Regression HW 7 Solutions

Jun Lu, Mingfan Meng

Fall 2023

Contents

Total: 75 pts

Reminder. Project progress.

Reminder. Labs will be on weekly basis

1. Problem 8.4.(skip g) Muscle Mass (p.335, 25 pts)

```
mm.data <- read.table("./CH01PR27.txt", header = F)
colnames(mm.data) <- c("mass", "age")
head(mm.data, 3)
```

```
##   mass age
## 1  106  43
## 2  106  41
## 3   97  47
```

a. Fit regression model and comment. Plot the data with the esimated function.

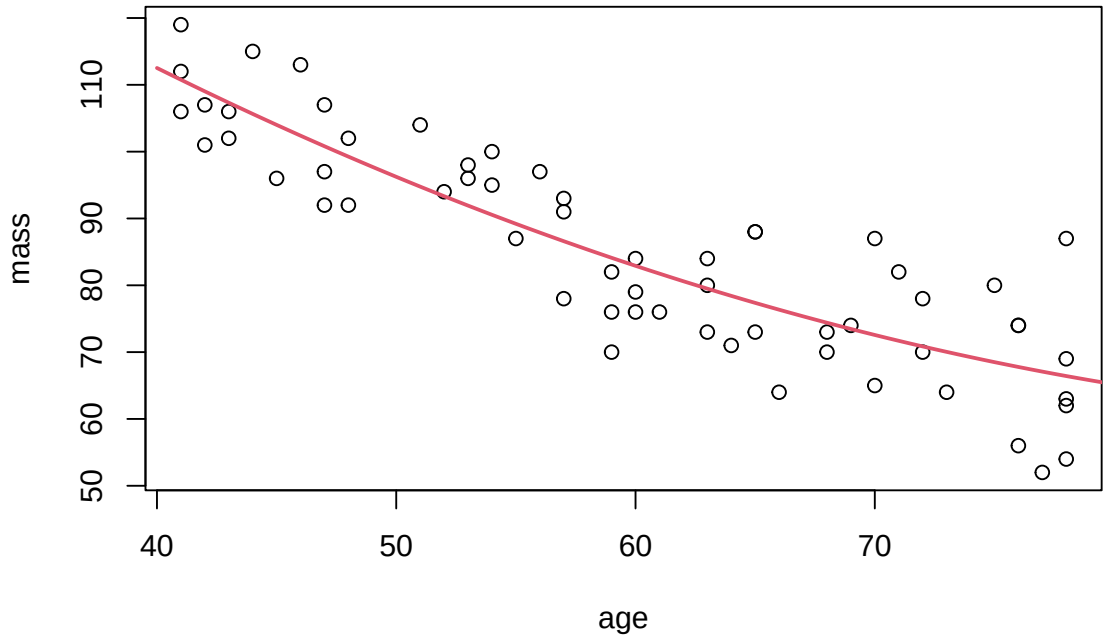
```
mm.reg <- lm(mass ~ age + I(age^2), data = mm.data)
summary(mm.reg)
```

```
##
## Call:
## lm(formula = mass ~ age + I(age^2), data = mm.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.086  -6.154  -1.088   6.220  20.578
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 207.349608  29.225118   7.095 2.21e-09 ***
## age          -2.964323   1.003031  -2.955  0.00453 **
## I(age^2)      0.014840   0.008357   1.776  0.08109 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.026 on 57 degrees of freedom
## Multiple R-squared:  0.7632, Adjusted R-squared:  0.7549
## F-statistic: 91.84 on 2 and 57 DF,  p-value: < 2.2e-16
```

- The estimated regression function is: $\widehat{mass} = 207.3496 - 2.9643 \cdot age + 0.01484 \cdot age^2$.
- If you center the variable `age`, you'll get $\widehat{mass} = 82.9357 - 1.1840 \cdot (centered.age) + 0.01484 \cdot (centered.age)^2$
- $R^2 = 0.7632$. About 76% of the total variation in muscle mass can be determined by the quadratic regression model using age as the predictor.
- Based on the R^2 and the following plot, the quadratic regression appear to be a good fit.

```
plot(mass ~ age, data=mm.data)
newx <- seq(40, 80, by=0.5)
yhat <- predict(mm.reg, newdata = data.frame(age=newx))
lines(newx, yhat, lwd=2, col=2)
```



b. Test the significance of the model.

