Chimdi_Homework_5

Chimdi Chikezie

5/23/2023

```
library(Sleuth3)
library(Sleuth3)
library(ggplot2)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
library(graphics)
```

Problem 1: Data taken from Exercise 13.19 Nature-Nurture. A 1989 study investigated the effect of heredity and environment on intelligence. From adoption registers in France, researchers selected samples of adopted children whose biological parents and adoptive parents came from either the very highest or the very lowest socio-economic status (SES) categories (based on years of education and occupation). The 38 selected children were given intelligence quotient (IQ) tests. The scores as well as the SES of biological and adoptive parents are reported in Display 13.24 of the Sleuth book. (Data from C. Capron and M. Duyme, "Children's IQs and SES of Biological and Adoptive Parents in a Balanced Cross-fostering Study," European Bulletin of Cognitive Psychology 11(3) (1991): 323-48.) The data frame consists on 38 observations on the following 3 variables: SES of adoptive parents: a factor with two levels "High" and "Low" SES of biological parents: a factor with two levels "High" and "Low" IQ Score: IQ score of children Each question carries four marks

a) Produce a two-way table of sample averages, along with row and column averages. Does it seem from the table that the difference in SES of biological parents affect the IQ of children?

```
?ex1319
#head(ex1319)
data1 <- ex1319
#View(data1)
```

```
table <- aggregate(IQ ~ Adoptive + Biological, data = data1, FUN = length)
xtabs(IQ ~ Adoptive + Biological, table)
```

```
##          Biological
## Adoptive High Low
##     High   10  10
##     Low     8  10
```

The data is unbalanced, so we cannot take the mean directly using the "xtabs" and "addmargins" functions.

```
aggregate(IQ ~ Adoptive, data = data1, FUN = mean)
```

```
##   Adoptive        IQ
## 1     High 111.60000
## 2      Low  99.11111
```

```
aggregate(IQ ~ Biological, data = data1, FUN = mean)
```

```
##   Biological       IQ
## 1       High 114.2222
## 2        Low  98.0000
```

From this 2-way mean tables, it appears that children with high-SES biological parents have higher average IQ scores (114.2222) than those with low-SES biological parents (98.0000) and same applies to children with adopted parents. But to know whether this difference is statistically significant, we would need to perform a statistical test using the raw data.

(b) Fit the additive model for IQ scores of children vs SES of adoptive and biological parents. Based on the model fit, which factors seem to have a significant effect on the scores?

```
fit_add <- lm(IQ ~ Adoptive + Biological, data = data1)
summary(fit_add)
```

```
##
## Call:
## lm(formula = IQ ~ Adoptive + Biological, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.188  -9.621  -1.788   7.973  23.812
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    119.388      3.603  33.132  < 2e-16 ***
## AdoptiveLow    -11.624      4.240  -2.741 0.009574 **
## BiologicalLow  -15.576      4.240  -3.674 0.000793 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.03 on 35 degrees of freedom
## Multiple R-squared:  0.3881, Adjusted R-squared:  0.3531
## F-statistic:  11.1 on 2 and 35 DF,  p-value: 0.000185
```

```
anova(fit_add)
```

```
## Analysis of Variance Table
##
## Response: IQ
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Adoptive    1 1477.6  1477.6   8.702 0.0056350 **
## Biological  1 2291.5  2291.5  13.495 0.0007935 ***
## Residuals  35 5943.1   169.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both SES of adoptive parents and of biological parents have a significant effect on intelligence.

β_1 measures the difference in mean between AdoptiveLow and Adoptivehigh. Since the p-value (0.009574) associated with β_1 is less than $\alpha = 0.05$, it means that there is difference in the mean of AdoptiveLow and Adoptivehigh. Therefore, we conclude that having adopted parents in either highest or the very lowest socio-economic status (SES) impacts a child's IQ level.

β_2 measures the difference in mean between BiologicalLow and Biologicalhigh. Since the p-value (0.000793) associated with β_2 is less than $\alpha = 0.05$, it means that there is difference in the mean of BiologicalLow and

Biologicalhigh. Therefore, we conclude that having biological parents both either highest or the very lowest socio-economic status (SES) impacts a child's IQ level.

c) Run an extra sum of squares F-test to compare the saturated model to the additive model. Is there evidence against the simpler additive model?

