MATH - 4360: Linear Statistical Models

Chapter 8 - Indicator Variables

Suthakaran Ratnasingam

- The variables employed in regression analysis are often quantitative variables, that is, the variables have a well - defined scale of measurement. Variables such as temperature, distance, pressure, and income are quantitative variables.
- In some situations it is necessary to use qualitative or categorical variables as predictor variables in regression.
- Examples of qualitative or categorical variables are operators, employment status (employed or unemployed), shifts (day, evening, or night), and sex (male or female).
- In general, a qualitative variable has no natural scale of measurement. We must assign a set of levels to a qualitative variable to account for the effect that the variable may have on the response.
- This is done through the use of indicator variables. Sometimes indicator variables are called dummy variables.

Example 8.1 The Tool Life Data

Suppose that a mechanical engineer wishes to relate the **effective life of a cutting tool** (y) used on a lathe to the **lathe speed** in revolutions per minute (x_1) and the **type of cutting tool** used.

$$x_2 = \begin{cases} 0 & \text{if the observation is from tool type A} \\ 1 & \text{if the observation is from tool type B} \end{cases}$$

- Assuming that a first - order model is appropriate, we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- To interpret the parameters in this model, consider first tool type A, for which $x_2 = 0$. The regression model becomes

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \epsilon \end{aligned}$$

- Thus, the relationship between tool life and lathe speed for tool type A is a straight line with intercept β_0 and slope β_1. For tool type B, we have $x_2 = 1$, and

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon \end{aligned}$$

```
rm(list = ls())
# I assume that you have installed the following R packages. If not, please install
# them using the R function: install.packages('package_name')
library(olsrr)
library(ggfortify)
library(ggplot2)
library(tidyverse)
library(car)
library(Rcpp)
```

```r
library(GGally)
library(leaps)
library(matlib)
data1 = read.table("D:\\CSUSB\\Fall 2021\\MATH 4360\\RNotes\\ex81.txt", header = TRUE)
head(data1)
```

```
##       y   x ToolType
## 1 18.73 610        A
## 2 14.52 950        A
## 3 17.43 720        A
## 4 14.54 840        A
## 5 13.44 980        A
## 6 24.39 530        A
```

```r
names(data1)
```

```
## [1] "y"        "x"        "ToolType"
```

```r
n = nrow(data1)
data1$ToolType1 = ifelse(data1$ToolType == 'A', 0, 1)
Fit1 = lm(y ~ x + ToolType1, data = data1)
summary(Fit1)
```

```
##
## Call:
## lm(formula = y ~ x + ToolType1, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6255 -1.6308  0.0612  2.2218  5.5044
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.208726   3.738882   9.417 3.71e-08 ***
## x           -0.024557   0.004865  -5.048 9.92e-05 ***
## ToolType1   15.235474   1.501220  10.149 1.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.352 on 17 degrees of freedom
## Multiple R-squared:  0.8787, Adjusted R-squared:  0.8645
## F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

```r
anova_fit = anova(Fit1)
tab = as.table(cbind(
  'SS' = c('Regression' = sum(anova_fit[1:2, 2]),
           'Residual' = anova_fit[3, 2],
           'Total' = sum(anova_fit[1:3, 2])),
  'Df' = c( sum(anova_fit[1:2, 1]),
            anova_fit[3, 1],
            sum(anova_fit[1:3, 1])),
  'MS' = c( sum(anova_fit[1:2, 2])/sum(anova_fit[1:2, 1]),
            anova_fit[3, 2] / anova_fit[3, 1],
            NA),
  'F-Test' = c( (sum(anova_fit[1:2, 2])/sum(anova_fit[1:2, 1]))/(anova_fit[3, 2] / anova_fit[3, 1]),
                NA,
                NA)
))
round(tab, 4)
```