- $H_0 : \beta_1 = \beta_{11} = 0$, H_a: at least one inequality.
- From the `summary()` output, the resulting F-statistic is 91.84, with $df_1 = 2$ and $df_2 = 57$. The resulting p-value is$< 2.2 \times 10^{-16}$, approximately 0.
- Since the p-value is $< \alpha = 0.05$, we reject H_0. The data provide significance eveidence to support that there is a regression relation.

c. CI for mean response

```
predict(mm.reg, newdata=data.frame(age=48), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 99.25461 96.28436 102.2249
```

- With 95% confidence, the mean muscle mass for women aged 48 is estimated to be between 96.2843 and 102.2249.

d. PI for one women

```
predict(mm.reg, newdata=data.frame(age=48), interval = "prediction", level = 0.95)
```

```
##        fit     lwr      upr
## 1 99.25461 82.9116 115.5976
```

- With 95% confidence, the muscle mass for **one** women aged 48 is estimated to be between 82.9116 and 115.5976.

e. Test the significance of the quadratic term.

- Let β_{11} denote the slope for the quadratic term (age^2). $H_0 : \beta_{11} = 0$, $H_a : \beta_{11} \neq 0$.
- From the `summary()` output (see part a), the resulting t-statistic is $t_{obs} = 1.776$, and the p-value= 0.08109.
- Since p-value= 0.08109 $> \alpha = 0.05$, we fail to reject H_0. There is not significance evidence to support the β_{11} is not 0. Hence, at 0.05 level, we can drop the quadratic terms from the regression model.

f. If you use centered age in part (a), express the fitted regression in terms of the original age.

- $\widehat{mass} = 207.3496 - 2.9643 \cdot age + 0.01484 \cdot age^2$.
- Again, this shows centering the variable may not be necessary.

2. Problem 8.25.(plus c, d, e, f) Grocery retailer. (p.249, 30 pts)

```
gr <- read.table("./CH06PR09.txt", header = F)
colnames(gr) <- c("labor", "case", "costs", "holiday")
```

a. Fit regression model (equation 8.58, p.334) using X_1 (case) and X_3 (holiday).

- The model is (without centering the variables):

$$labor = \beta_0 + \beta_1(case) + \beta_2(case^2) + \beta_3(holiday) + \beta_4(case \times holiday) + \beta_5(case^2 \times holiday) + \varepsilon$$

```
gr.pr825 <- lm(labor ~ case + I(case^2) + holiday + case:holiday
               + I(case^2):holiday, data = gr)

## The following code is equivalent (with variables in slightly different order)
# gr.pr825 <- lm(labor ~ case*holiday + I(case^2)*holiday, data = gr)

summary(gr.pr825)
```

```
##
## Call:
```

```
## lm(formula = labor ~ case + I(case^2) + holiday + case:holiday +
##     I(case^2):holiday, data = gr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -289.03  -95.85  -12.22   73.69  295.31
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.878e+03  7.193e+02   5.391 2.34e-06 ***
## case               1.858e-03  4.329e-03   0.429    0.670
## I(case^2)         -1.577e-09  6.328e-09  -0.249    0.804
## holiday            8.370e+02  1.380e+03   0.607    0.547
## case:holiday      -1.283e-03  8.692e-03  -0.148    0.883
## I(case^2):holiday  1.808e-09  1.309e-08   0.138    0.891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.8 on 46 degrees of freedom
## Multiple R-squared:  0.6867, Adjusted R-squared:  0.6526
## F-statistic: 20.16 on 5 and 46 DF,  p-value: 1.345e-10
```

- The fitted regression function is.

$$\widehat{labor} = 3877.9 + 0.001858(case) - 1.577 \times 10^{-9}(case^2) + 8.37(holiday) - 0.001283(case \times holiday) + 1.808 \times 10^{-9}(case^2 \times holiday)$$

- If you center age, the fitted regression function may look different using the centered variable. However, the estimates of the slopes for X_1^2 and $X_1^2X_3$ (high-order terms) will not be affected by centering or not. Further more, when you rewrite the fitted function in terms of the original variable, you will get the same results as above.

