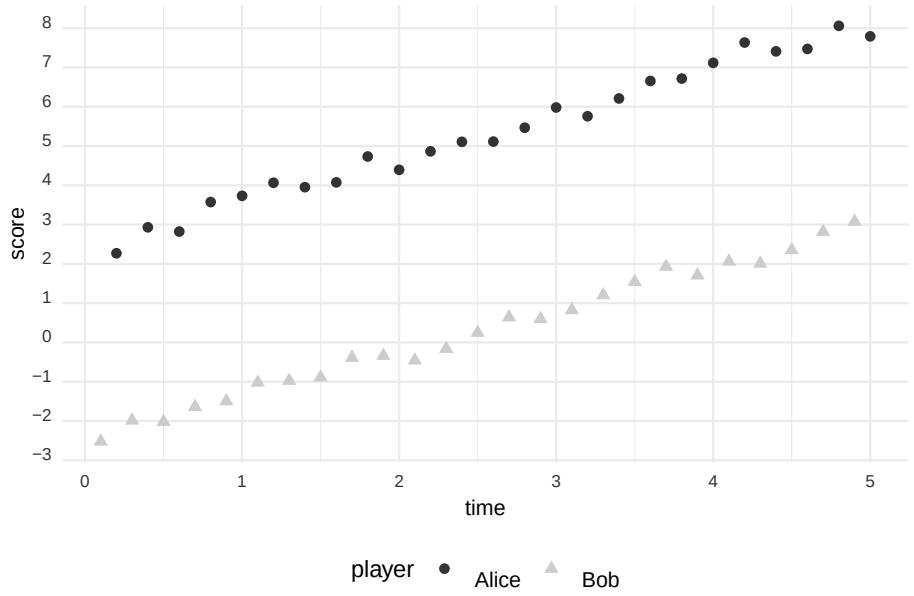

Figure 1: Scores of Alice and Bob as a function of the number of hours of play.

2.1 Bob and Alice decide to play board games during the Pandemic. They notice their scores (as a function of the number of hours they play) could be modelled using a linear model of the form

$$\texttt{score}_i = \beta_0 + \beta_1 \texttt{time}_i + \beta_2 \texttt{player}_i + \beta_3 \texttt{time}_i \texttt{player}_i + \varepsilon_i,$$

where ε_i is a mean-zero error term and $\texttt{player}_i$ is a binary indicator equal to 1 if the ith score belongs to Alice and 0 if it belongs to Bob.
In view of Figure 1, what can we say about the sign of $\widehat{\beta}_1, \ldots, \widehat{\beta}_3$?

2.2 We consider a regression model to explain the impact of education and the number of children on the salary of women, viz.

$$\texttt{logsalary}_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \beta_2 \mathrm{X}_{i2} + \varepsilon_i,$$

where

$$\mathrm{X}_1 = \begin{cases} 0, & \text{if the woman did not complete high school,} \\ 1, & \text{if the woman completed high school, but not college,} \\ -1, & \text{if the woman completed college.} \end{cases}$$

$$\mathrm{X}_2 = \begin{cases} 0, & \text{if the woman has no children,} \\ 1, & \text{if the woman has 1 or 2 children,} \\ -1, & \text{if the woman has 3 or more children.} \end{cases}$$

According to the model, what would be the mean **difference** in log-salary between (i) a woman who completed college and has three children and (ii) the average log-salary of all women in the sample, assuming the sub-sample size in each of the nine group is the same (balanced design)?

2.3 We consider log of housing price as a function of location (urban or rural), whether or not the house includes a

Compiled 06/08/2020 at 12:13

garage and the surface of the latter (in square feet). The postulated linear model is

$$\texttt{logprice} = \beta_0 + \beta_1 \texttt{garage} + \beta_2 \texttt{area} + \beta_3 \mathbf{1}_{\texttt{loc=urban}} + \varepsilon,$$

where ε is a mean-zero error term and `garage` is an indicator variable,

$$\texttt{garage} = \begin{cases} 0, & \text{if the house has a garage } (\texttt{area} > 0); \\ 1, & \text{if the house doesn't have a garage } (\texttt{area} = 0). \end{cases}$$

Suppose we fit the model via least squares and find $\widehat{\beta}_1 > 0$ and $\widehat{\beta}_2 > 0$. Which of the following statement is **always** correct?

