MA329 Statistical linear models

Ch2. Example

1. Two processes for hydraulic drilling of rock are dry drilling and wet drilling. In a dry hole, compressed air is forced down the drill rods to flush the cuttings and drive the hammer; in a wet hole, water is forced down. An experiment was conducted to determine whether the time *y* (in minutes) it takes to dry drill a distance of 5 feet in rock increases with depth *x* (in feet). [Data can also be found in the file: DRILLROCK.csv]

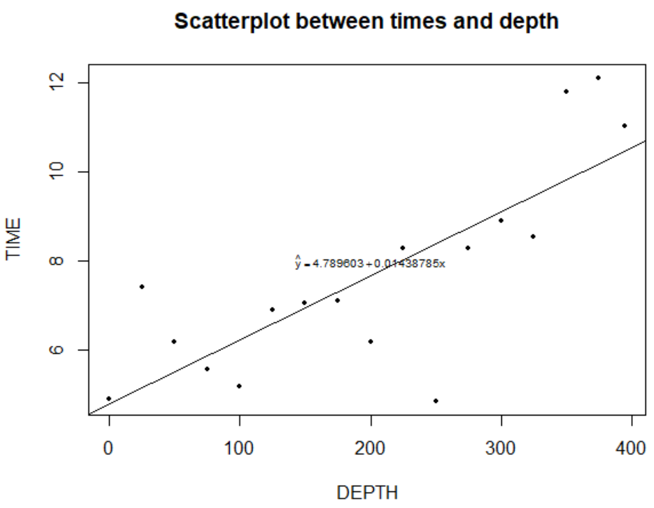
x	0	25	50	75	100	125	150	175	200	225	250	275	300	325	350	375	395
y	4.9	7.41	6.19	5.57	5.17	6.89	7.05	7.11	6.19	8.28	4.84	8.29	8.91	8.54	11.79	12.12	11.02

a. Construct and comment a scatterplot of the data.

```
xy <- read.csv("D:shi/DRILLROCK.csv",header=T)
# a. scatterplot

plot(xy,pch=16,cex=0.5) #scatter plot
title("Scatterplot between times and depth")

x=xy$DEPTH
y=xy$TIME
n=length(x)
lm.sol=lm(y~x)
abline(lm.sol) #add line to the scatterplot
text(200, 8, labels = bquote(hat(y) == .(beta0) + .(beta1) * x),cex=0.6)
```



Shows a quite strong linear relationship between time (y) and DEPTH (x).

b. Find the least squares line from the data and plot it on your scatterplot.

$\bar{X} = 199.7059 \qquad \bar{Y} = 7.662941 \qquad n = 17$

$S_{xx} = \Sigma(x_i - \bar{x})^2 = 253023.5$

$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 3640.465$

$S_{yy} = \Sigma(y_i - \bar{y})^2 = 83.14615$

$\hat{\beta}_1 = S_{xy}/S_{xx} = 0.01438785$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 4.78960$

c. What is your regression model? State the necessary assumptions.

The regression model is

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \hat{\beta}_0 = 4.78960$

$\hat{\beta}_1 = 0.01438785$

assumptions:

(1) y_i's are independent for $i = 1, \ldots, n$.

(2) $Var(\varepsilon_i) = \sigma^2$ — constant variance

(3) $\varepsilon_i \sim N(0, \sigma^2)$ — (not necessary for obtaining LSE)

d. Test the hypothesis that the depth of the rock provides no information for the prediction of the time required to drill a distance of 5 feet when a linear model is used (use $\alpha = 0.05$). State the null and alternative hypotheses. Draw the appropriate test conclusions.

$H_0: \beta_1 = 0 \quad$ v.s. $\quad H_1: \beta_1 \neq 0$

$SSE = \Sigma(y_i - \hat{y}_i)^2 = 30.767691$

$s^2 = \frac{SSE}{n-2} = 2.051179$

$t = \frac{\hat{\beta}_1 - \beta_1}{s / S_{xx}^{1/2}} \overset{H_0}{=} \frac{0.01438785}{(2.051179/253023.5)^{1/2}}$

$= 5.053214$

$\geq t_{0.025, 15} = 2.13145$

$\Rightarrow$ Reject H_0.

or P-value $= 2\Pr(t_{15} \geq 5.053214)$

$= 0.00014$

e. Find a 95% confidence interval for β_1 (the slope of the linear regression model). Interpret your results.

95% C.I. of β_1:

$\hat{\beta}_1 \pm t_{0.025, 15} \cdot s / S_{xx}^{1/2}$

$= (0.00831914, 0.02045656)$

there is 95% chance that β_1 would take values between 0.00831914 and 0.02045656.

f. Find the coefficient of determination for the linear regression model. Interpret your result.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \cdots = 0.6299$$

meaning: 63% of the variations of the time in the sample can be ~~excep~~ explained by the model.

g. What is the regression prediction equation? Find a prediction for the **mean** amount of time to drill a distance of 5 feet when depth is 6 feet and its 95% interval.

$X_h = 6$. $\quad \hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = 4.87593$

C.I. of $E(Y_h)$ is

$$\hat{Y}_h \pm t_{0.025,\,15} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}}$$

$$= (3.48663,\ 6.25197)$$

h. Find a 95% interval for the amount of time for a **single** drill (5 feet) when depth is 6 feet.

Predictive interval of Y_h at $x_h = 6$ is

$$\hat{y}_h \pm t_{0.025,15} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}}$$

$= (1.522017,\ 8.229842)$

i. Give the ANOVA table and interpret the result using the F test.

	SS	d.f	MS	F
Regression	SSR = SST − SSE = 52.37846	1	52.37846	MSR/MSE = 25.53578
Error	SSE = 30.76691	n−2 = 15	MSE = SSE/15 = 2.051179	
Total	SST = Syy = 83.14615	n−1 = 16		

p-value $= \Pr(F_{1,15} \geq 25.53578) = 0.0001428 5 \leq 0.05$

meaning the model with the independent variable fits the data better than the intercept-only model!