R_Project3_S24_

Zheng Yi Lai

2024-02-19

Question 1. Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar data in faraway library to answer the following:

a) Fit a regression model with taste as the response and the three chemical contents as predictors. What is the dimension of the design matrix X? What is the dimension of each sub-space (estimation space, error space)?

```
library(faraway)
data(cheddar)

#Fit in the model
lm_1<- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)

#The dimension of the design matrix X
cat("Dimension of the design matrix:",dim( model.matrix(lm_1))[1],",",dim( model.matrix(lm_1))[2],
    "\n")
```

```
## Dimension of the design matrix: 30 , 4
```

```
# Extract the coefficients matrix and find its dimensions
coefficients <- coef(lm_1)
dimensions <- length(coefficients)
cat("Dimension of the estimation space:", dimensions,"\n")
```

```
## Dimension of the estimation space: 4
```

```
# Extract the residuals and find their dimensions
residuals <- residuals(lm_1)
length1 <- length(residuals)
cat("Dimension of the error space:", length1 - dimensions,"\n")
```

```
## Dimension of the error space: 26
```

(b) Perform hypothesis tests to assess the individual statistical significance of the predictors at the 5% level for the model in (a).

```r
#Find the indiviudal p-values by looking the summary and put them into the dataframe.
summary(lm_1)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```r
variable_1 <- c("Acetic","H2S","Lactic")
p_value1 <-c(summary(lm_1)$coef[2,4],summary(lm_1)$coef[3,4],
             summary(lm_1)$coef[4,4])
df <- data.frame(variable_1,p_value1)
df
```

```
##   variable_1    p_value1
## 1     Acetic 0.941979774
## 2        H2S 0.004247081
## 3     Lactic 0.031079481
```

The p-value for the coefficient of the predictor variable "Acetic" is 0.942, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis. This indicates that there is insufficient evidence to conclude that the coefficient for "Acetic" is significantly different from zero, and hence, "Acetic" is not statistically significant in predicting the dependent variable.

The p-value for the coefficient of the predictor variable "H2S" is 0.004, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis. This suggests that the coefficient for "H2S" is significantly different from zero, and hence, "H2S" is statistically significant in predicting the dependent variable.

The p-value for the coefficient of the predictor variable "Lactic" is 0.03, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis. This indicates that the coefficient for "Lactic" is significantly different from zero, and hence, "Lactic" is statistically significant in predicting the dependent variable

(c) Acetic and H2S are measured on a natural log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

```r
#Fit the model
lm_2 <- lm(taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
summary(lm_2)
```

```
##
## Call:
## lm(formula = taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.209  -7.266  -1.651   7.385  26.335
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.897e+01  1.127e+01  -1.684   0.1042
## exp(Acetic)  1.891e-02  1.562e-02   1.210   0.2371
## exp(H2S)     7.668e-04  4.188e-04   1.831   0.0786 .
## Lactic       2.501e+01  9.062e+00   2.760   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 26 degrees of freedom
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.5264
## F-statistic: 11.75 on 3 and 26 DF,  p-value: 4.746e-05
```

```r
#Find the indiviudal p-values by looking the summary and put them into the dataframe.
variable_2 <- c("Acetic","H2S","Lactic")
p_value2 <-c(summary(lm_2)$coef[2,4],summary(lm_2)$coef[3,4],
             summary(lm_2)$coef[4,4])
df_1 <- data.frame(variable_2,p_value2)
df_1
```

```
##   variable_2   p_value2
## 1     Acetic 0.23711453
## 2        H2S 0.07856795
## 3     Lactic 0.01046242
```

For the adjusted model, the p-value for the coefficient of the predictor variable "Lactic" is 0.01, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis. This indicates that the coefficient for "Lactic" is significantly different from zero, and hence, "Lactic" is statistically significant in predicting the dependent variable.

(d) Can we use an F-test to compare these two models? Explain.

```r
#Annova to compare two models
anova(lm_2,lm_1)
```

```
## Analysis of Variance Table
##
## Model 1: taste ~ exp(Acetic) + exp(H2S) + Lactic
## Model 2: taste ~ Acetic + H2S + Lactic
```

```
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     26 3253.6
## 2     26 2668.4  0     585.2
```

We cannot use the F-test to compare these models because there is no nested relationship between the two models. This is because we have changed the values to exponent transformations of two variables. The F-test is typically used to compare nested models, where one model is a simplified version of the other, obtained by imposing constraints on the parameters.

Furthermore, when attempting to apply the F-test in this scenario, we encounter difficulties in obtaining a p-value. This further underscores the inappropriateness of using the F-test to compare these two models. Therefore, due to the lack of nestedness and the inability to obtain a p-value, it is evident that the F-test is not suitable for comparing these models.