```
##                    SS        Df        MS    F-Test
## Regression 1384.1084    2.0000  692.0542   61.6030
## Residual    190.9798   17.0000   11.2341
## Total      1575.0882   19.0000
```

```
summary(Fit1)$coefficients
```

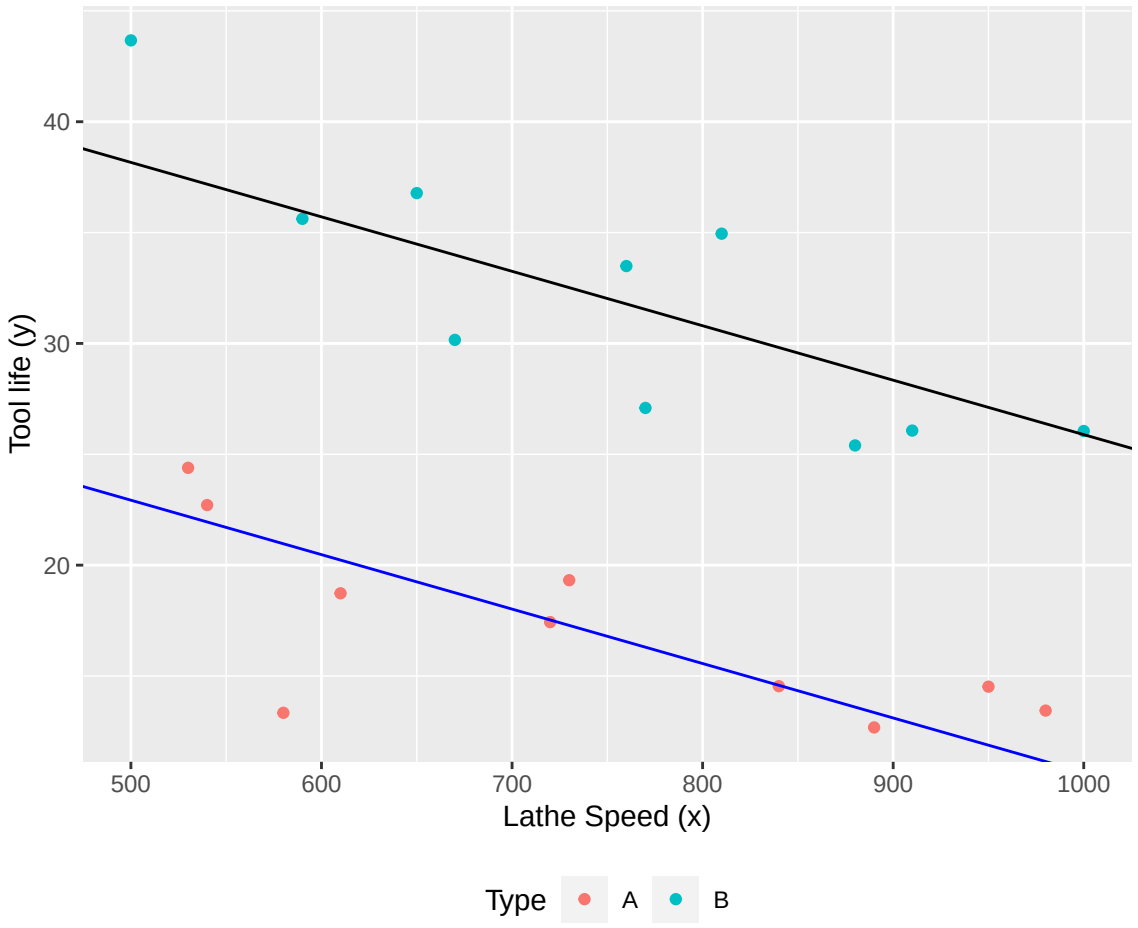
```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 35.20872630 3.738881646  9.416914 3.707376e-08
## x           -0.02455729 0.004864996 -5.047752 9.917114e-05
## ToolType1   15.23547401 1.501219870 10.148729 1.246464e-08
```

```
confint(Fit1)
```

```
##                   2.5 %      97.5 %
## (Intercept) 27.32037556 43.09707704
## x           -0.03482154 -0.01429305
## ToolType1   12.06817694 18.40277108
```

Let's visualize both regression lines. It can be clearly seen that two models describe two parallel regression lines, that is, two lines with a common slope β_1 and different intercepts.

