mock_quiz1_solution

February 12, 2025

1 STAT 301: Midterm 1

1.1 Data Wrangling

[2]:

```r
# The code will not be displayed in the quiz
library(tidyverse, quietly = T)
library(broom, quietly = T)
library(repr, quietly = T)
library(infer, quietly = T)
library(ggplot2)
library(cowplot)
library(contextual)

dat <- read.csv("Assessment_2015.csv")
dat <- dat %>% filter(ASSESSCLAS=="Residential")  %>% filter(ASSESSMENT < 1500000)

dat <- dat  %>% mutate(BLDG_AGE = cut(YEAR_BUILT,
                    breaks=c(-Inf,1980,2000,+Inf),
                    labels=c("plus40","mid40-20","new20")))

set.seed(123)
dat_tax <- rep_sample_n(dat, size = 1000)
```

1.2 Question 1

Answer TRUE/FALSE to the following statement:

Bootstrapping can be used to approximate the sampling distribution of the least square estimators in linear regression.

1.2.1 BEGIN SOLUTION

TRUE. You can resample with replacement, estimate a LR for each sample and obtain a list of LS estimates. The distribution of these estimates is an approximation of the sampling distribution. You can visualize the distribution using a histogram for each estimate.

1.2.2 END SOLUTION

1.3 Question 2

You have a dataset called `data_yvr` has information about 20,000 houses sold in Vancouver that includes the following variables:

`price` = selling price of the house

`bedrooms` = number of bedrooms

`bathrooms` = number of bathrooms

`sqft` = interior square footage

`waterfront` = 1 if the house has a view of the waterfront, 0 otherwise

`yr_built` = year the house was built

The following linear regression models are estimated using `lm` in R to study different factors associated with a house's selling price:

You believe that the expected change in price per additional square foot in size is different in houses with and without a waterfront view. Which model can you use to test this hypothesis?

A: `lm(price ~ yr_built, data = data_yvr)`

B: `lm(price ~ waterfront, data = data_yvr)`

C: `lm(price ~ sqft * waterfront, data = data_yvr)`

D: `lm(price ~ bedrooms + bathrooms + sqft + waterfront + yr_built, data = data_yvr)`

1.3.1 BEGIN SOLUTION

C: since you expect the association between price and size to be different in houses with and without a waterfront view, you need a model where these variable are interacted (`sqft * waterfront`)

1.3.2 END SOLUTION

1.4 Question 3

For the California School dataset and the plot provided,

Complete the code of the model used to estimate the lines illustrated in the plot below. Don't include spaces between variables and symbols:

1.4.1 BEGIN SOLUTION

(no spaces as requested to match Canvas response)

```
model.plot <- lm(read~grades*english,data=CASchools)
```

The plot shows non-parallel lines, i.e., different groups have different slopes.

Thus, this is **NOT** an additive model. We need to include an interaction between the covariates `grades` and `english` (`*`)

1.4.2 END SOLUTION

1.5 Question 4

Same dataset. Model:

`lm(read ~ grades + english, data = CASchools)`

a) The estimated coefficient for english is statistically significant and equal to -0.24. In your own words and within the context of the problem, interpret this coefficients.

b) Using the code above, how many estimates would you obtain for grades? Explain why and give an interpretation the estimate(s).

1.5.1 BEGIN SOLUTION

a) An additional 1 percent of students in the district that are English learners, is associated with an expected decrease in the reading score of 0.24.

NOTE 1: important to use the word *associated*. Don't use *cause* or *effect* since this is an observational study.

NOTE 2: the units for `english` are percent, so the interpretation is *an additional 1 percent*, otherwise, we don't refer to percent increases of the covariates.

b) The table will have one estimated coefficient called `gradesKK-08`. This will be the difference in the intercept of the line for `gradesKK-08`, relative to the reference line for `gradesKK-06`.

In the context of the problem, this estimate corresponds to the difference in average reading score when there are no English learners in `gradesKK-08` compared to `gradesKK-06` schools.

1.5.2 END SOLUTION

Question 6.1

Provide the R code to estimate a simple linear regression to model the relation between the response and a continuous input variable. Use the name of the variables given above.

1.5.3 BEGIN SOLUTION

lm_size<- lm(ASSESSMENT ~ BLDG_METRE, data = dat_tax)

1.5.4 END SOLUTION

Question 6.2

Using `dat_tax`, we estimate a simple linear regression using `lm()` to study the relation between the assessed value of a property and its size. Results are shown in the table below obtained with `tidy()`. In one sentence, interpret the estimate of the parameter for `BLDG_METRE`.

[2]:

```r
# The code will not be displayed in the quiz
lm_size<- lm(ASSESSMENT ~ BLDG_METRE, data = dat_tax)
tidy(lm_size)  %>% mutate_if(is.numeric, round, 3)
```

A tibble: 2 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	127317.865	10033.463	12.689	0
BLDG_METRE	2356.108	62.966	37.419	0

1.5.5 BEGIN SOLUTION

An additional meter in the size of the property is associated with an expected increase in the assessed value of $2356.108.