(e) Which model provides a better fit to the data? Explain your reasoning.

```r
#The original model of RSS
residual <- residuals(lm_1)
RSS_1 <- sum(residual^2)
RSS_1
```

```
## [1] 2668.411
```

```r
#The log transformation model of RSS
residuals <- residuals(lm_2)
RSS_2 <- sum(residuals^2)
RSS_2
```

```
## [1] 3253.608
```

We have observed that the original model, with a RSS (Residual Sum of Squares) value of 2668.41, exhibits a better fit to the data compared to the second model, which incorporates log transformations of two variables. The lower RSS value of the original model indicates that the observed data points deviate less from the model's predicted values, suggesting a better overall fit.

In contrast, the second model, with exponentiated values of two variables, yields a higher RSS value of 3253.608. This higher RSS value indicates that the deviations between the observed data points and the predicted values in this model are greater, implying a poorer fit compared to the original model.

Therefore, based on the RSS values alone, it appears that the original model provides a more suitable representation of the relationship between the variables in the dataset.

(f) If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

```r
# Extract coefficient for H2S from the model
coefficient_H2S <- coef(lm_1)["H2S"]

# Define the change in H2S
change_in_H2S <- 0.01

# Calculate the expected change in taste
expected_change_in_taste <- coefficient_H2S * change_in_H2S

# Print the expected change in taste
print(expected_change_in_taste)
```

```
##         H2S
## 0.03911841
```

From the regression output provided in part (a), we see that the coefficient for H2S is 3.9118. This coefficient represents the change in the taste variable for a one-unit increase in H2S, holding other predictors constant.

Therefore, if H2S is increased by 0.01, we can expect the taste variable to change by $0.01 \times 3.9118 = 0.039118$.

So, the expected change in taste for a 0.01 increase in H2S is approximately 0.0391.

Question 2. For the prostate data in faraway library, fit a model with lpsa as the response and the other variables as predictors:

(a) Compute 90% and 95% CIs for the parameter associated with age. Using just these intervals, what could we have deduced about the p-value for age in the regression summary?

```
#Import library and dataset
library(faraway)
data(prostate)

#Fit the model
fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data=prostate)
summary(fit)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
#Confidence interval
CI_90 <- confint(fit, "age", level = 0.90)
CI_95 <- confint(fit, "age", level = 0.95)

CI_90
```

```
##            5 %         95 %
## age -0.0382102 -0.001064151
```

```
CI_95
```

```
##           2.5 %      97.5 %
## age -0.04184062 0.002566267
```

In the regression summary, when the 95% confidence interval for a coefficient includes zero, it indicates that the coefficient is not statistically significant at the 5% level, as the null hypothesis cannot be rejected. Conversely, when the 90% confidence interval excludes zero, it suggests that the coefficient is statistically significant at the 10% level, as the null hypothesis is rejected.

Therefore, while the 95% confidence interval including zero indicates non-significance at the conventional 5% level, the 90% confidence interval excluding zero suggests significance at the 10% level.

(b) Compute and display a 95% joint confidence region for the parameters associated with age and lbph. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

```r
#Import library
library(ellipse)
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

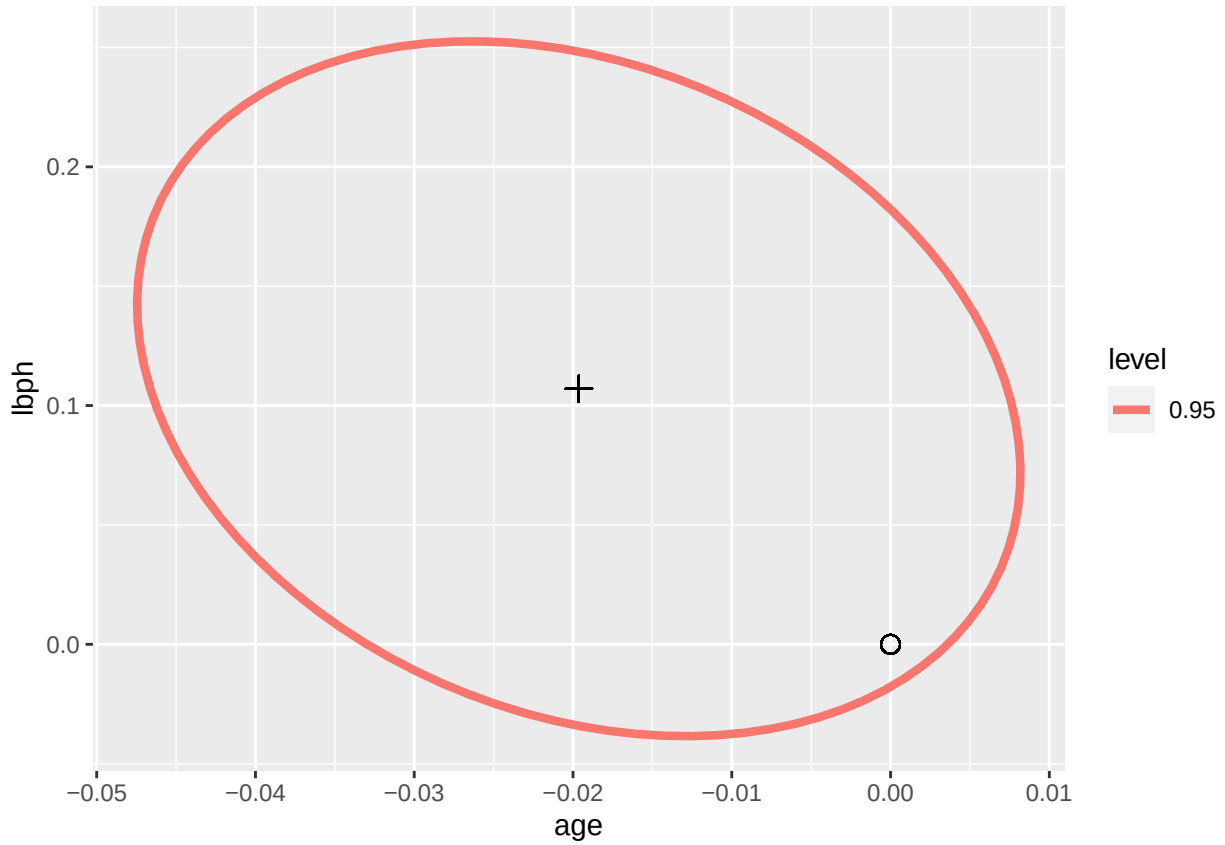
```r
library(ggplot2)

