Discussion Case

Interpreting Regression Models

Mike Nguyen

Permissible variables practice

The career services office in a business school on the west coast wants to identify the factors that are associated with the starting salary for its graduates. It has access to the following variables. Which would be permissible to include as independent variables in a regression model?

a) student id number
b) major (accountancy, finance, management, marketing)
c) whether student also had a minor (yes/no)
d) overall GPA
e) GPA in major
f) number of internships completed
g) participated in job fair (yes/no)
h) number of career center workshops participated in

Relationship between R-squared in Simple Regression and Correlation

```
n = 1000
x = sample(1:100, replace = T, size = n)
y = 0.8*x # + rnorm(n, mean = 0, sd = 0.01)

cor(x,y)
```

```
## [1] 1
```

```
summary(lm(y ~ x - 1))
```

```
##
## Call:
## lm(formula = y ~ x - 1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.777e-14 -3.630e-15 -1.290e-15  1.900e-16  1.449e-12
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## x 8.000e-01  2.472e-17 3.236e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.609e-14 on 999 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.047e+33 on 1 and 999 DF,  p-value: < 2.2e-16
```

```
summary(lm(x ~ y - 1))
```

```
##
## Call:
## lm(formula = x ~ y - 1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.906e-12  2.100e-15  5.000e-15  1.020e-14  2.400e-14
##
## Coefficients:
##    Estimate Std. Error   t value Pr(>|t|)
## y 1.250e+00  1.042e-16 1.199e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.555e-13 on 999 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.438e+32 on 1 and 999 DF,  p-value: < 2.2e-16
```

```
library(tidyverse)
library(faux)
dat <- rnorm_multi(n = 100,
                   mu = c(0, 20, 20),
                   sd = c(1, 5, 5),
                   r = c(0.5, 0.5, 0.25),
                   varnames = c("A", "B", "C"),
                   empirical = FALSE)

attach(dat)
cor(A,B)
```

```
## [1] 0.5228218
```

```
AB = summary(lm(A ~ B))
sqrt(AB$r.squared)
```

```
## [1] 0.5228218
```

```
BA = summary(lm(B ~ A))
sqrt(BA$r.squared)
```

```
## [1] 0.5228218
```

```
detach(dat)
```

Example 1

Customer service expense

A company that develops and maintains websites for businesses wants to develop estimates of how much developer time is required to maintain customer websites.

The dependent variable for the regression model is developer time per page on a customer's site (in hundreds of hours, 00)

The independent variables are

- number of years the client has been a customer of the company
- whether the website includes a payment portal (1 = yes, 0 = no)
- company size (annual revenue in millions of dollars–$000,000)

Hours = 6.38 - .47(customer years) + 1.14 (if payment portal) - .16 (annual revenue)

Generate a fictitious dataset

```
n = 200
# number of years the client has been a customer of the company
customers_year = sample(0:15, replace = T, size = n)

# 1 = yes, 0 = no
payment_portal = sample(c(0,1), replace = T, size = n)

# annual revenue in millions of dollars
firm_size = sample(0:10, replace = T, size = n)

# true model
hours = 6.38 - 0.47*customers_year + 1.14*payment_portal - 0.16*firm_size + rnorm(n)
data = as.data.frame(cbind(hours, customers_year, payment_portal,firm_size))

# rio::export(data, "discussion.xlsx")
```

Implement multiple regression

```
reg = lm(hours ~ customers_year + payment_portal + firm_size, data = data)

# see the result
summary(reg)
```

```
##
## Call:
## lm(formula = hours ~ customers_year + payment_portal + firm_size,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5108 -0.6913  0.1029  0.6889  2.3146
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.53600    0.17254  37.880  < 2e-16 ***
## customers_year -0.48022    0.01454 -33.038  < 2e-16 ***
## payment_portal  1.32939    0.14432   9.212  < 2e-16 ***
## firm_size      -0.17487    0.02114  -8.272 1.98e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9915 on 196 degrees of freedom
## Multiple R-squared:  0.8701, Adjusted R-squared:  0.8681
## F-statistic: 437.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

Example 2

What if we cannot capture every variable that is associated with the dependent variable

Generate a fictitious dataset

```
n = 200
# number of years the client has been a customer of the company
customers_year = sample(0:15, replace = T, size = n)

# 1 = yes, 0 = no
payment_portal = sample(c(0,1), replace = T, size = n)

# annual revenue in millions of dollars
firm_size = sample(0:10, replace = T, size = n)

# other variables we cannot capture
other_var = sample(0:10, replace = T, size = n)

# true model
hours = 6.38 - 0.47*customers_year + 1.14*payment_portal - 0.16*firm_size + other_var + rnorm(n)
data = as.data.frame(cbind(hours, customers_year, payment_portal,firm_size, other_var))
```

Implement multiple regression

```
reg = lm(hours ~ customers_year + payment_portal + firm_size, data = data)

# see the result
summary(reg)
```

```
##
## Call:
## lm(formula = hours ~ customers_year + payment_portal + firm_size,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3699 -2.5753  0.0319  2.8104  7.2749
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.61102    0.60369  19.234   <2e-16 ***
## customers_year -0.49037    0.05019  -9.771   <2e-16 ***
## payment_portal  1.05874    0.46993   2.253   0.0254 *
## firm_size      -0.17659    0.07137  -2.474   0.0142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.26 on 196 degrees of freedom
## Multiple R-squared:  0.331,  Adjusted R-squared:  0.3208
## F-statistic: 32.33 on 3 and 196 DF,  p-value: < 2.2e-16
```