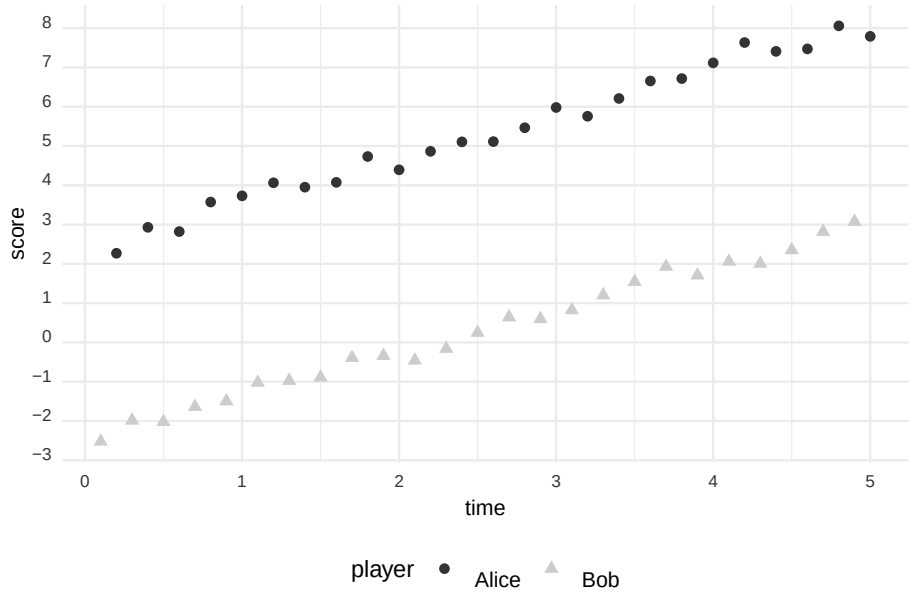

Figure 1: Scores of Alice and Bob as a function of the number of hours of play.

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 $\widehat{\beta}_1, \ldots, \widehat{\beta}_3$?

Solution
It suffices to check the respective intercept and slope and reparametrize the model. The equation of the slope for Alice is $2.5 + 1.1\texttt{time}$ and that of Bob is $-2.5 + 1.1\texttt{time}$. The parameter $\widehat{\beta}_0$ corresponds to the intercept of the baseline, namely -2.5 and the slope $\widehat{\beta}_1$ to the slope of the baseline, 1.1. The other parameters are mean difference between the intercept/slope of Alice minus that of Bob, meaning ($\widehat{\beta}_2 = 5, \widehat{\beta}_3 = 0$). It remains to consider the sign of the coefficients.

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

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

Compiled 24/09/2020 at 12:57

where

$$X_1 = \begin{cases} 0, & \text{if the woman did not complete high school,} \\ 1, & \text{if the woman completed high school, but not college,} \\ -1, & \text{if the woman completed college.} \end{cases}$$

$$X_2 = \begin{cases} 0, & \text{if the woman has no children,} \\ 1, & \text{if the woman has 1 or 2 children,} \\ -1, & \text{if the woman has 3 or more children.} \end{cases}$$

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)?

Solution

2.2 We model a different mean for each of the nine categories (two-way ANOVA additive model. Since we have the same number of women in each category (balanced design), the overall mean is the sum of each fitted value, namely $\widehat{\beta}_0$. The equation for the fitted mean of the reference in (i) is $\widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2$ and thus the difference is $-\widehat{\beta}_1 - \widehat{\beta}_2$.

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

Solution

None of the above. The residuals for least squares can be both negative and positive, so we cannot conclude houses with garages are "always less expensive" than ones without garage. Likewise, there is no information about the correlation between location and the other variables. The statement "everything else being constant" is meaningless: you cannot fix the dummy for garage without impacting area! In particular, if `garage` = 1 there is no garage and `area` is zero. For the same area, the difference between no garage and garage is $\beta_1 - \beta_2 \texttt{area}$. Since both $\beta_1 > 0$ and $\beta_2 > 0$, we cannot conclude anything because we don't know the average area size (for houses with garages).

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

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

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 $\widehat{\beta_1^i} = C\widehat{\beta_1^m}$?

