

Sample Answer Lab03: Multivariate Regression

Part I: Partial Regression Coefficients (3.5 points)

The SPSS dataset **STATESCHOOL.SAV** evaluates the impact of the average state's expenditure (variable **EXPEND**) in primary education per student onto their average state-wide performance in the SAT test (variable **SAT**). A potentially confounding variable is the state-wide participation rate of students in the SAT test (variable **PCTSAT**). Note some states encourage the participation of all their students in the SAT test while in other states only good students with university aspirations take the SAT test; therefore, a selection bias can be expected. See Gruber (1999) at

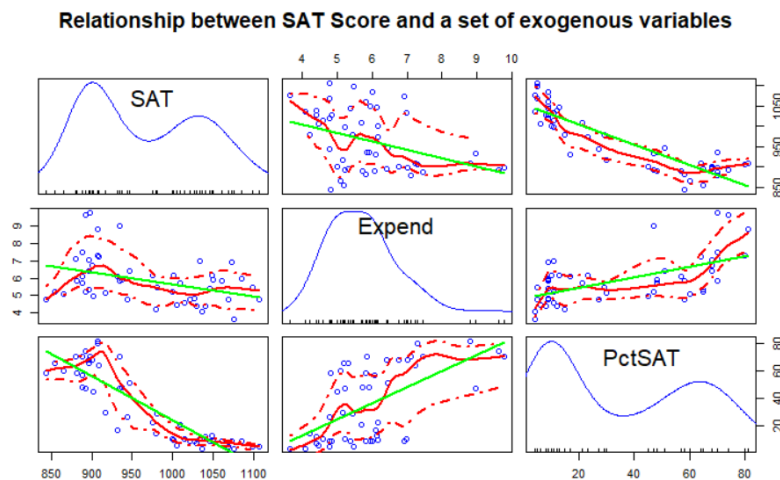
https://www.researchgate.net/publication/327474149_Getting_What_You_Pay_For_The_Debate_Over_Equity_in_Public_School_Expenditures for more information.

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

Task 1: Generate a scatterplot matrix of the three variables **SAT**, **EXPEND** and **PCTSAT**. Interpret their pairwise relationships. (0.5 points)

```
state_school <- foreign::read.spss('StateSchool.sav', to.data.frame = TRUE)

car::scatterplotMatrix(~SAT+Expend+PctSAT, data=state_school, main="Relationship between SAT
Score and a set of exogenous variables", pch=1, smooth=list(span = 0.35, lty.smooth=1,
col.smooth="red", col.var="red"), regLine=list(col="green"))
```



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**.

Use common sense arguments to **justify** your hypotheses. Explain, if you are unable to decide **a priori** into which the direction a particular independent variable may influence the dependent variable. (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)

$$[a] \text{ sat} = b_{0,\text{expend}} + b_{1,\text{expend}} \cdot \text{expend} + e_{\text{expend}} \quad \text{and}$$

$$[b] \text{ sat} = b_{0,\text{pctsat}} + b_{1,\text{pctsat}} \cdot \text{pctsat} + e_{\text{pctsat}}$$

Discuss the results in terms of the signs of the estimated regression coefficients $b_{1,\text{expend}}$ and $b_{1,\text{pctsat}}$.

<pre>lm1 <- lm(SAT~Expend,data = state_school) summary(lm1) lm(formula = SAT ~ Expend, data = state_school) Residuals:</pre>					<pre>lm2 <- lm(SAT~PctSAT,data = state_school) summary(lm2) lm(formula = SAT ~ PctSAT, data = state_school) Residuals:</pre>				
Min	1Q	Median	3Q	Max	Min	1Q	Median	3Q	Max
-145.074	-46.821	4.087	40.034	128.489	-79.158	-27.364	3.308	19.876	66.080

<p>Coefficients:</p> <table><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr><tr><td>(Intercept)</td><td>1089.294</td><td>44.390</td><td>24.539</td><td>< 2e-16 ***</td></tr><tr><td>Expend</td><td>-20.892</td><td>7.328</td><td>-2.851</td><td>0.00641 **</td></tr></table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 69.91 on 48 degrees of freedom</p> <p>Multiple R-squared: 0.1448, Adjusted R-squared: 0.127</p> <p>F-statistic: 8.128 on 1 and 48 DF, p-value: 0.006408</p>		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 **	<p>Coefficients:</p> <table><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr><tr><td>(Intercept)</td><td>1053.3204</td><td>8.2112</td><td>128.28</td><td><2e-16 ***</td></tr><tr><td>PctSAT</td><td>-2.4801</td><td>0.1862</td><td>-13.32</td><td><2e-16 ***</td></tr></table> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 34.89 on 48 degrees of freedom</p> <p>Multiple R-squared: 0.787, Adjusted R-squared: 0.7825</p> <p>F-statistic: 177.3 on 1 and 48 DF, p-value: < 2.2e-16</p>		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 ***
	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 **																											
	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 ***																											

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 $R^2 = 14\%$ 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 $R^2 = 0.78\%$ indicates a very good model fit.

Task 4: Perform a multiple regression analysis with the two independent variables

[c] $\text{sat} = b_0 + b_1 \cdot \text{expend} + b_2 \cdot \text{pctsat} + e$ and **fully interpret** the results. (1.0 point)

Pay in particular attention to any changes in the estimated slope parameters $b_{1,\text{expend}}$ and $b_{1,\text{pctsat}}$ of the bivariate models and the partial slope parameters b_1 and b_2 of the multiple regression model.