```
fit_sat <- lm(IQ ~ Adoptive * Biological, data = data1)
anova(fit_add, fit_sat)
```

```
## Analysis of Variance Table
##
## Model 1: IQ ~ Adoptive + Biological
## Model 2: IQ ~ Adoptive * Biological
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     35 5943.1
## 2     34 5941.2  1    1.9059 0.0109 0.9174
```

H_0 : Model 1 is significant H_A : Model 2 is significant

The F value from the anova comparison is 0.0109 and the p-value $= 0.9174$ is greater than $\alpha = 0.05$. This means that we will fail to reject the null hypothesis that Model 1 is significant and conclude that the combined effect of heredity and environment on intelligence does not significantly differ from what we would expect if these two factors acted independently. The main effects of heredity and environment still remain to be significant contributors to the variance in intelligence, indicating that both factors independently have an impact on intelligence.

(d) Using the additive model, produce a plot of the estimated mean responses for different SES of biological and adoptive parents (like the treatment curves shown in class). Also using the saturated model, produce a plot of the estimated mean responses for different SES of biological and adoptive parents. What conclusions can you draw from the two plots?

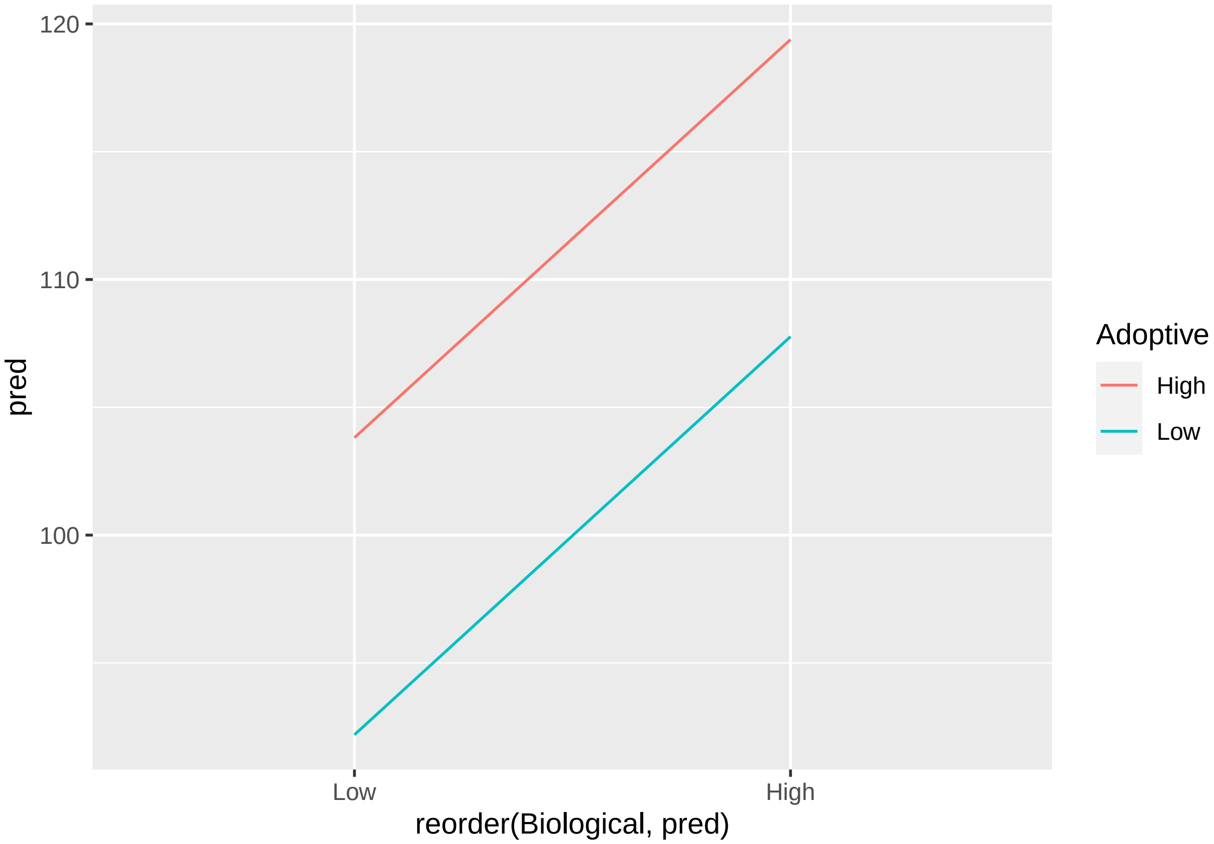
```
new_data <- expand.grid(
  Biological = unique(data1$Biological),
  Adoptive = unique(data1$Adoptive))
#View(new_data)
new_data$pred <- predict(fit_add, newdata = new_data)
```

```
round(xtabs(pred ~ Biological + Adoptive, new_data), 2) # a table
```

```
##           Adoptive
## Biological   High    Low
##       High 119.39 107.76
##       Low  103.81  92.19
```

