Exercise class 1, exercise 1, parts c) and e)

András Bálint, andras.balint@chalmers.se

January 22, 2020

Use "least squares coefficient estimates" in the formula sheet. Compute the sample means first:
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{x_1+x_2+x_3+x_4}{4} = \frac{70+30+10+90}{4} = 50,$
$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{4} y_i}{4} = \frac{y_1+y_2+y_3+y_4}{4} = \frac{20+60+100+20}{4} = 50.$

Having these available allows computation of the least squares slope:
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{4}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{4}(x_i-\bar{x})^2} =$
$= \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})+(x_4-\bar{x})(y_4-\bar{y})}{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2}$
$= \frac{(70-50)(20-50)+(30-50)(60-50)+(10-50)(100-50)+(90-50)(20-50)}{(70-50)^2+(30-50)^2+(10-50)^2+(90-50)^2}$
$= \frac{-600-200-2000-1200}{400+400+1600+1600} = -1.$

Now we are ready to estimate the intercept as well:
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 50 - (-1) \times 50 = 100.$

For the confidence intervals, we will need the standard errors $\mathrm{SE}(\hat{\beta}_0)$ and $\mathrm{SE}(\hat{\beta}_0)$ which in turn requires the standard residual error. That is computed using the residual sum of squares; to get that, we first compute the residuals. As the residuals are the difference between the observed and predicted values, we need to compute them as follows:
$e_1 = y_1 - \hat{y}_1 = y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1 = 20 - 100 - ((-1) \times 70) = -10;$
$e_2 = y_2 - \hat{y}_2 = y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2 = 60 - 100 - ((-1) \times 30) = -10;$
$e_3 = y_3 - \hat{y}_3 = y_3 - \hat{\beta}_0 - \hat{\beta}_1 x_3 = 100 - 100 - ((-1) \times 10) = 10;$
$e_4 = y_4 - \hat{y}_4 = y_4 - \hat{\beta}_0 - \hat{\beta}_1 x_4 = 20 - 100 - ((-1) \times 90) = 10.$
(Note: in the output from R, the residuals are reordered by x, hence we get the same residual values, but in a different order.)

Having the residuals available allows the computation of the residual sum of squares, the standard residual error and the standard error of the coefficient estimates:
$\mathrm{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2 = e_1^2 + e_2^2 + e_3^2 + e_4^2 = (-10)^2 + (-10)^2 + 10^2 + 10^2 = 400,$
$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{2}400} = 14.1421,$

$\mathrm{SE}(\hat{\beta}_0) = \mathrm{RSE} \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} =$
$= 14.1421 \times \sqrt{\frac{1}{4} + \frac{50^2}{(70-50)^2+(30-50)^2+(10-50)^2+(90-50)^2}} = 13.2288$
$\mathrm{SE}(\hat{\beta}_1) = \mathrm{RSE} \times \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}} =$
$= 14.1421 \times \sqrt{\frac{1}{(70-50)^2+(30-50)^2+(10-50)^2+(90-50)^2}} = 0.2236.$

Therefore, the confidence intervals for the coefficients using the simplified formulas are as follows:
$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0) = 100 \pm 2 \cdot 13.2288$, i.e. the interval $[73.5425, 126.4575]$;
$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1) = -1 \pm 2 \cdot 0.2236$, i.e. the interval $[-1.4472, -0.5528]$.

Note: for such a small sample, the simplified intervals are NOT good enough. For the precise confidence intervals, 2 should be replaced by the 97.5% quantile of the t distribution with $n-2$ degrees of freedom, i.e. in this case the t distribution with df=2. Tables or computers help us to find out that this value is 4.3027, i.e. much larger than 2. The precise confidence intervals for coefficients are therefore:
$\hat{\beta}_0 \pm 4.3027 \cdot \mathrm{SE}(\hat{\beta}_0) = 100 \pm 4.3027 \cdot 13.2288$, i.e. the interval $[43.0804, 156.9196]$;
$\hat{\beta}_1 \pm 4.3027 \cdot \mathrm{SE}(\hat{\beta}_1) = -1 \pm 4.3027 \cdot 0.2236$, i.e. the interval $[-1.9621, -0.0379]$.

The proportion of variability in completion time explained by the number of employees assigned to the project can be computed by using the formula for R^2 in the formula sheet (and remembering that we have computed RSS above):
$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}}{(y_1-\bar{y})^2+(y_2-\bar{y})^2+(y_3-\bar{y})^2+(y_4-\bar{y})^2} =$
$= 1 - \frac{400}{(20-50)^2+(60-50)^2+(100-50)^2+(20-50)^2} = 0.9091$