HARALD SCHMIDBAUER İstanbul Bilgi Üniversitesi

Bus 701: Advanced Statistics

Fall 2011

SOME PROBLEMS

Problem 1: Consider the multiple linear regression model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

a) Which kind of data set is required to fit this model? (That is, how would the data set be given in a spreadsheet program, such as MS-Excel?)

b) This model can be used from either a descriptive or an inductive point of view. Briefly discuss the main goals of either approach.

c) What exactly is meant when we say: "X_1 and X_2 are significant for Y"?

d) When diagnosing the regression: Which features should make us suspicious because the regression might be spurious or meaningless? Please give a few hints.

Problem 2: Data mining is a methodology for analyzing large data sets. Discuss the differences in the assumptions and goals between traditional statistics and data mining. The odds ratio is a typical tool in data mining. How is it defined, what can be achieved with it? Please explain some details.

Problem 3: Adolph Quetelet's ideas can be considered an important contribution to the development of statistical methodology. Discuss Quetelet's scientific outlook and the limitations inherent in his approach. When, and in what context, did these limitations become obvious? How were they overcome?

Problem 4: In an effort to investigate the determinants of four-star hotel room rates (price per night) in Prague in June 2011, the values of several variables were recorded for each hotel in a sample taken from the website www.hotels.com. The following table gives variable names and their explanations, minimum and maximum values observed in the sample, and sample means.

variable	explanation	minimum	maximum	average
`price`	room rate = price for 1 night (€)	52	166	104
`rating`	average customer rating	7.2	9.2	8.45
`number.ratings`	number of customer ratings	10	263	47.7
`ln.number.ratings`	natural logarithm of number.ratings	2.303	5.572	3.518
`distance`	distance from the city center (km)	0.1	11.0	2.6
`ln.distance`	natural logarithm of distance	−2.303	2.398	0.433

A higher rating indicates a higher degree of customer satisfaction. The number of customer ratings is the number of customers who gave a rating on www.hotels.com.
Computer output when fitting a regression model to the data, with `price` as dependent variable:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -49.041     55.260  -0.887  0.38027
rating              15.196      6.306   2.410  0.02079 *
ln.distance        -11.988      3.407  -3.519  0.00112 **
ln.number.ratings    8.509      4.398   1.934  0.06033 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.61 on 39 degrees of freedom
Multiple R-squared: 0.4943,Adjusted R-squared: 0.4554
F-statistic: 12.71 on 3 and 39 DF,  p-value: 6.159e-06
```

a) What is the sample size?

b) Write the regression equation of the fitted model.

c) Why do you think it was found better to use the natural logarithm of the number of ratings, rather than the number of ratings itself, as independent variable in the model?

d) What can be done with the fitted model, for what purposes is it useful?

e) Consider the following statement about the p-value (`Pr(>|t|)`) of 0.02079 in the second row of the table above:

> "The p-value is the probability that the regression coefficient of `rating` is actually zero."

Is this statement true? Explain.

f) How is the p-value in (e) computed? (You need not compute its numerical value, but explain the steps and assumptions necessary in some detail.)

g) Explain the meaning of `Multiple R-squared` (R^2) in the fitted model. What precisely is explained by this model?

h) Now suppose we fit another model to the data set, eliminating `ln.distance` from the list of independent variables. Which behaviour would you expect for (i) the residual standard error, (ii) the degrees of freedom, (iii) R^2? (Will they increase or decrease, or do you think this is impossible to say without actually computing them? Give reasons for your answer.)