Lab 10: Modeling Basics I

Sunghwu Song

April 16, 2021

Introduction

This lab is about multiple regression and model interpretation. Part 1 is about multiple regression and collinearity. Collinearity refers to the situation in which two or more predictor variables are closely related to one another. Collinearity reduces the accuracy of the estimates of the regression coefficients, causing the standard error for coefficients to grow. Consequently, collinearity resutls in a decline in the t-statistics.

Part 2 is about modeling with categorical variable. The interpretation of models contain categorical variables is different from models do not contain categorical variables.

You will need to modify the code chunks so that the code works within each of chunk (usually this means modifying anything in ALL CAPS). You will also need to modify the code outside the code chunk. When you get the desired result for each step, change `Eval=F` to `Eval=T` and knit the document to HTML to make sure it works. After you complete the lab, you should submit your HTML file of what you have completed to Sakai before the deadline.

Part 1: Multiple linear regression

Q1. Run the following code to create the vectors `x1`, `x2`, and `y`.

```
set.seed(1)
n <- 100
x1 <- runif(n)
x2 <- runif(n,10,20)
y <- 2+2*x1+0.3*x2+rnorm(n)
```

a. (2 points) What is the correlation coefficient between `x1` and `x2`?

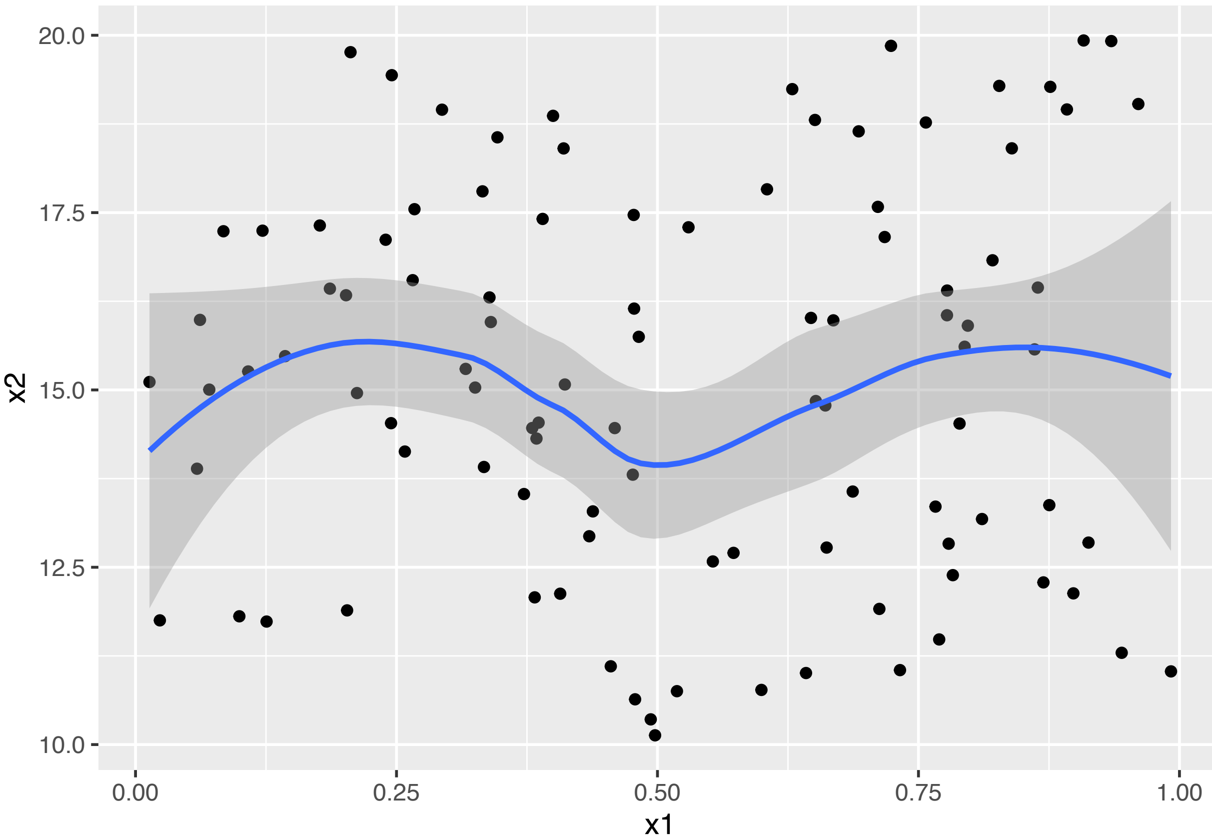
- Calculate the correlation between `x1` and `x2` with function `cor`.
- Create a scatter plot using `ggplot2` displaying the relationship between the variables `x1` and `x2` with scatter plot and smooth line.

```
cor(x1, x2)
```

```
## [1] 0.01703215
```

```
data = data.frame(x1=x1,x2=x2,y=y)
ggplot(data,aes(x=x1, y=x2)) +
  geom_point() +
  geom_smooth()
```

