

Simple Linear Regression in R Handout

1 Background

Rainbow trout, mountain whitefish and large-scale suckers taken from 4 different localities along the Spokane River (eastern Washington) during July, August and October of 1999 were analyzed for three heavy metals (lead, zinc and cadmium) by Manchester Laboratory for the Washington State Department of Ecology. Most of the analyses were done on filets, however a few whole specimens were tested as well. Metal contents are reported in milligrams of metal per kilogram of fish (mg/kg), which is an equivalent unit to parts per million (ppm). A few redundant analyses indicate crude uncertainties in reproducibility of at least 14 ppm for zinc, .07 ppm for lead, and .01 ppm for cadmium. The source of the heavy metals is upstream from Spokane in the Coeur d'Alene mining district of northern Idaho, one of the richest mining districts in the US and also one of the most heavily contaminated. Acid mine drainage directly from shafts and adits, and from leaching of metal rich mine tailings (waste rock), as well as metal-rich discharges from smelters, have contaminated many streams, rivers and lakes in the Spokane/Coeur d'Alene watershed. Metals are both dissolved in the river water and found as minute particles, and can enter the food web at various stages. The zinc and lead data are found in a tab-delimited text file (without a header) at <http://seattlecentral.edu/qelp/sets/022/s022.txt>. Note that the lead measurements are in the first column.

2 Initialization

```
> library(NCStats)

> sr <- read.table("http://seattlecentral.edu/qelp/sets/022/s022.txt",header=FALSE)
> str(sr)

'data.frame':      10 obs. of  2 variables:
 $ V1: num  0.73 1.14 0.6 1.59 4.34 1.98 3.12 1.8 0.65 0.56
 $ V2: num  45.3 50.8 40.2 64 150 106 90.8 58.8 35.4 28.4

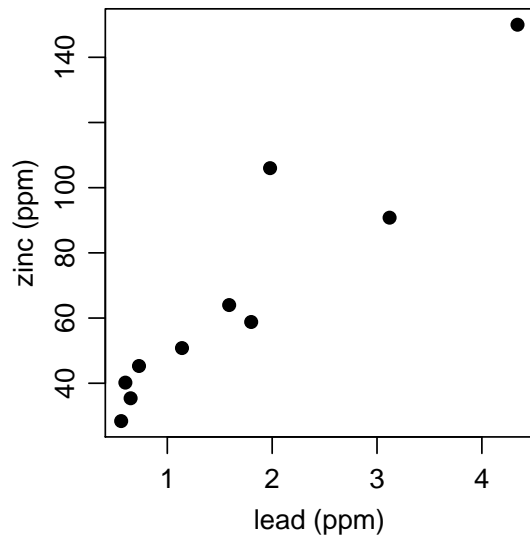
> names(sr) <- c("lead","zinc")
> str(sr)

'data.frame':      10 obs. of  2 variables:
 $ lead: num  0.73 1.14 0.6 1.59 4.34 1.98 3.12 1.8 0.65 0.56
 $ zinc: num  45.3 50.8 40.2 64 150 106 90.8 58.8 35.4 28.4

> attach(sr)
```

3 Bivariate EDA

```
> plot(zinc~lead,pch=19,xlab="lead (ppm)",ylab="zinc (ppm)")
```



```
> cor(zinc,lead)
```

```
[1] 0.9405332
```

4 Fitting the Regression

```
> lm.sr <- lm(zinc~lead)
```

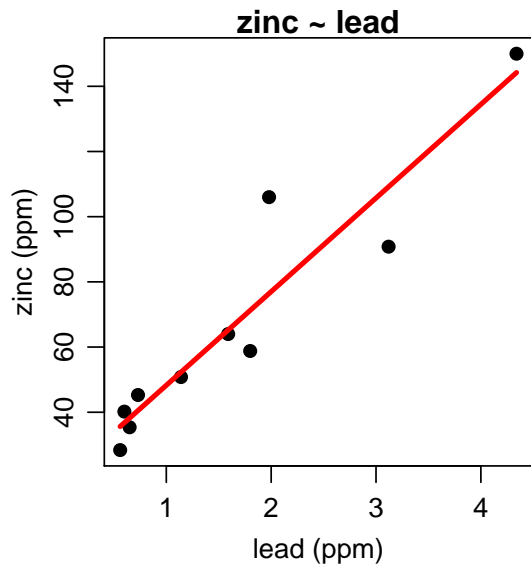
```
> coef(lm.sr)
```

```
(Intercept)      lead  
  19.54465      28.72523
```

```
> summary(lm.sr)$r.squared
```

```
[1] 0.8846028
```

```
> fit.plot(lm.sr,xlab="lead (ppm)",ylab="zinc (ppm)")
```



5 Using the Regression

Predict the zinc level if the lead level is known to be 2.5 ppm?

```
> predict(lm.sr, data.frame(lead=2.5))
      1
91.35772

> prediction.plot(lm.sr, data.frame(lead=2.5))

  obs lead   fit   lwr   upr
1   1  2.5 91.35772 57.47781 125.2376
```

What is the residual if that same individual had a zinc level of 100 ppm?

```
> yhat <- predict(lm.sr, data.frame(lead=2.5))
> 100-yhat

      1
8.64228
```

6 Good “Ending” Commands

```
> detach(sr)
```

7 Class Exercise

Recall the IBI and percent impervious surface data from the bivariate EDA handout. Use these data to answer the following questions.

1. What is the response variable?
2. What is the explanatory variable?
3. In terms of the variables of this problem, what is the equation of the best-fit line?
4. In terms of the variables of this problem, INTERPRET the value of the slope?
5. In terms of the variables of this problem, INTERPRET the value of the y-intercept?
6. What is the predicted IBI for an impervious surface percentage of 80%?
7. What is the predicted IBI for an impervious surface percentage of 20%?
8. What is the residual if the impervious surface percentage is 30 and the IBI is 20?
9. What is the correlation coefficient between pressure and index count?
10. What proportion of the variability in IBI is explained by knowing the percentage of impervious surface?
11. What aspect of this regression analysis concerns you (i.e., consider the regression assumptions)?