```
ggplot(data1, aes(x=x, y=y,  color=as.factor(ToolType))) +
  geom_point()+
  geom_abline(intercept = coef(Fit1)[[1]], slope = coef(Fit1)[[2]], col = "blue")+
  geom_abline(intercept = coef(Fit1)[[1]] + coef(Fit1)[[3]] , slope = coef(Fit1)[[2]], col = "black")+
  labs(color = "Type") +
  xlab("Lathe Speed (x)") + ylab("Tool life (y)")+
  theme(legend.position="bottom")
```



Example 8.2 The Tool Life Data

Now suppose that we expect the regression lines relating tool life to lathe speed to differ in both intercept and slope. We will fit the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

This new model is equivalent to

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon, & x_2 \in 0 \text{ (Type A)} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon, & x_2 \in 1 \text{ (Type B)} \end{cases}$$

```
Fit2 = lm(y ~ x * ToolType1, data = data1)
summary(Fit2)
```

```
##
## Call:
## lm(formula = y ~ x * ToolType1, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5534 -1.7088  0.3283  2.0913  4.8652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.176013   4.724895   6.387 9.01e-06 ***
## x           -0.017729   0.006262  -2.831  0.01204 *
## ToolType1   26.569340   7.115681   3.734  0.00181 **
## x:ToolType1 -0.015186   0.009338  -1.626  0.12345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.201 on 16 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8764
## F-statistic: 45.92 on 3 and 16 DF,  p-value: 4.37e-08
```

```
anova_fit = anova(Fit2)
tab1 = as.table(cbind(
  'SS' = c('Regression' = sum(anova_fit[1:3, 2]),
           'Residual' = anova_fit[4, 2],
           'Total' = sum(anova_fit[1:4, 2])),
  'Df' = c( sum(anova_fit[1:3, 1]),
            anova_fit[4, 1],
            sum(anova_fit[1:4, 1])),
  'MS' = c( sum(anova_fit[1:3, 2])/sum(anova_fit[1:3, 1]),
            anova_fit[4, 2] / anova_fit[4, 1],
            NA),
  'F-Test' = c( (sum(anova_fit[1:3, 2])/sum(anova_fit[1:3, 1]))/(anova_fit[4, 2] / anova_fit[4, 1]),
                NA,
                NA)
))
round(tab1, 4)
```

```
##                   SS        Df        MS    F-Test
## Regression 1411.1951    3.0000  470.3984   45.9225
## Residual    163.8930   16.0000   10.2433
## Total      1575.0882   19.0000
```