Solution
Conversion is counterintuitive: 10 USD = 13.9 CAD, so you need to multiply $\texttt{price}^{\texttt{USD}}$ by 1.39 to get the amount in CAD. Substituting the metric/Canadian measures in the first equation, we get

$$\texttt{price}^{\texttt{CAD}} = 1.39\texttt{price}^{\texttt{USD}} = 1.39\beta_0^i + \frac{1.39}{1.61}\beta_1^i \texttt{distance}^{\texttt{km}} + 1.39\varepsilon^i.$$

It follows that $\widehat{\beta_1^m} = (1.39/1.61)\beta_1^i$ and we deduce $C = 1.158 = 1.61/1.39$. The best way to make sure this is correct is to generate fake data and make the change of unit to verify your answer.

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.

Solution
The plots in Figure 2 show the adjusted regression line for both models and residual plots. There is some structure left in the model `output ~ velocity`, since the smallest values occur at the endpoint of the output. There is less visible structure in the model with the reciprocal, which also captures more of the variability since its R^2 value is 0.98 compared to 0.87 for the first model. Note that, in the second model, the intercept corresponds to infinite strength wind gusts.

(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.

Solution
The predicted output is 1.34 units of electrity for the first model, while the point forecast is 1.59 for the model with the reciprocal velocity. Both intervals overlap, but the second one [1.39, 1.79] is considerable narrower than the first one, given by [0.84, 1.84].

(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?

Solution
A consequence of the removal of the intercept is that the average of the residuals is not zero anymore and that R returns different values for the multiple R^2 — the F-test values returned are not meaningful either. Even if the coefficient is not statistically significant and we can think that setting the value of β_0 to zero is meaningful, we could justify the non-zero intercept by saying that there is a measurement error in the response. Likewise, there is no zero output measured and so the constraint would amount to extrapolation beyond the range of the explanatory variable.

(d) Produce a quantile-quantile plot of the residuals and comment on the normality assumption.