(a) Everything else being equal, houses with garages are on average more expensive than ones without a garage.
(b) Everything else being equal, houses with garages are always less expensive than ones without a garage.
(c) Everything else being equal, houses with garages are on average cheaper than ones without a garage.
(d) Location (urban versus rural) is negatively correlated with garage surface.
(e) None of the above

2.4 We consider a simple linear regression model for the price of an electric car as a function of its autonomy (distance); the model is

$$\texttt{price}^{\texttt{USD}} = \beta_0^i + \beta_1^i \texttt{distance}^{\texttt{mi.}} + \varepsilon^i,$$

where ε is a zero-mean error term. Your friends collects some data in which the price is expressed in American dollars (`USD`) and distance is measured in miles (`mi.`) and run the regression to get estimates $(\widehat{\beta_0^i}, \widehat{\beta_1^i})$.
You would like to know the estimates for the model with the price expressed in Canadian dollars (`CAD`) and distance expressed in kilometers (`km`), i.e.,

$$\texttt{price}^{\texttt{CAD}} = \beta_0^m + \beta_1^m \texttt{distance}^{\texttt{km}} + \varepsilon^m.$$

Knowing that 1 USD is 1.39 CAD and that 1 mile is 1.61km, what is the value of C in the equation $\widehat{\beta_1^i} = C\widehat{\beta_1^m}$?

2.5 The dataset `windturbine` contains measurements of electricity output of wind turbine over 25 separate fifteen minute periods. We are interested in the relation between direct output and the average wind speed (measured in miles per hour) during the recording.

(a) Fit a linear model with wind speed as covariate and plot the standardized residuals against the fitted values. Do you notice any residual structure missed by the model mean specification? Try fitting a model using the reciprocal of wind speed as covariate. Comment on the adequacy of the models.
(b) Predict, using both models in turn, the output of electricity given that the average wind speed in a given period is 5 miles per hour. Provide prediction intervals for your estimates.
(c) [⋆] The electricity production should be zero if there is no wind, yet this is not captured by the model linking output to velocity. Update your model to remove the intercept (in `prog reg`, using the option `/noint`). What are the consequences of the latter?
(d) Produce a quantile-quantile plot of the residuals and comment on the normality assumption.

2.6 In a study performed at Tech3Lab, subjects navigated a website that contained, among other things, an advertisement for candies. During the site navigation, an "eye-tracker" measured the location on the screen on which the subject's eyes were fixated and also recorded whether the subject saw the ad and for how long it was in sight. Additionally, facial expression analysis software (FaceReader) can be used to guess the subject's emotions when the ad was in sight. At the end of the study, a questionnaire measured the subject's intention to buy this type of candy and socio-demographic variables. Only the 120 subjects that had seen the ad in question are included in the data

`intention`, which contains the following variables The study objective is to evaluate whether there is a link between the duration of fixation on the advertisement and the intention to buy and whether perceived emotion is linked to the intention to buy.

- `intention`: discrete variable ranging between 2 and 14; larger values indicate higher interest in buying the product. Specifically, the score was constructed by summing the response of two questions, both measured using a Likert scale ranging from strongly disagree (1) to strongly agree (7).
- `fixation`: the total duration of fixation on the ad (in seconds).
- `emotion`: a measure of reaction during fixation; the ratio of the probability of showing a positive emotion to the probability of showing a negative emotion.
- `sex`: sex of subject, either man (0) or woman (1).
- `age`: age (in years).
- `revenue`: categorical variable indicating the subject's annual income; one of (1) $[0, 20\,000]$; (2) $[20\,000, 60\,000]$; (3) 60 000 and above.
- `educ`: categorical variable indicating the highest educational achievement, either (1) high school or lower; (2) college or (3) university degree.
- `marital`: civil status, either single (0) or in a relationship (1).

