Homework 5

Kevin Linares and Jamila Sani

2024-10-15

This exercise involves the Boston housing dataset in ISLR2. Assume that we are interested in median home values, medv.

```r
library(ISLR2)
library(dplyr)
library(gridExtra)
library(grid)
library(ggeffects)
library(ggplot2)
library(knitr)
library(tidyverse)

data(Boston)
```

```r
head(Boston, 4) |> kable()
```

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4

```r
tail(Boston, 4) |> kable()
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
503	0.04527	0	11.93	0	0.573	6.120	76.7	2.2875	1	273	21	9.08	20.6
504	0.06076	0	11.93	0	0.573	6.976	91.0	2.1675	1	273	21	5.64	23.9
505	0.10959	0	11.93	0	0.573	6.794	89.3	2.3889	1	273	21	6.48	22.0
506	0.04741	0	11.93	0	0.573	6.030	80.8	2.5050	1	273	21	7.88	11.9

```r
glimpse(Boston)
```

```
## Rows: 506
## Columns: 13
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.08829,~
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5, 1~
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87, 7.~
## $ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524,~
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.631,~
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9, 9~
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9505~
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4,~
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 31~
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15~
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17.10~
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9, 15~
```

```r
# convert chas into a factor variable, chas=0 as the reference group
Boston$chas_f <- factor(Boston$chas, levels=c("0", "1"))

class(Boston$chas_f)
```

```
## [1] "factor"
```

```r
levels(Boston$chas_f)
```

```
## [1] "0" "1"
```

1. Examine medv as a function of chas in a simple linear regression model with an intercept. Notice the nature of chas for its use in this and following models. What hypothesis are you testing with each coefficient in lay terms? Given the results of the hypothesis testing, what do the coefficients mean?

- Intercept (β_0)
  - H_0: $\beta_0 = 0$, The mean expected median home value for owner occupied homes in suburbs not bounding the Charles River equals zero.
  - H_A: $\beta_0 \neq 0$. The mean expected median home value for owner occupied homes in suburbs not bounding the Charles River does not equal zero.
    - Reject the null (p-value<0.05), the mean expected median home value of $22,094 for owner occupied homes in suburbs not bounding Charles River does not equal zero.
- Slope (β_1)
  - H_0: $\beta_1 = 0$ the expected mean difference in median home values for owner occupied homes in suburbs not bounding Charles River (reference group), and suburbs bounding the Charles River is equal to zero.
  - H_A: $\beta_1 \neq 0$ the expected mean difference in median home values for owner occupied homes in suburbs that bound or do not bound the charles river is not equal to 0.
    - Reject the null (p-value<0.05) the difference in the mean expected median home value for owner occupied homes in suburbs that bound or do not bound the Charles River is not equal to zero. Owner occupied homes in suburbs that bound the Charles River have on average an expected median value of $6,346 more than owner occupied homes in suburbs that do not bound the Charles River. The mean expected median home value for owner occupied homes that bound the Charles River = $\beta_0 + \beta_1 = \$28,440$.

```r
Boston$chas_f <- as.factor(Boston$chas)

# median home value as a function of Charles River bounds
model_medv_chas <- lm(formula=medv ~ chas_f, Boston)
summary(model_medv_chas)
```

```
##
## Call:
## lm(formula = medv ~ chas_f, 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_f1       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
```

2. Examine medv as a function of chas and indus in a multiple linear regression. What do the coefficients mean?

- β_0: $29,432 is the mean expected median value for owner-occupied homes in suburbs that do not bound the Charles River and whose proportion of non-retail business acres per town is zero.
- β_1: $7,478 is the expected mean difference in median value of owner-occupied homes in suburbs that that do not bound (reference group) versus bound the Charles River while holding indus (proportion of non-retail business aces per town) constant.
  - Therefore, $\beta_0+\beta_1) = \$36,910$ is the mean expected median value of owner-occupied homes in suburbs that bound the Charles River for suburbs with zero proportion of non-retail business acres per town.
- β_2: one unit increase in proportion of non-retail business acres per town is associated with $666 decrease in the mean expected median value of owner-occupied homes while holding indus (proportion of non-retail business aces per town) constant.

