Statistics for Engineering, 4C3/6C3

Assignment 6

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 15 March 2013**

Note: Assignment objectives

- Using and interpreting an MLR model with integer variables.
- Using an MLR with integer variables that are at more than 2 levels.

Question 1 [0]

This question is fully solved in the course textbook, Process Improvement using Data. So it is worth no credit, and will not be graded. However, you are strongly recommended to complete the question without looking at the answers.

In this question we will use the LDPE data which is data from a high-fidelity simulation of a low-density polyethylene reactor. LDPE reactors are very long, thin tubes. In this particular case the tube is divided in 2 zones, since the feed enters at the start of the tube, and some point further down the tube (start of the second zone). There is a temperature profile along the tube, with a certain maximum temperature somewhere along the length. The maximum temperature in zone 1, `Tmax1` is reached some fraction `z1` along the length; similarly in zone 2 with the `Tmax2` and `z2` variables.

We will build a linear model to predict the `SCB` variable, the short chain branching (per 1000 carbon atoms) which is an important quality variable for this product. Note that the last 4 rows of data are known to be from abnormal process operation, when the process started to experience a problem. However, we will pretend we didn't know that when building the model, so keep them in for now.

1. Use only the following subset of x-variables: `Tmax1`, `Tmax2`, `z1` and `z2` and the y variable = `SCB`. Show the relationship between these 5 variables in a scatter plot matrix.

   Use this code to get you started (make sure you understand what it is doing):

   ```
   LDPE <- read.csv('http://datasets.connectmv.com/file/ldpe.csv')
   subdata <- data.frame(cbind(LDPE$Tmax1, LDPE$Tmax2, LDPE$z1, LDPE$z2, LDPE$SCB))
   colnames(subdata) <- c("Tmax1", "Tmax2", "z1", "z2", "SCB")
   ```

   Using bullet points, describe the nature of relationships between the 5 variables, and particularly the relationship to the y-variable.

2. Let's start with a linear model between `z2` and `SCB`. We will call this the `z2` model. Let's examine its residuals:

   (a) Are the residuals normally distributed?

   (b) What is the standard error of this model?

   (c) Are there any time-based trends in the residuals (the rows in the data are already in time-order)?

   (d) Use any other relevant plots of the predicted values, the residuals, the x-variable, as described in class, and diagnose the problem with this linear model.

   (e) What can be done to fix the problem?

Question 2 [6]

Operators have noticed differences in the yield from our batch process [g/L] depending on the raw material supplier. You've collected data from the last 12 batches and coded the data from the city and country of origin:

```
# 1 = València, Spain
# 2 = Luxembourg, Luxembourg
# 3 = Utrecht, Netherlands

country <- c(3, 2, 1, 3, 1, 1, 2, 2, 2, 1, 3, 3)
yield <- c(72.9, 69.3, 70.8, 79.1, 66.3, 73.3, 65.1, 66.5, 54.9, 74.7, 80.8, 79.3)
```

Build a linear model that predicts the yield from the country of origin. Make sure you reassign the `country` variable as follows: `country <- as.factor(country)` before you use it in the model (and understand what the `as.factor(...)` function does).

1. Interpret the `Intercept` term, the `country2` slope coefficient and the `country3` slope coefficient in your written answer. If you haven't yet discovered and used the `model.matrix(...)` command, you will need it here.
2. What have you learned from this model?
3. Is what you have learned still valid when you consider the 95% confidence intervals for the slope coefficients? Explain clearly in your answer.

Solution

We can write in R:

```
country <- as.factor(country)
mod <- lm(yield ~ country)
model.matrix(mod)  # useful output

Call:
lm(formula = yield ~ country)

Residuals:
   Min     1Q Median     3Q    Max
-9.050 -1.600  1.212  2.606  5.350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   71.275      2.333  30.554 2.11e-10 ***
country2      -7.325      3.299  -2.220   0.0535 .
country3       6.750      3.299   2.046   0.0711 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.666 on 9 degrees of freedom
Multiple R-squared: 0.6693,     Adjusted R-squared: 0.5958
```

1. The `model.matrix(mod)` command shows that R has coded Spain as (`country2`, `country3`) = (0, 0), while Luxembourg is (1, 0), and Netherlands is (0, 1).

   So this implies the intercept of 71.3 g/L is the yield we expect, on average, when we source the raw material from Spain.

   The yield will decrease on average by 7.3 g/L, relative to Spain's yield, when we obtain it from Luxembourg. The yield will increase on average by 6.75 g/L, relative to Spain's yield, when we source the raw materials from Utrecht.

2. We've learned that better yields are obtained if using material from The Netherlands, followed by Spain, followed by Luxembourg.

3. Despite that the 95% confidence intervals span zero, they do so very asymmetrically:

```
> confint(mod)
                  2.5 %     97.5 %
(Intercept)  65.9979126 76.5520874
country2    -14.7879285  0.1379285
country3     -0.7129285 14.2129285
```

   so, our conclusions above are still very reasonable. They are statistically valid at the 90% level: `confint(mod, level=0.90)`.