We will perform a linear regression analysis to evaluate the impact of the variable `revenue` on `intention`.

(a) Adjust the model by creating binary indicators that you will include in the model, using category 3 as baseline. Write down formally the model you are fitting and interpret the model coefficients.

(b) For the model fitted in the part a), what is the predicted `intention` for an individual with a revenue superior to 60 000.

(c) Fit the model by specifying that the `revenue` variable is categorical (using the command `class` in SAS , or `as.factor` in R). Write down the regression equation and interpret the coefficients.

(d) Fit the model one last time by treating `revenue` as a continuous variable. Compare the results and hence comment on the conceptual difference between categorical and continuous variables.

(e) Fit a regression model to buying intention using all the other variables in the database as covariates and interpret the effect of the latter.

(f) Test the global effect of the variables `revenue` and `educ` conditional on the other variables in the model.

2.7 The dataset `auto` contains measurements for various characteristics of 392 cars. Consider a linear model linking fuel consumption (in miles per gallon) of cars as a function of horsepower (in watts).

(a) Draw a scatterplot to assess graphically the relation between fuel consumption (`mpg`) and horsepower (`horsepower`) and comment.

(b) Fit a simple linear regression and comment on the results.

(c) Fit the quadratic model

$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{horsepower} + \beta_2 \texttt{horsepower}^2 + \varepsilon$$

and comment on the model fit and the statistical significance of the coefficients. In SAS , the model can be fitted using the following code:

```
proc glm data=infe.auto;
model mpg=horsepower horsepower*horsepower/ss3 solution;
run;
```

and in R , using

```
lm(mpg~horsepower+I(horsepower^2), data = auto)
```

Do an analysis of residuals for the linear and the quadratic models and compare the two.

Compiled 06/08/2020 at 12:13

line	`fixation`	`educ`	**Fitted average buying intention**
1	x	1	
2	x	2	
3	x	3	
4	$x+1$	1	
5	$x+1$	2	
6	$x+1$	3	
7	0	1	
8	0	2	
9	0	3	

Table 1: Fitted values for the buying intention for nine scenarios

(d) Repeat the previous question, this time with a cubic model. Conclude as to which model is the most appropriate for the data.

2.8 **Interactions between continuous and categorical variables** We covered in class the modelling and interpretation of interactions between binary and continuous variables. The goal of this exercise is to explain how to fit and interpret a model including an interaction between a **categorical** variable and a continuous variable. We will work with the `intention` data, but will only use the variables `educ` and `fixation` to model `intention`. Recall that `educ` has three levels and will be modelled with two binary indicators, `educ1` and `educ2`. The variable `educ1` (respectively `educ2`) is one if `educ=1` (`educ=2`) and zero otherwise.

(a) Fit a linear regression model including `educ` and `fixation`, without interaction. Use the third category of `educ` as baseline.

i. Write down the equation corresponding to the fitted model.

ii. Write down the equation describing the linear relation between `intention` and `fixation` when `educ=1`, `educ=2` and `educ=3`, respectively.

iii. The SAS graphical output shows the effect of `fixation` on `intention` as a function of the three education groups (color coded). What do you think of the goodness-of-fit? According to you, which characteristic if any should be included in the model?

(b) Include an interaction term between `educ` and `fixation` to the model. If you use `SAS`, make sure to do this directly using the command `class`. The fitted model is

$$\texttt{intention} = \beta_0 + \beta_1 \texttt{educ1} + \beta_2 \texttt{educ2} + \beta_3 \texttt{fixation} + \beta_4 \texttt{educ1} \times \texttt{fixation} + \beta_5 \texttt{educ2} \times \texttt{fixation} + \varepsilon \tag{E1}$$

i. Write down the equation for the fitted values. Are the interaction terms significant?

ii. Calculate the equation of the linear relation between `intention` and `fixation` when `educ=1`, `educ=2` and `educ=3`, respectively.

iii. Looking at the graph produced by SAS showing the effect of `fixation` on `intention` as a function of the three education groups (color coded), compare the fit of the model with and without interaction and comment.