```
qplot(reorder(Biological, pred), pred, data = new_data,
colour = Adoptive, geom = "line", group = Adoptive)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

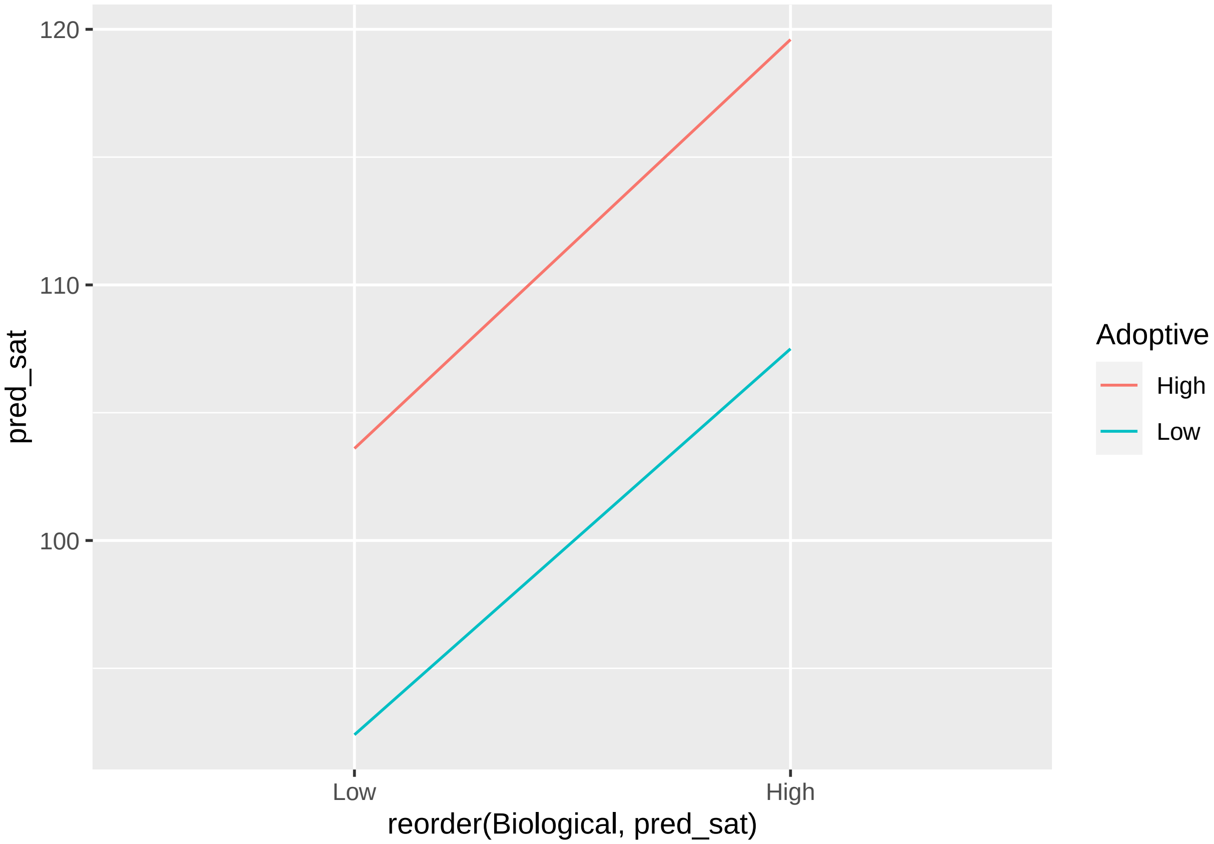


```
new_data$pred_sat <- predict(fit_sat, newdata = new_data)
```

```
round(xtabs(pred_sat ~ Biological + Adoptive, new_data), 2) # a table
```

```
##             Adoptive
## Biological   High   Low
##        High 119.6 107.5
##        Low  103.6  92.4
```

```
qplot(reorder(Biological, pred_sat), pred_sat, data = new_data,
colour = Adoptive, geom = "line", group = Adoptive)
```



The plots from the additive and saturated models look the same. This means that the interaction between the explanatory variables do not have an effect on IQ scores. Therefore, we use the additive model.

Problem 2: Clever Hans Effect. Read the description of the experiment and dataset from Problem 19 of Chapter 14 of the Sleuth book. In the dataset, consider the overall success rate of each mouse (Success) as the response variable. Also, consider the assigned block of the students (Block) as one explanatory variable, and the deceitful labeling of the mouses (Treatment) as the other explanatory variable. For now, leave rest of the variables alone. Each question carries four marks.

(a) Fit the additive model for Success vs Block and Treatment. Based on the model fit, does the deceitful labeling of the mouses (Treatment) seems to have an effect of the success rate of the mouse?

```
?ex1419
 #head(ex1419)
 data2 <- ex1419
 View(data2)
```

```
data2$fac.Block <- factor(data2$Block)

fit_add2 <- lm(Success ~ fac.Block + Treatment, data = data2)
summary(fit_add2)
```

```
##
## Call:
## lm(formula = Success ~ fac.Block + Treatment, data = data2)
##
## Residuals:
##      1      2      3      4      5      6      7      8      9     10     11
## -0.035  0.035 -0.080  0.080  0.040 -0.040  0.035 -0.035  0.020 -0.020 -0.020
##     12
##  0.020
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.41500    0.05123   8.100 0.000465 ***
## fac.Block2     0.02500    0.06708   0.373 0.724664
## fac.Block3     0.07500    0.06708   1.118 0.314373
## fac.Block4    -0.01000    0.06708  -0.149 0.887323
## fac.Block5     0.08500    0.06708   1.267 0.260928
## fac.Block6     0.04500    0.06708   0.671 0.532070
## Treatmentdull -0.16000    0.03873  -4.131 0.009075 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06708 on 5 degrees of freedom
## Multiple R-squared:  0.8033, Adjusted R-squared:  0.5672
## F-statistic: 3.402 on 6 and 5 DF,  p-value: 0.1001
```

$H_0 : \beta_6 = 0$ $H_A : \beta_6 \neq 0$

Thee parameter representing treatment is β_6 which has a t-value = -4.131 and p-value = 0.009075 which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis that $\beta_6 = 0$ and conclude that the deceitful labeling of the mouses (Treatment) has an effect of the success rate of the mouse.

(b) Can you fit a saturated model for Success vs Block and Treatment in this dataset? Explain. If you replace the explanatory variable Block with the numerical variable of amount of prior experience of the students (PriorExp, a numerical variable), can you fit a model with all possible interactions for Success vs PriorExp and Treatment for this dataset? If yes, based on the model fit, does the interaction effect seem to be significant? Note: The data set can be read in R using the following commands

```r
data2$fac.Block <- factor(data2$Block)

fit_sat2 <- lm(Success ~ fac.Block * Treatment, data = data2)
summary(fit_sat2)
```

```
##
## Call:
## lm(formula = Success ~ fac.Block * Treatment, data = data2)
##
## Residuals:
## ALL 12 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.45        NaN     NaN      NaN
## fac.Block2                   -0.09        NaN     NaN      NaN
## fac.Block3                    0.08        NaN     NaN      NaN
## fac.Block4                   -0.08        NaN     NaN      NaN
## fac.Block5                    0.07        NaN     NaN      NaN
## fac.Block6                    0.03        NaN     NaN      NaN
## Treatmentdull                -0.23        NaN     NaN      NaN
## fac.Block2:Treatmentdull      0.23        NaN     NaN      NaN
## fac.Block3:Treatmentdull     -0.01        NaN     NaN      NaN
## fac.Block4:Treatmentdull      0.14        NaN     NaN      NaN
## fac.Block5:Treatmentdull      0.03        NaN     NaN      NaN
## fac.Block6:Treatmentdull      0.03        NaN     NaN      NaN
```

```
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 11 and 0 DF,  p-value: NA
```

We cannot fit a saturated model for Success vs Block and Treatment in this dataset because the model is over parameterised, this means that it has too many parameters relative to the amount of data available.

```
fit_sat3 <- lm(Success ~ PriorExp * Treatment, data = data2)
summary(fit_sat3)
```

```
##
## Call:
## lm(formula = Success ~ PriorExp * Treatment, data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.090772 -0.034561 -0.006667  0.040253  0.074362
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.4458723  0.0258916  17.221 1.32e-07 ***
## PriorExp                0.0049666  0.0046882   1.059  0.32036
## Treatmentdull          -0.1642538  0.0382790  -4.291  0.00265 **
## PriorExp:Treatmentdull -0.0009473  0.0068331  -0.139  0.89316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06199 on 8 degrees of freedom
## Multiple R-squared:  0.7312, Adjusted R-squared:  0.6304
## F-statistic: 7.254 on 3 and 8 DF,  p-value: 0.01138
```

Yes, we can fit a model with all possible interactions for Success vs PriorExp and Treatment for this dataset because Block has been replaced with PriorExp (numerical), thereby reducing the number of parameters in the model.

From the p-value (0.89316) of β_3 (the parameter that measures the significance of the interaction term) is less than $\alpha = 0.05$. We conclude that the interaction interactions between PriorExp and Treatment does not have an effect on Success.