#Draw the graph
CR95 = ellipse(fit, c(4,5))

myCR = rbind(CR95);
myCR = data.frame(myCR);
names(myCR) = c("age","lbph");
myCR[, 'level']=as.factor(c(rep(0.95, dim(CR95)[1])));


ggplot(data=myCR, aes(x=age, y=lbph, colour=level)) +
  geom_path(aes(linetype=level), size=1.5) +
  geom_point(x=coef(fit)[4], y=coef(fit)[5], shape=3, size=3, colour='black') +
  geom_point(x=0, y=0, shape=1, size=3, colour='black')
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Since the point representing the estimated coefficients for age and lbph falls within the 95% joint confidence region, which is defined by the ellipse, we fail to reject the null hypothesis. This suggests that there is insufficient evidence to conclude that the coefficients for both variables are significantly different from zero.

(c) In the class, we discussed a permutation test corresponding to the F-test for the significance of a set of predictors. Execute the permutation test corresponding to the t-test for age in this model. (Hint: summary(g)$coef[4,3] gets you the t-statistic you need if the model is called g.)

```
n.iter = 2000;
tstats = numeric(n.iter);
tstat = summary(fit)$coef[4,3]
for(i in 1:n.iter){
   newprostate=prostate;
   newprostate$age=sample(prostate$age)
   ge = lm(lpsa ~., data=newprostate);
   tstats[i] = summary(ge)$coef[4,3]
}
#Estimated p-value
length(tstats[abs(tstats) > abs(tstat)])/n.iter
```

```
## [1] 0.0835
```

Upon conducting a permutation test corresponding to the t-test for age within this model, we obtained a p-value of approximately 0.08. This value closely aligns with the p-value obtained from the summary statistics of the regression model. Both results suggest a similar level of significance regarding the effect of age on the outcome variable.

(d) Remove all the predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

```
#Fit the model
fit1 <- lm(lpsa ~ lcavol + lweight + svi, data=prostate)
summary(fit1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
#Using annova to compare two models
anova(fit1,fit)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

Since the p-value (0.2167) exceeds the conventional significance level of 0.05, we fail to reject the null hypothesis. Consequently, we support the alternative model instead of the full model.

Question 3. Using the sat data from the faraway library.

(a) Fit a model with total sat score as the response and expend, ratio, and salary as predictors. Test the hypothesis that all of the betas = 0. Do any of these predictors have an effect on the response?

```
library(faraway)
data(sat)
#Fit the model
new_model <- lm(total ~ expend + ratio + salary, data = sat)
new_model
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Coefficients:
## (Intercept)       expend        ratio       salary
##    1069.234       16.469        6.330       -8.823
```

```
summary(new_model)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

The p-value associated with the F-test is 0.01209, indicating a statistically significant result. With a p-value below the conventional significance level of 0.05, we reject the null hypothesis that all regression coefficients are zero. This leads to the conclusion that at least one of the regression coefficients is not zero, suggesting that the predictors collectively have some explanatory power in relation to the dependent variable.

(b) Now add takers to the model. Test the hypothesis that beta_takers = 0. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

```r
#Fit in the model
new_model_1 <- lm(total ~ expend + ratio + salary + takers, data = sat)
summary(new_model_1)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

```r
#Use annova to compare to model
anova(new_model, new_model_1)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both the F-test and t-test yield extremely small p-values, approximately 2.607e-16, indicating a highly significant result.

From the t-test perspective, this small p-value suggests rejecting the null hypothesis, we favor the alternative hypothesis indicating that the takers should be in the model.

Similarly, from the F-test perspective, the p-value being less than 0.05 leads us to reject the null hypothesis. This implies rejecting the alternative model and favoring the full model, which includes all predictors.

In summary, both tests indicate strong evidence against the null hypothesis, reinforcing the decision to retain the full model with all predictors