Compiled 24/09/2020 at 12:57

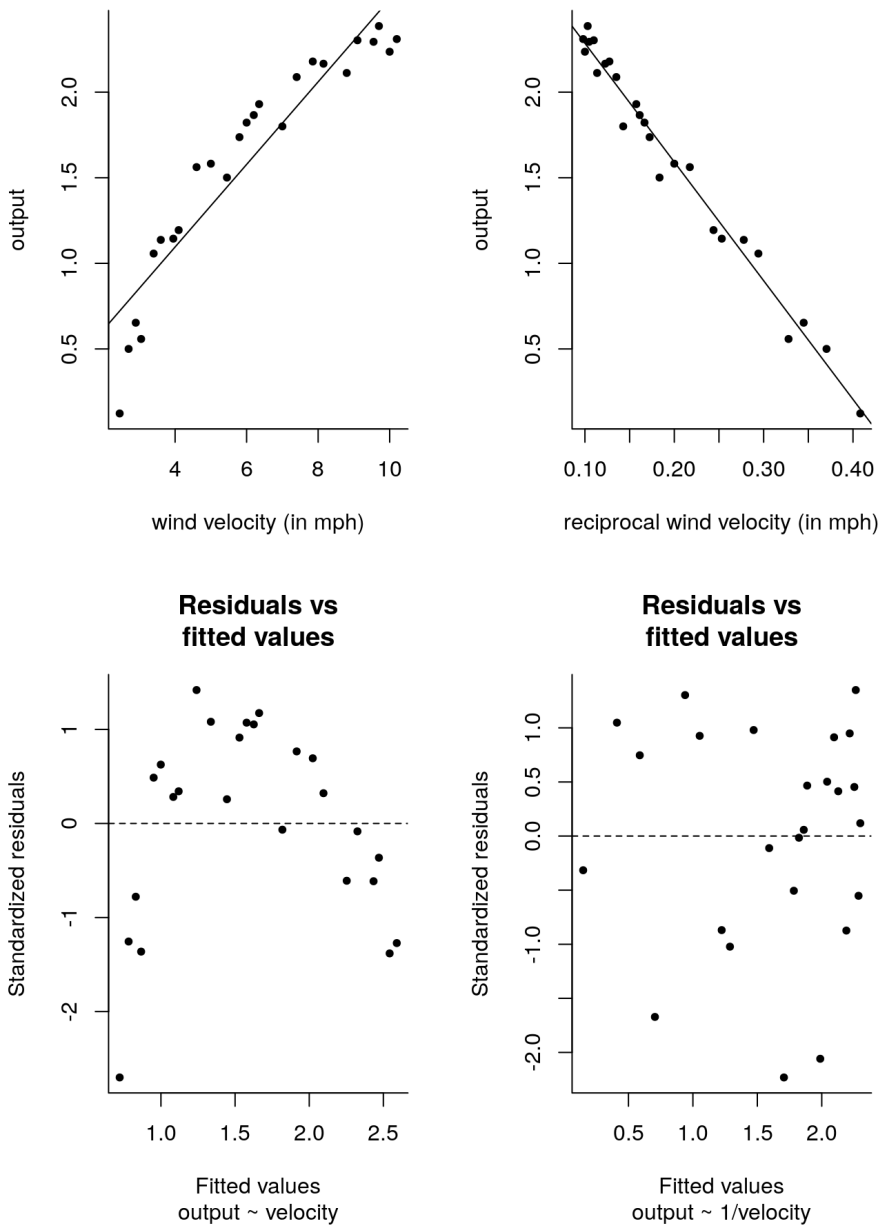

Figure 2: Top panel: fitted regression line for electricity output as a function of wind velocity (left) and reciprocal. Bottom panel: plot of residuals versus fitted values.

Solution

There is no visual evidence against normality of the errors in Figure 3, even if the smallest and largest observations are a bit smaller than expected; they remain well within the simulated confidence intervals.

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. Ad-

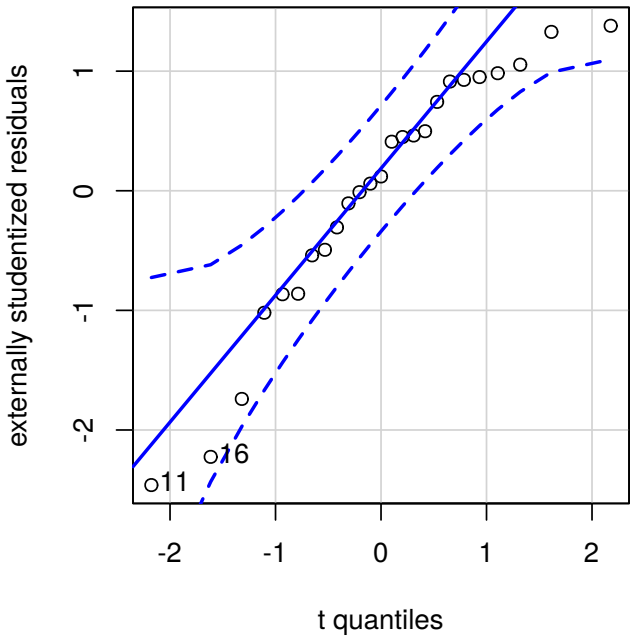

Figure 3: Student quantile-quantile plot of externally studentized residuals, with simulated 95% pointwise confidence intervals.

ditionally, 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.

Solution
Let $\texttt{revenue}_i$ ($i = 1,2$) be dummies equal to 1 if $\texttt{revenue} = i$ and zero otherwise. The linear model is

$$\texttt{intention} = \beta_0 + \beta_1 \texttt{revenue}_1 + \beta_2 \texttt{revenue}_2 + \varepsilon,$$

where β_0 is the average of the baseline group ($\texttt{revenue} = 3$) and β_1, β_2 are contrasts for group 1 and 2, i.e., the average difference n mean between people in revenue group i ($i = 1,2$) relative to revenue group 3.

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

Solution
The predicted intention 7.116 is the average of revenue group 3, which is the intercept.

(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.

Solution
The model is identical to that of part 4.1(a) if the baseline is the same.

(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.

Solution
The variable `revenue` is treated as a continuous variable. The intercept is meaningless. Since the lower revenue groups have a higher intention to buy, the slope β_1 represents the difference between groups, -1.24, which is assumed constant. Rather than $k-1$ additional parameters for a factor with k levels, there is a single one representing the change from group. If we changed the numerical labels of the variable, the interpretation would possibly change.

(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.

Solution
Everything else constant, the β's for the categorical variables represent the change in average buying intention for people in revenue class i ($i = 1,2$) relative to group 3. The coefficients are 1.7 and 0.24 (in the model that includes only `revenue`, these were rather larger, respectively 2.48 and 1.19).

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

Solution
The F-test for `revenue` has value 4.86 and the associated p-value is 0.0095, so we reject the null hypothesis at the 5% level that the revenue is not a useful predictor. On the contrary, `educ` is not statistically significant (test statistic of 1.45, p-value of 0.24).

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.

Solution
The relationship is seemingly nonlinear, with a plateau of `mpg` for high values of horsepower. The cars with less horsepower tend to do travel higher distances for the same fuel consumption.

