Homework 5

Evan Cranmer and Xiaoqing Liu

2024-10-12

Problem 1

```
library(ISLR2)
data("Boston")

#medv as a function of chas
medv_model <- lm(medv ~ chas, data=Boston)

summary(medv_model)
```

```
##
## Call:
## lm(formula = medv ~ chas, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## chas          6.3462     1.5880   3.996 7.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

Null Hypothesis: Our beta coefficient for chas is equal to 0

Alternative Hypothesis: chas is not equal to zero, and therefore being near the charles river (when chas = 1) has an impact on median home value.

The intercept represents the expected median home value when chas = 0. The estimated median home value is approximately 22.09 thousand dollars when a home does not bound the charles river.

The coefficient of chas indicates that homes located near the Charles River have a median home value that is approximately 6.35 thousand dollars higher than homes not near the river (when chas = 1, we see this expected increase).

Problem 2

```
#creating a new model
multi_model <- lm(medv ~ chas + indus, data=Boston)

summary(multi_model)
```

```
##
## Call:
## lm(formula = medv ~ chas + indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.379  -5.069  -1.406   3.295  33.607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.43170    0.66754  44.090  < 2e-16 ***
## chas         7.47818    1.37605   5.435 8.58e-08 ***
## indus       -0.66592    0.05095 -13.071  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.839 on 503 degrees of freedom
## Multiple R-squared:  0.2765, Adjusted R-squared:  0.2736
## F-statistic:  96.1 on 2 and 503 DF,  p-value: < 2.2e-16
```

The intercept this time represents the expected median home value when both chas and indus are zero, meaning when the home is not bound by the charles river and have no non-retail business acres. Based on our output, we expect the median home value to be 29.43 thousand dollars.

Interpreting chas: Chas coefficient represents how the proximity to the Charles River impacts the median home value, controlling for the proportion of non-retail business acreage. Homes near the Charles River, when chas = 1, are expected to have median home values that are 7.48 thousand dollars higher than homes not near the river, holding indus constant. This effect is statistically significant, with a p-value $< .001$.

Interpreting indus: This coefficient represents the impact of non-retail business acres on the median home value, controlling for proximity to the Charles River. Based on our model, for each 1-unit increase in the proportion of indus, the median home value is expected to decrease by approximately 0.67 thousand dollars, holding chas constant. This effect is also statistically significant becuase the p-value is $< .001$.

Problem 3: Visualizing our model

```
indus_visual <- ggplot(Boston, aes(x = indus, y = medv, color = as.factor(chas))) +
  geom_point() +
  labs(x = "Non-Retail Business Acres",
       y = "Median Home Value",
       color = "Charles River") +
  ggtitle("Median Home Value and Non-retail Business Acres")

indus_visual
```

In this graph, we see that as non-retail business acres increase (indus) median home value tends to decrease. Additionally, most homes near the Charles river in this graph have inherently less non-retail business acres, and usually have higher median home values. Therefore, median home value tends to increase when chas = 1.

Question 4