b. Test whether the interaction terms and quadratic terms can be dropped.

```
gr.pr825b <- lm(labor ~ case + holiday, data = gr)
anova(gr.pr825b, gr.pr825)
```

```
## Analysis of Variance Table
##
## Model 1: labor ~ case + holiday
## Model 2: labor ~ case + I(case^2) + holiday + case:holiday + I(case^2):holiday
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     49 992204
## 2     46 990762  3    1441.9 0.0223 0.9954
```

- Recall the full model is $labor = \beta_0+\beta_1(case)+\beta_2(case^2)+\beta_3(holiday)+\beta_4(case\times holiday)+\beta_5(case^2\times holiday)+\varepsilon$
- $H_0 : \beta_2 = \beta_4 = \beta_5 = 0$, H_a : At least one inquality.
- The resulting partial F-statistic is 0.0223, with $df_1 = 3$, $df_2 = 46$. The resulting p-value $= 0.9954$.
- Since p-value$= 0.9954 > \alpha = 0.05$, we fail to reject H_0. At 0.05 level, we can drop the quadratic and interactions terms from the model.

c. Why include variables that are not of main interest?

- We want to examine the effect of "holidays" on the response variable "labor hours" after considering/controlling/adjusting the effects of other predictors.
- If we do NOT include the number of cases shipped (X_1) in this problem, the estimated effect of holiday on labor hours (i.e., the difference in hours between holiday and non-holiday) may be due to different X_1 values. For example, holidays may have more cases, hence needs longer hours. So the labor hour could be related to cases, not holiday.
- By including X_1 in the regression model, we know the estimated effect (if there is any) of holiday on labor hour has been adjusted for the effect of X_1 on the labor hours.

d. The desgin matrix for the model in (a).

- The model in part (a) is

$$labor = \beta_0 + \beta_1(case) + \beta_2(case^2) + \beta_3(holiday) + \beta_4(case \times holiday) + \beta_5(case^2 \times holiday) + \varepsilon$$

  Hence, the design matrix for the model in part (a) has 52 rows (sample size = 52) and 6 columns (6 parameters now). The first 3 rows are:

$$\mathbf{X} = \begin{pmatrix} 1 & 305657 & 305657^2 & 0 & 0(=305657 \cdot 0) & 0(=305657^2 \cdot 0) \\ 1 & 328476 & 328476^2 & 0 & 0(=328476 \cdot 0) & 0(=328476^2 \cdot 0) \\ 1 & 317164 & 317164^2 & 0 & 0(=317164 \cdot 0) & 0(=317164^2 \cdot 0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}_{52 \times 6}$$

- For R users, the design matrix can be saved to the regression output when you set argument `x = T` in the `lm()` function. (The row index is NOT part of the design matrix.)

```
gr.pr825 <- lm(labor ~ case + I(case^2) + holiday + case:holiday
               + I(case^2):holiday, data = gr, x = T)
dim(gr.pr825$x)
```

```
## [1] 52  6
```

```
head(gr.pr825$x)  # Ignore the row index.
```

```
##   (Intercept)   case    I(case^2) holiday case:holiday I(case^2):holiday
## 1           1 305657  93426201649       0            0                 0
## 2           1 328476 107896482576       0            0                 0
## 3           1 317164 100593002896       0            0                 0
## 4           1 366745 134501895025       0            0                 0
## 5           1 265518  70499808324       1       265518       70499808324
## 6           1 301995  91200980025       0            0                 0
```

e. Reponse vector, paramter vector, and the estimate of the parameter vector.

- The response vector is:

$$\underline{y} = \begin{pmatrix} 4264 \\ 4496 \\ 4317 \\ \vdots \\ 4499 \\ 4186 \\ 4342 \end{pmatrix}$$

- The parameter vector and its estimate are provide below. Note that the "parameter vector" should be just in β notation, since the "parameters" refers to the unknown "true values" about the population or the model. The "estimates" are the values we calculate from the sample data, and the values are in the software output. (Use `gr.pr8.25$coef` or `summary()`.)

$$\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}, \quad \underline{\hat{\beta}} = \underline{\mathbf{b}} = \begin{pmatrix} 3877.90 \\ 0.001858 \\ -1.577 \times 10^{-9} \\ 8.37 \\ -0.001283 \\ 1.808 \times 10^{-9} \end{pmatrix}.$$

f. Compare the matrix notations for this model (with interaction) and the ones for the 1st-order model (previous assignment).