(c) The next part pertains to interpretation of the model coefficients in the presence of an interaction.

i. Fill the rightmost column of Table 1 with the fitted buying intention for the nine given scenarios.

ii. Using lines 3 and 6 of Table 1, interpret the coefficient β_3 from model (slope of `fixation`).

iii. Using lines 7 and 9 of Table 1, interpret the coefficient β_1 from (E1) (slope of the variable `educ1`).

iv. Using lines 8 and 9 of Table 1, interpret the coefficient β_2 from (E1) (slope of the variable `educ2`).

2.9 The time series `airpassengers` provides the monthly totals of international airline passengers from 1949 to 1960 (in thousands).

(a) Fit a linear model linking the number of passengers to years. What is the interpretation of the intercept and of the trend? Consider a model in which the time covariate is shifted by 1949, i.e., $t-1949$. How does this affect the interpretation of the coefficients?

(b) Consider adding a monthly effect using binary indicators; write down the equation of the model. Do you notice a significant improvement in fit?

(c) Use the model with seasonal and linear effects to predict the number of passengers in December 1962.

(d) Provide diagnostic plots for the fit. What do you notice?

(e) There may be evidence that the growth in aerial traffic is exponential during the time period under study. Try fitting a linear model with the log of the number of passengers as response. Produce the following diagnostic: (1) a scatterplot of fitted values versus ordinary residuals (2) a scatterplot of studentized residuals against time (3) a quantile-quantile plot of the jackknife studentized residuals and (4) a scatterplot of lagged residuals, i.e. plot e_i against e_{i+1} for $i = 1,\ldots,n-1$. Is any of the hypothesis of the linear model seemingly violated?

2.10 The `Ratemyprofessor` data set provides the ratings of 366 instructors (159 women, 207 men) at one large campus in the Midwest. Each instructor included in the data had at least 10 ratings over several years. Students provided numerical ratings on a scale of 5: `helpfulness`, `clarity` and `easiness` are average rating for sub-categories taking values in $[1,5]$, and range between 1 (worst) to 5 (best). The data provide these average ratings and additional characteristics of the instructors. The goal is to predict the overall quality rating from the available covariates. Table 2 provides the summary of different linear models fitted to the data.

(a) Give the average overall quality rating for women faculty in the sample.

(b) Based on Model 8, predict the average overall quality rating for a men whose `helpfulness`, `clarity` and `easiness` ratings are all equal to 4.

(c) What are the null and alternative hypotheses associated to the F statistic with value 62228.971 in Model 4? Give the conclusion of that hypothesis test. *Hint: the 95% of the null distribution is 3.021.*

(d) Give an approximate 95% confidence interval for the parameter `clarity` in Model 4, of the form $\widehat{\beta}_j \pm t_{\nu,1-\alpha/2}\mathsf{se}(\widehat{\beta}_j)$. Is Model 2 an adequate simplification of Model 4?

(e) Contrast the estimated coefficient values of Model 2 and Model 4. Are the finding consistent with Figure 2?

(f) Explain why one should not consider Model 7, regardless of whether the coefficient associated to the interaction `men:helpfulness` is statistically significant.

(g) What are the assumptions of the multiple linear model? Comment on the appropriateness of these assumptions based on the diagnostic plots presented in Figures 2 and 3.

	Model 1	Model 2	Model 3	Model 4
`constant`	3.532 (0.066)	0.033 (0.038)	0.221 (0.040)	−0.020 (0.011)
`men (sex)`	0.077 (0.088)			
`helpfulness`		0.975 (0.010)		0.538 (0.007)
`clarity`			0.952 (0.011)	0.466 (0.007)
R^2	0.002	0.962	0.952	0.997
degrees of freedom	364	364	364	363
F statistic (global significance)	0.755	9322.673	7299.061	62228.971
residual sum of squares (RSS)	255.479	9.620	12.161	0.745
AIC	913.088	-287.129	-201.361	-1221.679

