# Lab02: Multiple Regression Analysis, Factors and Interaction Effects

**Handed out:** Monday, February 10, 2020

**Return date:** Monday, February 24, 2020, at the beginning of the class.

**Grading:** This lab counts 12 % towards your final grade

**Objectives:** In this lab you will explore the meaning of the partial regression coefficients, and the partial *F*-test and you will build and analyze a full-fledge multiple regression model which includes a factor.

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab02, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task and sub-task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code the use mathematical type-setting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

## Task 1. Partial Regression Coefficient [2 points]

Use the **Concord1.sav** file for this task. You will demonstrate that in multiple regression the partial effect of an independent variable is free from any linear effects of the remaining independent variables in a regression model.

**Task 1.1:** Run the multiple model **water81~income+water80+educat** and *interpret* its regression coefficients. [0.5 points]

model.1 <- lm(water81~income+water80+educat,data = concord)

summary(model.1)

Coefficients:

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

(Intercept) 811.18243 196.29879 4.132 4.22e-05 \*\*\*

income 24.71652 3.54917 6.964 1.06e-11 \*\*\*

water80 0.59130 0.02477 23.872 < 2e-16 \*\*\*

educat -49.88881 14.18671 -3.517 0.000478 \*\*\*

---

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

Residual standard error: 914.9 on 492 degrees of freedom

Multiple R-squared: 0.6233, Adjusted R-squared: 0.621

F-statistic: 271.3 on 3 and 492 DF, p-value: < 2.2e-16

The coefficients in multiple regression represent the relationship between the dependent and one independent variable while statistically holding the effects of all other variables constant. Here, the regression coefficients of income are 24.7, which means one unit of income increase would bring 24.7 unit increase in water consumption when education and the water consumption of last year stay the same.

**Task 1.2:** Calculate the residuals of the two models [a] **water81~income+water80** and [b] **educat~income+water80** . *What are these residuals specifically measuring*? [0.5 points]

model.2 <- lm(water81~income+water80,data = concord)

sum(r1 <- residuals(model.2))

model.3 <- lm(educat~income+water80,data = concord)

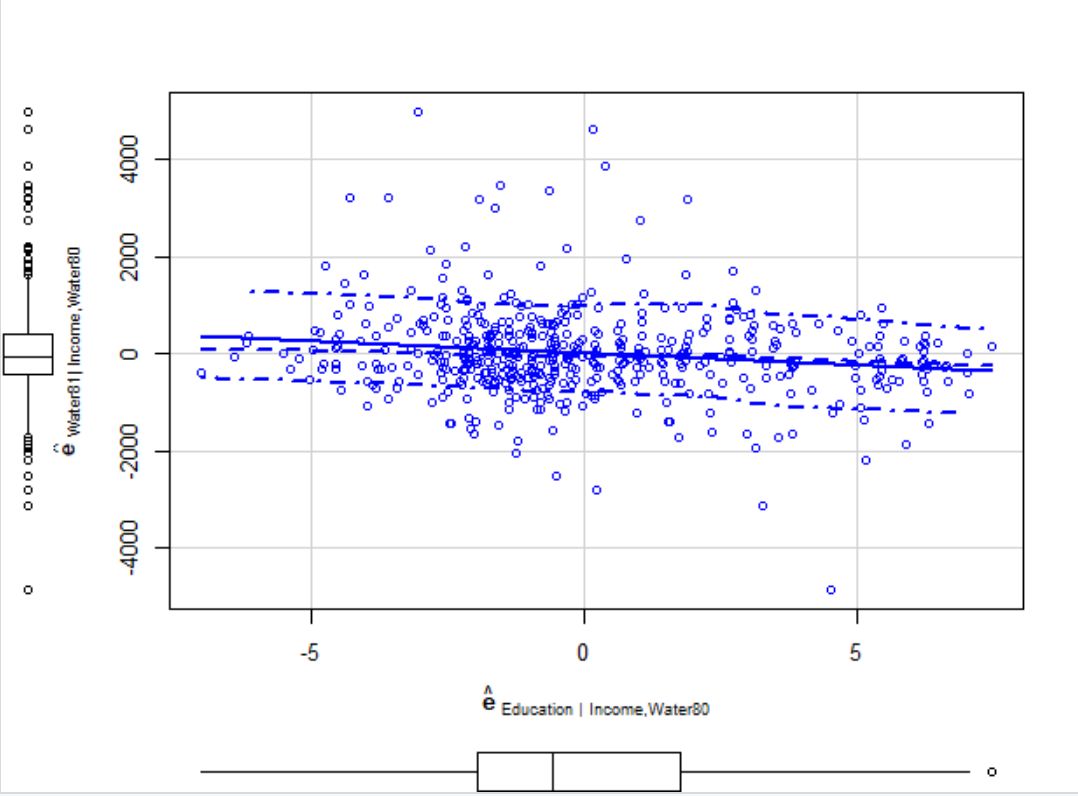
sum(r2 <- residuals(model.3))

They measure the remaining (unexplained) variation of the dependent variable y after accounting for the included independent variables.

**Task 1.3:** Generate the partial regression leverage scatterplot of the water residuals against the education residuals. Make sure to use properly labeled axes. *Briefly interpret the scatterplot*. [0.5 points]

scatterplot(r1~r2, xlab=bquote(hat(bold(e))[" Water81 | Income,Water80"]),

ylab=bquote(hat(bold(e))[" Education | Income,Water80"]))



The relationship of water residuals against the education residuals is negative, means education has a negative impact on water consumption

**Task 1.4:** Estimate a regression model of the *water residuals* on the *education residuals* and *compare* its estimate slope coefficient against the slope coefficient for **educat** of the multiple model from task 1.1. *Why are you allowed to suppress the intercept in this model*? [0.5 points]

summary(lm(r1~r2-1))

Coefficients:

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

r2 -49.89 14.14 -3.527 0.000459 \*\*\*

It is same with the one got at task 1.1.

Because regression lines must go through , and mean of error is 0. So even you do not suppress the intercept, it is also very close to 0.

## Task 2: A Multiple Regression Model with Factors and Partial *F*-test [4 points]

Use the dBase data file **provinces.dbf** (read it into  with the function **foreign::read.dbf**)

You will experiment with regression models that aim at explaining the 1994 total fertility rate **TOTFERTRAT** (number of children born by a woman during her lifetime) within the 95 Italian provinces. The following independent variables are:[a] the metric illiteracy rate **ILLITERRAT** , [b] the metric average woman’s age at first marriage **FEMMARAGE**, [c] the metric divorce rate **DIVORCERAT**, [d] the metric televisions per household **TELEPERFAM** and [e] a regional factor **REGION** denoting whether a province is located on the islands of Sicily or Sardina, or in the southern, central or northern parts of Italy’s mainland.

Note: please do ***not*** perform variable transformations in the task.

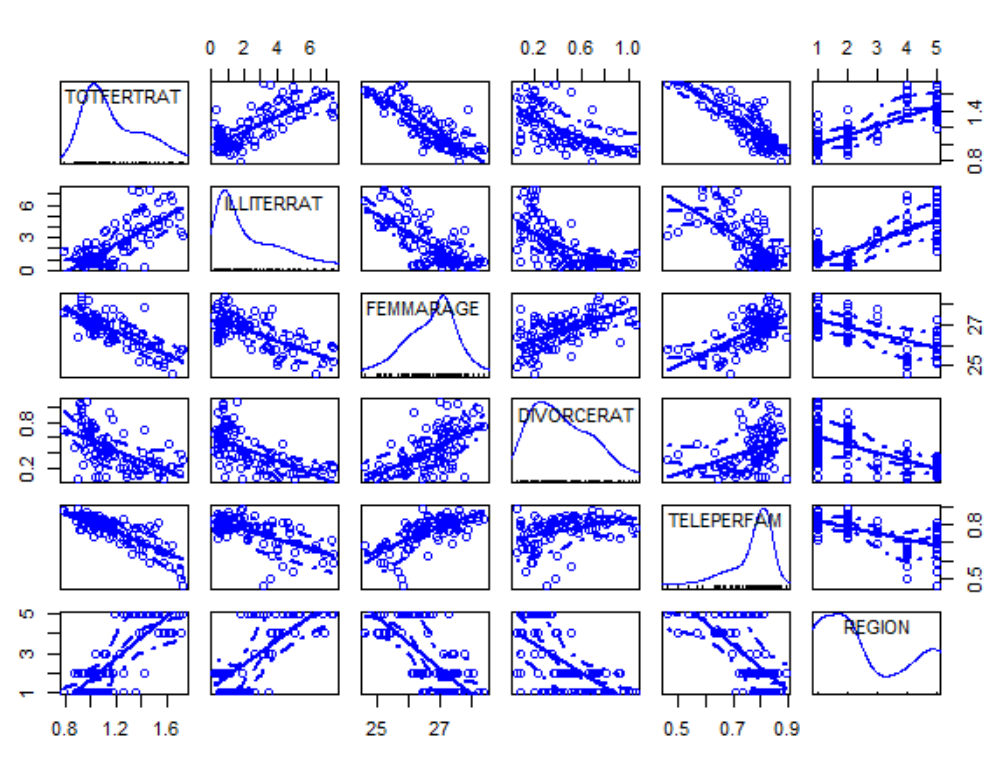
**Task 2.1:** Use common sense arguments ***how*** these four metric variables will influence the provincial fertility rates. Use one or two sentences per explanation, and formulate *one* or *two*-sided null and alternative hypotheses based on your explanation. The statistical hypotheses should be *type-set* properly, for instance, as against . Format everything in the table shown below. [0.5 points]

|  |  |  |
| --- | --- | --- |
| Variable | Common Sense Arguments | Statistical Hypotheses |
| ILLITERRAT | Higher education rate may lower fertility rate |  |
| FEMMARAGE | Higher married age may lower fertility rate |  |
| DIVORCERAT | Higher divorce rate may lower fertility rate |  |
| TELEPERFAM | Higher televisions penetration rate may lower fertility rate |  |

**Task 2.2:** Generate a scatterplot matrix showing the dependent variable and the four metric independent variables. *Briefly interpret the scatterplot matrix.* [0.5 points]

The relation of all independent variables with fertility rate are seemly like to the alternative hypothesis IS proposed in task 1.

ILLITERRAT are positive related, DIVORCERAT and TELEPERFAM are negative related. FEMMARAGE has negative impact.



**Task 2.3:** 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 2.1, [b] the significances of the estimate regression coefficients and [c] the goodness of fit. [0.5 points]

model.1 <- lm(TOTFERTRAT~ILLITERRAT+FEMMARAGE+DIVORCERAT+TELEPERFAM, data=province)

summary(model.1)

Coefficients:

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

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

ILLITERRAT 0.020377 0.008735 2.333 0.0219 \*

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

DIVORCERAT -0.112265 0.055648 -2.017 0.0466 \*

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

The coefficient of the average woman’s age at first marriage is different from what I suggested in task.

All those coefficients are significant and could explain 81% variation in the dependent variable.

**Task 2.4:** Calculate the standardized *beta-coefficients* for the multiple model in task 2.3. 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. Use proper options for the **coefplot( )** function. [1 point]

col\_lst <- c('TOTFERTRAT','ILLITERRAT','FEMMARAGE','DIVORCERAT','TELEPERFAM')

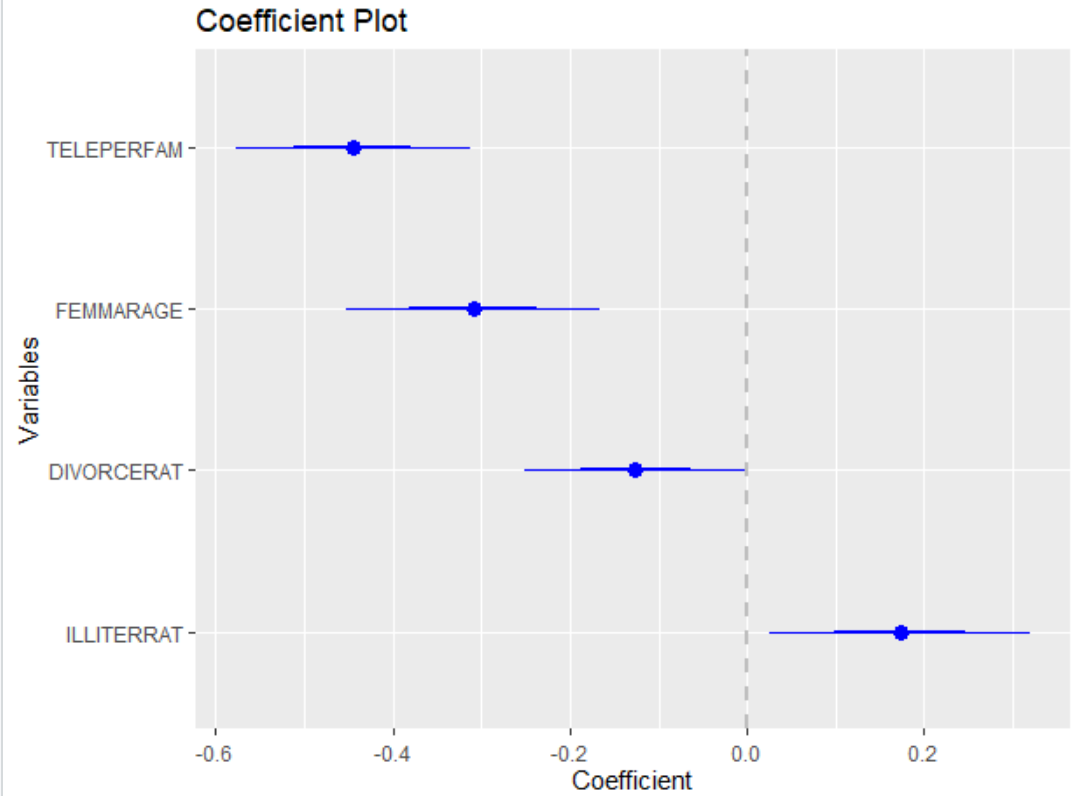
province\_new <- province[,col\_lst]

province\_scale <- as.data.frame(scale(province\_new))

model.2 <- lm(TOTFERTRAT~-1+ILLITERRAT+FEMMARAGE+DIVORCERAT+TELEPERFAM, data=province\_scale)

summary(model.2)

coefplot(model.2,xlab = 'Coefficient',ylab = 'Variables',decreasing = TRUE,sort = "magnitude")



**Task 2.5:** Run five separate regressions on the [a] independent variables as well as [b] the dependent fertility rate using the factor **REGION** as independent variable.   
Does the **REGION** factor *explain variation* of the four independent variables as well as the fertility rate, i.e., what are the factor’s ’s? [0.5 points]

Hint: To calibrate all five models with one function call you can use the regression formula syntax **cbind(TOTFERTRAT,ILLITERRAT,FEMMARAGE,DIVORCERAT,TELEPERFAM)~REGION**.  
The **summary** function gives you the results for all five models.

model.combine <- lm(cbind(TOTFERTRAT,ILLITERRAT,FEMMARAGE,DIVORCERAT,TELEPERFAM)~REGION,data=province)

summary(model.combine)

|  |  |
| --- | --- |
|  | R square from REGION |
| TOTFERTRAT | 0.7409 |
| ILLITERRAT | 0.7485 |
| FEMMARAGE | 0.6288 |
| DIVORCERAT | 0.423 |
| TELEPERFAM | 0.5306 |

**Task 2.6:** 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 why some independent metric variables are no longer significant? [0.5 points]

model.3 <- lm(TOTFERTRAT~+ILLITERRAT+FEMMARAGE+DIVORCERAT+TELEPERFAM+province$REGION, data=province\_scale)

summary(model.3)

because the high correlation among independent variables has existed, so there was a lot of redundancy variables, then the dependent variable becomes only correlated with part of independent variables.

**Task 2.7:** Use a partial *F*-test to check whether the model in task 2.6 has improved the model fit of the base model in task 2.3 significantly. [0.5 points]   
That is, test the null hypothesis: against the alternative hypothesis is *for at least one* .

summary(model.3)

anova(model.2,model.3)

Analysis of Variance Table

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

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

REGION

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

1 90 0.96406

2 86 0.79079 4 0.17327 4.7107 0.001743 \*\*

---

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

Yes, the model in task 2.6 has improved significantly by adding region variables

## Task 3. Identification of the Underlying Model Structure [6 points]

Use the workspace **ModelSpecs.RData** for this task. It contains the six data-frames **mod1** to **mod6**. Each data-frame is comprised of three variables: **y** for the dependent variables, **g** for a binary ***factor***, and **x** for a ***metric*** variable. Each of these data-frames is best ***statistically*** described by one of these competing models:

|  |  |
| --- | --- |
| Name | Models Structure |
| Full interaction model | **lm(y~g+x+g:x, data=mod?)  lm(y~g\*x, data=mod?)** |
| Intercept model | **lm(y~g+x, data=mod?)** |
| Slope model | **lm(y~g:x, data=mod?)** |
| Means model | **lm(y~g, data=mod?)** |
| Plain regression model | **lm(y~x, data=mod?)** |

For each of the data-frame generate an informative scatterplot showing the regression regimes for both groups of observations. You can employ the syntax:

**car::scatterplot(y~x|g,smoother=F,boxplots="xy",data=mod???,main="Model???")**

Then identify which of the competing model structures best describes the given data-frame. If several competing model structures seem to be reasonably relevant then try to eliminate inferior models using by looking for statistically superior , non-significant coefficients’ -tests and ***nested partial F-tests***. Provide a ***rational*** for your model selection.

summary\_model <- function(mod){

full.model <- lm(y~g\*x, data=mod)

Intercept.model <- lm(y~g+x, data=mod)

slope.model <- lm(y~g:x, data=mod)

means.model <- lm(y~g,data=mod)

plain.model <- lm(y~x, data=mod)

model\_lst <- list(full.model,Intercept.model,slope.model,means.model,plain.model)

model\_summary <- lapply(model\_lst, summary)

r\_square\_lst <- function(a) {return(a$adj.r.squared)}

model\_summary[[6]] <- as.data.frame(lapply(model\_summary, r\_square\_lst),

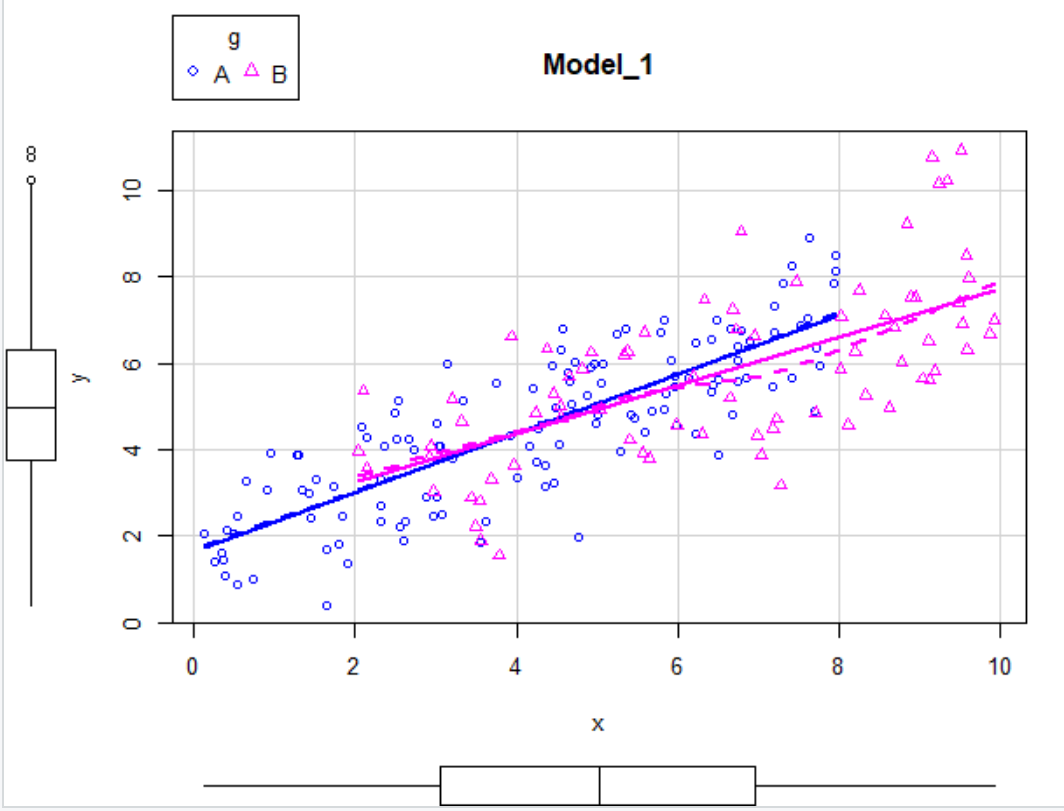
col.names = c('full.model', 'Intercept.model','slope.model','means.model','plain.model'))

model\_summary[[7]] <- model\_lst

return(model\_summary)

}

**Task 3.1:** Identify the underlying model structure for **mod1**. [1 point]



full.model Intercept.model slope.model means.model plain.model

1 0.6060689 0.6032791 0.6061828 0.0816716 0.6030411

lm(formula = y ~ g \* x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-3.0298 -0.9569 0.0743 0.9122 3.5398

Coefficients:

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

(Intercept) 1.64686 0.24101 6.833 1.02e-10 \*\*\*

gB 0.47813 0.49235 0.971 0.333

x 0.68324 0.05026 13.595 < 2e-16 \*\*\*

gB:x -0.12382 0.08001 -1.548 0.123

---

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

lm(formula = y ~ g:x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-3.0276 -0.9537 0.0884 0.8988 3.4489

Coefficients:

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

(Intercept) 1.76142 0.21013 8.383 9.75e-15 \*\*\*

gA:x 0.66210 0.04529 14.619 < 2e-16 \*\*\*

gB:x 0.60904 0.03555 17.130 < 2e-16 \*\*\*

Analysis of Variance Table

Model 1: y ~ g:x

Model 2: y ~ g \* x

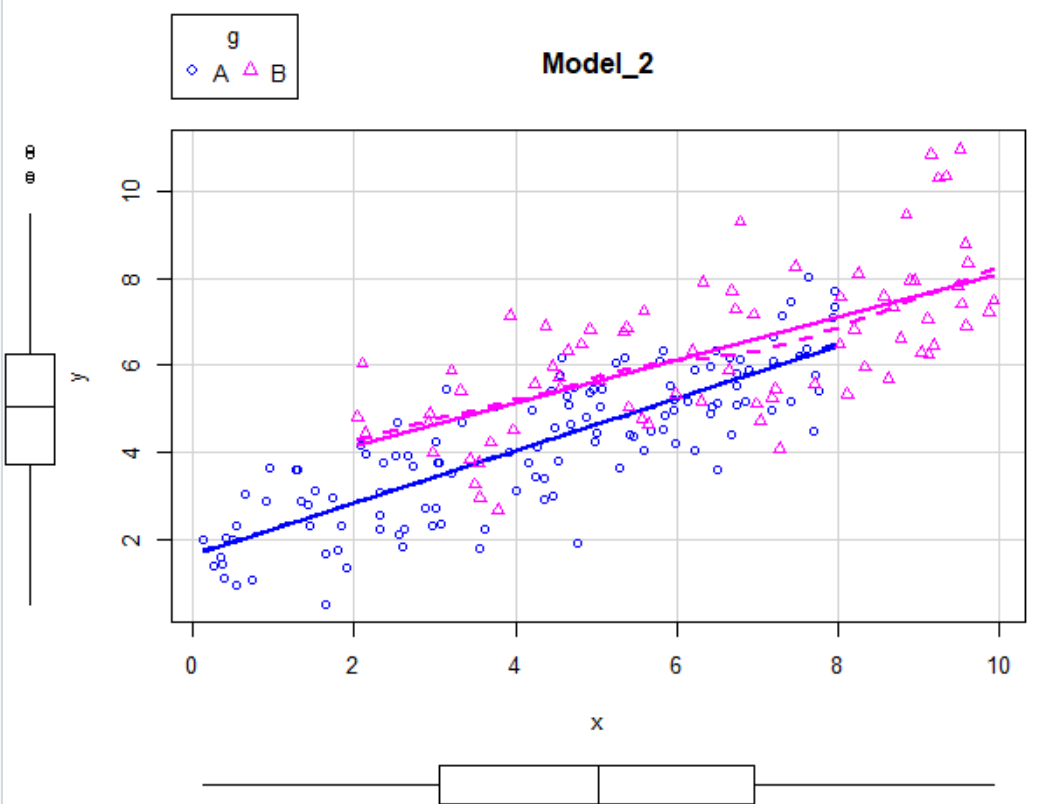
Res.Df RSS Df Sum of Sq F Pr(>F)

1 197 310.33

2 196 308.84 1 1.486 0.943 0.3327

The adjusted r square of slope model and full model are obviously higher than others. And there is not significant difference between 2 models. However, all variables in slope model are significant, due to reducing complexity of model, we select slope model here.

**Task 3.2:** Identify the underlying model structure for **mod2**. [1 point]



full.model Intercept.model slope.model means.model plain.model

1 0.6923208 0.6901418 0.6738623 0.2827411 0.6487722

lm(formula = y ~ g \* x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-2.67768 -0.84564 0.06566 0.80622 3.12836

Coefficients:

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

(Intercept) 1.61085 0.21299 7.563 1.49e-12 \*\*\*

gB 1.55788 0.43513 3.580 0.000433 \*\*\*

x 0.60383 0.04442 13.595 < 2e-16 \*\*\*

gB:x -0.10943 0.07071 -1.548 0.123325

lm(formula = y ~ g + x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-2.72996 -0.83083 0.04285 0.81349 2.95172

Coefficients:

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

(Intercept) 1.79407 0.17769 10.097 < 2e-16 \*\*\*

gB 0.94458 0.18034 5.238 4.15e-07 \*\*\*

x 0.56065 0.03468 16.166 < 2e-16 \*\*\*

Analysis of Variance Table

Model 1: y ~ g \* x

Model 2: y ~ g + x

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

1 196 241.22

2 197 244.17 -1 -2.9478 2.3951 0.1233

The adjusted r square of full.model and Intercept.model are obviously higher than others. And there is not significant difference between 2 models. However, all variables in Intercept.model are significant, due to reducing complexity of model, we select Intercept.model here.

**Task 3.3:** Identify the underlying model structure for **mod3**. [1 point]

full.model Intercept.model slope.model means.model plain.model

1 0.1428114 0.1367409 0.09138647 0.1410937 0.02148611

lm(formula = y ~ g \* x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-4.4694 -1.4115 0.1096 1.3457 5.2216

Coefficients:

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

(Intercept) 4.11480 0.35551 11.574 < 2e-16 \*\*\*

gB 2.60030 0.72628 3.580 0.000433 \*\*\*

x 0.06974 0.07413 0.941 0.348009

gB:x -0.18265 0.11802 -1.548 0.123325

lm(formula = y ~ g, data = mod)

Residuals:

Min 1Q Median 3Q Max

-4.5585 -1.3911 0.0671 1.3607 4.9206

Coefficients:

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

(Intercept) 4.4107 0.1658 26.605 < 2e-16 \*\*\*

gB 1.5714 0.2707 5.804 2.53e-08 \*\*\*

Analysis of Variance Table

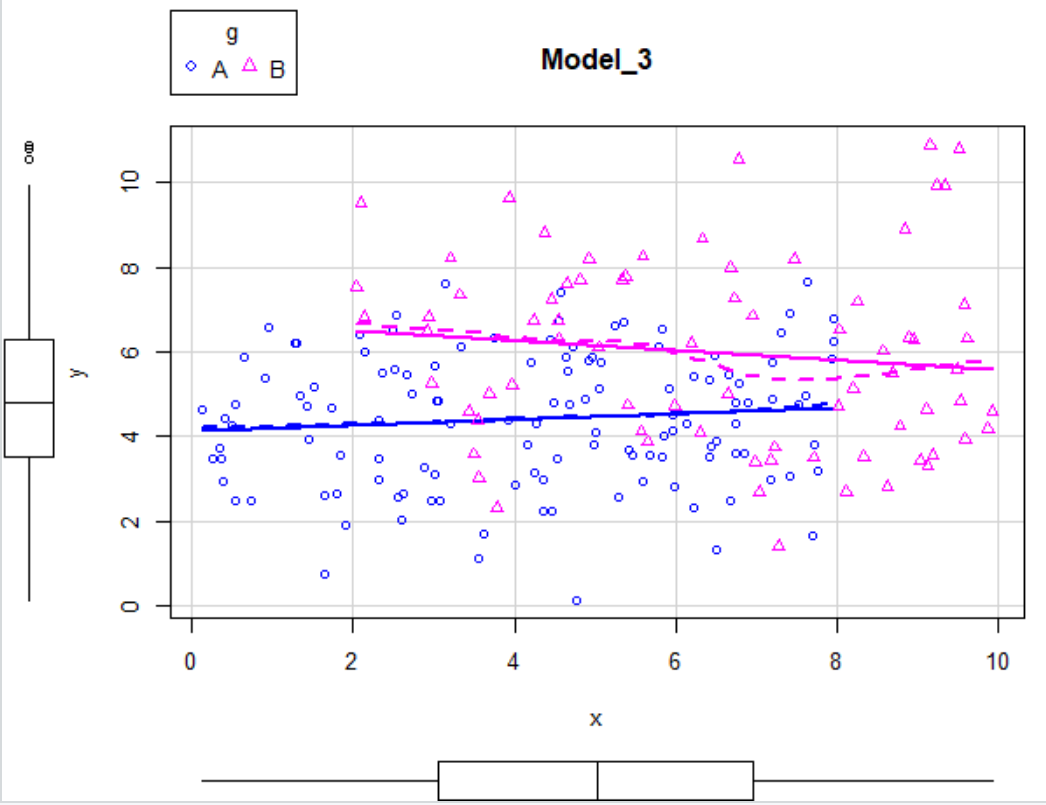
Model 1: y ~ g \* x

Model 2: y ~ g

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

1 196 672.04

2 198 680.25 -2 -8.218 1.1984 0.3039



The adjusted r square of full model and means model are obviously higher than others. And there is not significant difference between 2 models. However, all variables in means model are significant, due to reducing complexity of model, we select means model here.

**Task 3.4:** Identify the underlying model structure for **mod4**. [1 point]

full.model Intercept.model slope.model means.model plain.model

1 0.9025466 0.8675698 0.9025747 0.5889637 0.6506641

lm(formula = y ~ g \* x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-1.50698 -0.47592 0.03695 0.45373 1.76062

Coefficients:

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

(Intercept) 2.41024 0.11987 20.107 < 2e-16 \*\*\*

gB 0.23781 0.24489 0.971 0.333

x 0.33034 0.02500 13.215 < 2e-16 \*\*\*

gB:x 0.33697 0.03979 8.468 5.84e-15 \*\*\*

lm(formula = y ~ g:x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-1.50585 -0.47436 0.04397 0.44703 1.71543

Coefficients:

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

(Intercept) 2.46722 0.10451 23.61 <2e-16 \*\*\*

gA:x 0.31983 0.02253 14.20 <2e-16 \*\*\*

gB:x 0.69199 0.01768 39.13 <2e-16 \*\*\*

Analysis of Variance Table

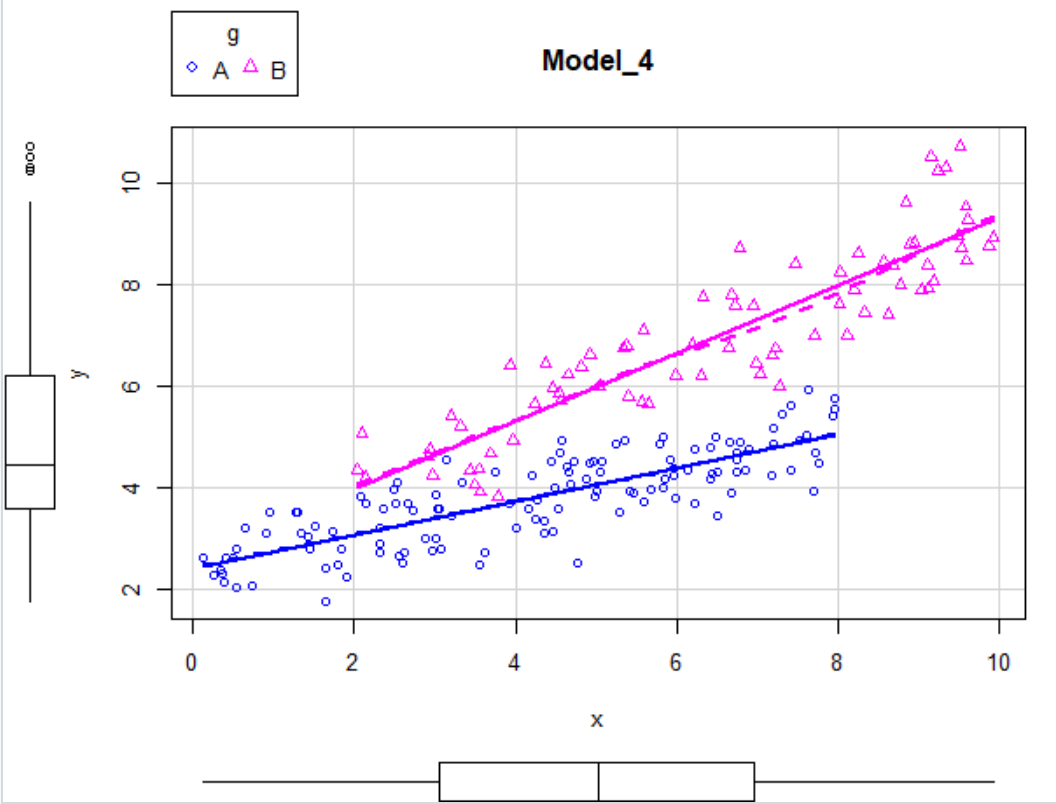
Model 1: y ~ g:x

Model 2: y ~ g \* x

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

1 197 76.771

2 196 76.403 1 0.36761 0.943 0.3327



The adjusted r square of slope model and full model are obviously higher than others. And there is not significant difference between 2 models. However, all variables in slope model are significant, due to reducing complexity of model, we select slope model here.

**Task 3.5:** Identify the underlying model structure for **mod5**. [1 point]

full.model Intercept.model slope.model means.model plain.model

1 0.9587611 0.9129509 0.9562871 0.7387832 0.5569946

lm(formula = y ~ g \* x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-0.98031 -0.30959 0.02404 0.29516 1.14531

Coefficients:

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

(Intercept) 2.75873 0.07798 35.38 < 2e-16 \*\*\*

gB 0.57035 0.15930 3.58 0.000433 \*\*\*

x 0.21489 0.01626 13.21 < 2e-16 \*\*\*

gB:x 0.38382 0.02589 14.83 < 2e-16 \*\*\*

lm(formula = y ~ g + x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-1.50016 -0.39030 -0.02668 0.39717 1.83903

Coefficients:

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

(Intercept) 2.11610 0.09418 22.47 <2e-16 \*\*\*

gB 2.72146 0.09558 28.47 <2e-16 \*\*\*

x 0.36634 0.01838 19.93 <2e-16 \*\*\*

lm(formula = y ~ g:x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-0.97759 -0.32436 -0.00265 0.30116 1.04563

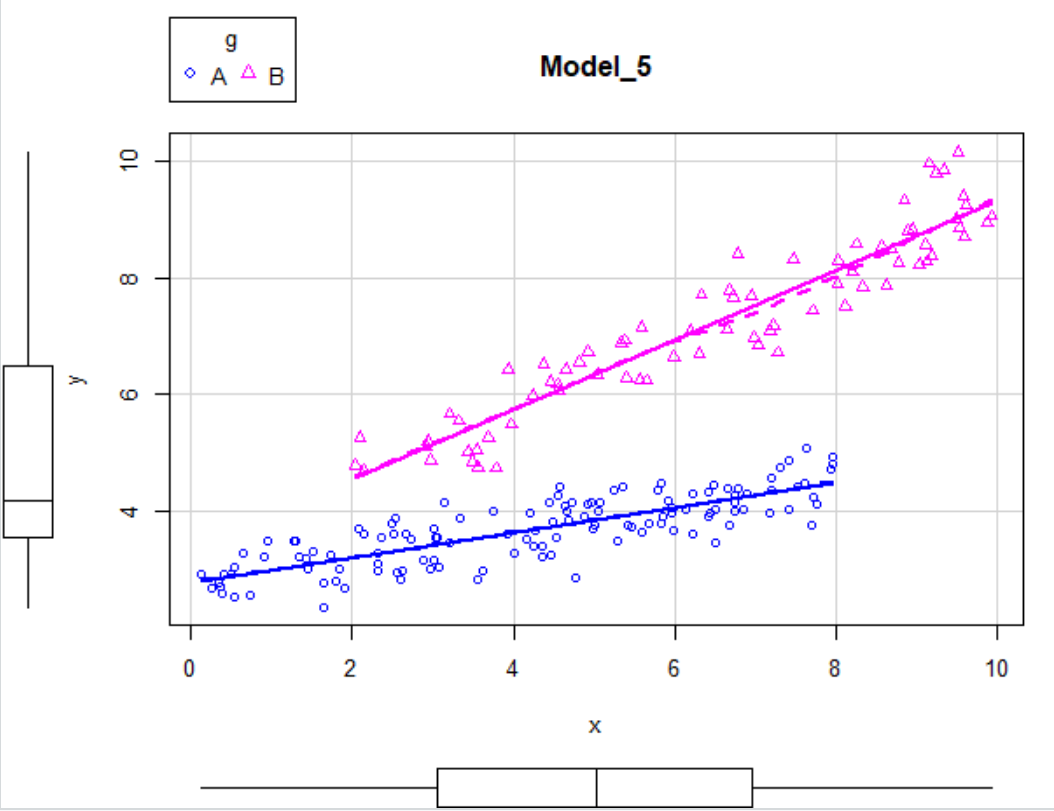
Coefficients:

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

(Intercept) 2.89539 0.07001 41.36 <2e-16 \*\*\*

gA:x 0.18967 0.01509 12.57 <2e-16 \*\*\*

gB:x 0.65790 0.01185 55.54 <2e-16 \*\*\*



Analysis of Variance Table

Model 1: y ~ g:x

Model 2: y ~ g \* x

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

1 197 34.446

2 196 32.331 1 2.1145 12.819 0.0004327 \*\*\*

Analysis of Variance Table

Model 1: y ~ g + x

Model 2: y ~ g \* x

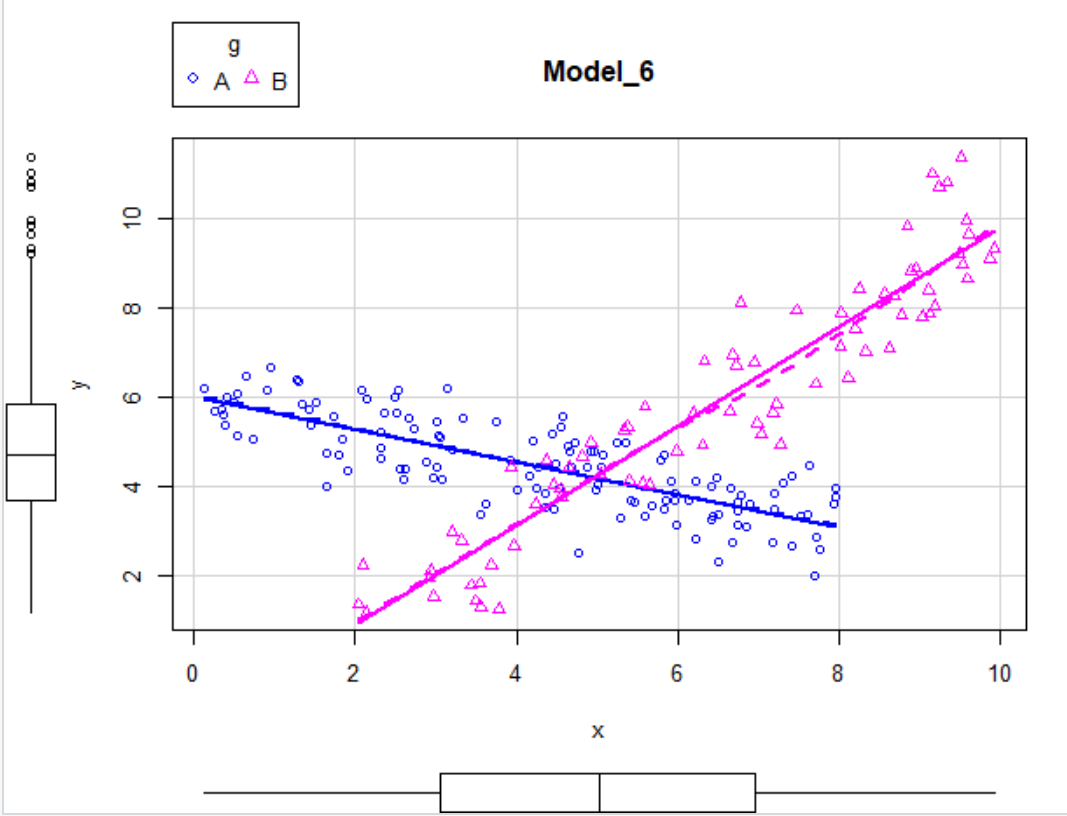
Res.Df RSS Df Sum of Sq F Pr(>F)

1 197 68.595

2 196 32.331 1 36.263 219.84 < 2.2e-16 \*\*\*

In this model, we could visually see that g has impact on both slope and intercept. According to F test, full model also improves our model significantly, so we select full model here.

**Task 3.6:** Identify the underlying model structure for **mod6**. [1 point]



full.model Intercept.model slope.model means.model plain.model

1 0.8570267 0.1726441 0.410064 0.1148816 0.1349401

lm(formula = y ~ g \* x, data = mod)

Residuals:

Min 1Q Median 3Q Max

-1.82531 -0.57645 0.04476 0.54958 2.13253

Coefficients:

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

(Intercept) 6.01986 0.14519 41.46 <2e-16 \*\*\*

gB -7.36695 0.29662 -24.84 <2e-16 \*\*\*

x -0.36614 0.03028 -12.09 <2e-16 \*\*\*

gB:x 1.48092 0.04820 30.73 <2e-16 \*\*\*

In this situation, we could visually see that g has impact on both slope (opposite) and intercept. All coefficients in full model are significant, and the r square of full model are obviously higher than others. So we do not need to do F test here, just choose full model as the best.