```
confint(Fit2)
```

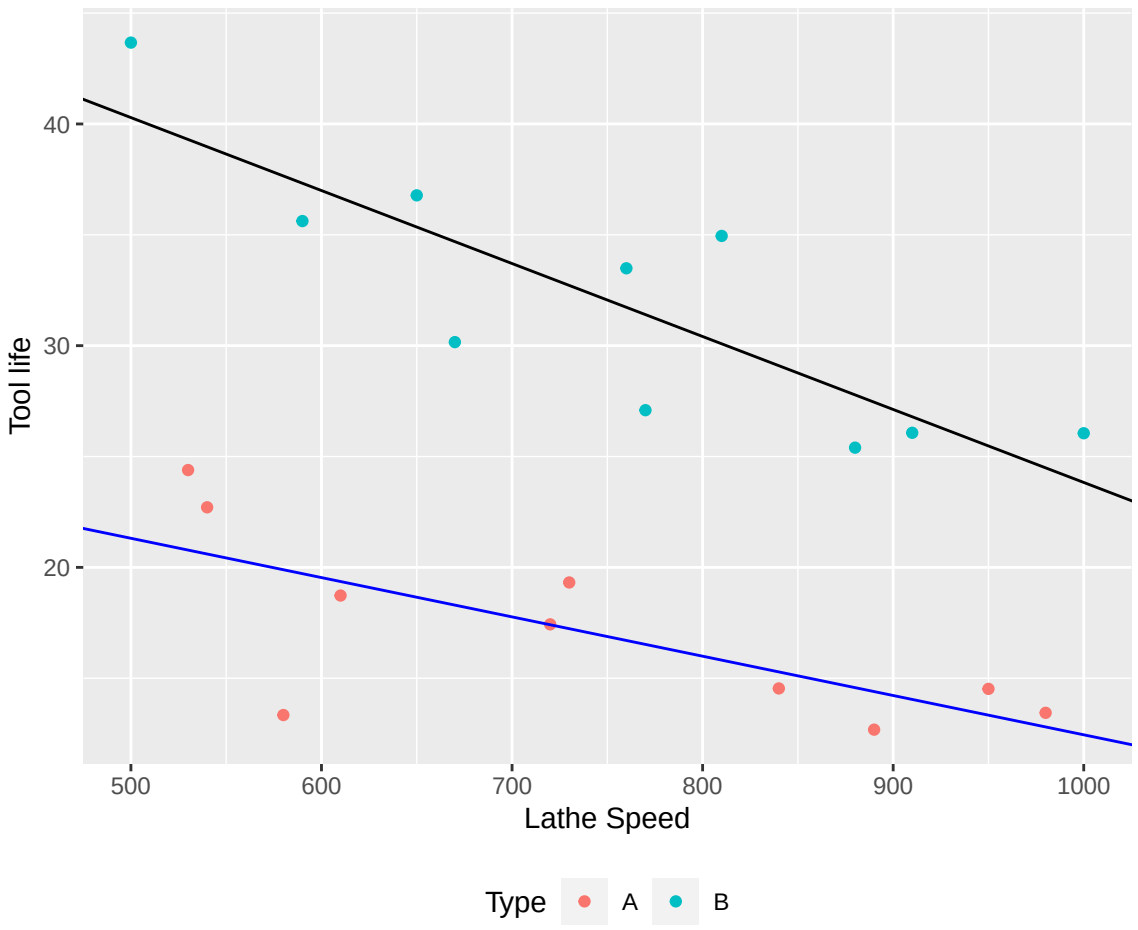
```
##                    2.5 %       97.5 %
## (Intercept) 20.15968375 40.192342590
## x           -0.03100387 -0.004453425
## ToolType1   11.48476940 41.653910786
## x:ToolType1 -0.03498222  0.004610988
```

```
summary(Fit2)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 30.17601317 4.724894751  6.386600 9.011164e-06
## x           -0.01772865 0.006262177 -2.831068 1.204343e-02
## ToolType1   26.56934009 7.115681394  3.733914 1.808135e-03
## x:ToolType1 -0.01518561 0.009338437 -1.626141 1.234509e-01
```

I plotted both regression lines below. Note that these two regression lines have different slopes and intercepts.

