PSTAT 126 - Assignment 6

Fall 2022

Due: Tuesday, November 15 at 11:59 pm

Note: ***Submit both your `Rmd` and generated pdf file to Gradescope.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

1. Using the prostate data from the faraway package with lpsa (log prostate specific antigen) as response and lcavol (log cancer volume) as predictor, the fitted model is

$$\text{lpsa } = 1.507 + 0.719 \text{ lcavol}$$

Provide an interpretation of the estimated coefficient for lcavol based on the fact that both variables are log-transformed.

Solution:

Given lpsa and lcavol are both log-transformed, and with the lcavol coefficient of 0.719, then the expected value of lpsa is changed by 100[(1+p)^0.719-1]% where p is the % change in lcavol. For example, if lcavol is changed by 10%, then lpsa changes by 100[(1+.10)^0.719-1] = 7.09306%

2. In a study of faculty salaries in a small college in the Midwest, a linear regression model was fit, giving the fitted mean function

$$E(\text{ Salary } \mid \text{ Sex }) = 24697 - 3340 \text{ Sex},$$

where Sex equals 1 if the faculty member was female and 0 if male. The response Salary is measured in dollars (the data are from the 1970s).

   (a) Give a sentence that describes the meaning of the two estimated coefficients.

   Solution:

24697 is the expected salary for faculty members if male (0) and -3340 is the decrease in salary if the faculty member is a female (1).

   (b) An alternative mean function fit to these data with an additional term, Years, the number of years employed at this college, gives the estimated mean function

$$E(\text{ Salary } \mid \text{ Sex, Years }) = 18065 + 201 \text{ Sex } + 759 \text{ Years }.$$

   The important difference between these two mean functions is that the coefficient for Sex has changed signs. Provide an explanation as to how this could happen.

Solution:

Now, the salary for employees are not only considered by their sexuality but also their employment time in the company. So, an explanation as to why the coefficient has changed signs could be that the male faculty members on average have been employed longer than female faculty members. So if we were to control for the number of years employed, the salary of female faculty members is actually higher than that of males. Thus in the first model, we do not account for years employed and so it seems that the salary of male faculty members is larger. Hence the coefficient for sex has changed signs.

3. This problem uses the data set cakes from the alr4 package, which contains the results of a baking experiment on n = 14 packaged cake mixes. The variables X1 and X2 data are the predictors representing baking time in minutes and baking temperature in degrees Fahrenheit, respectively. The response Y is a palatability score indicating quality of the cake.

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

(a) Fit the model

$$E(\mathrm{Y} \mid \mathrm{X1}, \mathrm{X2}) = \beta_0 + \beta_1 \mathrm{X1} + \beta_2 \mathrm{X2} + \beta_{11} \mathrm{X1}^2 + \beta_{22} \mathrm{X2}^2 + \beta_{12} \mathrm{X1X2}$$

and verify that the p-values for the quadratic terms and the interaction are all less than 0.005.

Solution:

```
data(cakes)
attach(cakes)
fit <- lm(Y~ X1+X2+I(X1^2)+I(X2^2)+I(X1*X2), data = cakes)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2), data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
```

```
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
## I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## I(X1 * X2)  -4.163e-02  1.072e-02  -3.883 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1  4.3232  4.3232  23.515 0.0012730 **
## X2          1  7.4332  7.4332  40.432 0.0002186 ***
## I(X1^2)     1  2.1308  2.1308  11.591 0.0092987 **
## I(X2^2)     1 10.5454 10.5454  57.361 6.462e-05 ***
## I(X1 * X2)  1  2.7722  2.7722  15.079 0.0046537 **
## Residuals   8  1.4707  0.1838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see by the summary and anova tables that the p-values are indeed all less than 0.005.

(b) The cake experiment was carried out in two blocks of seven observations each. It is possible that the response might differ by block, due to differences in air temperature or humidity, for example. Add a main effect for the Block variable to model in part a), fit the model, and summarize results.

Solution:

```
fit1 <- lm(Y~ X1+X2+I(X1^2)+I(X2^2)+I(X1*X2)+ block, data = cakes)
summary(fit1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2) + block,
##     data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## X1           2.592e+01  4.903e+00   5.287 0.001140 **
## X2           9.918e+00  1.228e+00   8.080 8.56e-05 ***
```

