7 Heavy Metals

Metals.txt online under labs contains information on 17 core samples from Louisiana. Each sample's depth was recorded in meters, and then each core was tested for zinc (ppm) and iron concentration (percentage). You will be exploring the data to determine if depth can be used to predict either zinc or iron concentration.

1. What does a basic data analysis reveal about the variables? Are there interesting features to the data? depth/zinc are skewed right. Iron is more normal but has possible outliers, Depth & iron have high outliers.

2. Identify the response and explanatory variables. Note there are 2 different regressions to consider here. Exp: Depth Resp: Zinc, Iron

3. Estimate the correlations and find exact values using Rcmdr. Does a linear regression appear appropriate for predicting either zinc or iron? It may not appear appropriate for all examples.

D, I ≈ mod. strong .7473 D, Z ≈ weak, nonexistent −.0165

4. Fit a regression model if appropriate. Report the fitted regression line. If someone wanted predictions for a depth of 50 m, what would you tell them? can't fit for zinc but can iron

$\hat{y} = 3.078 + .0116\,x$ in range for prediction

$x = 50 \Rightarrow \hat{y} = 3.658$

5. Does the model look like it fits well? Explain.

$R^2 = .5585$ and $s_e = .307$ y-values range 2.9-4.8 so about 1/10

It is a decent fit, not poor but not great either.

6. Test to see whether or not the predictor is a significant predictor of the response.

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ $t = 4.356$ p-value = $.000564$ Reject H_0

We have evidence of a significant linear relationship b/w depth and iron concentration,

7. Make a confidence interval for the slope. 95% t^*_{15} 2.131

$b_1 \pm t^* se(b_1) \Rightarrow$

$.0116 \pm 2.131\,(.0027) \Rightarrow .0116 \pm .0058 \Rightarrow (.0058, .0174)$

8. What are the four regression assumptions? Do the regression assumptions appear valid? Explain.

1. Linear relationship. Yes. Scatterplot shows /.
2. Residuals are a random sample. Must assume.
3. Errors have constant variance. Residual plot looks decent, no glaring problems or patterns.
4. Errors are normally 4 dist. QQ plot of residuals looks good → all pts are in bounds though upper tail is little off.

8 Forest Fires

A data set on forest fires (different than the previous one we had during hypothesis testing) lists number of fires in thousands along with number of acres burned in millions. Does the number of fires appear to be a significant predictor for number of acres burned? (Data from Triola). Online as FFx2.txt.

1. What does a basic data analysis reveal about the variables? Are there interesting features to the data? Acres is right-skewed but no outliers. Med. around 2.75.

Fires is maybe bimodal - peaks @ 45,65 but no outliers Med @ 62-63

2. Identify the response and explanatory variables.

Exp: Fires, Resp: Acres

3. Estimate the correlation and find the exact value using Rcmdr. Does a linear regression appear appropriate? It may not appear appropriate for all examples. yes ok but not very strong

moderate .4 maybe .5173

4. Fit a regression model if appropriate. Report the fitted regression line. What is your best estimate of the number of acres burned if there were 50,000 fires?

$\hat{y} = -1.1294 + .06769x$ $x = 50$ in range

$\hat{y} = 2.2551$ 2 1/4 million acres

5. Does the model look like it fits well? Explain.

$R^2 = .2677$ and $S_e = 1.416$ with y values 1.5-6.2

so no, poor fit

6. Test to see whether or not the predictor is a significant predictor of the response.

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ $t = 1.814$ p-value = .103

Do Not Reject H_0

We do not have evidence of a sign linear relationships btw fires and acres.

7. Make a confidence interval for the slope.

$b_1 \pm t^*_9 \, se(b_1) \Rightarrow .06769 \pm (2.262)(.0372) \Rightarrow (-.01673, .15211)$

8. Do the regression assumptions appear valid? Explain.

1. LR ⇒ yes, weak
2. RS of residuals, not stated so assume
3. Residual plot has a +/- pattern but variance looks constant
4. QQ plot checks out nicely so errors appear normally distributed.