If a regression coefficient changes substantially, why may this be the case?

```
lm3 <- lm(SAT~Expend+PctSAT,data = state_school)
summary(lm3)
lm(formula = SAT ~ Expend + PctSAT, data = state_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: Plot the regression residuals e from the multiple model [4] against the independent variable PCTSAT (placed on the x-axis). (1.0 points)

Does the scatterplot indicate a deviation from what we would expect if the model would be properly specified? Would it be advisable to add the SAT participation rate square to the model, i.e., $\text{sat} = b_0 + b_1 \cdot \text{expend} + b_2 \cdot \text{pctsat} + b_3 \cdot I(\text{pctsat}^2) + e$?

Tips:

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

```

> preSAT <- predict(lm3)
> cor(preSAT, School$SAT)^2
[1] 0.8194726

```

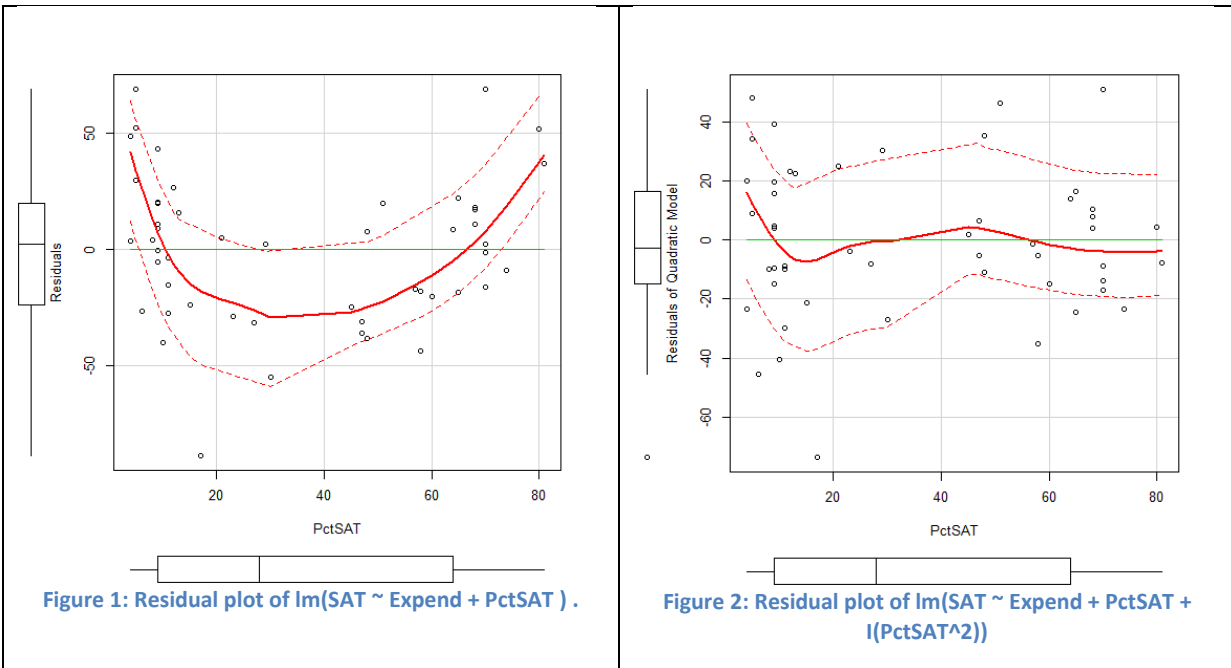
Comment: the squared correlation coefficient is equivalent to the multiple R^2 in model (c), which means how close the predicted values are to the observed values. In other words, it measures the proportion of variability in the dependent variable that can be explained by the independent variables.

Task 6: Plot the regression residuals e 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"))


```



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 $\text{lm}(\text{SAT} \sim \text{Expend} + \text{PctSAT} + \text{I}(\text{PctSAT}^2), \text{data} = \text{School})$

Part II: A Multiple Regression Model with Factors and Partial *F*-test (3.5 points)

You will find the data needed for this part in the zipped data file **Italy.zip** from Lab01. (read the attribute file **PROVINCES.DBF** into  with the function `foreign::read.dbf`)

You will experiment with regression models that aim at explaining the 1994 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 Sardinia, or in the southern, central or northern parts of Italy's mainland.