```
## I(X1^2)     -1.569e-01  4.151e-02  -3.779 0.006898 **
## I(X2^2)     -1.195e-02  1.660e-03  -7.197 0.000178 ***
## I(X1 * X2)  -4.163e-02  1.128e-02  -3.690 0.007754 **
## block1       1.143e-01  2.412e-01   0.474 0.650014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9077
## F-statistic: 22.31 on 6 and 7 DF,  p-value: 0.0003129
```

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1  4.3232  4.3232 21.2361 0.0024600 **
## X2          1  7.4332  7.4332 36.5132 0.0005198 ***
## I(X1^2)     1  2.1308  2.1308 10.4670 0.0143465 *
## I(X2^2)     1 10.5454 10.5454 51.8009 0.0001779 ***
## I(X1 * X2)  1  2.7722  2.7722 13.6177 0.0077544 **
## block       1  0.0457  0.0457  0.2246 0.6500138
## Residuals   7  1.4250  0.2036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the table that the block variable is not significant.

4. The data BGSall in the alr4 package contains information on n = 136 children in the Berkeley Guidance study, including heights at ages 9 and 18 (HT9 and HT18), and gender (Sex = 0 for male, 1 for female). Consider the regression of HT18 on HT9 and the grouping factor Sex.

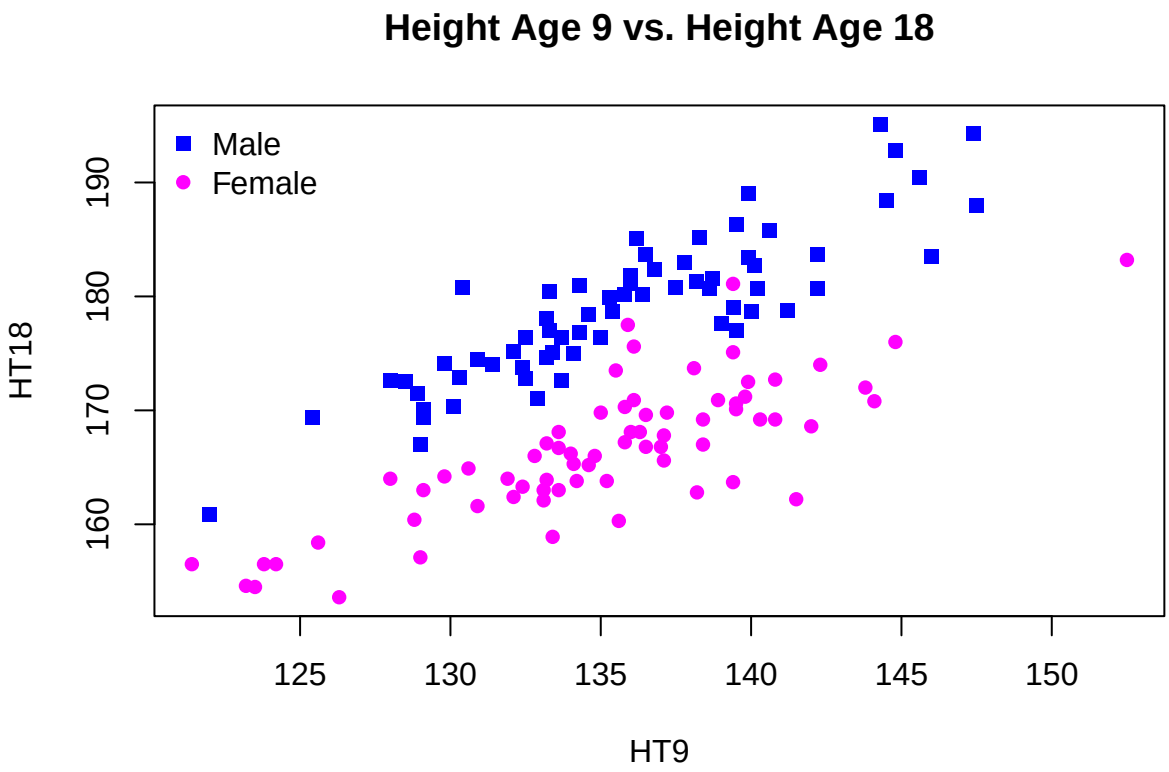
```
data('BGSall')
attach(BGSall)
```

(a) Draw the scatterplot of HT18 versus HT9, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.

Solution:

```
HT18 <- BGSall$HT18
HT9 <- BGSall$HT9
sex <- BGSall$Sex

plot(HT9,HT18,type = 'n', main = 'Height Age 9 vs. Height Age 18')
points(HT9[Sex==0], HT18[Sex==0],col = 'blue', pch =15)
points(HT9[Sex==1], HT18[Sex==1],col = 'magenta', pch = 16)
legend('topleft',bty='n',col=c('blue','magenta'),c('Male','Female'),pch =c(15,16))
```



The appropriate mean function for this data would be fitting a parallel model. Here, we see that there is no interaction between gender and age groups.

(b) Obtain the appropriate test for a parallel regression model.

Solution:

```
fit2 <- lm(HT18 ~ HT9 + factor(Sex), data = BGSall)
fit3 <- lm(HT18 ~ HT9*factor(Sex), data = BGSall)
anova(fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ HT9 + factor(Sex)
## Model 2: HT18 ~ HT9 * factor(Sex)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    133 1566.9
## 2    132 1532.5  1    34.409 2.9638 0.08749 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.08749, greater than 0.05; thus we accept the null hypothesis that the coefficient for interaction is 0 at a 5% confidence interval. Hence the parallel regression model is adequate.