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

Solution
While the clear negative trend is significant, the slower decay in distance for high horsepower is not captured by the simple linear model. The value of R^2 is 0.6, indicating a strong negative dependence.

(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=statmod.auto;
model mpg=horsepower horsepower*horsepower/ss3 solution;
run;
```

and in R , using

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

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

Solution
The quadratic term is statistically significant, hence we conclude that the quadratic model is more adequate. The plot of residual versus fitted values in Figure 4 is also closer to linear and there is no discernible pattern (some residual heteroscedasticity).

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

Solution
Visually, there is little difference between the quadratic and the cubic model (at the endpoints of horsepower, where the cubic model is flatter; the cubic coefficient, -2.1×10^{-6}, is negligible and not significant at the 5% level (p-value of 0.36). Since the quadratic model is simpler, we prefer it over the more complicated model.

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.

Solution

$$\widehat{\texttt{intention}} = 5.53 + 1.41\texttt{educ1} - 1.4\texttt{educ2} + 1.097\texttt{fixation}.$$

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

Compiled 24/09/2020 at 12:57

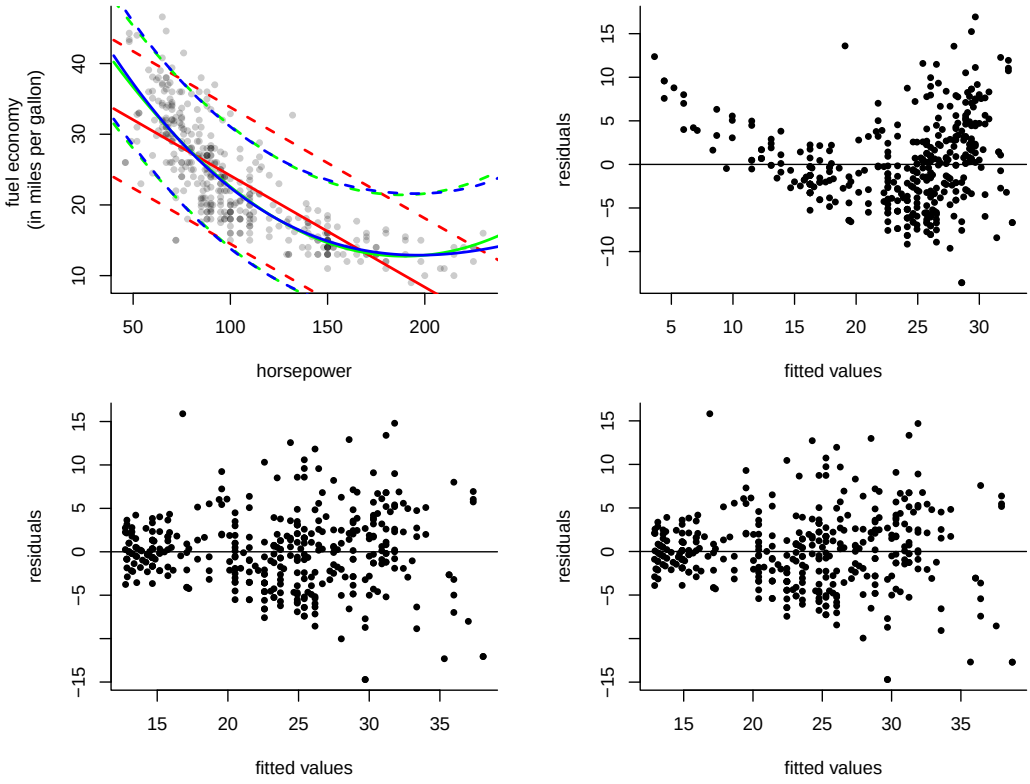

Figure 4: Top left: scatterplot of mileage as a function of power for the `auto` data with fitted linear (red), quadratic (green) and cubic (blue) mean model, along with 95% pointwise prediction intervals. The other three plots show ordinary residuals against fitted values for the linear (top right), quadratic (bottom left) and cubic (bottom right).

Solution

When `educ=1`, the variable `educ1` is 1 and `educ2` zero (and vice-versa when `educ=2`). The fitted model for each of the three groups is

$$\widehat{\texttt{intention}} = \begin{cases} 6.92 + 1.09\texttt{fixation}, & \text{if } \texttt{educ=1} \\ 6.93 + 1.09\texttt{fixation}, & \text{if } \texttt{educ=2} \\ 5.53 + 1.09\texttt{fixation}, & \text{if } \texttt{educ=3}. \end{cases}$$

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?

Solution

We can see that the fit for each category of `educ` is not ideal because the slope is the same; the slope of group 2 seems too high, whereas that of group 3 is too small. The model with an interaction (three different slopes) would be perhaps preferable.

(b) Include an interaction term between `educ` and `fixation` to the model. If you use SAS, make sure to do this