```
ggplot(data1, aes(x=x, y=y,  color=as.factor(ToolType))) +
  geom_point()+
  geom_abline(intercept = coef(Fit2)[[1]], slope = coef(Fit2)[[2]], col = "blue")+
  geom_abline(intercept = coef(Fit2)[[1]] + coef(Fit2)[[3]] ,
              slope = coef(Fit2)[[2]] + coef(Fit2)[[4]], col = "black")+
  labs(color = "Type") +
  xlab("Lathe Speed") + ylab("Tool life")+
  theme(legend.position="bottom")
```



An Indicator Variable with More Than Two Levels

An electric utility is investigating the effect of the size of a single - family house and the type of air conditioning used in the house on the total electricity consumption during warm - weather months. Let y be the total electricity consumption (in kilowatt - hours) during the period June through September

- Let x_1 be the size of house (square feet of floor space).
- There are four types of air conditioning systems:
  - no air conditioning,
  - window units,
  - heat pump,
  - central air conditioning.
- The four levels of this factor can be modeled by three indicator variables, x_2, x_3, and x_4, defined as follows:

Type of Air Conditioning	x_2	x_3	x_3
No air conditioning	0	0	0
Window units	1	0	0
Heat pump	0	1	0
Central air conditioning	0	0	1

- The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

- If the house has no air conditioning, becomes

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- If the house has window units, then

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

- If the house has a heat pump, the regression model is

$$y = (\beta_0 + \beta_3) + \beta_1 x_1 + \epsilon$$

- If the house has central air conditioning, then

$$y = (\beta_0 + \beta_4) + \beta_1 x_1 + \epsilon$$

- The parameters β_2, β_3, and β_4 modify the height (or intercept) of the regression model for the different types of air conditioning systems.
- That is, β_2, β_3, and β_4 measure the effect of window units, a heat pump, and a central air conditioning system, respectively, compared to no air conditioning.
- Furthermore, other effects can be determined by directly comparing the appropriate regression coefficients. For example, $\beta_3 - \beta_4$ reflects the relative efficiency of a heat pump compared to central air conditioning.
- It is also possible to use different slopes by adding interaction terms between the quantitative variable x_1 and each of the three indicator variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \epsilon$$

More Than One Indicator Variable

- Frequently there are several different qualitative variables that must be incorporated into the model.
- Suppose that a second qualitative factor, the type of cutting oil used, must be considered.
- Assuming that this factor has two levels, "low-viscosity oil" and "medium-viscosity oil", we may define a second indicator as follows;

$$x_3 = \begin{cases} 0 & \text{if low-viscosity oil is used} \\ 1 & \text{if medium-viscosity oil is used} \end{cases}$$

- A regression model relating tool life (y) to cutting speed (x_1), tool type (x_2), and type of cutting oil (x_3) is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{if type A tool, low viscosity oil} \\ (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon & \text{if type B tool, low viscosity oil} \\ (\beta_0 + \beta_3) + \beta_1 x_1 + \epsilon & \text{If type A tool, medium viscosity oil} \\ (\beta_0 + \beta_2 + \beta_3) + \beta_1 x_1 + \epsilon & \text{If type B tool, medium viscosity oil} \end{cases}$$

- This defines four parallel regression lines corresponding to the four pairs of levels of the two categorical variables
- To allow the regression lines to have different slopes, we can add all the interaction effects (between the quantitative variable and the two categorical variables):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$$

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{if type A tool, low viscosity oil} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \epsilon & \text{if type B tool, low viscosity oil} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \epsilon & \text{If type A tool, medium viscosity oil} \\ (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_4 + \beta_5) x_1 + \epsilon & \text{If type B tool, medium viscosity oil} \end{cases}$$

- Suppose that we add a cross - product term involving the two indicator variables x_2 and x_3 to the model, resulting in

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \epsilon$$

- We then have the following:

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{if type A tool, low viscosity oil} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \epsilon & \text{if type B tool, low viscosity oil} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \epsilon & \text{If type A tool, medium viscosity oil} \\ (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) x_1 + \epsilon & \text{If type B tool, medium viscosity oil} \end{cases}$$

- The addition of the cross - product term $\beta_6 x_2 x_3$ results in the effect of one indicator variable on the intercept depending on the level of the other indicator variable.

References

- Introduction to Linear Regression Analysis, 5th Edition, by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (Wiley), ISBN: 978-0-470-54281-1.
- R Core Team (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- RStudio Team (2020). RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC.