Sample Answer Lab03: Multivariate Regression

**Handed out:** Thursday, September 26, 2019

**Return date:** Thursday, October 10, 2019 as hardcopy.

**Grading:** This lab counts 10 % towards your final grade

# Part I: Partial Regression Coefficients (4 points)

Task 1: Generate a scatterplot matrix of the three variables **sat**, **expend** and **pctsat**. Interpret their pairwise relationships. (0.5 points)

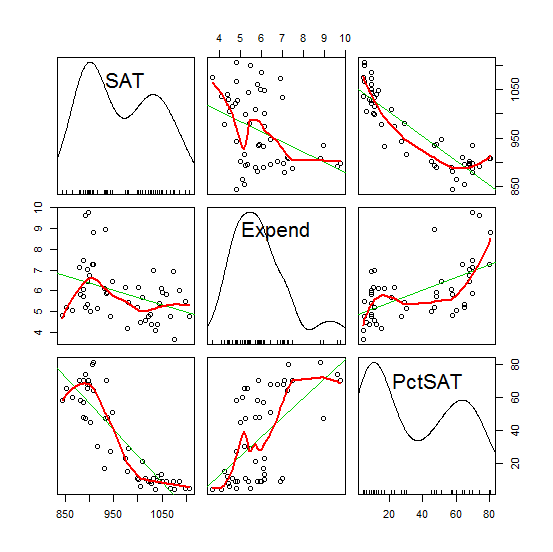
> library(foreign)

> library(car)

> School <- read.spss("StateSchool.sav", to.data.frame=TRUE)

> scatterplotMatrix(~SAT + Expend + PctSAT, reg.line=lm, smooth=TRUE,

+ spread=FALSE, span=0.5, id.n=0, diagonal = 'density', data=School)



Comments: The density plots shows all three variables (SAT, Expend and PctSAT) have bimodal distributions.

**Pairwise Relationships:**

**SAT – Expend:** Moderate negative relationship. When the average state’s expenditure increases, the average state-wide SAT score decreases. This disobeys our common sense that academic performance should increase with increasing investments into education.

**SAT – PctSAT:** Strong negative relationship. If the state-wide participation rate of students in the SAT test increases, overall more students will participate in the SAT test. If the participation would be voluntary only better students who plan to go to university would take the test. However, the excellent students are only in a small proportion of the overall student body. The base number of students increases far more rapidly than the number of excellent students so that the average state-wide SAT score decreases.

**Expend – PctSAT:** Moderate positive relationship. The state-wide SAT participation rate increases as an effect of an increased state’s education expenditure. It seems as states invest more into education they mandate that their students take the SAT test

*Note: for immediate comparison purposes please make sure that the dependent variable “SAT” is plotted in the upper left corner.*

Task 2: Formulate ***explicit hypotheses*** about the ***direction*** in which the two independent variables **expend** and **pctsat** may influence the SAT performance **sat**. (0.5 points)

Comments:

[a] Hypothesis: SAT score versus Expenditure:

When average state’s educational expenditure increases, the average state-wide SAT score should increase because the more the state spends on education the better trained its students will become. “Qualification” is measured here by the SAT score.

[b] Hypothesis: SAT score versus Participation Rate

As the state-wide participation rate of students in the SAT test increases, the average state-wide SAT scores decreases. In states with a low participation rate only the good students, who are interested in perusing a university degree take the SAT test. This is known in the literature as *selection bias*. In contrast, in states with a mandatory participation in the SAT test, academically stronger and weaker ones are taking the test. This will drag the state’s overall score down.

Task 3: Evaluate each hypothesis individually with a bivariate regression models (0.5 points)  
 [a] and   
 [b] .   
Discuss the results in terms of the signs of the estimated regression coefficients and .

> lm1 <- lm(SAT ~ Expend, data = School)

> summary(lm1)

**Model a:**

Call:lm(formula = SAT ~ Expend, data = School)

Residuals:

Min 1Q Median 3Q Max

-145.074 -46.821 4.087 40.034 128.489

Coefficients:

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1089.294 44.390 24.539 < 2e-16 \*\*\***

**Expend -20.892 7.328 -2.851 0.00641 \*\***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.91 on 48 degrees of freedom

**Multiple R-squared: 0.1448, Adjusted R-squared: 0.127**

F-statistic: 8.128 on 1 and 48 DF, p-value: 0.006408

> lm2 <- lm(SAT ~ PctSAT, data = School)

> summary(lm2)

**Model b:**

Call:lm(formula = SAT ~ PctSAT, data = School)

Residuals:

Min 1Q Median 3Q Max

-79.158 -27.364 3.308 19.876 66.080

Coefficients:

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1053.3204 8.2112 128.28 <2e-16 \*\*\***

**PctSAT -2.4801 0.1862 -13.32 <2e-16 \*\*\***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.89 on 48 degrees of freedom

**Multiple R-squared: 0.787, Adjusted R-squared: 0.7825**

F-statistic: 177.3 on 1 and 48 DF, p-value: < 2.2e-16

Comment: The intercepts and slopes of both models are significant. The regression coefficient of “Expend” is -20.892 and that of “PctSAT” is -2.4801. Both of them are negative.   
[a] In other words, when “Expend” increases by $1000, “SAT” decreases 20.892 point. This is a ***counterintuitive finding*** for the impact of expenditure on the educational outcome. If this would be the true effect then an investment into education in fact would leave the students less educated. The is relatively low, which indicates that expenditure, while significant, does not have a strong explanatory power.

[b] When “PctSAT” increases 1%, “SAT” decreases 2.4801 point. This is in line with our common-sense expectations. The participation rate is highly significant and the

The indicates a very good model fit.

Task 4: Perform a multiple regression analysis with the two independent variables   
 [c] and ***fully interpret*** the results. (1.0 point)  
Pay in particular attention to any changes in the estimated slope parameters and of the bivariate models and the partial slope parameters and of the multiple regression model.   
If a regression coefficient changes substantially why may this be the case?

> lm3 <- lm(SAT ~ Expend + PctSAT, data = School)

> summary(lm3)

**Model c:**

Call:lm(formula = SAT ~ Expend + PctSAT, data = School)

Residuals:

Min 1Q Median 3Q Max

-88.400 -22.884 1.968 19.142 68.755

Coefficients:

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 993.8317 21.8332 45.519 < 2e-16 \*\*\***

**Expend 12.2865 4.2243 2.909 0.00553 \*\***

**PctSAT -2.8509 0.2151 -13.253 < 2e-16 \*\*\***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.46 on 47 degrees of freedom

**Multiple R-squared: 0.8195, Adjusted R-squared: 0.8118**

F-statistic: 106.7 on 2 and 47 DF, p-value: < 2.2e-16

Comment:

[a] The coefficient of “Expend” changes its sign and becomes positive. The reason is that effect of “Expend” is confounded by “PctSAT” (i.e., they are positively correlated). After controlling for the effect of “PctSAT” the partial expenditure effect indicates as expected that an investment into education overall has a positive effect on the educational outcome of the students.

[b] The absolute value of regression coefficient of “PctSAT” in multivariate model became larger than in the model bivariate model. It now measures the sole effect of the state-wide SAT participation rate without interference from “Expend”.

*Both coefficients are highly significant and are now unbiased. Overall the model explains approximately 81% of the total variation in the SAT score performance.*

Task 5: Calculate the squared correlation between predicted value of multiple model [c] the observed value . Which statistic of the multiple regression model output from [c] is equivalent to the calculated ***squared*** correlation coefficient? (0.5 points)

**Tips:**

1. **Use cor()^2 to calculate the squared correlation.**

> preSAT <- predict(lm3)

> cor(preSAT, School$SAT)^2

[1] **0.8194726**

Comment: the squared correlation coefficient is equivalent to the multiple in model (c), which means how close the predicted values are to the observed values. In other words, it measures the proportion of variability in the dependent variable that can be explained by the independent variables.

Task 6: Plot the regression residuals from the multiple model [c] against the independent variable **pctsat** (placed on the *x*-axis). (1.0 points)

> scatterplot(School$PctSAT, residuals(lm3), xlab = c("PctSAT"), ylab = c("Residuals"))

> lm4 <- **lm(SAT ~ Expend + PctSAT + I(PctSAT^2), data = School)**

> scatterplot(School$PctSAT, residuals(lm4), xlab = c("PctSAT"), ylab = c("Residuals of Quadratic Model"))

|  |  |
| --- | --- |
| Figure 1: Residual plot of lm(SAT ~ Expend + PctSAT ) . | Figure 2: Residual plot of lm(SAT ~ Expend + PctSAT + I(PctSAT^2)) |

Comment: OLS guarantees that the residuals of the exogenous variable and the endogenous variables are linearly independent. While both variables are linearly uncorrelated, a clearly **quadratic** relationship between the residuals and **pctsat** is visible (see Figure 1). While the linear relationship component is captured by **pctsat**, the U-shaped pattern between residuals and **pctsat** in the scatterplot suggests a better fit with a non-linear model. After we use the quadratic function, the regression residuals and **pctsat** exhibit no longer a systematic pattern (see Figure 2).

This suggested that we have a model misspecification and the proper model should be   
**lm(SAT ~ Expend + PctSAT + I(PctSAT^2), data = School)**

## Part II: A Multiple Regression Model with Factors and Partial *F*-test (4 points)

You will find the data needed for this part in the dBase data file **provinces.dbf** (read this file into  with the function **foreign::read.dbf**)

Task 7: Use common sense arguments ***how*** the four metric variables will influence the provincial fertility rates. Use one or two sentences per explanation, and formulate one or two-sided null and alternative hypotheses based on your explanation. Format everything in at table. (0.5 points)

|  |  |  |
| --- | --- | --- |
| Variable | Common Sense Arguments | Statistical Hypotheses |
| ILLITERRAT | A higher illiteracy rate leads to higher fertility rate due to lack of education. |  |
| FEMMARAGE | The latter a woman marries the lower will be her likelihood to have many children. |  |
| DIVORCERAT | A higher divorce rate leads to lower chance of having many children. |  |
| TELEPERFAM | An increased number of televisions will lead to more distractions and a decreased fertility rate. |  |

Task 8: Generate a scatterplot matrix showing the dependent variable and the four metric independent variables. *Briefly interpret the scatterplot matrix.* (0.5 points)

|  |
| --- |
|  |

Comments:

[a] Distributional characteristics: The distributions of the dependent variable and the four independent variables are unimodal. **TOTFERTRAT**, **DIVORCERAT**, and **ILLITERRAT** are positively skewed, and **FEMMARAGE** and **TELEPERFAM** are negatively skewed.

[b] Y-X relationships: **FEMMARAGE**, **DIVORCERAT**, and **TELEPERFAM** have strong negative effects on **TOTFERTRAT**. However, **ILLITERRAT** has a positive relationship with **TOTFERTRAT**.

[c] Positive X-X relationships: **FEMMARAGE- DIVORCERAT**, **FEMMARAGE- TELEPERFAM**, and **DIVORCERAT- TELEPERFAM** have positive relationships.

[d] Negative X-X relationships: **FEMMARAGE- ILLITERRAT**, **DIVORCERAT- ILLITERRAT**, and **ILLITERRAT- TELEPERFAM** have negative relationships.

Task 9: Run a base model multiple regression with the four metric variables to explain the variation of the fertility rates. Interpret this model [a] in the light of your earlier stated hypotheses in task 7, [b] the significances of the estimate regression coefficients and [c] the goodness of fit. (0.5 points)

Call:

lm(formula = TOTFERTRAT ~ FEMMARAGE + DIVORCERAT + ILLITERRAT +

TELEPERFAM, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-0.21906 -0.06267 -0.00966 0.05425 0.41272

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.496337 0.513726 8.752 1.13e-13 \*\*\*

FEMMARAGE -0.088837 0.020771 -4.277 4.71e-05 \*\*\*

DIVORCERAT -0.112265 0.055648 -2.017 0.0466 \*

ILLITERRAT 0.020377 0.008735 2.333 0.0219 \*

TELEPERFAM -1.226364 0.183037 -6.700 1.76e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1035 on 90 degrees of freedom

Multiple R-squared: 0.8096, Adjusted R-squared: **0.8012**

F-statistic: 95.69 on 4 and 90 DF, p-value: < 2.2e-16

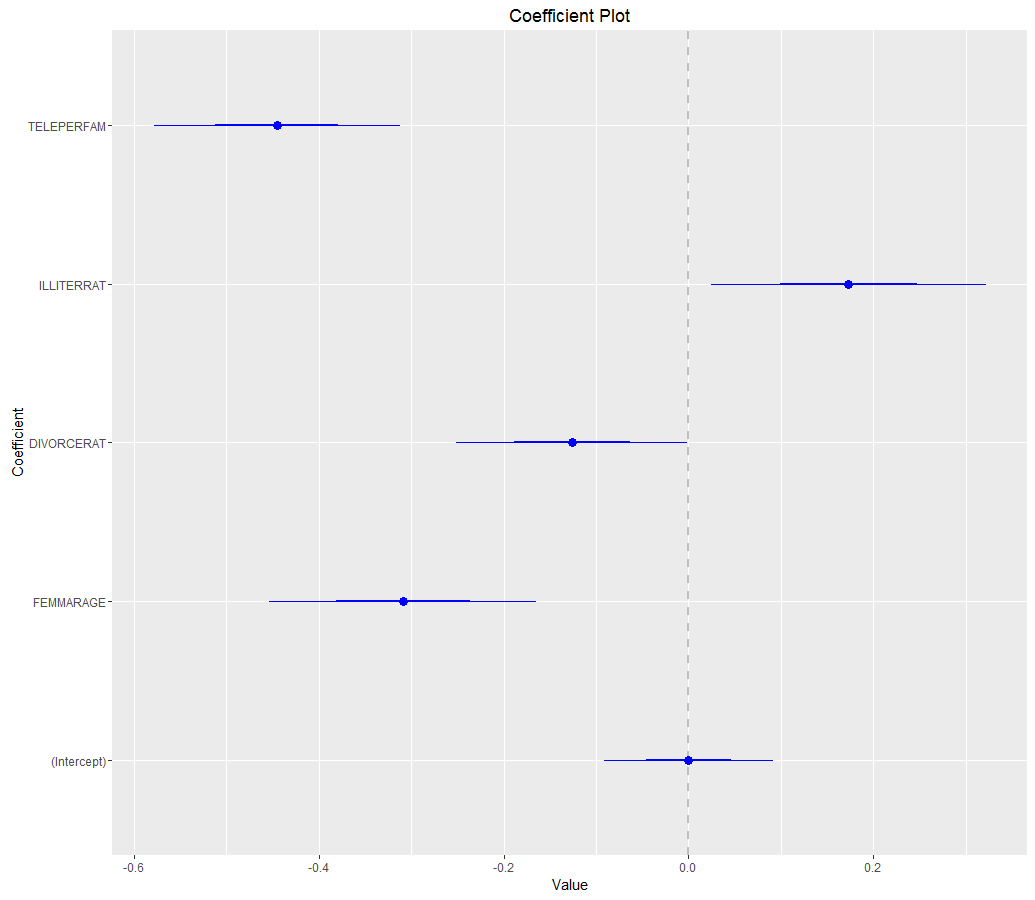
Comment: All independent variables exhibit a relationship with the dependent variable as stated by the one-sided alternative hypotheses in task 2.1. All regression coefficients are significantly different from zero at an error probability of . Since the reported error probabilities are associated with two-sided tests; for one-sided tests they need to be divided by 2. The overall goodness of fit of this model is high ().

Task 10: Calculate the *beta-coefficients* for the multiple model in task 9. Rank the independent variables *according to the absolute strength* *of their effects* on the fertility rates and plot the beta coefficients with the **coefplot( )** function. (1 point)

> Italy.scale <- as.data.frame(scale(Italy[ ,13:17]))

> lm.scale <- lm(TOTFERTRAT~FEMMARAGE+DIVORCERAT+ILLITERRAT+TELEPERFAM, data=Italy.scale)

> coefplot::coefplot(lm.scale)



Comment: The influence strengths of the independent variables on the variation of the dependent variable are: **DIVORCERAT** < **ILLITERRAT** < **FEMMARAGE** < **TELEPERFAM**.

Task 11: Run five separate regressions on the [a] independent variables and [b] the fertility rates using the factor **REGION** as independent variable.   
Does the **REGION** factor *explain the variation* of the four independent variables as well as the fertility rates, i.e., is this factor highly correlated with other variables? (0.5 points)

> lm2 <- lm(cbind(TOTFERTRAT,ILLITERRAT,FEMMARAGE,DIVORCERAT,TELEPERFAM)~REGION, data=Italy)

> summary(lm2)

**[1]** Response **TOTFERTRAT** :

Call:

lm(formula = TOTFERTRAT ~ REGION, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-0.23300 -0.09275 -0.01300 0.07167 0.36333

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.02300 0.02205 46.405 <2e-16 \*\*\*

REGIONNorth 0.03367 0.03118 1.080 0.283

REGIONSardinia 0.09950 0.06427 1.548 0.125

REGIONSicily 0.53700 0.04589 11.702 <2e-16 \*\*\*

REGIONSouth 0.39473 0.03389 11.647 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1207 on 90 degrees of freedom

Multiple R-squared: **0.7409**, Adjusted R-squared: 0.7294

F-statistic: 64.34 on 4 and 90 DF, p-value: < 2.2e-16

**[2]** Response **ILLITERRAT** :

Call:

lm(formula = ILLITERRAT ~ REGION, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-2.82455 -0.50394 -0.07267 0.34756 2.83545

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.4157 0.1844 7.676 1.89e-11 \*\*\*

REGIONNorth -0.7630 0.2608 -2.925 0.00435 \*\*

REGIONSardinia 1.8068 0.5377 3.360 0.00114 \*\*

REGIONSicily 3.2466 0.3839 8.457 4.64e-13 \*\*\*

REGIONSouth 3.1689 0.2835 11.176 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.01 on 90 degrees of freedom

Multiple R-squared: **0.7485**, Adjusted R-squared: 0.7373

F-statistic: 66.97 on 4 and 90 DF, p-value: < 2.2e-16

**[3]** Response **FEMMARAGE** :

Call:

lm(formula = FEMMARAGE ~ REGION, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-0.96636 -0.31017 -0.04033 0.29057 1.21000

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 27.25033 0.09179 296.888 < 2e-16 \*\*\*

REGIONNorth -0.27333 0.12981 -2.106 0.038 \*

REGIONSardinia 0.18217 0.26760 0.681 0.498

REGIONSicily -1.88033 0.19107 -9.841 6.11e-16 \*\*\*

REGIONSouth -1.19397 0.14111 -8.461 4.54e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5027 on 90 degrees of freedom

Multiple R-squared: **0.6288**, Adjusted R-squared: 0.6124

F-statistic: 38.12 on 4 and 90 DF, p-value: < 2.2e-16

**[4]** Response **DIVORCERAT** :

Call:

lm(formula = DIVORCERAT ~ REGION, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-0.51233 -0.10767 -0.01267 0.13591 0.46767

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.59233 0.03703 15.994 < 2e-16 \*\*\*

REGIONNorth -0.04967 0.05237 -0.948 0.346

REGIONSardinia -0.36733 0.10797 -3.402 0.001 \*\*\*

REGIONSicily -0.34789 0.07709 -4.513 1.93e-05 \*\*\*

REGIONSouth -0.38324 0.05694 -6.731 1.53e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2028 on 90 degrees of freedom

Multiple R-squared: **0.423**, Adjusted R-squared: 0.3973

F-statistic: 16.49 on 4 and 90 DF, p-value: 3.596e-10

**[5]** Response **TELEPERFAM** :

Call:

lm(formula = TELEPERFAM ~ REGION, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-0.247360 -0.025351 0.008611 0.033786 0.106420

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.812704 0.010775 75.422 < 2e-16 \*\*\*

REGIONNorth -0.006805 0.015239 -0.447 0.656

REGIONSardinia -0.049512 0.031416 -1.576 0.119

REGIONSicily -0.180715 0.022431 -8.057 3.12e-12 \*\*\*

REGIONSouth -0.106844 0.016566 -6.449 5.48e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05902 on 90 degrees of freedom

Multiple R-squared: **0.5306**, Adjusted R-squared: 0.5098

F-statistic: 25.43 on 4 and 90 DF, p-value: 4.118e-14

Comment: **Region** is a significant variable in all five models. It can explain the 42.3% variation of **DIVORCERAT**, 74.85% variation of **ILLITERRAT**, 62.88% variation of **FEMMARAGE**, 53.06% variation of **TELEPERFAM**, and74.09% variation of **TOTFERTRAT**. In other words, **Region** is highly correlated with the dependent variable and all four independent variables.

Task 12: Run the multiple regression model with the four metric variables plus the **REGION** factor to explain the variation of the fertility rates.   
*Speculate* in an informed way by looking at the relationships between the **REGION** factor and other metric variables, why some independent metric variables are no longer significant? (0.5 points)

> lm3 <- lm(TOTFERTRAT~FEMMARAGE+DIVORCERAT+ILLITERRAT+TELEPERFAM+REGION, data=Italy)

> summary(lm3)

Call:

lm(formula = TOTFERTRAT ~ FEMMARAGE + DIVORCERAT + ILLITERRAT +

TELEPERFAM + REGION, data = Italy)

Residuals:

Min 1Q Median 3Q Max

-0.17487 -0.06724 0.00231 0.04516 0.39168

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.538560 0.609124 5.809 1.03e-07 \*\*\*

FEMMARAGE -0.060186 0.023278 -2.585 0.011409 \*

DIVORCERAT -0.086453 0.055473 -1.558 0.122795

ILLITERRAT -0.001634 0.010915 -0.150 0.881362

TELEPERFAM -1.011377 0.182312 -5.547 3.14e-07 \*\*\*

REGIONNorth 0.004793 0.027447 0.175 0.861794

REGIONSardinia 0.031584 0.059267 0.533 0.595473

REGIONSicily 0.216287 0.060134 3.597 0.000537 \*\*\*

REGIONSouth 0.186853 0.046051 4.057 0.000109 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09589 on 86 degrees of freedom

Multiple R-squared: **0.8438**, Adjusted R-squared: 0.8293

F-statistic: 58.09 on 8 and 86 DF, p-value: < 2.2e-16

Comment: **DIVORCERAT**, and **ILLITERRAT** are no longersignificant in this model. And the significance of **FEMMARAGE** also decreases dramatically while **TELPERFAM** stays approximately the same. This drop in the significance is induced by the high correlation of these variables with the factor **REGION** which now also captures most of the variability in the dependent variable **TOTFERTRAT**. A high degree of multicollineary is present in the independent variables.

Task 13: Use a partial *F*-test to check whether the model in task 12 has improved the model fit of the base model in task 9 significantly. (0.5 points)  
That is, test the null hypothesis: against the alternative hypothesis is *for at least one* .

> anova(lm3, lm1)

Analysis of Variance Table

Model 1: TOTFERTRAT ~ FEMMARAGE + DIVORCERAT + ILLITERRAT + TELEPERFAM +

REGION

Model 2: TOTFERTRAT ~ FEMMARAGE + DIVORCERAT + ILLITERRAT + TELEPERFAM

Res.Df RSS Df Sum of Sq F Pr(>F)

1 86 0.79079

2 90 0.96406 -4 -0.17327 4.7107 **0.001743** \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Comment: The *p*-value (0.001743) is substantially smaller than 0.05, thus the null hypothesis can be rejected. We can conclude **REGION** is a significantly different from zero and the model in task 4.6 has improved the model fit of the base model in task 4.3 significantly.

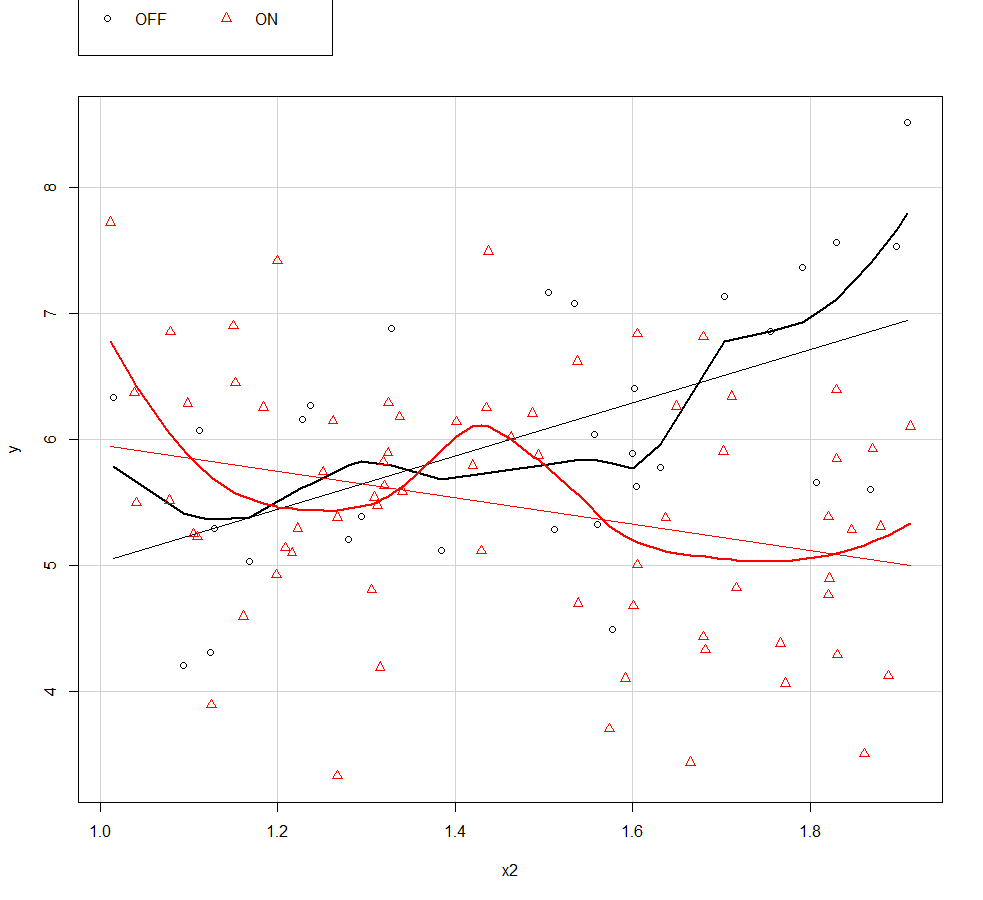
## Part III. Identification of the Underlying Model Structure (2 points)

Use the four **dBase** datasets **Data01.dbf** to **Data04.dbf** for this task. These datasets comprise of three variables: **y** for the dependent variables, **x1** for a binary ***factor***, and **x2** for a ***metric*** variable. Each of these datasets is best described by one of these competing models:

|  |  |
| --- | --- |
| Name | Models Structure |
| Full interaction model | **mod.full <- lm(y~x1\*x2, data=df)** |
| Intercept model | **mod.int <- lm(y~x1+x2, data=df)** |
| Slope model | **mod.slope <- lm(y~x1:x2, data=df)** |
| Means model | **mod.means <- lm(y~x1, data=df)** |
| Plain model | **mod.plain <- lm(y~x2, data=df)** |

For each of the dataset generate an informative scatterplot showing the regression regimes for both groups of observations. You can employ the syntax **car::scatterplot(y~x2|x1, data=df)**. Then identify which of the competing model structures best describes the given datasets. You can use the significance of the individual regression coefficients, the , or if several competing model structures seem to be reasonably relevant then try to eliminate inferior models using ***nested partial F-tests***. Provide a ***rational*** for your model selection.

Task 14: Identify the underlying model structure for **Data01.dbf**. [1 point]



(1) Full interaction model

> summary(mod.full <- lm(y~x1\*x2, data=df))

Call:

lm(formula = y ~ x1 \* x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.3473 -0.6750 0.0057 0.6790 1.9950

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.9056 0.9703 2.995 0.003496 \*\*

x1ON 4.1006 1.1539 3.554 0.000591 \*\*\*

x2 2.1153 0.6418 3.296 0.001376 \*\*

x1ON:x2 -3.1647 0.7676 -4.123 7.95e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9371 on 96 degrees of freedom

Multiple R-squared: 0.2071, Adjusted R-squared: 0.1823

F-statistic: 8.357 on 3 and 96 DF, p-value: 5.416e-05

(2) Intercept model

> summary(mod.int <- lm(y~x1+x2, data=df))

Call:

lm(formula = y ~ x1 + x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.16453 -0.69488 -0.01309 0.75932 2.50704

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.19794 0.59479 10.420 < 2e-16 \*\*\*

x1ON -0.58108 0.22098 -2.630 0.00994 \*\*

x2 -0.09727 0.37997 -0.256 0.79849

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.011 on 97 degrees of freedom

Multiple R-squared: 0.06669, Adjusted R-squared: 0.04744

F-statistic: 3.465 on 2 and 97 DF, p-value: 0.03519

(3) Slope model

> summary(mod.slope <- lm(y~x1:x2, data=df))

Call:

lm(formula = y ~ x1:x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.15572 -0.71472 -0.00485 0.76321 2.27986

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.8050 0.5557 10.446 <2e-16 \*\*\*

x1OFF:x2 0.2274 0.3812 0.597 0.552

x1ON:x2 -0.2527 0.3771 -0.670 0.504

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9917 on 97 degrees of freedom

Multiple R-squared: 0.1028, Adjusted R-squared: 0.08426

F-statistic: 5.555 on 2 and 97 DF, p-value: 0.0052

(4) Means model

> summary(mod.means <- lm(y~x1, data=df))

Call:

lm(formula = y ~ x1, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.14586 -0.71375 0.01236 0.77875 2.46598

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.0532 0.1838 32.938 < 2e-16 \*\*\*

x1ON -0.5783 0.2197 -2.633 0.00984 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.007 on 98 degrees of freedom

Multiple R-squared: 0.06605, Adjusted R-squared: 0.05652

F-statistic: 6.931 on 1 and 98 DF, p-value: 0.009843

(5) Plain model

> summary(mod.plain <- lm(y~x2, data=df))

Call:

lm(formula = y ~ x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.32892 -0.63523 -0.01894 0.62188 2.89190

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.71852 0.58299 9.809 3.15e-16 \*\*\*

x2 -0.04777 0.39078 -0.122 0.903

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

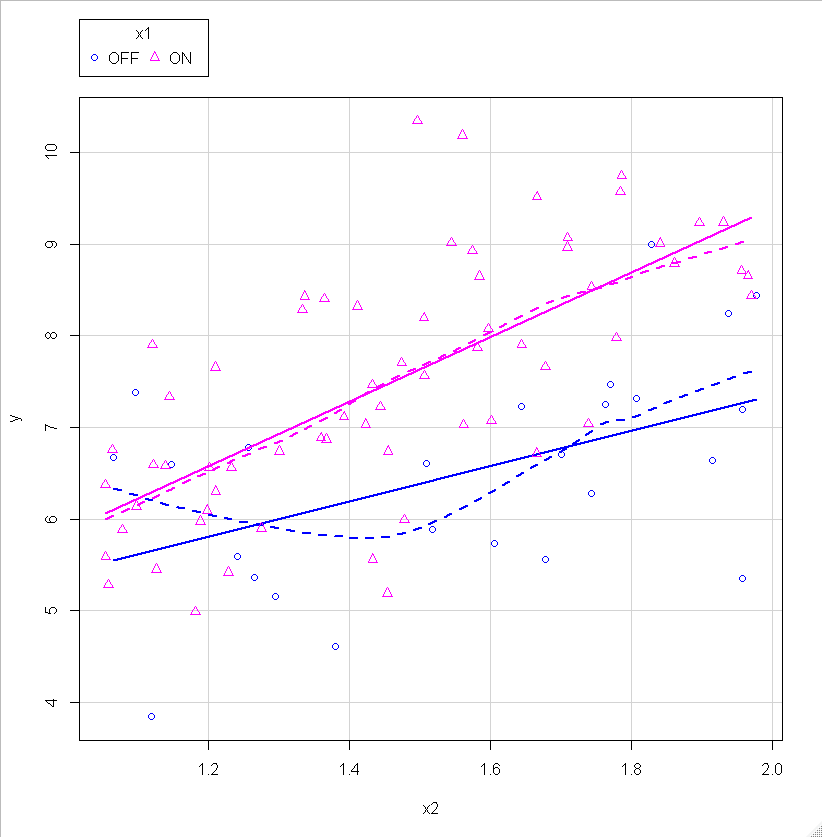
Residual standard error: 1.041 on 98 degrees of freedom

Multiple R-squared: 0.0001524, Adjusted R-squared: -0.01005

F-statistic: 0.01494 on 1 and 98 DF, p-value: 0.903

Comment: Based on the scatterplot and the highest value, the full interaction model fits Data01 best. Thus, the underlying model structure should be different slopes and different intercepts. For this model all coefficients are significant.

Task 15: Identify the underlying model structure for **Data02.dbf**. [1 point]



(1) Full interaction model

> summary(mod.full <- lm(y~x1\*x2, data=df))

Call:

lm(formula = y ~ x1 \* x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.2803 -0.5894 -0.0767 0.6894 2.7247

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.5129 1.0587 3.318 0.00133 \*\*

x1ON -1.1490 1.2694 -0.905 0.36789

x2 1.9177 0.6634 2.890 0.00487 \*\*

x1ON:x2 1.5954 0.8143 1.959 0.05333 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9994 on 86 degrees of freedom

Multiple R-squared: 0.4846, Adjusted R-squared: 0.4666

F-statistic: 26.95 on 3 and 86 DF, p-value: 2.203e-12

(2) Intercept model

> summary(mod.int <- lm(y~x1+x2, data=df))

Call:

lm(formula = y ~ x1 + x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.3269 -0.6045 -0.1066 0.6784 2.7445

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.8533 0.6454 2.872 0.00513 \*\*

x1ON 1.2935 0.2427 5.330 7.63e-07 \*\*\*

x2 2.9767 0.3909 7.615 2.98e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.016 on 87 degrees of freedom

Multiple R-squared: 0.4616, Adjusted R-squared: 0.4492

F-statistic: 37.29 on 2 and 87 DF, p-value: 2.01e-12

(3) Slope model

> summary(mod.slope <- lm(y~x1:x2, data=df))

Call:

lm(formula = y ~ x1:x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.2926 -0.6186 -0.1027 0.7127 2.7223

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.7135 0.5835 4.651 1.17e-05 \*\*\*

x1OFF:x2 2.4096 0.3802 6.337 1.00e-08 \*\*\*

x1ON:x2 3.2810 0.3961 8.284 1.31e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9984 on 87 degrees of freedom

Multiple R-squared: 0.4797, Adjusted R-squared: 0.4677

F-statistic: 40.1 on 2 and 87 DF, p-value: 4.543e-13

(4) Means model

> summary(mod.means <- lm(y~x1, data=df))

Call:

lm(formula = y ~ x1, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.66794 -0.92973 0.08894 0.93032 2.85409

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.5181 0.2607 25.000 < 2e-16 \*\*\*

x1ON 0.9739 0.3068 3.175 0.00207 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.304 on 88 degrees of freedom

Multiple R-squared: 0.1028, Adjusted R-squared: 0.09256

F-statistic: 10.08 on 1 and 88 DF, p-value: 0.002069

(5) Plain model

> summary(mod.plain <- lm(y~x2, data=df))

Call:

lm(formula = y ~ x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-3.09233 -0.75539 0.00649 0.80869 3.10630

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.3243 0.6681 4.975 3.20e-06 \*\*\*

x2 2.6163 0.4409 5.934 5.73e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

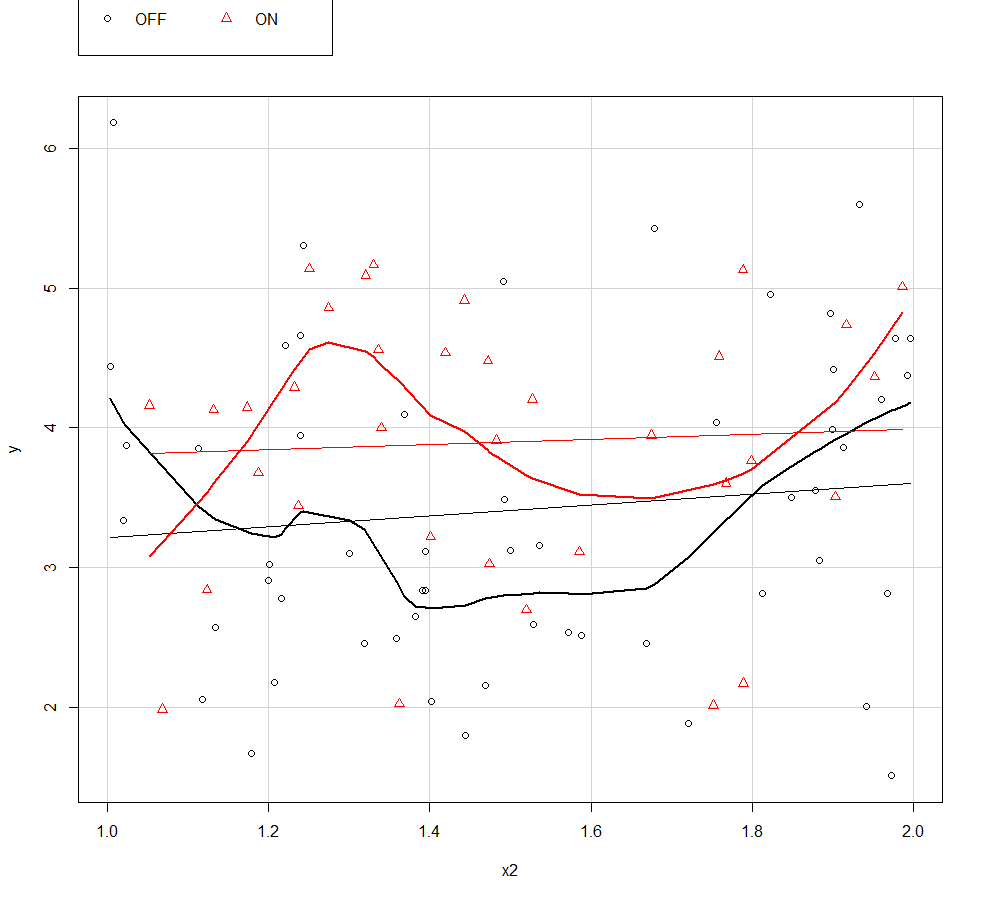
Residual standard error: 1.163 on 88 degrees of freedom

Multiple R-squared: 0.2858, Adjusted R-squared: 0.2777

F-statistic: 35.21 on 1 and 88 DF, p-value: 5.729e-08

Comment: While the full interaction model has a high not all of its coefficients are significant. The slope model has higher value than the intercept model. Although, the scatterplot displays almost parallel slope with small difference, the underlying model structure of Data02 is the different slopes with the same intercept. For this model all regression coefficients are significant. The slope and the intercept model cannot be compared using a partial *F*-test because both models are not nested.

Task 16: Identify the underlying model structure for **Data03.dbf**. [1 point]



(1) Full interaction model

> summary(mod.full <- lm(y~x1\*x2, data=df))

Call:

lm(formula = y ~ x1 \* x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.08470 -0.79532 0.00543 0.72624 2.97172

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.8269 0.7112 3.975 0.000146 \*\*\*

x1ON 0.8016 1.2350 0.649 0.518047

x2 0.3875 0.4575 0.847 0.399333

x1ON:x2 -0.2072 0.8119 -0.255 0.799143

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.068 on 86 degrees of freedom

Multiple R-squared: 0.05584, Adjusted R-squared: 0.02291

F-statistic: 1.695 on 3 and 86 DF, p-value: 0.174

(2) Intercept model

> summary(mod.int <- lm(y~x1+x2, data=df))

Call:

lm(formula = y ~ x1 + x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.05505 -0.79907 0.00115 0.70767 2.93790

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.9271 0.5899 4.962 3.43e-06 \*\*\*

x1ON 0.4919 0.2302 2.137 0.0354 \*

x2 0.3217 0.3759 0.856 0.3944

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.062 on 87 degrees of freedom

Multiple R-squared: 0.05513, Adjusted R-squared: 0.03341

F-statistic: 2.538 on 2 and 87 DF, p-value: 0.08487

(3) Slope model

> summary(mod.slope <- lm(y~x1:x2, data=df))

Call:

lm(formula = y ~ x1:x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.02014 -0.80891 0.00766 0.69917 2.87477

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.0928 0.5795 5.337 7.43e-07 \*\*\*

x1OFF:x2 0.2201 0.3766 0.584 0.560

x1ON:x2 0.5304 0.3972 1.336 0.185

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.064 on 87 degrees of freedom

Multiple R-squared: 0.05122, Adjusted R-squared: 0.02941

F-statistic: 2.348 on 2 and 87 DF, p-value: 0.1016

(4) Means model

> summary(mod.means <- lm(y~x1, data=df))

Call:

lm(formula = y ~ x1, data = df)

Residuals:

Min 1Q Median 3Q Max

-1.91282 -0.81352 0.03256 0.75855 2.77258

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.4169 0.1430 23.892 <2e-16 \*\*\*

x1ON 0.4787 0.2293 2.087 0.0398 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.061 on 88 degrees of freedom

Multiple R-squared: 0.04717, Adjusted R-squared: 0.03634

F-statistic: 4.357 on 1 and 88 DF, p-value: 0.03976

(5) Plain model

> summary(mod.plain <- lm(y~x2, data=df))

Call:

lm(formula = y ~ x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.22114 -0.90060 0.00563 0.87282 2.71970

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.1998 0.5875 5.446 4.61e-07 \*\*\*

x2 0.2677 0.3826 0.700 0.486

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 88 degrees of freedom

Multiple R-squared: 0.005532, Adjusted R-squared: -0.005768

F-statistic: 0.4896 on 1 and 88 DF, p-value: 0.486

> anova(mod.int, mod.means)

Analysis of Variance Table

Model 1: y ~ x1 + x2

Model 2: y ~ x1

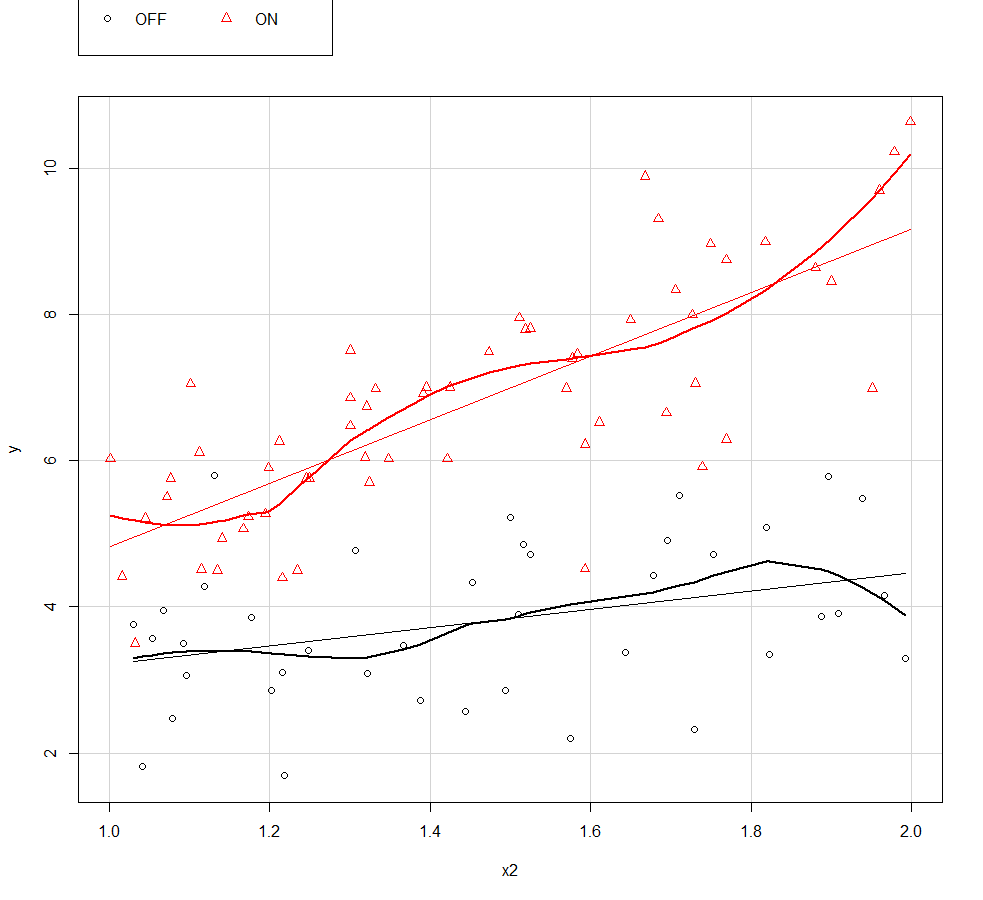
Res.Df RSS Df Sum of Sq F Pr(>F)

1 87 98.163

2 88 98.990 -1 -0.82643 0.7324 0.3944

Comment: Among the low values of all four models, the relatively high values of the intercept model and the means model are very close. After conducting the nested partial *F*-test, we can conclude the means model is not significantly different from the intercept model due to such large *p*-value. Based on the parsimony rule, we should choose the means model for Data03.

Task 17: Identify the underlying model structure for **Data04.dbf**. [1 point]



(1) Full interaction model

> summary(mod.full <- lm(y~x1\*x2, data=df))

Call:

lm(formula = y ~ x1 \* x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.87651 -0.61819 0.08394 0.70364 2.40823

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.9749 0.7921 2.493 0.0144 \*

x1ON -1.5119 1.0472 -1.444 0.1521

x2 1.2466 0.5294 2.355 0.0206 \*

x1ON:x2 3.1082 0.7043 4.413 2.66e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.014 on 96 degrees of freedom

Multiple R-squared: 0.7567, Adjusted R-squared: 0.7491

F-statistic: 99.54 on 3 and 96 DF, p-value: < 2.2e-16

(2) Intercept model

> summary(mod.int <- lm(y~x1+x2, data=df))

Call:

lm(formula = y ~ x1 + x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.6799 -0.8496 0.1314 0.7501 2.9948

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.5986 0.5850 -1.023 0.309

x1ON 3.0184 0.2260 13.359 < 2e-16 \*\*\*

x2 3.0029 0.3809 7.883 4.76e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.106 on 97 degrees of freedom

Multiple R-squared: 0.7074, Adjusted R-squared: 0.7013

F-statistic: 117.2 on 2 and 97 DF, p-value: < 2.2e-16

(3) Slope model

> summary(mod.slope <- lm(y~x1:x2, data=df))

Call:

lm(formula = y ~ x1:x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-2.8375 -0.6569 0.1064 0.7043 2.6328

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.1098 0.5210 2.130 0.0357 \*

x1OFF:x2 1.8128 0.3576 5.069 1.91e-06 \*\*\*

x1ON:x2 3.9243 0.3581 10.959 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 97 degrees of freedom

Multiple R-squared: 0.7515, Adjusted R-squared: 0.7463

F-statistic: 146.6 on 2 and 97 DF, p-value: < 2.2e-16

(4) Means model

> summary(mod.means <- lm(y~x1, data=df))

Call:

lm(formula = y ~ x1, data = df)

Residuals:

Min 1Q Median 3Q Max

-3.2650 -0.9458 -0.0333 0.9301 3.8682

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.8016 0.2229 17.05 <2e-16 \*\*\*

x1ON 2.9650 0.2878 10.30 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.41 on 98 degrees of freedom

Multiple R-squared: 0.5199, Adjusted R-squared: 0.515

F-statistic: 106.1 on 1 and 98 DF, p-value: < 2.2e-16

(5) Plain model

> summary(mod.plain <- lm(y~x2, data=df))

Call:

lm(formula = y ~ x2, data = df)

Residuals:

Min 1Q Median 3Q Max

-4.0359 -1.4569 0.2535 1.5108 3.7027

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.4347 0.9469 1.515 0.133

x2 2.8502 0.6384 4.465 2.15e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.855 on 98 degrees of freedom

Multiple R-squared: 0.169, Adjusted R-squared: 0.1605

F-statistic: 19.93 on 1 and 98 DF, p-value: 2.147e-05

> anova(mod.full, mod.slope)

Analysis of Variance Table

Model 1: y ~ x1 \* x2

Model 2: y ~ x1:x2

Res.Df RSS Df Sum of Sq F Pr(>F)

1 96 98.717

2 97 100.861 -1 -2.1436 2.0846 0.1521

Comment: The scatterplot shows different slopes of the two groups. Although both of the slope model and the full interaction model have high and similar values, not all regression coefficients of the full interaction model are significant. Moreover, the *p*-value of the nested partial *F*-test is larger than 0.05, which means that there is no significant between these two models. Therefore, the slope model should be chosen based on the parsimony rule and the two groups have the same intercept but different slopes.