	Model 5	Model 6	Model 7	Model 8
`constant`	−0.029 (0.011)	−0.030 (0.012)	0.323 (0.057)	−0.054 (0.016)
`men (sex)`		0.002 (0.005)	−0.397 (0.076)	0.048 (0.021)
`helpfulness`	0.536 (0.007)	0.535 (0.007)		0.541 (0.008)
`clarity`	0.465 (0.007)	0.465 (0.007)	0.863 (0.016)	0.466 (0.007)
`easiness`	0.007 (0.004)	0.007 (0.004)	0.062 (0.014)	0.007 (0.004)
`men:helpfulness`			0.116 (0.020)	−0.013 (0.006)
R^2	0.997	0.997	0.959	0.997
degrees of freedom	362	361	361	360
F statistic (global significance)	41739.797	31236.209	2107.165	25272.111
residual sum of squares (RSS)	0.738	0.738	10.515	0.727
AIC	-1222.912	-1221.120	-248.592	-1224.244

Table 2: Coefficients and standard errors (in parenthesis) for the coefficients of linear models for the `Ratemyprofessor` data, along with goodness-of-fit statistics.

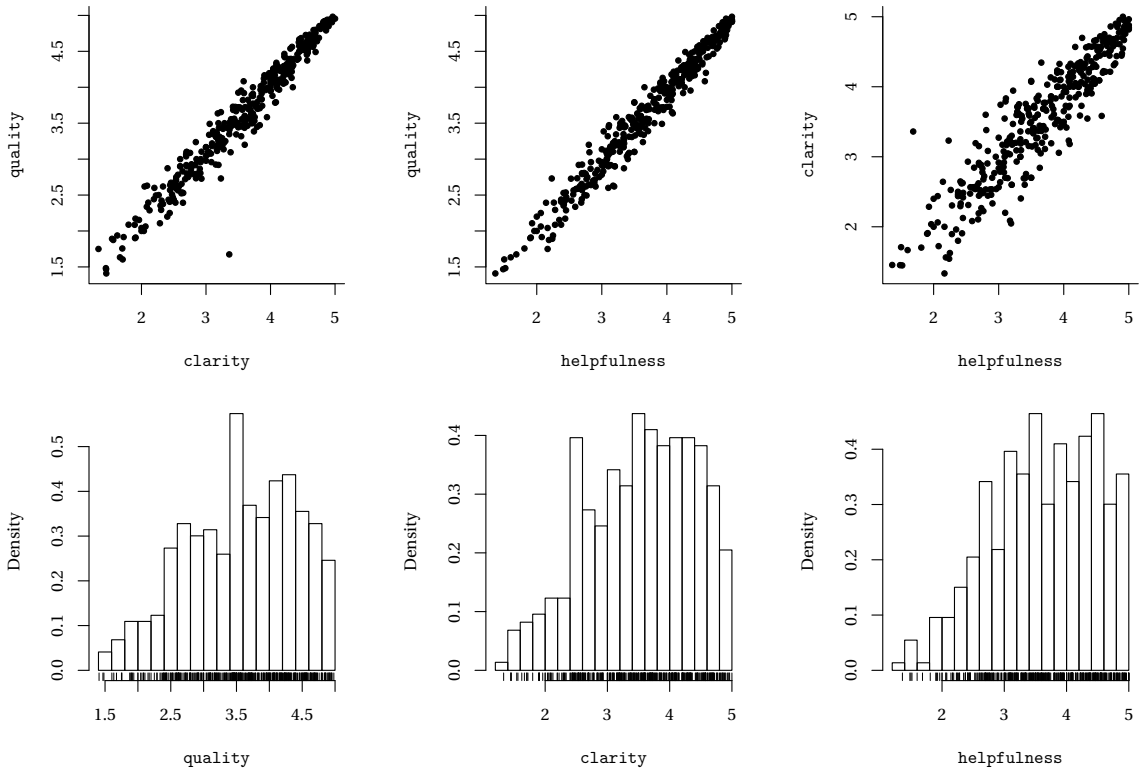

Figure 2: Top: pair plots (linear correlation from left to right of 0.98,0.98, 0.92). Bottom: histograms of average `quality`, `helpfulness` and `clarity` indicators.