Question 3 [5]

In a previous assignment you compared the `TK104` reactor to the `TK105` using the Brittleness Index dataset.

1. Repeat the confidence interval calculation for the comparison between the `TK104` and `TK105` reactors, assuming the variances can be pooled. Report your answer as:

$$\text{LB} \leq \mu_{105} - \mu_{104} \leq \text{UB}$$

2. Now build a linear model that uses a single integer variable coded as `1` when running the batch in `TK105`, and coded as `0` when running the batch in `TK104`. The y-variable is the brittleness index value.

   Prove to yourself that you get the same confidence interval for the integer variable, as you do with the regular confidence interval in the first part of the question. Make sure you can explain why this is the case.

Solution

1. The regular method for building a confidence (must be shown for full grade) can be used to obtain $-31.4 \leq \mu_{105} - \mu_{104} \leq 134$.

2. The R code below was used to code the integer variable, d_R, as required. The confidence interval for the slope coefficient, b_R, found using the `confint(...)` was $-31.4 \leq b_R \leq 133.7781$, where $y = b_0 + b_R d_R$. The `model.matrix(...)` function confirms our raw data was coded as we expected.

```
brittle <- read.csv('http://datasets.connectmv.com/file/brittleness-index.csv')

# Report the t-test bounds: LB=-31.4 to UB=134.
# Make sure you know how to use t.test function, and to interpret it.
# Or do the work by hand.
t.test(brittle$TK105, brittle$TK104, alternative = "two.sided", var.equal = TRUE)

# Linear model. There are many ways to create the factor variable. Here's one:
TK104 <- seq(from=0, to=0, length.out=length(brittle$TK104))
TK105 <- seq(from=1, to=1, length.out=length(brittle$TK105))
reactor <- as.factor(c(TK104, TK105))
brittleness <- c(brittle$TK104, brittle$TK105)

# Delete the NAs
reactor <- reactor[!is.na(brittleness)]
brittleness <- brittleness[!is.na(brittleness)]
mod <- lm(brittleness ~ reactor)
summary(mod)
confint(mod)
model.matrix(mod)
```

Question 4 [6]

1. Using the data from the previous question, code the integer variable in the linear model as `0` when running the batch in `TK105`, and code it as `1` when running the batch in `TK104`. The y-variable is the brittleness index value. Report the slope coefficient and confidence interval. (*This question is mostly a repeat of the previous one*).
2. Now code the integer variable in the linear model as `1` when running the batch in `TK105`, and code it as `2` when running the batch in `TK104`. The y-variable is the brittleness index value. Report the slope coefficient and confidence interval. How do the answers compare? Explain any differences or similarities you observe.
3. Now code the integer variable in the linear model as -1 when running the batch in `TK105`, and code it as $+1$ when running the batch in `TK104`. The y-variable is the brittleness index value. Report the slope coefficient and confidence interval. How do the answers compare? Explain any differences or similarities you observe.

Solution

1. The same slope coefficient and confidence interval as in the previous question will be obtained.
2. The same slope coefficient and confidence interval as in the previous question will be obtained, because the slope coefficient is interpreted as a 1 unit increase in the variable, keeping all other factors constant. Any coding we chose where the spacing between the coding is 1 unit, will give the same slope coefficient and confidence interval.
3. The numeric value of the factor is halved, since the coding range has doubled. The confidence interval decreases to half the range as well. Note that the interpretation of the slope and confidence interval is still identical.

```
brittle <- read.csv('http://datasets.connectmv.com/file/brittleness-index.csv')

# Linear model. There are many ways to create the factor variable. Here's one:
TK104 <- seq(from=1, to=1, length.out=length(brittle$TK104))
TK105 <- seq(from=0, to=0, length.out=length(brittle$TK105))
reactor <- c(TK104, TK105)
brittleness <- c(brittle$TK104, brittle$TK105)

# Delete the NAs
reactor <- reactor[!is.na(brittleness)]
brittleness <- brittleness[!is.na(brittleness)]
mod <- lm(brittleness ~ reactor)
summary(mod)
confint(mod)

# Linear model. There are many ways to create the factor variable. Here's another:
TK104 <- seq(from=2, to=2, length.out=length(brittle$TK104))
TK105 <- seq(from=1, to=1, length.out=length(brittle$TK105))
reactor <- c(TK104, TK105)
brittleness <- c(brittle$TK104, brittle$TK105)

# Delete the NAs
reactor <- reactor[!is.na(brittleness)]
brittleness <- brittleness[!is.na(brittleness)]
mod.12 <- lm(brittleness ~ reactor)
summary(mod.12)
confint(mod.12)

# Linear model. There are many ways to create the factor variable. Last one:
TK104 <- seq(from=1, to=1, length.out=length(brittle$TK104))
TK105 <- seq(from=-1, to=-1, length.out=length(brittle$TK105))
reactor <- c(TK104, TK105)
brittleness <- c(brittle$TK104, brittle$TK105)
```

```
# Delete the NAs
reactor <- reactor[!is.na(brittleness)]
brittleness <- brittleness[!is.na(brittleness)]
mod <- lm(brittleness ~ reactor)
summary(mod)
confint(mod)
```

END