HW7

Jinhong Du - 12243476

2019/10/08

Contents

1. (Faraway 13.1) Using the `teengamb` data, model `gamble` as the response and the other variables as predictors. Take care to investigate the possibility of interactions between `sex` and the other predictors. Interpret your final model.

First, we fit a full model with all 2-way interaction terms. As we can see, the t tests of coefficients of only two terms `income` and `sex:income` are significant. Then we try to remove terms not related to `sex:income`. Finally, we get a reduced model `gamble`$\sim$`sex*income`. And the ANOVA does not reject the null hypothesis. The final model means that for males,

$$\texttt{gamble} = -2.65963 + 6.51812 * \texttt{income}$$

while for females,

$$\begin{aligned}\texttt{gamble} &= (-2.65963 + 5.79960) + (6.51812 - 6.34320) * \texttt{income} \\ &= 3.13997 + 0.17492 * \texttt{income}\end{aligned}$$

the average difference of response for a unit change of `income` for males is much larger than that for females.

```
library(faraway)
data("teengamb")
full_model <- lm(gamble~.+sex*status+sex*income+sex*verbal, teengamb)
sumary(full_model)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  27.63540   17.62176  1.5683   0.12490
## sex         -33.01324   35.05297 -0.9418   0.35209
## status       -0.14557    0.33158 -0.4390   0.66308
## income        6.02908    1.05385  5.7210 1.264e-06
## verbal       -2.97476    2.42654 -1.2259   0.22758
## sex:status    0.35288    0.54923  0.6425   0.52431
## sex:income   -5.34777    2.42436 -2.2059   0.03335
## sex:verbal    2.83552    4.59730  0.6168   0.54097
##
## n = 47, p = 8, Residual SE = 20.97834, R-Squared = 0.62
```

```
reduced_model <- lm(gamble~sex*income, teengamb)
sumary(reduced_model)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.65963    6.31642 -0.4211  0.675804
## sex          5.79960   11.20025  0.5178  0.607245
## income       6.51812    0.98808  6.5967 4.951e-08
## sex:income  -6.34320    2.14456 -2.9578  0.005018
##
## n = 47, p = 4, Residual SE = 20.98167, R-Squared = 0.59
```

```
anova(reduced_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ sex * income
## Model 2: gamble ~ sex + status + income + verbal + sex * status + sex *
##     income + sex * verbal
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     43 18930
## 2     39 17164  4    1766.4 1.0034 0.4175
```

2. Consider a data set with a response Y and covariates A and B, all quantitative variables. Suppose that your data set has the following sample correlations:

```
> cor(cbind(A,B,Y))
       A      B      Y
A 1.0000 0.4395 0.3141
B 0.4395 1.0000 0.9587
Y 0.3141 0.9587 1.0000
```

(a) The correlations above are calculated on the data set where all values are observed. Now suppose that some fairly large fraction of the values of A are missing (you can assume they're missing completely at random), You impute these values using the mean of A. What will happen to the correlation values, if the correlation coefficients are now computed on the new data set, with the imputed values filled in? Specifically, will you expect to see an increase, a decrease, or no change, for each of the following values? Explain your answers briefly.

i. **The correlation of A with B?**

ii. **The correlation of A with Y ?**

iii. **The magnitude of the coefficient β_A, in the regression `lm(Y~A+B)`?**

The correlation of A with B will decrease in general. If the values of B in the missing data points are not the same, it will have low correlation with the imputed values of A, which is a constant. And therefore, it will reduce the correlation of A and B in the whole samples including imputed samples.

The correlation of A with Y will decrease because of the same reason as above.

The magnitude of the coefficient β_A will shrink toward zero since the mean fill-in method introduces bias and pull the regression line to get near to $y =$mean(A).

(b) Next, suppose that you instead impute the values of A by regressing onto B, i.e. `lm(A~B)`. Will you expect to see an increase, a decrease, or no change, for each of the following values? Explain your answers briefly.

i. **The correlation of A with B?**

ii. **The correlation of A with Y ?**

iii. **The magnitude of the coefficient β_A, in the regression `lm(Y~A+B)`?**

The correlation of A with B will increase since imputing by regression relies on the collinearity of A and B and it will incease collinearity since we add more imputed data points.

The correlation of A with Y will increase since B and Y have high correlation. By imputing by regression A on B, A get higher correlated to B and hence higher correlated to Y.

The magnitude of the coefficient β_A will shrink toward zero since the regression fill-in method introduces bais as well.

3. This problem will work with the `barley` dataset which you can download from the `lattice` library. This data set has $n = 120$ data points, each giving the crop `yield` of barley with covariates `variety` (10 types), `site` (6 types), and `year` (2 different years).

(a) How many degrees of freedom would be used by the model with all interactions, (i.e. the regression `yield`∼`variety * site * year`)? Would we be able to do significance testing on this model?

The numbers of dummy variables for `variety`, `site` and `year` are 9,5 and 1, respectively. The number of covariates (including the intercept term) for the model with all interactions is

$$p = 1 + (9 + 5 + 1) + (9 * 5 + 9 * 1 + 5 * 1) + (9 * 5 * 1) = 120.$$

So the degrees of freedom are $n - p = 0$.

(b) How many d.f. would be used by the model with all factors and two-way interactions, but not three-way interactions, (i.e. the regression `yield`∼`(variety+site+year) ** 2`)? For both this part and the part above, show your d.f. calculation by hand, not by running the model in R.

For the model without three-way interactions, the number of covariates is $p = 1+(9+5+1)+(9*5+9*1+5*1) = 75$ and the degrees of freedom are $n - p = 45$.

(c) From this point on, we will use the data set with data points 23 and 83 removed (these are identified as outliers, perhaps mistaken data entry, by your textbook), so your sample size is now 118. Run the model with all two-way interactions, `yield`∼`(variety+site+year) ** 2`. Use ANOVA to check each of the two-way interactions for significance at the 0.05 level, removing them one at a time if appropriate. Show your R code/output as you perform each step of this procedure.

Since the t test for the coefficient of `variety:year` is not significant, we remove it and refit the model. The ANOVA of the reduced model shows that all t tests are significant and so we can stop.

```
data(barley)
model <- lm(yield~(variety+year+site)**2,barley[-c(23,83),])
anova(model)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## variety       9 1029.6  114.40   9.8935 4.271e-08 ***
## year          1  912.1  912.10  78.8815 2.271e-11 ***
## site          5 6607.1 1321.43 114.2814 < 2.2e-16 ***
## variety:year  9  189.9   21.10   1.8244  0.090593 .
## variety:site 44 1161.8   26.40   2.2835  0.003615 **
## year:site     5 2164.7  432.94  37.4421 8.767e-15 ***
## Residuals    44  508.8   11.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model <- lm(yield~variety*site+year*site,barley[-c(23,83),])
anova(model)
```

```
## Analysis of Variance Table
##
## Response: yield
```

```
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## variety       9 1029.6  114.40   8.6716 7.427e-08 ***
## site          5 6607.1 1321.43 100.1663 < 2.2e-16 ***
## year          1  912.1  912.10  69.1387 3.525e-11 ***
## variety:site 44 1161.8   26.40   2.0015  0.008104 **
## site:year     5 2164.1  432.83  32.8090 4.377e-15 ***
## Residuals    53  699.2   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Pairwise comparisons. The `pulp` data set contains 20 data points, 5 each in groups A, B, C, D. The response `bright` is paper brightness and the covariate `operator` is some treatment applied during paper production, either A or B or C or D. We will investigate whether there are significant differences between any pair, e.g. brightness is significantly higher for production method A than for D, or statements of this type.

(a) Begin by calculating the sample mean of `bright` in each group, $\hat{\alpha}_A, \hat{\alpha}_B, \hat{\alpha}_C$, and $\hat{\alpha}_D$.

$\hat{\alpha}_A = 60.24, \hat{\alpha}_B = 60.06, \hat{\alpha}_C = 60.62, \hat{\alpha}_D = 60.68.$

```
data(pulp)
pulp$operator = as.factor(pulp$operator)
alpha_hat <- aggregate(pulp$bright, by=list(pulp$operator), FUN=mean)$x
cat(alpha_hat)
```

```
## 60.24 60.06 60.62 60.68
```

(b) Next calculate $\hat{\sigma}$, assuming that each observation is normally distributed as $Y_i \sim N(\alpha_{\ldots}, \sigma^2)$ (where $\ldots$ denotes the group, A or B or C or D, that data point i is assigned to)

$\hat{\sigma} = \frac{1}{n-L}\sum_{l=1}^{L}\sum_{i=1}^{n_l}(x_{li} - \overline{x}_l)^2 = 0.3259601$, where $L = 4$, x_{li} denote the ith sample in the lth group, $\overline{x}_l$ is the mean in the lth group and $n = \sum_{l=1}^{L} n_l$.

```
n <- dim(pulp)[1]
sigma_hat <- sqrt(sum(aggregate(pulp$bright, by=list(pulp$operator),
                                FUN=function(x)sum((x-mean(x))^2))$x)/(n-4))
cat(sigma_hat)
```

```
## 0.3259601
```

(c) Supposing that σ were known, what's the square root of the variance of $\hat{\alpha}_A - \hat{\alpha}_B$ in terms of σ? Now plug in $\hat{\sigma}$ in place of σ, this is now the standard error, SE($\hat{\alpha}_A - \hat{\alpha}_B$). Repeat for every possible pairwise comparison.

Let $n_1 = \ldots = n_4 = 5$ be the numebr of samples in group A,B,C,D respectively. Since $\hat{\alpha}_A \sim N(\alpha_A, \frac{\sigma^2}{n_1})$ and $\hat{\alpha}_B \sim N(\alpha_B, \frac{\sigma^2}{n_2})$ are independent, we have $\sqrt{Var(\hat{\alpha}_A - \hat{\alpha}_B)} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. So $SE(\hat{\alpha}_A - \hat{\alpha}_B) = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \hat{\sigma}\sqrt{\frac{2}{5}}$. Similarly, $SE(\hat{\alpha}_A - \hat{\alpha}_C) = SE(\hat{\alpha}_A - \hat{\alpha}_D) = SE(\hat{\alpha}_B - \hat{\alpha}_C) = SE(\hat{\alpha}_B - \hat{\alpha}_D) = SE(\hat{\alpha}_C - \hat{\alpha}_D) = \hat{\sigma}\sqrt{\frac{2}{5}}$.