- Both the design matrix and the parameter vector (hence the estimate of the parameter vector) depend on how we build the model, even when we use the same data set.
- The response vector will remain the same as long as the repose variable is not changed. The response vector will not be affected by what predictors we use.

3. Problem 8.39.(plus d) CDI (p.341. Data in Appendix C.2. 20 pts)

The number of active physicians (Y, 8th column) is to be regressed against total population (5th column), total personal income (16th column), and geographic region (17th column).

```
cdi.data <- read.table("APPENCO2.txt", header=F)

# Select the columns.
cdi.data <- cdi.data[ , c(8, 5, 16, 17)]
colnames(cdi.data) <- c("physician", "population", "income", "region")

# Declare "Region" as a categorical variable
cdi.data$region1 <- as.factor(cdi.data$region)
cdi.data$region2 <- factor(cdi.data$region, labels = c("NE", "NC", "S", "W"))

head(cdi.data)
```

```
##   physician population income region region1 region2
## 1     23677    8863164 184230      4       4       W
## 2     15153    5105067 110928      2       2      NC
## 3      7553    2818199  55003      3       3       S
## 4      5905    2498016  48931      4       4       W
## 5      6062    2410556  58818      4       4       W
## 6      4861    2300664  38658      1       1      NE
```

- Note that you you can use `region1` as long as you know how 1=NE, 2=NC, 3=S and 4=W.
- Alternatively, `region2` has lables that make the results easier to read.

a. Fit a 1st-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise.

- Note that Region W is used as the "reference level".

```
cdi.data$region2 <- relevel(cdi.data$region2, ref="W")
cdi.regW <- lm(physician ~ population + income + region2, data=cdi.data)
summary(cdi.regW)
```

```
##
## Call:
## lm(formula = physician ~ population + income + region2, data = cdi.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.075e+02  7.028e+01  -2.952  0.00332 **
## population   5.515e-04  2.835e-04   1.945  0.05243 .
## income       1.070e-01  1.325e-02   8.073  6.8e-15 ***
## region2NE    1.490e+02  8.683e+01   1.716  0.08685 .
## region2NC    1.455e+02  8.515e+01   1.709  0.08817 .
## region2S     1.912e+02  8.003e+01   2.389  0.01731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

- The estimated regression function is

$$\widehat{physcian} = -207.5 + 0.0005515(population) + 0.107(income) + 149.0(NE) + 145.5(NC) + 191.2(S)$$

b. Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the northcentral region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.

This question is asking for the estimation of the difference on "physicians" between NE and NC regions. We have two options here. (Option 2 is easier.)

- Method 1: Revise the reference level and rerun the analysis. (Easier.)
  - If we set "NC" as the reference level, then the slope for the dummy variable of "NE" is the difference on physicians between NE and NC regions.
  - We can then use the `confint()` function to get the CI for the slope.

```
cdi.regNC <- lm(physician ~ population + income + relevel(region2, ref="NC"),
                data=cdi.data)
confint(cdi.regNC, level = 0.90)
```

```
##                                             5 %          95 %
## (Intercept)                       -1.559984e+02  32.059815828
## population                         8.407549e-05   0.001018844
## income                             8.516238e-02   0.128860685
## relevel(region2, ref = "NC")W     -2.858902e+02  -5.162732980
## relevel(region2, ref = "NC")NE    -1.264158e+02 133.402068257
## relevel(region2, ref = "NC")S     -7.199636e+01 163.376078063
```

- Method 2: Use the results from part (a), where W is the reference level. (Students in Math Stat or Data Science should know about this.)
  - The difference between NE and NC regions is $(\beta_{NE} - \beta_{NC})$, which can be estimated by $(\hat{\beta}_{NE} - \hat{\beta}_{NC})$.
  - In order to compute the 90% confidence interval, we need to find the standard error: $se(\hat{\beta}_{NE} - \hat{\beta}_{NC})$. This requires addtional statistical knowledge and output results. (Students in Math Stat or Data Science should know about this.)

$$se(\hat{\beta}_{NE} - \hat{\beta}_{NC}) = \sqrt{(se(\hat{\beta}_{NE}))^2 + (se(\hat{\beta}_{NC}))^2 - 2 \cdot cov(\hat{\beta}_{NE},\ \hat{\beta}_{NC})}$$