scenario	`fixation`	`educ`	**Fitted average buying intention**
1	x	1	$(\beta_0+\beta_1)+(\beta_3+\beta_4)x$
2	x	2	$(\beta_0+\beta_2)+(\beta_3+\beta_5)x$
3	x	3	$\beta_0+\beta_3 x$
4	$x+1$	1	$(\beta_0+\beta_1)+(\beta_3+\beta_4)(x+1)$
5	$x+1$	2	$(\beta_0+\beta_2)+(\beta_3+\beta_5)(x+1)$
6	$x+1$	3	$\beta_0+\beta_3(x+1)$
7	0	1	$\beta_0+\beta_1$
8	0	2	$\beta_0+\beta_2$
9	0	3	β_0

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

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?

Solution
The equation of the fitted model is

$$\widehat{\texttt{intention}} = 4.58 + 1.45\texttt{educ1} + 3.32\texttt{educ2} + 1.75\texttt{fixation} - 0.11\texttt{educ1}\cdot\texttt{fixation} - 1.25\texttt{educ2}\cdot\texttt{fixation}.$$

The interaction is significative at the 5% level (p-value of 0.02); we thus conclude that the model with the interaction is more adequate that the one without.

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

Solution

$$\widehat{\texttt{intention}} = \begin{cases} 6.03 + 1.64\texttt{fixation}, & \text{if } \texttt{educ=1} \\ 7.91 + 0.5\texttt{fixation}, & \text{if } \texttt{educ=2} \\ 4.59 + 1.75\texttt{fixation}, & \text{if } \texttt{educ=3}. \end{cases}$$

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.

Solution
We can see that the fit is more flexible; in the first graph, the slopes for groups 1 and 2 where overlaid, but the slope coefficient for group 2 is now smaller.

(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.

Solution

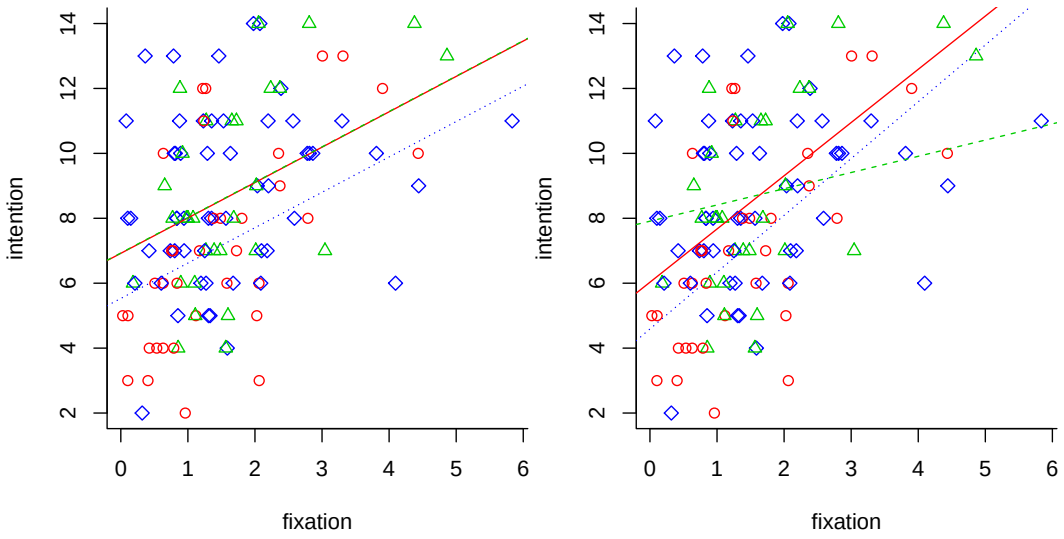

Figure 5: Fitted lines for the model without interaction (left) and the model with interaction (right). The observations and the lines are colored according to the education group (red circles and full line for `educ=1`, green triangles and dashed line for `educ=2`, bleu diamonds and dotted lines for `educ3`).

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

Solution

If we subtract line (6) from (3), we obtain β_3; it thus represents the increase in mean intention to buy when fixation time increases by one second for people in the third education category (university degree).

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

Solution

If we subtract lines (9) from (7), we obtain β_1: it represents the mean difference in intention to buy between individuals in education categories 1 (high school and lower) and 3 (university), when they do not fixate the advertisement (`fixation` equals zero).

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

Solution