1.5.6 END SOLUTION

[3]:

```
# The code will not be displayed in the quiz

lm_size_age_add <- lm(ASSESSMENT ~ BLDG_METRE + BLDG_AGE, data = dat_tax)
tidy(lm_size_age_add)  %>% mutate_if(is.numeric, round, 3)
```

A tibble: 4 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	125789.449	9916.236	12.685	0
BLDG_METRE	2214.924	66.918	33.099	0
BLDG_AGEmid40-20	35158.027	7953.204	4.421	0
BLDG_AGEnew20	43237.149	8369.910	5.166	0

Question 7.1

Based on the estimates and results in the table given by `tidy()`, determine if the model contains any interaction term and in one or two sentences justify your answer.

1.5.7 BEGIN SOLUTION

The model does not contain an interaction term. There are only 4 parameters to characterize 3 lines with equal slopes.

The model does not contain an interaction term. There is only one common slope given by the estimate of `BLDG_METRE`.

1.5.8 END SOLUTION

Question 7.2

In one or two sentences, explain which assumption was made about the relation of the variables to propose such a model (i.e., with or without interactions). In other words, what was assumed about the relation between the assessed value and the size for properties of different ages?

1.5.9 BEGIN SOLUTION

An additive model was fit since it was assumed that the association between the assessed value and the size of the property is the same for properties of all ages. The slope is the same for all levels of `BLDG_AGE`.

1.5.10 END SOLUTION

Question 7.3

Which of the following plots matches the model fit and summarized in the previous table? In one or two sentences, justify your choice and explain how the estimates given in the table relate with the line(s) in your selected plot.

1.5.11 BEGIN SOLUTION

PLOT B: An additive model is represented by 3 paralell lines slope. The common slope is equal to 2214.924. The intercept of the reference red line is 125789.449. The intercept of the green line is (2214.924 + 35158.027). The intercept of the blue line is (2214.924 + 43237.149).

1.5.12 END SOLUTION

[7]:

```r
# The code will not be displayed in the quiz
options(repr.plot.width = 15, repr.plot.height = 7)

#SLR
dat_tax$pred_slm <- predict(lm_size)

lm_size_plot <- ggplot(dat_tax, aes(
  x = BLDG_METRE,
  y = ASSESSMENT,
  color = BLDG_AGE
)) +
  geom_point() +
  geom_line(aes(y = pred_slm), size = 1, color = "black") +
  labs(
    title = "PLOT A",
    x = "Building Size (mts)",
    y = "Assessed Value ($)"
  ) +
  theme(
    text = element_text(size = 16.5),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
  ) +
  labs(color = "Building Age")

#Additive
dat_tax$pred_lm_add <- predict(lm_size_age_add)

lm_size_age_add_plot <- ggplot(dat_tax, aes(
  x = BLDG_METRE,
  y = ASSESSMENT,
  color = BLDG_AGE
)) +
  geom_point() +
  geom_line(aes(y = pred_lm_add), size = 1) +
  labs(
    title = "PLOT B",
    x = "Building Size (mts)",
    y = "Assessed Value ($)"
  ) +
  theme(
    text = element_text(size = 16.5),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
  ) +
  labs(color = "Building Age")

#Interaction
lm_size_age_int <- lm(ASSESSMENT ~ BLDG_METRE * BLDG_AGE, data = dat_tax)
dat_tax$pred_lm_int <- predict(lm_size_age_int)
png(filename="faithful.png")
lm_size_age_int_plot <- ggplot(dat_tax, aes(
  x = BLDG_METRE,
  y = ASSESSMENT,
  color = BLDG_AGE
)) +
  geom_point() +
  geom_line(aes(y = pred_lm_int), size = 1) +
  labs(
    title = "PLOT C",
    x = "Building Size (mts)",
    y = "Assessed Value ($)"
  ) +
  theme(
    text = element_text(size = 16.5),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
  ) +
  labs(color = "Building Age")

png(filename="/home/lourenzutti/question10.png", width = 1000, height = 750)
plot_grid(lm_size_plot, lm_size_age_add_plot, lm_size_age_int_plot)
dev.off()
```

png: 2

[5]:

```
# The code will not be displayed in the quiz
tidy(lm_size_age_int, conf.int = TRUE)  %>% mutate_if(is.numeric, round, 3) %>%
    subset(term == "BLDG_METRE")
```

A tibble: 1 × 7

term	estimate	std.error	statistic	p.value	conf.low	conf.high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
BLDG_METRE	1768.566	120.009	14.737	0	1533.065	2004.067

A. at significance level of 5%, the assessed value of a property more than 40 years old is statistically different from that of houses less than 20 years old.

B. at significance level of 5%, the assessed value of any property is significantly associated with the size of the property.

C. at significance level of 5%, we have enough evidence to reject the null hypothesis that the assessed value of any property are not associated with the size of the property.

D. at significance level of 5%, we have enough evidence to reject the null hypothesis that the assessed value of properties more than 40 years old are not associated with the size of the property.

E. with 95% confidence, we expect that the assessed value of a property increases between \$1533.065 and \$2004.067 for every 1 additional metre in size.

F. with 95% confidence, we expect that the assessed value of a property increases between \$1533.065 and \$2004.067.

G. with 95% confidence, the expected assessed value of a property more than 40 years old is between \$1533.065 and \$2004.067.

1.5.13 BEGIN SOLUTION

D is the only correct one. All other options do not state "for properties with more than 40 years" (reference value) or refer to a wrong parameter in a model *with* interaction terms.

1.5.14 END SOLUTION