```
est <- 149.0 - 145.5  # From part (a)
vcov(cdi.regW)  # compute the covariance
```

```
##                (Intercept)    population        income     region2NE
## (Intercept)  4.939505e+03 -3.714907e-03  1.191948e-01 -4.510082e+03
## population  -3.714907e-03  8.039664e-08 -3.709872e-06  4.293737e-03
## income       1.191948e-01 -3.709872e-06  1.756947e-04 -1.852445e-01
## region2NE   -4.510082e+03  4.293737e-03 -1.852445e-01  7.539998e+03
## region2NC   -4.468275e+03  1.243014e-03 -3.591859e-02  4.289966e+03
## region2S    -4.447947e+03  5.539746e-06  2.436229e-02  4.232971e+03
##                region2NC      region2S
## (Intercept) -4.468275e+03 -4.447947e+03
## population   1.243014e-03  5.539746e-06
## income      -3.591859e-02  2.436229e-02
## region2NE    4.289966e+03  4.232971e+03
## region2NC    7.251021e+03  4.279499e+03
## region2S     4.279499e+03  6.405283e+03
```

```
se <- sqrt((86.83)^2 + (85.15)^2 -2*4.289966e+03)
tcrit <- qt(1 - 0.1/2, df = cdi.regW$df.residual)
data.frame(Estimate = est, Std.Err = se, CI90.L = est - tcrit*se, CI90.U = est+tcrit*se )
```

```
##   Estimate Std.Err   CI90.L  CI90.U
## 1      3.5 78.8038 -126.398 133.398
```

- Results from both methods are the same (subject to rounding error.)
  - At 90% confidence, the mean number of active physicians in NE region differs from NC region by (-126.4, 133.4).
  - Note that 0 is included in the above interval. Thus, we conclude that there is NOT a significant difference between the NE and NC regions.

c. Test wheter "Region" can be dropped.

```
cdi.noregi <- lm(physician ~ population + income, data=cdi.data)
anova(cdi.noregi, cdi.regW)
```

```
## Analysis of Variance Table
##
## Model 1: physician ~ population + income
## Model 2: physician ~ population + income + region2
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    437 140967081
## 2    434 139093455  3   1873626 1.9487  0.121
```

```
# anova(cdi.noregi, cdi.regNC)  # will lead to same conclusion
```

- $H_0: \ \beta_{NE} = \beta_{NC} = \beta_S = 0$. (No geographic effects. We can drop "region".)
- H_A : At least one inequality. (Significant geographic effects. Keep "region" in the model.)
- The resulting partial F-statistic is $F_{partial} = 1.9487$, with $df_1 = 3$, $df_2 = 437$. The p-value=0.121.
- Since the p-value= $0.121 > \alpha = 0.1$, do not reject H_0. The geographic effects is not significant, after adjusting for the effects of other predictors (population, income).

d. Design matrix in follow part (a) (using W as the reference level)

- The design matrix should be 440 rows and 6 columns (1 column of 1's, 1 column for population, 1 column for income, 3 columns for the dummy variables of region).

$$X = \begin{pmatrix} 1 & 8863164 & 184230 & 0 & 0 & 0 \\ 1 & 5105067 & 110928 & 0 & 1 & 0 \\ 1 & 2818199 & 55003 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}_{440\times 6}$$

- You can also use the `x=T` option inside the `lm()` function. (The row index is NOT part of the design matrix.)

```
cdi.regW <- lm(physician ~ population + income + region2, data=cdi.data, x=T)
dim(cdi.regW$x)
```

```
## [1] 440   6
```

```
head(cdi.regW$x)
```

```
##   (Intercept) population income region2NE region2NC region2S
## 1           1    8863164 184230         0         0        0
## 2           1    5105067 110928         0         1        0
## 3           1    2818199  55003         0         0        1
## 4           1    2498016  48931         0         0        0
## 5           1    2410556  58818         0         0        0
## 6           1    2300664  38658         1         0        0
```

This is the end of HW 7.