If we subtract lines (9) from (8), we obtain β_2: it represents the mean difference in intention to buy between individuals in education categories 2 (college) and 3 (university), when they do not fixate the advertisement (`fixation` equals zero).

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?

Solution

With the original time (in years), the intercept β_0 represents the average monthly air traffic in AD1, which is nonsensical. If we shift the covariate by 1949, the intercept is now the average monthly traffic in 1949 (in thousands of passengers). The slope β_1 represents in both cases the increase in monthly traffic per year.

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

Solution
The model is of the form

$$Y_i = \beta_0 + \beta_1 \texttt{time}_i + \sum_{j=2}^{12} \mathrm{S}_{ij}\beta_j + \varepsilon_i, \qquad i = 1, \ldots, n$$

where $\mathrm{S}_1, \ldots, \mathrm{S}_{12}$ are dummy variables for the twelve months of the year; for example, S_2 is equal to unity if the month is February and zero otherwise.
To assess improvement in fit, we can perform an F test for the model with and without the categorical explanatory with 12 levels. The F-statistic has value 27.589 with null distribution $\mathcal{F}(11, 131)$; the corresponding p-value is negligible, meaning that the improvement in fit for the monthly effect is statistically significant.

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

Solution
The average monthly predicted traffic for December 1962 is 501 263 passengers.

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

Solution
We notice that the model does an overall good job at getting the big features. The residuals versus fitted value plot shows a somewhat quadratic relation between the fitted values and the residuals; looking at the fitted curve relative to the observations, we see that the amplitude of the seasonal pattern increases over time: the predicted values are too large in early 1950s periods and too small near 1960. Since the variance is increasing, a log-transformation may help stabilize it; other alternatives is to include an interaction term between time and months. The last few points have large leverage and drive the curve up.
Looking at the studentized residuals as a function of time (as opposed to fitted values) shows marked heteroscedasticity. The independence assumption here is implausible because of the nature of the time series structure.

(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?

Solution
