 - t_{9;0.975} \sqrt{\widehat{\text{var}}(\hat\beta_2)} \leq \beta_2 \leq \hat\beta_2 + t_{9;0.975} \sqrt{\widehat{\text{var}}(\hat\beta_2)} \right) = 0.95$

With the data, $\hat\beta_2 = 88.95$
$\widehat{\text{var}}(\hat\beta_2) = s^2 \left(\frac{1}{n} + \frac{\bar x^2}{\sum_{i=1}^n (x_i - \bar x)^2} \right) = 7.36 \left(\frac{1}{11} + \frac{159.76^2}{1134.764} \right) = 159.22$

$\sqrt{\widehat{\text{var}}(\hat\beta_2)} = 12.61$

Hence $\beta_2 \in \left(\hat\beta_2 \pm t_{9;0.975} \cdot \sqrt{\widehat{\text{var}}(\hat\beta_2)} \right)$
$\in (88.95 \pm 2.26 \cdot 12.61)$
$\in (60.45 ; 117.44)$

For β_1 we follow an analogous procedure.

f) $SST = \sum_{i=1}^n (y_i - \bar y)^2 = \sum_{i=1}^n y_i^2 - n\bar y^2 = 287179.3 - 11 \cdot 161.49^2 = 306.809$

$SSE = \sum_{i=1}^n (y_i - \hat y_i)^2 = (n-2)\, s^2 = 9 \cdot 7.36 = 66.24$

$SSR = \sum_{i=1}^n (\hat y_i - \bar y)^2 = SST - SSE = 240.56$

$R^2 = \frac{SSR}{SST} = \frac{240.56}{306.809} = 0.784$

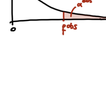
$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sqrt{\sum_{i=1}^n (x_i - \bar x)^2} \sqrt{\sum_{i=1}^n (y_i - \bar y)^2}} = \frac{539.039}{\sqrt{1134.764 \cdot 306.809}} = 0.994$

and $r_{xy}^2 = 0.79 = R^2$

g) $\begin{cases} H_0 : R^2 = 0 \\ H_1 : R^2 > 0 \end{cases}$

We use the test statistic
$F = \frac{R^2}{1 - R^2}(n-2) \overset{H_0}{\sim} F_{1,n-2} = F_{1,9}$

With the data we obtain the observed value of the test,
$f^{obs} = \frac{0.784}{1 - 0.784} \cdot 9 = 32.81$



The p-value is the probability of observing "more extreme" values than f^{obs}, "more against H_0".

The p-value is $\alpha^{obs} = \mathbb{P}_{H_0}(F \geq f^{obs})$
$= \mathbb{P}_{H_0}(F \geq 32.81) < 0.001$
We reject H_0 for $\alpha = 0.05, 0.01, 0.001$.
The model is useful to explain the variability of y

We know that the tests
$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$ and $\begin{cases} H_0 : R^2 = 0 \\ H_1 : R^2 \neq 0 \end{cases}$
are equivalent.

h) In the first case, we use the test statistic

$T = \frac{\hat\beta_2}{\sqrt{\frac{s^2}{\sum_{i=1}^n (x_i - \bar x)^2}}} \overset{H_0}{\sim} t_{n-2}$

in the second, we use
$F = \frac{R^2}{1 - R^2}(n-2) = \left(\frac{SST}{SSE} - 1 \right)(n-2) \overset{H_0}{\sim} F_{1,n-2}$

the equivalence holds because
$F = T^2$
With the data
$f^{obs} = 32.81$

$t^{obs} = \frac{0.45}{\sqrt{0.0063}} = 5.66 \quad \Rightarrow (t^{obs})^2 = 32.14$