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

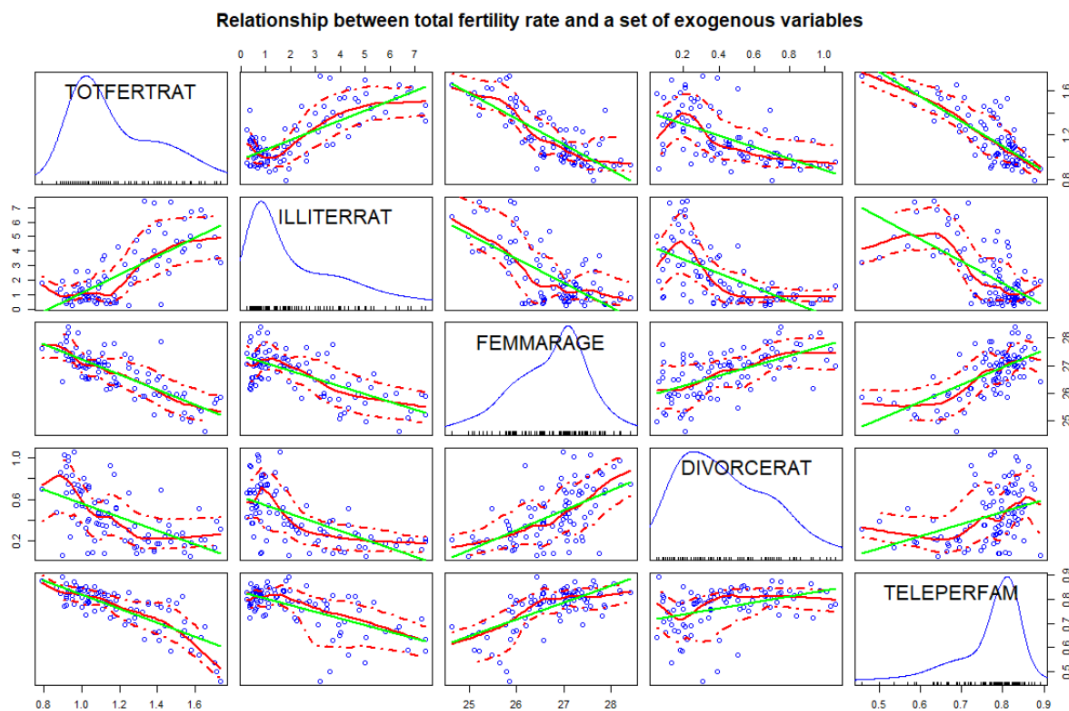
Task 6: Use common sense arguments **how** the four metric variables will influence the provincial fertility rates. Use one or two sentences per explanation and formulate one or two-sided null and alternative hypotheses based on your explanation. Format everything in a table. (0.5 points)

Variable	Common Sense Arguments	Statistical Hypotheses
ILLITERAT	A higher illiteracy rate leads to higher fertility rate due to lack of education.	$H_0: \beta \leq 0$ $H_1: \beta > 0$
FEMMARAGE	The latter a woman marries the lower will be her likelihood to have many children.	$H_0: \beta \geq 0$ $H_1: \beta < 0$
DIVORCERAT	A higher divorce rate leads to lower chance of having many children.	$H_0: \beta \geq 0$ $H_1: \beta < 0$
TELEPERFAM	An increased number of televisions will lead to more distractions and decreased fertility rate.	$H_0: \beta \geq 0$ $H_1: \beta < 0$

Task 7: Generate a scatterplot matrix showing the dependent variable and the four metric independent variables. Also generate a parallel boxplot of **TOTFERTRAT** ~ **REGION**. Briefly interpret the scatterplot matrix and the boxplot. (0.5 points)

```
provItaly <- foreign::read.dbf("provinces.dbf")
```

```
car::scatterplotMatrix(~TOTFERTRAT+ILLITERAT+FEMMARAGE+DIVORCERAT+TELEPERFAM,
  data=provItaly, main="Relationship between total fertility rate and a set
  of exogenous variables", pch=1, smooth=list(span = 0.35, lty.smooth=1,
  col.smooth="red", col.var="red"), regLine=list(col="green"))
```



Comments:

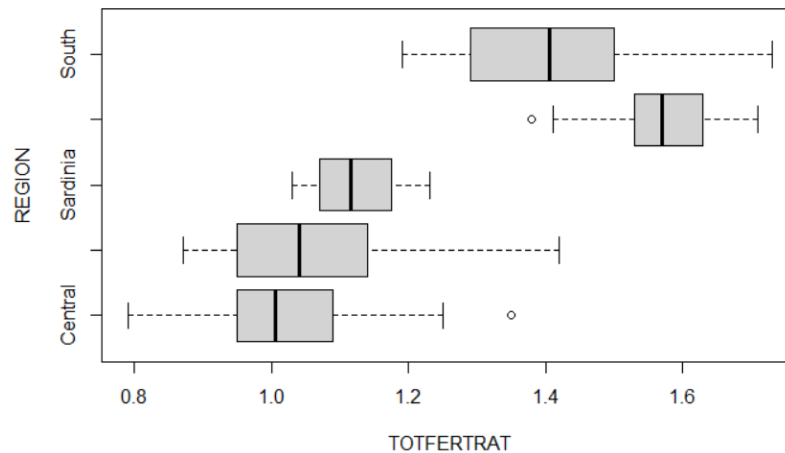
[a] Distributional characteristics: The distributions of the dependent variable and the four independent variables are unimodal. **TOTFERTRAT**, **DIVORCERAT**, and **ILLITERAT** are positively skewed, and **FEMMARAGE** and **TELEPERFAM** are negatively skewed.

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

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

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

```
boxplot(TOTFERTRAT~ REGION,data = provItaly,horizontal = TRUE)
```



A significant difference exists among the mean ratio of different regions. So we could assume TOTFERTRAT is related to variable REGION.