The log-transform fixes the non-normality diagnostics and the model now captures non-linearity detected with the original data and the heteroscedasticity. One hypothesis of the linear model that is clearly violated here is the independence of the errors: the residuals are positively correlated. Ignoring the serial dependence in the error has consequences: the standard errors are too small (since errors are correlated, there is less units of information so we are overconfident in our uncertainty quantification).

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.

Solution
The intercept of Model 1, 3.532 points, gives the average overall quality for women.

(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.

Solution
The model is parametrized in terms of contrasts, with women as base category. The intercept for men is $(-0.054+0.048)$, to which we include contribution for the scores with the interaction `men:helpfulness`. This gives an average predicted rating of

$$(-0.054+0.048)+(0.541+0.466+0.0007-0.013)\times 4 = 3.9728.$$

(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.*

Solution
The null hypothesis is that the parameters for the non-intercept covariates are zero, i.e., $(\beta_{\texttt{helpfulness}}, \beta_{\texttt{clarity}}) = \mathbf{0}_2$ against the alternative $(\beta_{\texttt{helpfulness}}, \beta_{\texttt{clarity}}) \neq \mathbf{0}_2$. The null hypothesis that the intercept-only model is adequate is soundly rejected (the 95% of the Fisher $\mathcal{F}(2,363)$ distribution is 3.021, much smaller than the observed value of the statistic, 62228.971).

(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?

Solution
The confidence intervals for the parameter β is based on the t-test and its distribution; since $n-p$ is large, we can replace the 97.5% of the Student distribution with that of the Normal distribution, 1.96. We get the asymptotic confidence interval

$$\widehat{\beta}_2 \pm \mathsf{se}(\widehat{\beta}_2)\times 1.96 = 0.466 \pm 0.007 \times 1.96 = [0.4523, 0.4797].$$

The confidence interval does not contain zero, so the coefficient is significative and Model 2 is not an adequate simplification of Model 4 — recall that the t-test for a single coefficient and the F-test lead to similar inferences.

(e) Contrast the estimated coefficient values of Model 2 and Model 4. Are the finding consistent with Figure 6?

Solution
We rejected the null hypothesis that Model 2 is an adequate simplification of Model 4. The two covariates `helpfulness` and `clarity` are positively correlated, as can be seen from top right panel of Figure 6. Suppose Model 4 is the true data generating mechanism; it is clear that, in Model 2, `helpfulness` will capture most of the effect of `clarity` because they nearly perfectly linearly correlated — this seemingly happens with the coefficients. In both cases, the coefficients are significantly different from zero. The coefficients are the effect of either of `helpfulness` or `clarity`, accounting for the effect of the other.