```
interaction_model <- lm(medv ~ chas * indus, data=Boston)
summary(interaction_model)
```

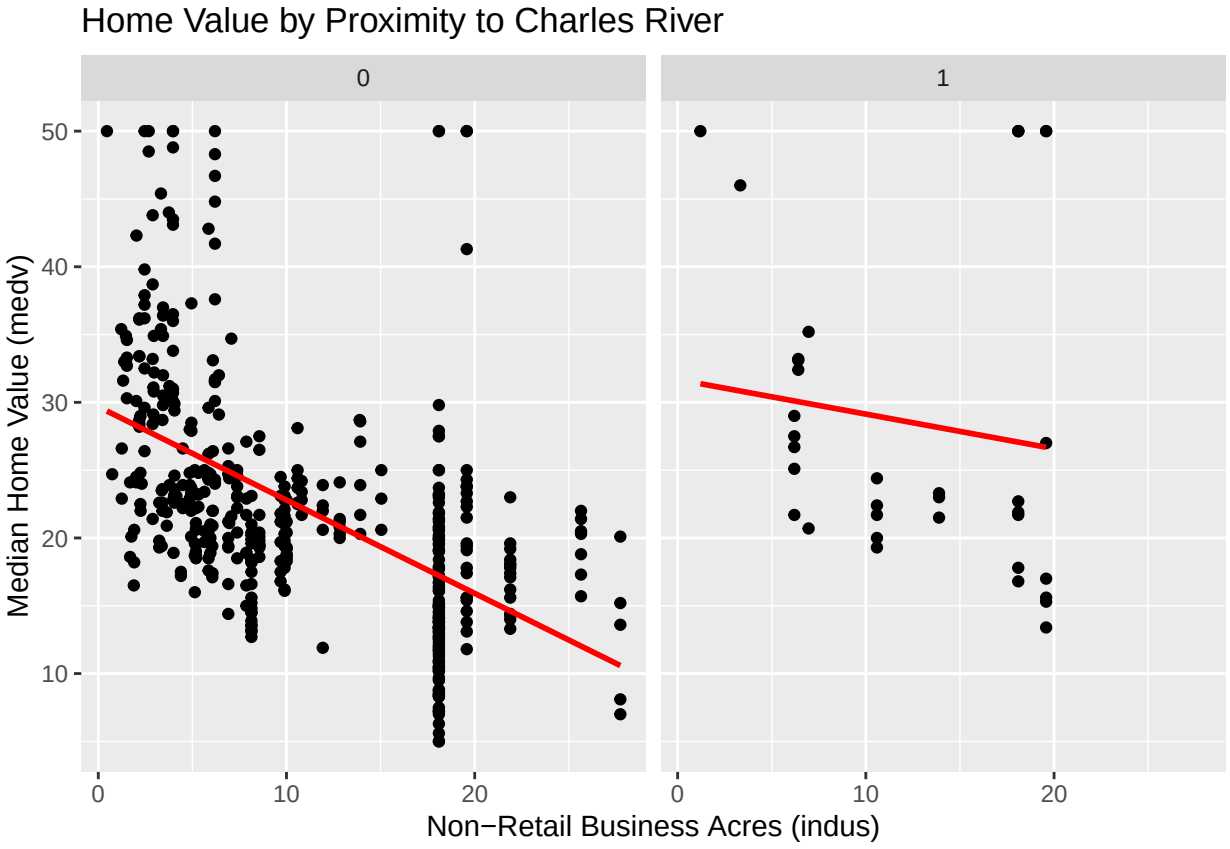
```
##
## Call:
## lm(formula = medv ~ chas * indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.291  -5.049  -1.453   3.393  33.796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.67493    0.67841  43.742   <2e-16 ***
## chas         2.00791    3.22539   0.623   0.5339
## indus       -0.68799    0.05217 -13.188   <2e-16 ***
## chas:indus   0.43303    0.23105   1.874   0.0615 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.819 on 502 degrees of freedom
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2772
## F-statistic: 65.56 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
#visualizing interaction using facet grid
interaction_plot <- ggplot(Boston, aes(x = indus, y = medv)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  facet_grid(. ~ chas) +
  labs(x = "Non-Retail Business Acres (indus)",
       y = "Median Home Value (medv)",
       title = "Home Value by Proximity to Charles River")

interaction_plot
```

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



The interaction term suggests that the effect of indus on medv is different for homes near the Charles River compared to those not near the river. A value of 0.43 coefficient implies that the negative effect of indus on medv is moderated for homes near the river (when both chas = 1). The p-value for this interction term is at, 0.0615, which is slightly above the standard alpha value of 0.05. So, even though it is not statistically significant at our alpha value, there is still some evidence of an interaction, but it is not conclusive at this level.

Our facet plot supports our findings from the interaction variable. On the left, where chas = 0, the negative relationship between indus and medv is a stronger negative relationship. The graph on the right shows that when chas = 1 (meaning our interaction variable is present) the decrease in median home values is less steep, suggesting the decrease in home values as indus increases is less for homes by the Charles River.