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



b. (2 points) Fit a least squares regression to predict `y` using `x1` and `x2`.

```
tidy(lm(y~x1+x2))
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.98     0.580       3.41 9.51e- 4
## 2 x1             1.93     0.363       5.31 6.89e- 7
## 3 x2             0.301    0.0358      8.42 3.33e-13
```

Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$? (alpha=0.05)

ANSWER_HERE:We can reject the null hypothesis for both beta values because of their low p values.

c. (2 points) Now fit a least squares regression to predict `y` using only `x1`.

```
tidy(lm(y~x1))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     6.52     0.277      23.5  3.94e-42
## 2 x1              1.98     0.476      4.17 6.64e- 5
```

Can you reject the null hypothesis $H_0 : \beta_1 = 0$? (alpha=0.05)

ANSWER_HERE: We can reject the null hypothesis because of the low p.value.

d. (2 points) Now fit a least squares regression to predict `y` using only `x2`.

```
tidy(lm(y~x2))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    2.93     0.623       4.70 8.64e- 6
## 2 x2             0.305    0.0404      7.53 2.46e-11
```

Can you reject the null hypothesis $H_0 : \beta_2 = 0$? (alpha=0.05)

ANSWER_HERE: We can reject the null hypothesis because of the low p.value.

2. Run the following code to create the vectors `x1`, `x2`, and `y`.

```
set.seed(1)
n <- 100
x1 <- runif(n)
x2 <- 0.5*x1+rnorm(n,0,.01)
y <- 2+2*x1+0.3*x2+rnorm(n)
```

a) (4 points) Repeat parts a, b, c, and d of Exercise 1 using the new vectors `x1`, `x2` and `y`.

