Stack loss

The built-in R dataset `stackloss` has measurements on 21 days of operation of a plant for the oxidation of ammonia (NH3) to nitric acid (HNO3).

Air.Flow represents operation rate of plant
Water.Temp temperature of cooling water
Acid.Conc. parts per 1000 minus 500
stack.loss 10 times percentage of ingoing ammonia that escapes

\[
\text{stack.loss} = \beta_0 + \beta_1 \text{Air.Flow} + \beta_2 \text{Water.Temp} + \beta_3 \text{Acid.Conc.} + \varepsilon
\]

```
> lsl = lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc., data=stackloss)
> summary(lsl)
Call:
  lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
      data = stackloss)
Residuals:
      Min     1Q   Median     3Q    Max
-7.238 -1.712  -0.455  2.361  5.698
Coefficients:
                Estimate  Std. Error   t value Pr(>|t|)
(Intercept)   -39.9200     11.8960    -3.360    0.0038 **
Air.Flow       0.7159     0.1350     5.308    5.8e-05 ***
Water.Temp    1.2949     0.3680     3.520    0.0026 **
Acid.Conc.    -0.1523     0.1560    -0.970    0.3440
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.24 on 17 degrees of freedom
Multiple R-squared: 0.914,  Adjusted R-squared: 0.898
F-statistic: 59.9 on 3 and 17 DF,  p-value: 3.02e-09
```

The regression line is

\[
\text{stack.loss} = -39.92 + .72 \text{Air.Flow} + 1.30 \text{Water.Temp} - .15 \text{Acid.Conc.}
\]

Two important hypothesis tests:

\[
H_0 : \beta_1 = \cdots = \beta_p = 0 \quad H_0 : \beta_i = 0
\]

\[
R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \quad \text{adj.-} R^2 = R^2 - \left( \frac{p}{n - p - 1} \right) (1 - R^2)
\]
Ice

Exercise 8.3.18 of Navidi gives data on the thickness of ice in 13 Minnesota lakes. The data are in `ex8_3_18`. The variables are

- **y**: maximum ice thickness in mm
- **x1**: average number of days per year of ice cover
- **x2**: average number of days low temperature is lower than 8 Centigrade
- **x3**: average snow depth in mm

Fit a model of the form \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \) to this data.

1. Write the equation of the fitted model.
2. Do lakes with greater snow cover tend to have greater or less maximum ice thickness? Explain.
3. Are all the explanatory variables useful in this model?

Polynomials

The variables on the right hand side of the model need not be independent. We can use linear regression to fit a polynomial to the data.

`ex8_3_15` has data on the relationship between extrusion pressure (\( x \), in KPa) and wear (\( y \), in mg).

We can use `lm` to fit a quadratic to the data.

```r
> l=lm(y~x+I(x^2),data=ex8_3_15)
```

4. What is \( R^2 \) for this relationship?
5. Plot the residuals and comment.
6. Fit a third-degree polynomial to these data. Give an argument that the second degree polynomial is a better model.
7. Something is quite peculiar about the third degree model. Look at the tests of the various hypotheses and say what it is.

Madras traffic

Data on traffic characteristics at 10 intersections in Madras, India are reported. The data are in `ex8_4_9`.

- **y**: median speed in km/h
- **x1**: width in m
- **x2**: traffic volume in vehicles per lane hour

8. Fit the model \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \) and test the hypotheses that the coefficients are 0.
9. Do the same for the models \( y = \beta_0 + \beta_1 x_1 + \varepsilon \).
10. Do the same for the model \( y = \beta_0 + \beta_1 x_2 + \varepsilon \).
11. Which of the above three models do you think is best and why?
Be careful

The dataframe EducExp in the M241 package has data on secondary school education in the states.

12. Write average SAT score of graduates (SAT) as a linear function of expenditures per student ((Exp)).
13. This relationship suggests that SAT scores go up as expenditures go ________________.
14. Now write average SAT score as a linear function of both the expenditures per student and the percentage of students in the state that take the SAT (PctSAT).
15. Explain what this last equation says about the relationship between educational expenditures and SAT scores.

**Full quadratic model**

ex8_4_11 has data on the resilience modulus of pavement. The variables are

- **y**  resilience modulus in 10^6 kPa,
- **x1**  surface area of aggregate in m^2/kg,
- **x2**  softening point of asphalt in degrees Centigrade

The full quadratic model in this case has the form

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon \]

We can fit the full quadratic model with lm.

```r
> lq=lm(y~ x1+x2+x1:x2+I(x1^2)+I(x2^2),data=ex8_4_11)
```

The term `x1:x2` corresponds to the term for \(x_1 x_2\) in the model and of course `x1^2` corresponds to the \(x_1^2\) term.

While the full quadratic model might be a place to start, a model with too many terms is not useful and even sometimes misleading.

16. Write the equation for the full quadratic model for these data. Notice the large difference between \(R^2\) and adj \(-R^2\).
17. What does the \(F\)-test for this model say?
18. Find a simpler model that seems better than the full quadratic model.