Task 8: 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 6, [b] the significances of the estimate regression coefficients and [c] the goodness of fit. (0.5 points)

```
lm1<- lm(TOTFERTRAT~ ILLITERRAT+ FEMMARAGE+ DIVORCERAT+ TELEPERFAM,
data=provItaly)
```

```
summary(lm1)
```

Call:

```
lm(formula = TOTFERTRAT ~ ILLITERRAT + FEMMARAGE + DIVORCERAT + TELEPERFAM,
data = provItaly)
```

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 ***
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 ***
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

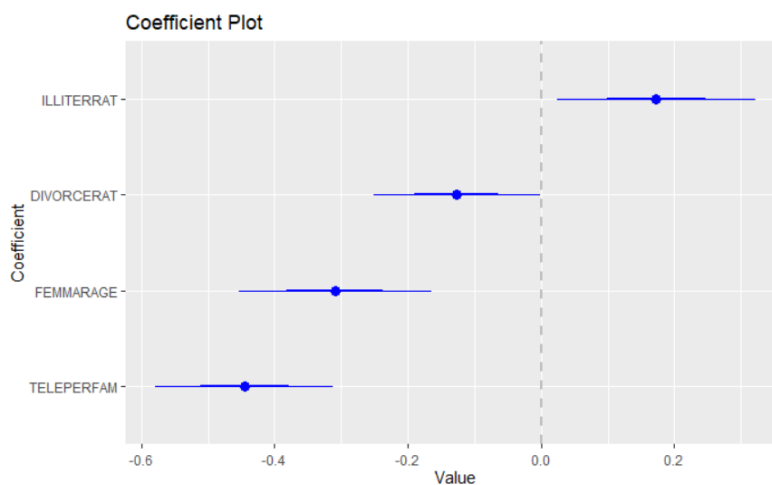
F-statistic: 95.69 on 4 and 90 DF, p-value: < 2.2e-16

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 $\alpha = 0.05$. 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 ($R_{adj}^2 = 0.8012$).

Task 9: Calculate the *beta-coefficients* for the multiple model in task 8. 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)

```
prov <- as.data.frame(scale(provItaly[,13:17]))
beta.lm <- lm(TOTFERTRAT~., data=prov)
coef(beta.lm)
coefplot::coefplot(beta.lm, sort="magnitude", intercept=F)
```

Variables	Coefficients (absolute value)	Rank
TELEPERFAM	0.44537	1
FEMMARAGE	0.30905	2
ILLITERRATE	0.17303	3
DIVORCERAT	0.12638	4



Task 10: Run the multiple regression model with the four metric variables plus the **REGION** factor to explain the variation of the fertility rates. (0.5 points)

From the perspective of interpreting the augmented total fertility rate model, which model, i.e., task 8 or task 10, is more informative?

```
lm2 <-
lm(TOTFERTRAT~FEMMARAGE+DIVORCERAT+ILLITERRAT+TELEPERFAM+REGION,data=provItaly)

summary(lm2)

Call:
lm(formula = TOTFERTRAT ~ FEMMARAGE + DIVORCERAT + ILLITERRAT +
    TELEPERFAM + REGION, data = provItaly)

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**, **ILLITERRAT**, and **TELEPERFAM** are no longer significant in this model. And the significance of **FEMMARAGE** also decreases dramatically. 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 multicollinearity is present in the independent variables.

Task 11: Use a partial *F*-test to check whether the model in task 10 has improved the model fit of the base model in task 8 significantly. (0.5 points)

That is, test the null hypothesis: $H_0: \beta_{Region\ 1} = \beta_{Region\ 2} = \dots = \beta_{Region\ J} = 0$ against the alternative hypothesis is $H_0: \beta_{Region\ j} \neq 0$ for at least one $j \in \{1, 2, \dots, J\}$.

```
anova(lm2, lm1)
```

Analysis of Variance Table

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

```
Model 2: TOTFERTRAT ~ ILLITERRAT + FEMMARAGE + DIVORCERAT + 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 **
```

Comment: The p -value (0.001743) is substantially smaller than 0.05, thus the null hypothesis can be rejected. We can conclude that the effect of the factor **REGION** is a significantly different from zero and the model in task 2.6 has improved the model fit of the base model in task 2.3 significantly.

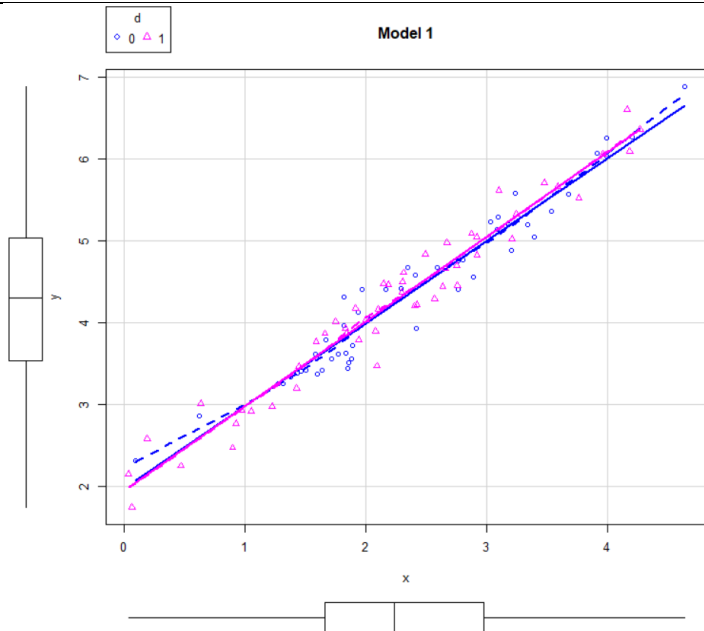
Part III. Identification of the Underlying Model Structure (1 point)

Use the five data-frames **df1** to **df5** in the workspace **DummyInteraction.RData** for this task (the workspace can be imported with the `load()` function). These data-frames comprise of three variables: **y** for the dependent variables, **d** for a binary dummy variable, and **x** for a *metric* variable. Each of these data-frames is best described by one of these competing models:

Name	Models Structure
Full interaction model	<code>mod.full <- lm(y~d*x, data=df)</code>
Different intercept model	<code>mod.int <- lm(y~d+x, data=df)</code>
Different slope model	<code>mod.slope <- lm(y~d:x, data=df)</code>
Just different means model	<code>mod.means <- lm(y~d, data=df)</code>
Plain bivariate model	<code>mod.plain <- lm(y~x, data=df)</code>

For each of the data-frame generate an informative scatterplot showing the regression regimes for both groups. You can employ the syntax `car::scatterplot(y~x|d, data=df)`. Then identify which of the competing model structures best describes the given datasets.

Task 12: Plot the data and identify the underlying model structure for **df1**. (0.2 points)



Plain regression model