```r
model_medv_chas_indus <- lm(formula=medv ~ chas_f + indus, Boston)
summary(model_medv_chas_indus)
```

```
##
## Call:
## lm(formula = medv ~ chas_f + 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_f1      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
```

```r
class(Boston$indus)
```

```
## [1] "numeric"
```

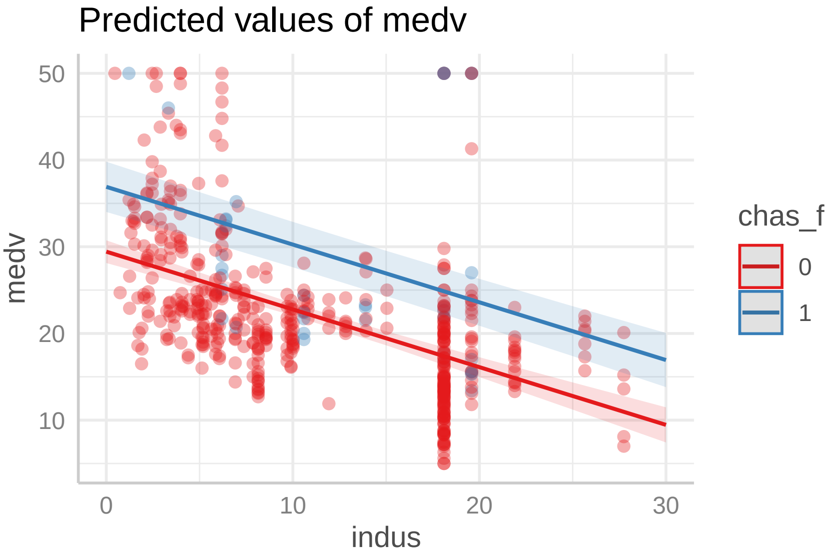
3. Given the results from #1 and #2, what do you conclude about chas and indus in relation to medv? Support your conclusion with a visualization.

- The full model (with chas and indus) has an $R^2 = 0.2765$, meaning that this set of predictors explain 27.7% of the variability in our outcome's values, while the reduced model (with chas only) has an $R^2 = 0.03072$ and explains only 3.1% of the variability in our outcome's values. The full model with 503 degrees of freedom reduces the residual sum of squares by 10,498 (F-test=170.85, p<.05) compared to the reduced model with 504 degrees of freedom. Taken together, the more complicated model is a better fit and explains more variation in our Y variable and significantly minimizes the least squares, and we conclude that variables chas and indus has a linear relationship with to median home values.