(c) Assuming the parallel regression model is adequate, estimate a 95% confidence interval for the difference between males and females. For the parallel regression model, this is the difference in the intercepts of the two groups.

Solution:

```
summary(fit2)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + factor(Sex), data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4694  -2.0952  -0.0136   1.7101  10.4467
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   48.51731    7.33385   6.616 8.27e-10 ***
## HT9            0.96006    0.05388  17.819  < 2e-16 ***
## factor(Sex)1 -11.69584    0.59036 -19.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.432 on 133 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8494
## F-statistic: 381.7 on 2 and 133 DF,  p-value: < 2.2e-16
```

Observing the summary, we see that The difference in the intercepts of the two groups (male and female) is 11.69584. SO, the estimated difference between males and females is 11.69584 and the standard error is 0.59036 and degree of freedom of residual is equal to 133.Now constructing a t-value for the test,

```
abs(qt(0.05/2, 133))
```

```
## [1] 1.977961
```

```
(11.69584 - 1.98 * 0.59036)
```

```
## [1] 10.52693
```

```
(11.69584 + 1.98 * 0.59036)
```

```
## [1] 12.86475
```

Hence, our 95% confidence interval between males and females is : (10.52693, 12.86475)

5. The data set infmort from the faraway package contains information on the mortality of infants for 105 nations. The variable mortality gives the number of deaths per 1000 live births, while income is the per capita income in US dollars and region indicates the geographic area of the nation. Consider the model

   $E(\log(\text{ mortality }) \mid \text{ income,region }) = \beta_0 + \beta_1 \log(\text{ income }) + \beta_2 \text{ region } + \beta_{12} \text{ region } * \log(\text{ income })$

```
library("faraway")
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:alr4':
##
##     cathedral, pipeline, twins
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
data("infmort")
```

(a) State the null and alternative hypotheses for the overall F-test for this model. Perform the test and summarize results.

Solution:

Null Hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_{12} = 0$$

Alternative hypothesis:

$$H_1 : \text{At least one of } \beta_{ai} \neq 0$$

Perform the test:

```
fit4 <- lm(log(mortality)~region+log(income)+log(income)*region, data = infmort)
summary(fit4)
```

```
##
## Call:
## lm(formula = log(mortality) ~ region + log(income) + log(income) *
##     region, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46809 -0.26530 -0.02148  0.27478  3.14219
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.9385     0.6362   7.763 1.06e-11 ***
## regionEurope                  2.0882     1.8422   1.134   0.2599
## regionAsia                    1.2634     0.8561   1.476   0.1434
## regionAmericas                1.5661     1.1856   1.321   0.1898
## log(income)                  -0.0112     0.1235  -0.091   0.9280
## regionEurope:log(income)     -0.5205     0.2516  -2.069   0.0413 *
## regionAsia:log(income)       -0.3798     0.1580  -2.404   0.0182 *
## regionAmericas:log(income)   -0.3978     0.1979  -2.010   0.0473 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5971 on 93 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6464, Adjusted R-squared:  0.6198
## F-statistic: 24.29 on 7 and 93 DF,  p-value: < 2.2e-16
```

From Summary, observe that F= 24.29 on 7 and 93 DF and p-value is 2.2e-16. Therefore, we reject the null hypothesis since p-value is less than the 0.05, and we can conclude that the fitted model is significant.

(b) Explain the practical meaning of the hypothesis $\mathcal{H}_0 : \beta_{12} = \beta_2 = 0$ in the context of the above model.

Solution: Of the above model, $\mathcal{H}_0 : \beta_{12} = \beta_2 = 0$ implies that the region has no impact on the relationship between income and mortality, so when given income and region, log(mortality) is independent of the region and interaction between the region and log(income).

0.1 c).

(c) Perform a test for the hypothesis in part b) and summarize your results.

Solution:

```
m1 <- lm(log(mortality)~log(income), data = infmort)
m2 <- lm(log(mortality)~log(income)+ region +log(income)*region, data = infmort)
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: log(mortality) ~ log(income)
## Model 2: log(mortality) ~ log(income) + region + log(income) * region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     99 46.685
## 2     93 33.152  6    13.533 6.3274 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the p-value is 1.31e-05, which is much smaller than 0.05. Therefore, we reject the null hypothesis that beta12 and beta2 are 0. Therefore, the region and interaction between region and log(income) are significant in determining the log(mortality).