```
lm(formula = y ~ x, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.62406	-0.18763	0.00383	0.18842	0.49935

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.95310	0.05955	32.80	<2e-16 ***
x	1.02206	0.02374	43.05	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.239 on 98 degrees of freedom

Multiple R-squared: 0.9498, Adjusted R-squared: 0.9493

F-statistic: 1853 on 1 and 98 DF, p-value: < 2.2e-16

Means model

```
lm(formula = y ~ d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.48922	-0.77859	0.01215	0.74824	2.50984

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.3735	0.1505	29.067	<2e-16 ***
d	-0.1443	0.2128	-0.678	0.499

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.064 on 98 degrees of freedom

Multiple R-squared: 0.004671, Adjusted R-squared: -0.005486

F-statistic: 0.4599 on 1 and 98 DF, p-value: 0.4993

Intercept model

```
lm(formula = y ~ x + d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.64122	-0.19228	0.01645	0.18365	0.51755

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.93215	0.06629	29.147	<2e-16 ***
x	1.02357	0.02389	42.847	<2e-16 ***
d	0.03495	0.04810	0.727	0.469

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2396 on 97 degrees of freedom

Multiple R-squared: 0.95, Adjusted R-squared: 0.949

F-statistic: 922.5 on 2 and 97 DF, p-value: < 2.2e-16

Full interaction model

```
lm(formula = y ~ x * d, data = df)
```

Residuals:

Slope model

```
lm(formula = y ~ x:d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.63997	-0.18149	0.01431	0.18177	0.51019
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96324	0.09200	21.339	<2e-16 ***
x	1.01054	0.03584	28.195	<2e-16 ***
d	-0.01949	0.12128	-0.161	0.873
x:d	0.02360	0.04823	0.489	0.626

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.2405 on 96 degrees of freedom				
Multiple R-squared: 0.9502, Adjusted R-squared: 0.9486				
F-statistic: 610.2 on 3 and 96 DF, p-value: < 2.2e-16				

Min	1Q	Median	3Q	Max
-2.2720	-0.5914	-0.1298	0.6710	2.8897
Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.99366	0.12969	30.795	< 2e-16 ***
x:d	0.27843	0.07491	3.717	0.000336 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.9984 on 98 degrees of freedom				
Multiple R-squared: 0.1236, Adjusted R-squared: 0.1146				
F-statistic: 13.82 on 1 and 98 DF, p-value: 0.0003357				

Comment: Except the mean model, the relatively R_{adj}^2 values of the rest four models are very close. After conducting the nested partial F -test, we can conclude the plain regression model is not significantly different from the other three models due to such large p -value. Based on the parsimony rule, we should choose the **plain regression** model for df_1

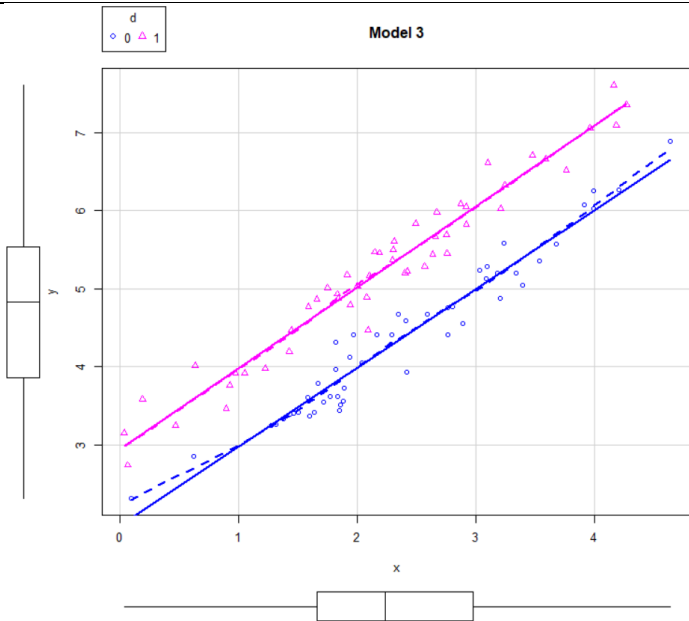
Task 13: Plot the data and identify the underlying model structure for df_2 . (0.2 points)

<p>Model 2</p>	<h3>Plain regression model</h3> <pre>lm(formula = y ~ x, data = df)</pre> <p>Residuals:</p> <table><tr><th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr><tr><td>-1.78290</td><td>-0.55348</td><td>0.04286</td><td>0.64134</td><td>1.95617</td></tr></table> <p>Coefficients:</p> <table><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr><tr><td>(Intercept)</td><td>2.48718</td><td>0.20968</td><td>11.862</td><td>< 2e-16 ***</td></tr><tr><td>x</td><td>0.52628</td><td>0.08359</td><td>6.296</td><td>8.65e-09 ***</td></tr></table> <p>---</p> <p>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p> <p>Residual standard error: 0.8415 on 98 degrees of freedom</p> <p>Multiple R-squared: 0.288, Adjusted R-squared: 0.2807</p> <p>F-statistic: 39.64 on 1 and 98 DF, p-value: 8.652e-09</p>	Min	1Q	Median	3Q	Max	-1.78290	-0.55348	0.04286	0.64134	1.95617		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	2.48718	0.20968	11.862	< 2e-16 ***	x	0.52628	0.08359	6.296	8.65e-09 ***
Min	1Q	Median	3Q	Max																						
-1.78290	-0.55348	0.04286	0.64134	1.95617																						
	Estimate	Std. Error	t value	Pr(> t)																						
(Intercept)	2.48718	0.20968	11.862	< 2e-16 ***																						
x	0.52628	0.08359	6.296	8.65e-09 ***																						
<h3>Means model</h3> <pre>lm(formula = y ~ d, data = df)</pre>	<h3>Intercept model</h3> <pre>lm(formula = y ~ x + d, data = df)</pre>																									

<pre> Residuals: Min 1Q Median 3Q Max -2.05916 -0.32837 0.02214 0.28399 2.50984 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 4.3735 0.1026 42.618 < 2e-16 *** d -1.3543 0.1451 -9.332 3.42e-15 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.7256 on 98 degrees of freedom Multiple R-squared: 0.4705, Adjusted R-squared: 0.4651 F-statistic: 87.08 on 1 and 98 DF, p-value: 3.418e-15 </pre>	<pre> Residuals: Min 1Q Median 3Q Max -1.04338 -0.47360 0.06784 0.36295 1.44879 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 3.24929 0.15198 21.380 < 2e-16 *** x 0.47134 0.05477 8.606 1.36e-13 *** d -1.27175 0.11026 -11.534 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.5492 on 97 degrees of freedom Multiple R-squared: 0.6997, Adjusted R-squared: 0.6936 F-statistic: 113 on 2 and 97 DF, p-value: < 2.2e-16 </pre>
<p>Full interaction model</p> <pre> lm(formula = y ~ x * d, data = df) Residuals: Min 1Q Median 3Q Max -0.63997 -0.18149 0.01431 0.18177 0.51019 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 1.96324 0.09200 21.339 < 2e-16 *** x 1.01054 0.03584 28.195 < 2e-16 *** d 0.98051 0.12128 8.085 1.88e-12 *** x:d -0.97640 0.04823 -20.244 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.2405 on 96 degrees of freedom Multiple R-squared: 0.943, Adjusted R-squared: 0.9412 F-statistic: 529.6 on 3 and 96 DF, p-value: < 2.2e-16 </pre>	<p>Slope model</p> <pre> lm(formula = y ~ x:d, data = df) Residuals: Min 1Q Median 3Q Max -1.8357 -0.5551 -0.1056 0.4438 2.7333 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 4.14998 0.10782 38.491 < 2e-16 *** x:d -0.41054 0.06228 -6.592 2.19e-09 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.83 on 98 degrees of freedom Multiple R-squared: 0.3072, Adjusted R-squared: 0.3002 F-statistic: 43.46 on 1 and 98 DF, p-value: 2.194e-09 </pre>

Comment: The R_{adj}^2 of the full interaction model is much higher than the other models, so we should choose the full interaction model for `df2`.

Task 14: Plot the data and identify the underlying model structure for `df3`. (0.2 points)



Plain regression model

```
lm(formula = y ~ x, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.99644	-0.50395	-0.04568	0.52658	1.02657

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.55236	0.14271	17.89	<2e-16 ***
x	0.97886	0.05689	17.21	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5727 on 98 degrees of freedom

Multiple R-squared: 0.7513, Adjusted R-squared: 0.7488

F-statistic: 296 on 1 and 98 DF, p-value: < 2.2e-16

Means model

```
lm(formula = y ~ d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.48922	-0.77859	0.01215	0.74824	2.50984

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.3735	0.1505	29.067	< 2e-16 ***
d	0.8557	0.2128	4.021	0.000114 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.064 on 98 degrees of freedom

Multiple R-squared: 0.1416, Adjusted R-squared: 0.1329

F-statistic: 16.17 on 1 and 98 DF, p-value: 0.0001138

Intercept model