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

Solution
This is a model in which there is an interaction between `men:helpfulness` without main effect, i.e., the slope for `helpfulness` is zero for women. There is no physical reason for this constraint and, if we change the baseline category from women to men, we would get different inferences.

(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 6 and 7.

Solution

Assumptions are (1) independence of the errors (2) linearity, (3) homoscedasticity and (4) normality of the errors.

i. Specification of the mean model: the overall fit is excellent, the relation between `quality` and `clarity` or `quality` and `helpfulness` is linear based on Figure 6. The top right panel of Figure 7 shows that there is no structure in the mean, whereas the bottom right plot shows a lack of relationship with easiness (a regression line would have non-zero slope by virtue of a few outliers).

ii. Homoscedasticity: there is no difference in variance between women and men faculty members (box-and-whiskers plot at bottom left of Figure 7). There is some heteroscedasticity in the ratings for Model 4; students are more unanimous for high scores than for low ones (residual vs fitted values in top right plot of Figure 7).

iii. Independence: plausible by design provided individuals are included at random or all the instructors are included.

iv. Gaussianity: the top-left plot of Figure 7 is a Student quantile-quantile plot and can be used to assess the normality assumption, there is some heavy left tail due to the skewness of the response and the boundedness of its support, but the impact is negligible because the same size is large. Pointwise 95% confidence intervals are superposed, and we expect 1/20 points to fall outside these for independent data, but quantiles are ordered.

	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.

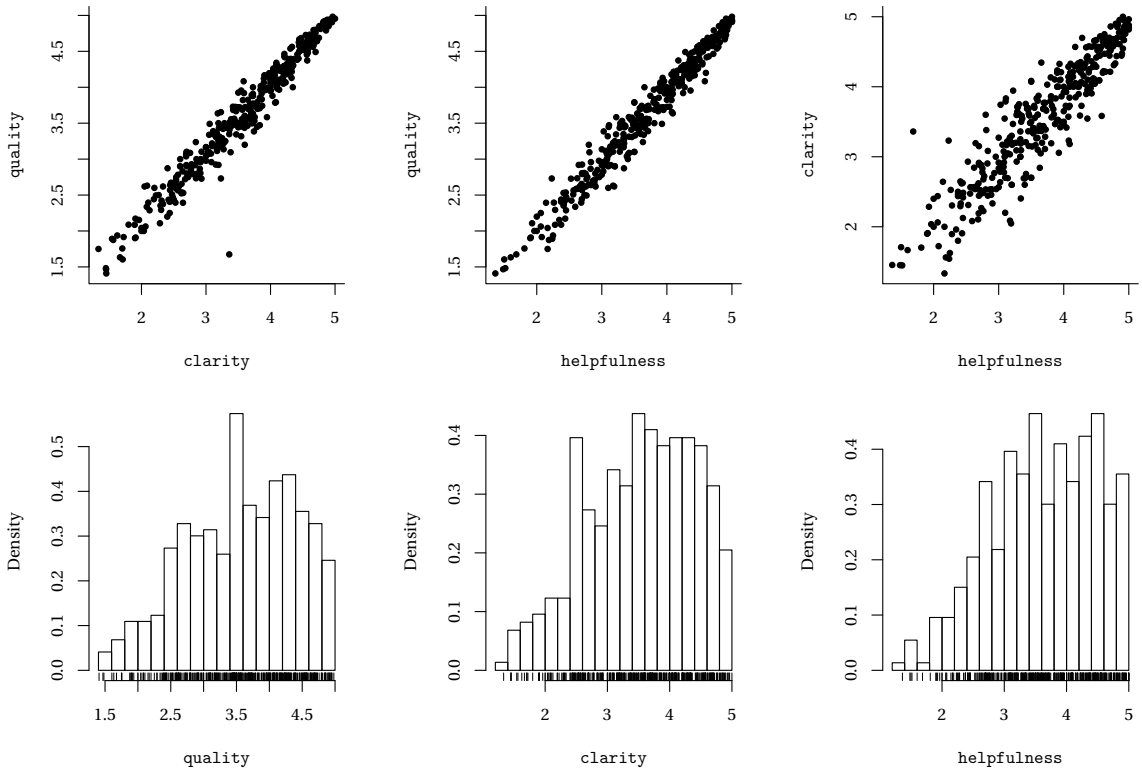

Figure 6: Top: pair plots (linear correlation from left to right of 0.98,0.98, 0.92). Bottom: histograms of average `quality`, `helpfulness` and `clarity` indicators.

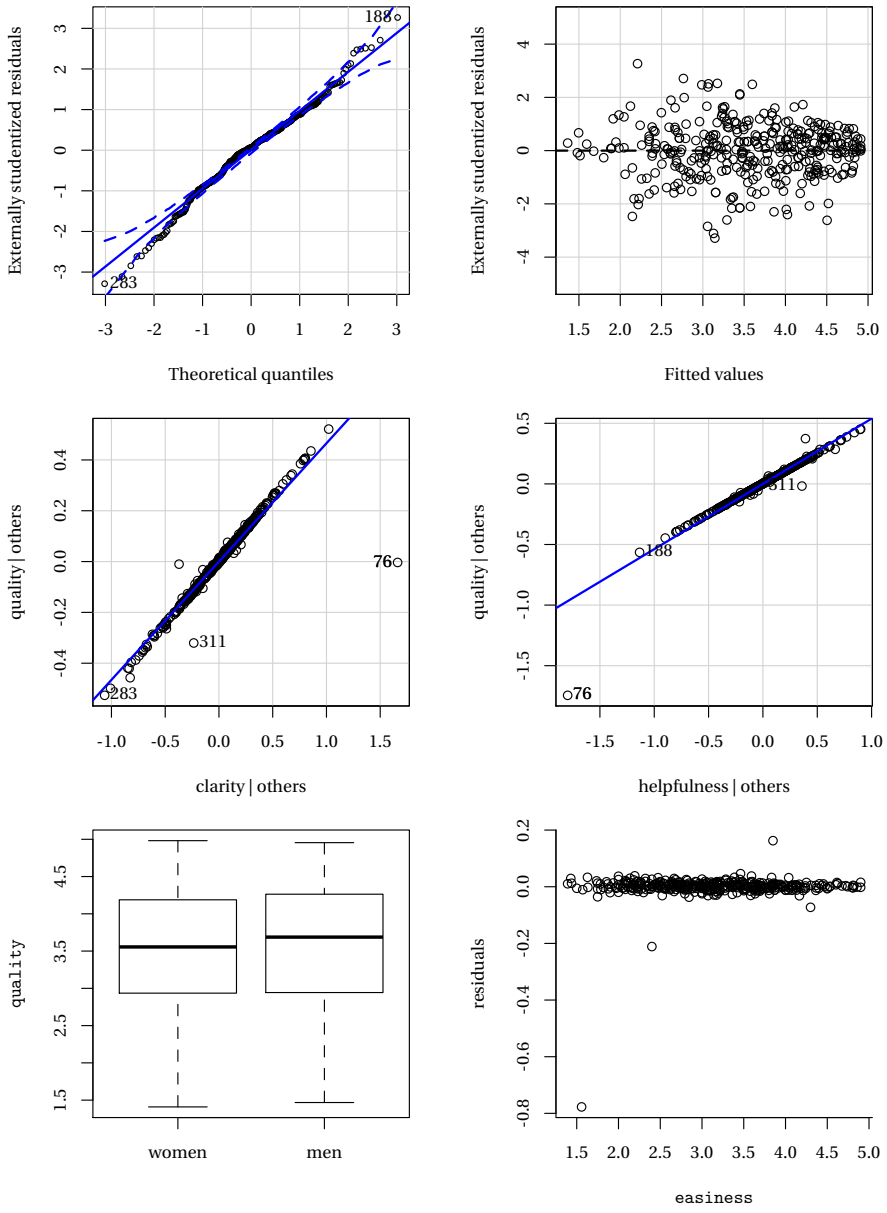

Figure 7: Diagnostic plots for the Model 4 fitted to the `Ratemyprofessor` data. Top left: quantile-quantile plot of externally studentized residuals with pointwise 95% confidence intervals (dashed lines), excluding observation 76. Top right: residual vs fitted values plot. Middle: added-variable plots 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`.