HW 04: Multiple linear regression

Jwalin Patel

03/24/2021

```
library(tidyverse)
library(broom)
library(knitr)
library(patchwork)
```

```
sitting <- read.csv("data/sitting.csv")
```

Part 1

Question 1

We will use SLR to model MET and sitting as predictor and response variables, respectively.

$$Simple\ Linear\ Regression\ Model: sitting = \beta_0 + \beta_1 MET$$

```
m1 <- lm(sitting ~ MET, data = sitting)
m1 %>%
  tidy(conf.int = TRUE) %>%
  kable (digits = 3, caption = "Prediction of the Reported hours per day spent sitting for subjects giv
```

Table 1: Prediction of the Reported hours per day spent sitting for subjects given reported metabolic equivalent unit minutes per week

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.502	0.915	8.203	0.000	5.641	9.362
MET	0.000	0.000	-0.421	0.676	-0.001	0.001

Now we calculate R^2 to consider how well the model fits the relationship between reported hours per day spent sitting and the reported metabolic equivalent unit minutes per week.

```
anova(m1) %>%
  kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MET	1	2.008	2.008	0.177	0.676
Residuals	33	373.592	11.321	NA	NA

We calculate R^2 using the formula:

$$R^2 = \frac{SS_{Model}}{SS_{Total}}$$

```r
2.008/(2.008+373.592)
```

```
## [1] 0.005346113
```

Thus, $R^2 = 0.005346113$. This means only around 0.53 percent of the variation in the reported hours per day spent sitting is explained by the reported metabolic equivalent unit minutes per week. Since the model only considers one predictor variable to explain variation in sitting hours, this is not a good representation of sitting hours.

Question 2

```r
m2 <- lm(MTL ~ sitting, data = sitting)
m2 %>%
  tidy(conf.int = TRUE) %>%
  kable (digits = 3, caption = "Prediction of the Medial temporal lobe thickness in mm for subjects giv
```

Table 3: Prediction of the Medial temporal lobe thickness in mm for subjects given Reported hours per day spent sitting

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.700	0.073	36.933	0.000	2.551	2.848
sitting	-0.023	0.009	-2.476	0.019	-0.042	-0.004

The coefficient is estimated as -0.023, with expected value ranging from lower bound of -0.042 and upper bound of -0.004. This means we are 95% confident that with every 1 hour increase in Reported hours per day spent sitting for subjects, we can expect the Medial temporal lobe thickness to reduce by 0.004 to 0.042 mm, and reduce by 0.023 mm on average.

Question 3

```r
sitting %>%
  summarize (mean = mean(age))
```

```
##       mean
## 1 60.37143
```

```r
sitting <- sitting %>%
  mutate(age_cent = age - 60.37143)
m3 <- lm(MTL ~ sitting + MET + age_cent, data = sitting)
m3 %>%
  tidy(conf.int = TRUE) %>%
  kable (digits = 3, caption = "Prediction of the Medial temporal lobe thickness in mm for subjects giv
```

Table 4: Prediction of the Medial temporal lobe thickness in mm for subjects given Reported hours per day spent sitting, Reported metabolic equivalent unit minutes per week and Mean centered age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.682	0.088	30.551	0.000	2.503	2.861
sitting	-0.021	0.010	-2.189	0.036	-0.040	-0.001
MET	0.000	0.000	0.078	0.939	0.000	0.000
age_cent	0.004	0.004	1.113	0.274	-0.004	0.012

Sitting: The coefficient of sitting is estimated as -0.021, with value ranging from lower bound of -0.040 and upper bound of -0.001. This means we are 95% confident that with every 1 hour increase in Reported hours per day spent sitting for subjects, we can expect the Medial temporal lobe thickness to reduce by 0.001 to 0.040 mm, and reduce by 0.021 mm on average, ceteris paribus. Age: The coefficient of age is estimated as 0.004, with value ranging from lower bound of -0.004 and upper bound of 0.012. This means we are 95% confident that with every 1 year increase in mean age, we can expect the Medial temporal lobe thickness to change by -0.004 to + 0.012 mm, and + 0.004 mm on average, ceteris paribus. However, the p-value of our predictor age_cent is higher than what our confidence interval of 95% allows (<0.05), thus we should be wary of the explanatory power of age_cent for predicting Medial temporal lobe thickness.

Question 4

```
glance(m2) %>%
  select(r.squared, adj.r.squared, AIC, BIC)
```

```
## # A tibble: 1 x 4
##   r.squared adj.r.squared   AIC   BIC
##       <dbl>         <dbl> <dbl> <dbl>
## 1     0.157         0.131 -17.1 -12.5
```