```
lm(formula = y ~ x + d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.64122	-0.19228	0.01645	0.18365	0.51755

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.93215	0.06629	29.15	<2e-16 ***
x	1.02357	0.02389	42.85	<2e-16 ***
d	1.03495	0.04810	21.52	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2396 on 97 degrees of freedom

Multiple R-squared: 0.9569, Adjusted R-squared: 0.956

F-statistic: 1077 on 2 and 97 DF, p-value: < 2.2e-16

Full interaction model

```
lm(formula = y ~ x * d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

Slope model

```
lm(formula = y ~ x:d, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

```
-0.63997 -0.18149 0.01431 0.18177 0.51019
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96324	0.09200	21.339	< 2e-16 ***
x	1.01054	0.03584	28.195	< 2e-16 ***
d	0.98051	0.12128	8.085	1.88e-12 ***
x:d	0.02360	0.04823	0.489	0.626

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2405 on 96 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9557

F-statistic: 712.7 on 3 and 96 DF, p-value: < 2.2e-16

```
-1.8357 -0.5551 -0.1056 0.4438 2.7333
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.14998	0.10782	38.491	< 2e-16 ***
x:d	0.58946	0.06228	9.465	1.75e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

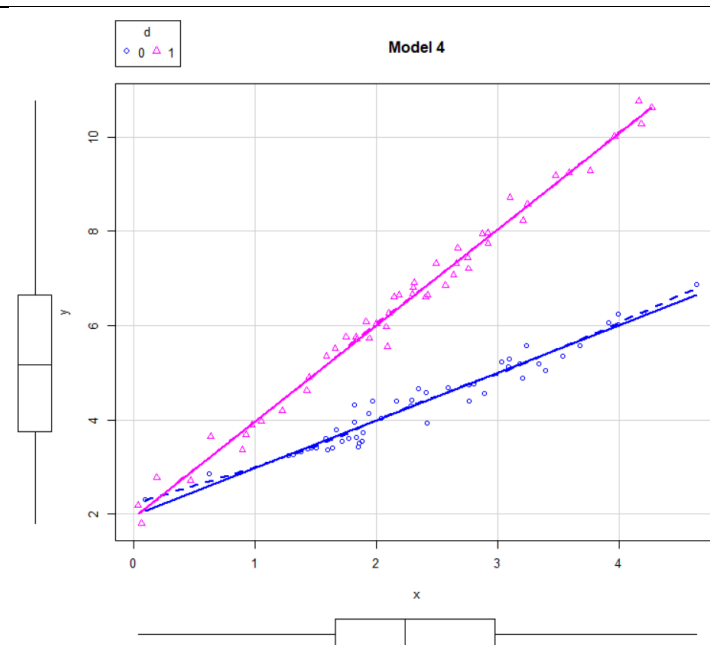
Residual standard error: 0.83 on 98 degrees of freedom

Multiple R-squared: 0.4776, Adjusted R-squared: 0.4723

F-statistic: 89.59 on 1 and 98 DF, p-value: 1.753e-15

Comment: The intercept and full interaction models have relatively the highest R^2_{adj} values among all models. After conducting the nested partial F -test, we can conclude the intercept model is not significantly different from the full interaction model. Based on the parsimony rule, we should choose the **intercept model** for $df3$.

Task 15: Plot the data and identify the underlying model structure for **df4**. (0.2 points)



Plain regression model

```
lm(formula = y ~ x, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.98402	-1.10130	-0.03571	1.14695	2.61302

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.0183	0.3259	6.193	1.39e-08 ***
x	1.4746	0.1299	11.349	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.308 on 98 degrees of freedom

Multiple R-squared: 0.5679, Adjusted R-squared: 0.5635

F-statistic: 128.8 on 1 and 98 DF, p-value: < 2.2e-16

Means model