```
aggregate(pulp$bright, by=list(pulp$operator), FUN=length)$x
```

```
## [1] 5 5 5 5
```

```
L <- 4
n_i <- 5
J <- c(rep(n_i,L)) # number of samples in each group

# Pairwise di erences & SE
diffmat = outer(alpha_hat,alpha_hat,'-')
SEdiffmat = sigma_hat * sqrt(outer(1/J,1/J,'+'))
```

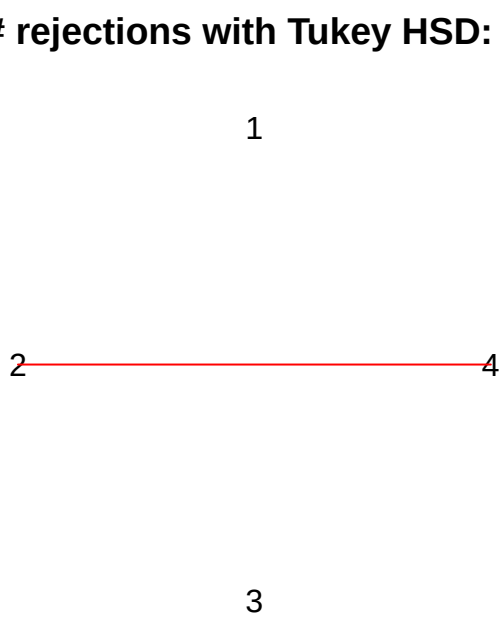
(d) Finally, calculate the Tukey honest significant difference (Tukey HSD) confidence interval for each possible pairwise comparison. At the 0.05 level, what conclusions can you draw about the four production methods?

As we can see from the plot below, at the level 0.05, only the p-value for testing $\alpha_B = \alpha_D$ or not is significant. So we can conclude that the hypothesis $\alpha_B = \alpha_D$ is rejected while we cannot reject hypotheses for other pairwise tests.

```
Tstat = diffmat/SEdiffmat
pvals = 2*(1-pt(abs(Tstat),n-L))

Tukey_stat = abs(Tstat)*sqrt(2)
pvals_Tukey = 1-ptukey(Tukey_stat,L,n-L)

plot(-1:1,-1:1,type='n',axes=FALSE,xlab='',ylab='',asp=1)
title(main=paste0('# rejections with Tukey HSD: ',sum(pvals_Tukey[upper.tri(pvals_Tukey)]<=0.05)))
for(i in 1:L){
  theta = i/L * 2*pi
  text(cos(theta),sin(theta),toString(i))
}
for(i in 1:(L-1)){
  for(j in (i+1):L){
    theta_i = i/L * 2*pi ; theta_j = j/L * 2*pi
    if(pvals_Tukey[i,j]<=0.05){
      segments(cos(theta_i),sin(theta_i),cos(theta_j),sin(theta_j),col='red')
    }
  }
}
```



5. Suppose that combinations of three drugs, called A and B and C, are being examined for their ability to lower blood pressure. Suppose that, without any medication, expected systolic blood pressure (the response Y) in the population being studied is 150. Any one drug on its own has no effect on blood pressure. However, drug A in combination with B or C will reduce blood pressure to 140. Drugs B and C are chemically very similar and it doesn't matter which one is used in combination with drug A. There's no benefit to using both—it's equivalent to just using one.
Now suppose we want to write down a linear model, using treatment coding, to describe this scenario. What are the values of all the coefficients in the model:.

- **The intercept term** β_0
- **The one-way terms** β_{A1}, β_{B1}, β_{C1}
- **If needed, the two-way interaction terms** $\beta_{A1:B1}$, $\beta_{A1:C1}$, $\beta_{B1:C1}$
- **If needed, the three-way interaction term** $\beta_{A1:B1:C1}$

$$Y = \beta_0 + \beta_{A1} + \beta_{B1} + \beta_{C1} + \beta_{A1:B1} + \beta_{A1:C1} + \beta_{B1:C1} + \beta_{A1:B1:C1}$$

Since without any medication, expected systolic blood pressure (the response Y) in the population being studied is 150, we have $\beta_0 = 150$.

Since any one drug on its own has no effect on blood pressure, we have $\beta_{A1} = \beta_{B1} = \beta_{C1} = 0$.

Since drug A in combination with B or C will reduce blood pressure to 140, we have $\beta_{A1:B1} = \beta_{A1:C1} = -10$.

Since drugs B and C are chemically very similar when used in combination with drug A, and there's no benefit to using both B and C, we have $\beta_{B1:C1} = 0$.

Since there is no evidence that using all A, B and C has extra effect than just using A and B or A and C, we have $\beta_{A1:B1:C1} = 10$ to cancel out either $\beta_{A1:B1}$ or $\beta_{A1:C1}$.