```
glance(m3) %>%
  select(r.squared, adj.r.squared, AIC, BIC)
```

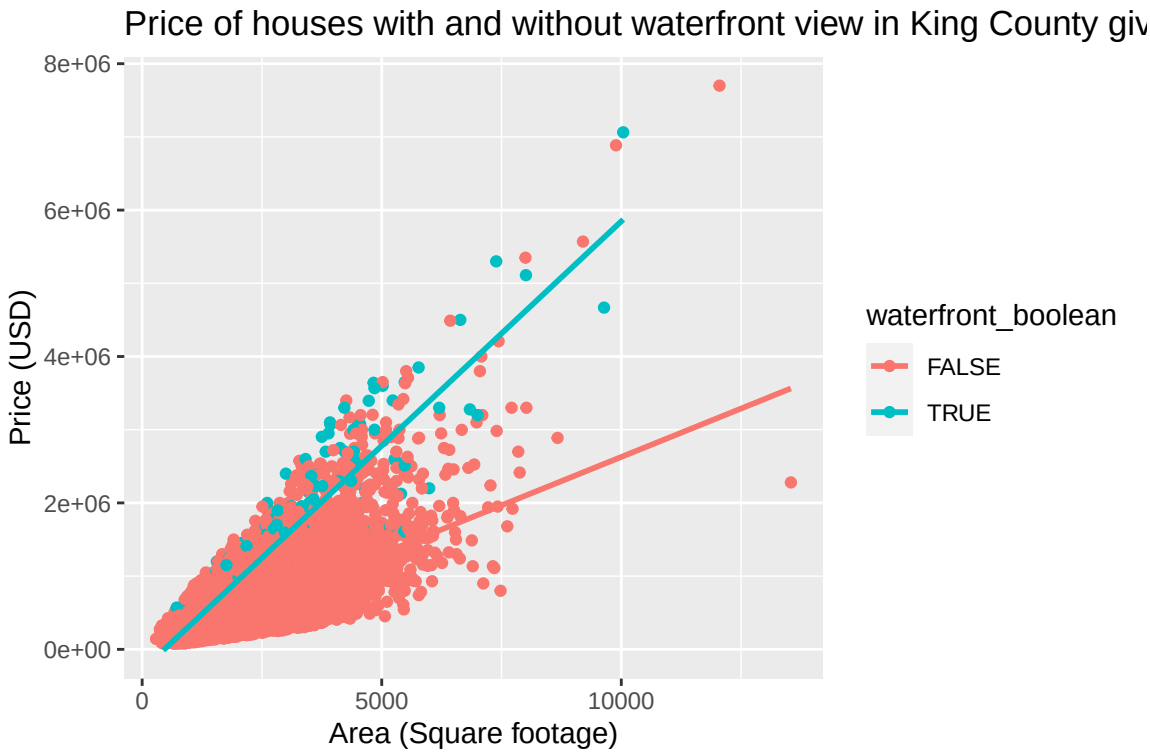
```
## # A tibble: 1 x 4
##   r.squared adj.r.squared   AIC   BIC
##       <dbl>         <dbl> <dbl> <dbl>
## 1     0.189         0.111 -14.5 -6.73
```

1. Based on adjusted R^2 we would choose the first model. Even though the second model has higher R^2, it has a lower adjusted R^2 implying that there the tradeoff/penalty of adding MET and age_cent to our model is greater than their contribution to explanatory power.
2. Based on AIC, we choose the model with the smaller value of AIC which indicates better fit, which is the first model. This means adding MET and age_cent to our model did not reduce the sum of squares error enough to give us a lower AIC.
3. Based on BIC, we choose the model with the smaller value of BIC which indicates better fit, which is again, the first model. This means adding MET and age_cent to our model did not reduce the sum of squares error enough to give us a lower BIC. Note, BIC value is much higher in the second model also because BIC has a greater penalty for the addition of predictor variables to the regression model.

Question 5

```
houses <- read.csv("data/KingCountyHouses.csv")
houses <- houses %>%
  mutate(waterfront_boolean = as.logical(waterfront))
ggplot(data = houses, aes(y = price, x = sqft, color=waterfront_boolean,)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)+
  labs(x = "Area (Square footage)",
       y = "Price (USD)", title = "Price of houses with and without waterfront view in King County given
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the visualization, while there is a general positive correlation between price and area of house, we can also see that houses without waterfront (waterfront = false; red color) seem to gain price slower with increasing area, as compared to houses with waterfront (waterfront = true; blue color). Therefore the lines of fit would not appear to be parallel, indicating the presence of an interaction effect happening, such that price in response to the square footage changes differently due to another variable, in this case: whether the house has waterfront view or not.