```
lm(formula = y ~ d, data = df)
```

Residuals:

Intercept model

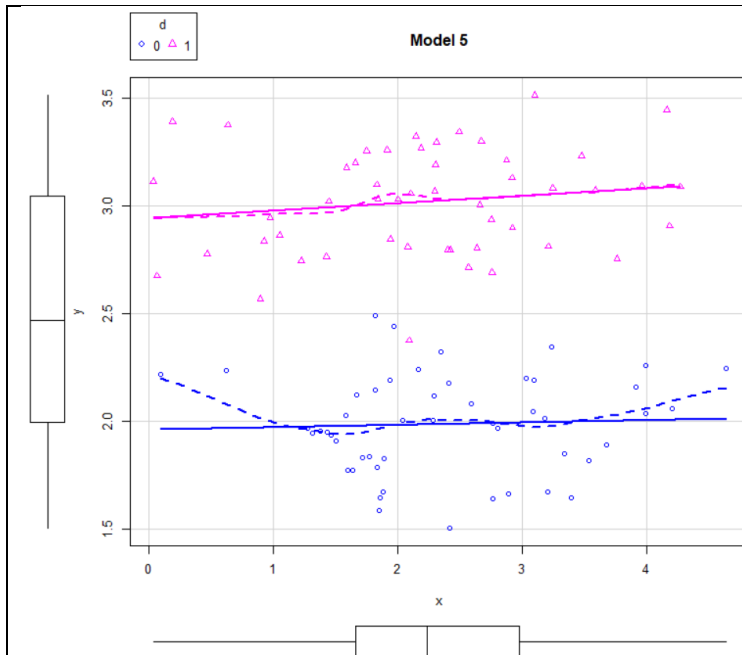
```
lm(formula = y ~ x + d, data = df)
```

Residuals:

<pre> Min 1Q Median 3Q Max -4.6336 -0.8897 0.0346 0.8869 4.3318 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 4.3735 0.2396 18.252 < 2e-16 *** d 2.0657 0.3389 6.096 2.15e-08 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 1.694 on 98 degrees of freedom Multiple R-squared: 0.2749, Adjusted R-squared: 0.2675 F-statistic: 37.16 on 1 and 98 DF, p-value: 2.154e-08 </pre>	<pre> Min 1Q Median 3Q Max -1.25447 -0.36820 0.01613 0.41469 1.54886 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 0.61502 0.15795 3.894 0.000181 *** x 1.57580 0.05692 27.683 < 2e-16 *** d 2.34164 0.11460 20.433 < 2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.5708 on 97 degrees of freedom Multiple R-squared: 0.9185, Adjusted R-squared: 0.9169 F-statistic: 546.9 on 2 and 97 DF, p-value: < 2.2e-16 </pre>
<p>Full interaction model</p> <pre> lm(formula = y ~ x * d, data = df) Residuals: Min 1Q Median 3Q Max -0.63997 -0.18149 0.01431 0.18177 0.51019 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 1.96324 0.09200 21.339 <2e-16 *** x 1.01054 0.03584 28.195 <2e-16 *** d -0.01949 0.12128 -0.161 0.873 x:d 1.02360 0.04823 21.223 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.2405 on 96 degrees of freedom Multiple R-squared: 0.9857, Adjusted R-squared: 0.9852 F-statistic: 2204 on 3 and 96 DF, p-value: < 2.2e-16 </pre>	<p>Slope model</p> <pre> lm(formula = y ~ x:d, data = df) Residuals: Min 1Q Median 3Q Max -2.2720 -0.5914 -0.1298 0.6710 2.8897 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 3.99366 0.12969 30.80 <2e-16 *** x:d 1.27843 0.07491 17.07 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.9984 on 98 degrees of freedom Multiple R-squared: 0.7483, Adjusted R-squared: 0.7457 F-statistic: 291.3 on 1 and 98 DF, p-value: < 2.2e-16 </pre>

Comment: The slope and full interaction models have relatively the highest R_{adj}^2 values among all models. The d variable is not significant in the full interaction model. Based on the parsimony rule, we should choose the **Intercept model** for df4.

Task 16: Plot the data and identify the underlying model structure for df5. (0.2 points)



Plain regression model

```
lm(formula = y ~ x, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.99644	-0.50395	-0.04568	0.52658	1.02657

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.55236	0.14271	17.885	<2e-16 ***
x	-0.02114	0.05689	-0.372	0.711

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5727 on 98 degrees of freedom

Multiple R-squared: 0.001408, Adjusted R-squared: -0.008782

F-statistic: 0.1381 on 1 and 98 DF, p-value: 0.7109

Means model

```
lm(formula = y ~ d, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.64402	-0.18766	0.01327	0.18953	0.50425

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.98837	0.03388	58.70	<2e-16 ***
d	1.03082	0.04791	21.52	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2395 on 98 degrees of freedom

Multiple R-squared: 0.8253, Adjusted R-squared: 0.8235

F-statistic: 463 on 1 and 98 DF, p-value: < 2.2e-16

Intercept model

```
lm(formula = y ~ x + d, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.64122	-0.19228	0.01645	0.18365	0.51755

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.93215	0.06629	29.147	<2e-16 ***
x	0.02357	0.02389	0.987	0.326
d	1.03495	0.04810	21.518	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2396 on 97 degrees of freedom

Multiple R-squared: 0.827, Adjusted R-squared: 0.8235

F-statistic: 231.9 on 2 and 97 DF, p-value: < 2.2e-16

Full interaction model

```
lm(formula = y ~ x * d, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.63997	-0.18149	0.01431	0.18177	0.51019

Slope model

```
lm(formula = y ~ x:d, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.63295	-0.24898	-0.03704	0.22224	1.18991

Coefficients:					Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)		Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96324	0.09200	21.339	< 2e-16 ***	(Intercept)	2.13771	0.04675	45.73	<2e-16 ***
x	0.01054	0.03584	0.294	0.769	x:d	0.33128	0.02700	12.27	<2e-16 ***
d	0.98051	0.12128	8.085	1.88e-12 ***	---				
x:d	0.02360	0.04823	0.489	0.626	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					Residual standard error: 0.3599 on 98 degrees of freedom				
Residual standard error: 0.2405 on 96 degrees of freedom					Multiple R-squared: 0.6057, Adjusted R-squared: 0.6017				
Multiple R-squared: 0.8275, Adjusted R-squared: 0.8221					F-statistic: 150.5 on 1 and 98 DF, p-value: < 2.2e-16				
F-statistic: 153.5 on 3 and 96 DF, p-value: < 2.2e-16									

Comment: The mean and the intercept model have similar R_{adj}^2 -values, however, the slope coefficient of the intercept model is not significant. Based on the parsimony rule, we should choose the **Mean model** for df5.