```r
anova(model_medv_chas, model_medv_chas_indus)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ chas_f
## Model 2: medv ~ chas + indus
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 41404
## 2    503 30906  1     10498 170.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Relationship between median home value, river bounds, & industrial acres
ggpredict(model_medv_chas_indus, terms = c("indus","chas_f")) %>%
  plot(add.data=TRUE, ci=TRUE)
```



4. You think that the proportion of non-retail business acres may have a different impact on the expected average median home values in towns bordering the Charles River vs. those that don't. Fit a linear regression model with an interaction term to assess this. What do you observe? Support your observations with a visualization.

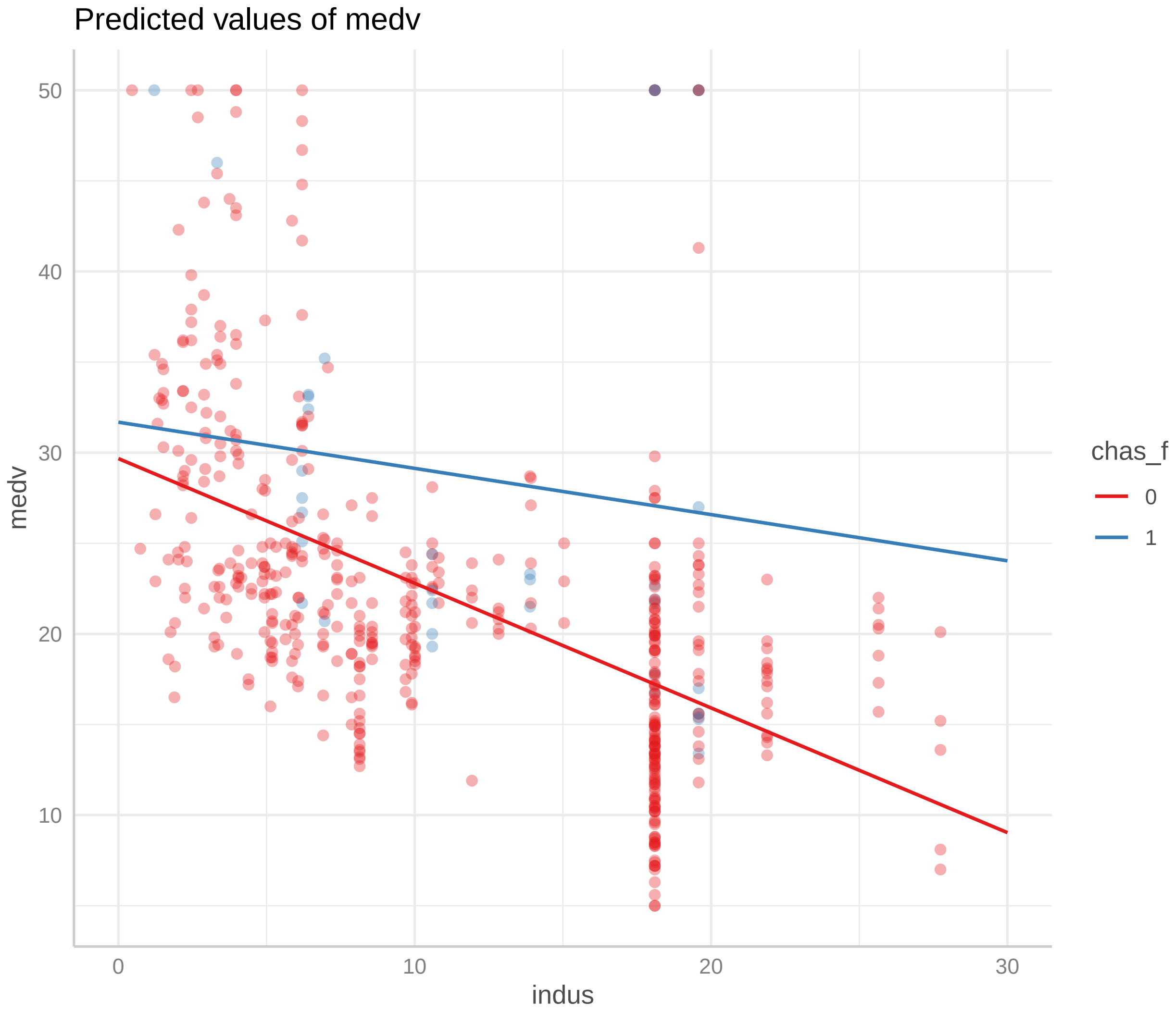
- The proportion of non-retail business acres does have a different impact (higher rate of change) on the expected average median home values for owner occupied homes in suburbs bordering the Charles River as proportion of non-retail business acres increase as indicated on the graphs below.

```
model_chas_indus <- lm(medv ~ chas_f*indus, Boston)
summary(model_chas_indus)
```

```
##
## Call:
## lm(formula = medv ~ chas_f * 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_f1       2.00791    3.22539   0.623   0.5339
## indus        -0.68799    0.05217 -13.188   <2e-16 ***
## chas_f1: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
```

```
# violin_y<-ggplot(aes(x=as.factor(chas), y=medv,
#                        color=as.factor(chas)), data=Boston)+
#   geom_violin(trim=T)+
#   stat_summary(fun=mean, geom="point", size=1, color="brown")+
#   theme(legend.position="none")
# violin_y
```

```
 #interaction of medv, river boundary, and industrial acres
ggpredict (model_chas_indus, terms = c("indus", "chas_f")) %>%
plot(add.data=TRUE, ci=FALSE)
```



```
#
# indus<-ggplot(aes(x=indus,y=medv),data=Boston)+
#   geom_smooth(method="lm", se=FALSE)
#
# chas<-ggplot(aes(x=chas, y=medv, color=chas),data=Boston)+
#   geom_violin(trim=T)+
#   stat_summary(fun=mean, geom="point", size=3, color="purple")+
#   theme(legend.position="none")
#
# int<-ggplot(aes(x=indus,y=medv),data=Boston)+
#   facet_grid(~chas)+
#   geom_smooth(method="lm", se=FALSE)
# grid.arrange(indus,chas,int, layout_matrix=rbind(c(1,2),c(3,3)))
# int
```