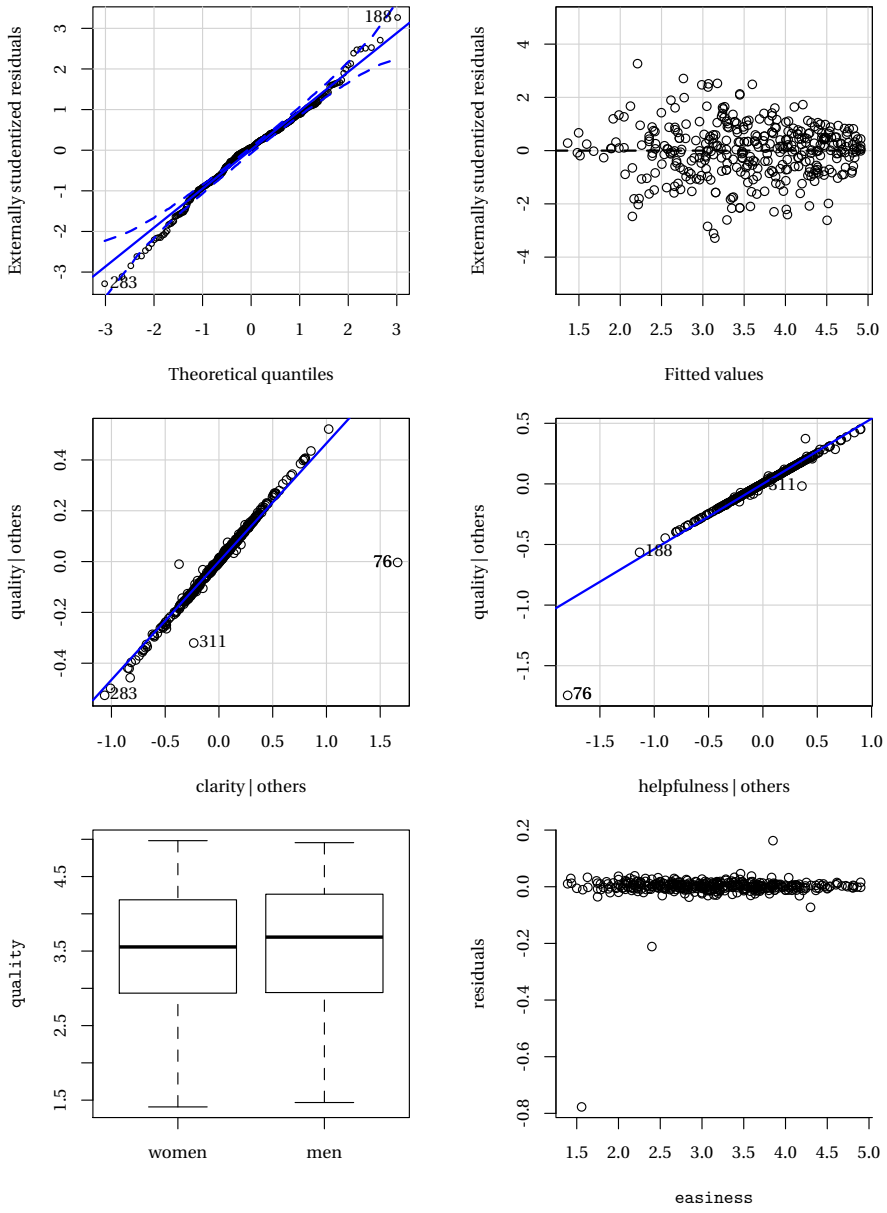

Figure 3: Diagnostic plots for the Model 4 fitted to the `Ratemyprofessor` data. Top left: quantile-quantile plot of externally studentized residuals with pointwise 95% confidence intervals (dashed lines), excluding observation 76. Top right: residual vs fitted values plot. Middle: added-variable plots for `clarity` and `helpfulness`. Bottom left: box and whiskers plot of `quality` as a function of sex. Bottom right: ordinary residuals $\boldsymbol{e}$ against the omitted variable `easiness`.