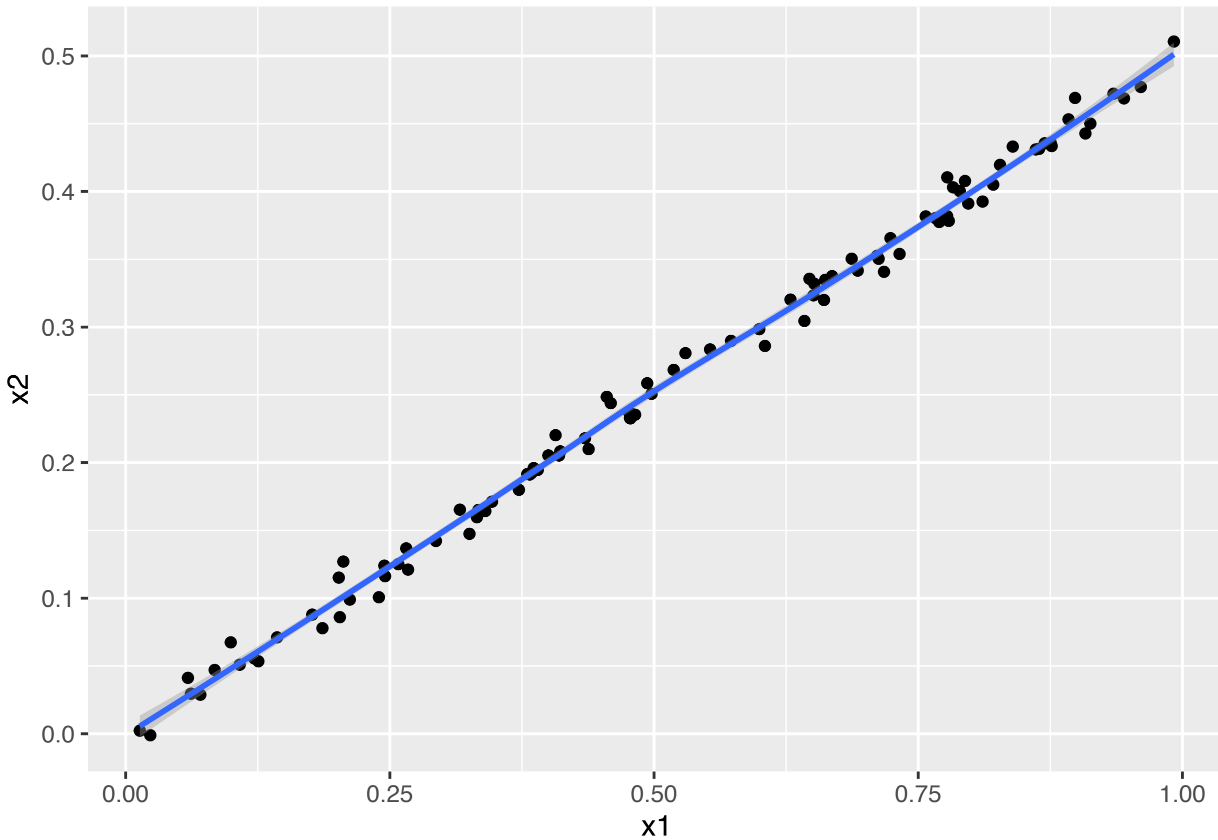
correlation and plot:

```
cor(x1, x2)
```

```
## [1] 0.9975904
```

```
data = data.frame(x1=x1,x2=x2,y=y)
ggplot(data,aes(x=x1, y=x2)) +
  geom_point() +
  geom_smooth()
```

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



three models:

```
tidy(lm(y~x1+x2))
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     2.13     0.232     9.19  7.61e-15
## 2 x1             -1.75     5.72     -0.307 7.60e- 1
## 3 x2              7.40    11.3       0.652 5.16e- 1
```

```
tidy(lm(y~x1))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     2.12     0.230      9.19 6.83e-15
## 2 x1              1.97     0.396      4.97 2.79e- 6
```

```
tidy(lm(y~x2))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     2.12     0.228      9.29 4.24e-15
## 2 x2              3.93     0.783      5.02 2.35e- 6
```

What differences do you see between Exercise 1 and Exercise 2? Explain why these differences occur.

ANSWER_HERE:The differences between ex1 and ex2 are evident in the p.values. For ex2 we can see that the p.values are significantly higher than that of ex1, and thus we can conclude that the differences between each exercise (shown in the graph/plots) are from this difference.

Part 2: Model with Categorical Variable

3. This part should be answered using the `Carseats` data set.

```
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

a. (1 point) Fit a multiple regression model to predict `Sales` using `Price`,`Urban`, `US` and get summary of the model.

```
summary(lm(Sales ~ Price+Urban+US, data = Carseats))
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b. (3 points) Provide an interpretation of each coefficient in the model. Be careful, some of the variables in the model are categorical variables. (Note that `Sales` variable represents unit sales **in thousands** at each location.)

ANSWER_HERE:

- `Price`: Sales went down by 54.459 when price was increased by a $1000.
- `Urban`: Sales went down by 21.916 when the store is in an urban area.
- `US`: Sales went up by 1200.573 when the store is in the US.

c. (1 point) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

ANSWER_HERE: We can reject UrbanYes because of its p.value.

d. (1 point) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome, and get summary of the model.

```r
summary(lm(Sales ~ Price+US, data = Carseats))
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

e. (2 points) How well do the models in (a) and (d) fit the data?

ANSWER_HERE: The models do not fit the data that well because of the low multiple r-squared values.