Question 6

```
houses <- houses %>%
  mutate(log_price = log(price))
```

$$Model: log(price) = \beta_0 + \beta_1 sqft + \beta_2 waterfront + \beta_3 sqft \times waterfront + \epsilon, such\ that\ \epsilon \sim N(0, \sigma_\epsilon^2)$$

Now we will regress.

```
houses_model <- lm(log_price ~ sqft + waterfront + sqft*waterfront, data = houses)
houses_model %>%
  tidy() %>%
  kable(digits = 5)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.22424	0.00638	1915.59064	0.00000
sqft	0.00039	0.00000	139.49653	0.00000
waterfront	0.78058	0.06545	11.92625	0.00000
sqft:waterfront	-0.00005	0.00002	-2.57719	0.00997

Question 7

$$Model : log(\hat{price}) = 12.22424 + 0.00039(sqft) + 0.78058(waterfront) - 0.00005(sqft \times waterfront))$$

Regression equation for no waterfront view: implies $waterfront = 0$

$$Model : log(\hat{price}) = 12.22424 + 0.00039(sqft)$$

This means for every 1 sqft increase in area of house, we expect price to increase by $100Ö(e^0.00039 - 1) = 0.039$ percent, on average.

Regression equation for yes waterfront view: implies $waterfront = 1$

$$Model : log(\hat{price}) = 12.22424 + 0.00039(sqft) + 0.78058 - 0.00005(sqft)$$

This implies that for houses with waterfront view:

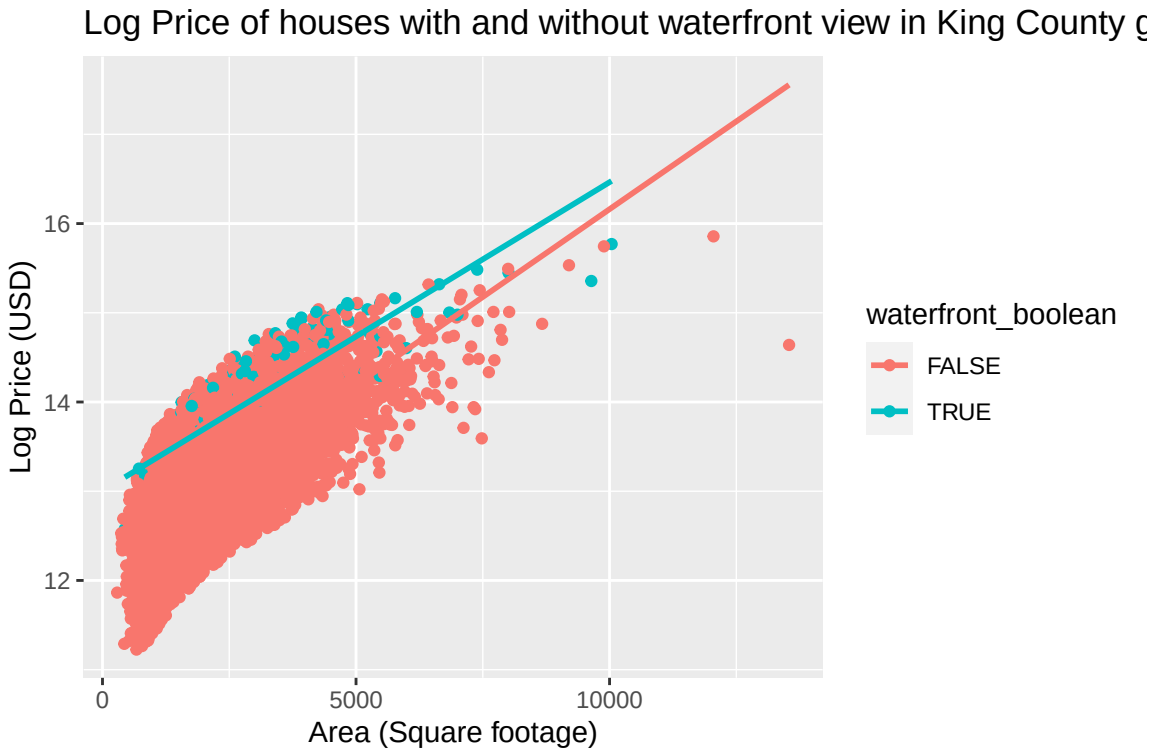
$$Model : log(\hat{price}) = 13.00482 + 0.00034(sqft)$$

This means for every 1 sqft increase in area of house, we expect price to increase by $100Ö(e^0.00034 - 1) = 0.034$ percent on average.

Let us confirm our observation by plotting log(price) of houses with and without waterfront view in King county given area in sqft.

```
ggplot(data = houses, aes(y = log_price, x = sqft, color=waterfront_boolean,)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)+
  labs(x = "Area (Square footage)",
       y = "Log Price (USD)", title = "Log Price of houses with and without waterfront view in King Cou
```

```
## `geom_smooth()` using formula 'y ~ x'
```



This confirms our estimated values of the coefficient estimates for sqft area such that when plotting log(price), the slope is steeper for houses without waterfront houses, compared to houses with waterfront view.

Part 2

... All in all, in general, we can expect houses in King County to increase in price with increase in area by sqft. On average, houses in our dataset that do not have waterfront view tend to appear cheaper. As their size increases, houses without waterfront view gain value in price at a slower rate than houses with a waterfront view. For Houses with waterfront view, as their size increases, they get expensive much faster than if they did not have a waterfront view.