University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC33 (K. Butler), Final Exam
April 25, 2023

Aids allowed (on paper, no computers):

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Non-programmable, non-communicating calculator

This exam has 11 numbered pages of questions plus this cover page.
In addition, you have an additional booklet of Figures to refer to during the exam.
The maximum marks available for each part of each question are shown next to the question part.
If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.
The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

1. A chemical analysis was carried out of 178 Italian wines, from three different cultivars. (A cultivar is a grape plant deliberately bred to have certain desirable characteristics.) The aim of the chemical analysis was to see whether wine made from grapes grown from the three different cultivars differs in any important ways.

   Some of the data file is shown in Figure 2, and the file is saved in your current R Project as `wine.txt`. The dataframe after being read in is shown in Figure 3.

   The questions below ask for Tidyverse R code to accomplish the task described. Explanations are only needed where explicitly asked for.

   (a) [3] Read the data from the file into a dataframe called `wine`. Maximum points are for the simplest method, with a brief explanation of why it will work.

   (b) [2] Produce a suitable plot of the alcohol content of the wines made from each of the three different cultivars.

   (c) [3] Calculate the median and inter-quartile range of alcohol content for each of the cultivars.

   (d) [2] Display the columns whose names begin with the letter A (uppercase or lowercase), without naming the columns.

(e) [3] Display all the columns that are quantitative, without naming any of the columns.

(f) [3] Display the wines that are made from the `barolo` cultivar and that have an ash content of strictly less than 2.2.

(g) [3] Display how many wines of each cultivar have an ash content of strictly less than 2.2. (Do not count any wines with any other ash content.)

(h) [4] Make a display containing boxplots of each of the quantitative variables for each of the cultivars, using one `ggplot`, and with the side-by-side boxplot for each quantitative variable in a separate subplot within the overall display. Bear in mind that some of the quantitative variables have typically bigger values than others.

2. We are planning a study that will use the sign test to test whether the population median is 100 (against a two-sided alternative that it is not equal to 100). The data that our study will provide are assumed to come from a normal distribution with SD 20.

   (a) [2] How do you know that the mean and median of a normal distribution are equal to each other? Explain briefly. You may assume that the mean and *mode* of the normal distribution are equal.

   (b) [4] We want to estimate, by simulation, the power of the sign test to reject the null hypothesis that the median is 100, against a two-sided alternative, when the median is actually 110, using a sign test, under the conditions given in the question. What code would do that? Hint: the `smmr` package also has a function `pval_sign0` that obtains the two-sided P-value for a sign test. It has two inputs: the null median and a column of data, in that order, and returns only the P-value.

   (c) [3] The code of the previous part gives an estimated power of 0.677. A different power analysis, also applying to this situation, is shown in Figure 4. What code produced this power analysis, bearing in mind what you know so far?

   (d) [3] According to the information so far, which two tests are we comparing, and which test is the more powerful here? How do you know?

3. In an earlier question, we encountered some data from a chemical analysis of some Italian wines. The data, as read into a dataframe `wine`, are shown in Figure 3. In this question, we will see whether the malic acid level differs among wines from the different cultivars.

   (a) [2] Figure 5 shows a plot of malic acid for each cultivar. Why do you think the analyst on this project decided to run a Mood's median test? Explain briefly. (There are two points you need to make).

   (b) [2] Mood's median test is run on these data, as shown in Figure 6. What do you conclude from this Figure, in the context of the data?

   (c) [3] Are we justified in looking at Figure 7 for these data? Explain briefly why or why not. If appropriate, what do we conclude from this Figure?

   (d) [3] Some mystery code is shown in Figure 8. What task does running this code accomplish, and why might you be interested in the results? Hint: you are looking for the overall purpose of running the code, *not* a line-by-line description of what the code does. We are working at a higher level than that.

4. In each part of this question, you will be given a dataframe (shown in a Figure). You will either be given some code and asked what output it will produce, or you will be given the output and asked what code will produce it.

   (a) [3] Given the dataframe `d1` shown in Figure 9, and the code in Figure 10, what output will be produced?

   (b) [3] A dataframe `d2` is shown in Figure 11. Some output from operating on `d2` is shown in Figure 12. What code would produce that output, starting from `d2`? Hint: the start of your code should be `d2 %>%`.

   (c) [3] Given the dataframe `d3` shown in Figure 13, and the code in Figure 14, what output will be produced?

   (d) [2] Given the dataframe `d4` shown in Figure 15, and the code in Figure 16, what output will be produced?

(e) [3] A dataframe `d5` is shown in Figure 17. Some output from operating on `d5` is shown in Figure 18. What code would produce that output, starting from `d5`?

(f) [4] Given the dataframe `d6` shown in Figure 19, and the code in Figure 20, what output will be produced?

5. In 1982, a researcher named Engel studied the relationship between income and food expenditure in Belgium, with the aim of seeing whether food expenditure changed as income increased. 234 working-class households were studied, and annual income and food expenditure were measured in Belgian francs. Some of the data are shown in Figure 21.

(a) [2] What code would draw a suitable plot of these data, with, if appropriate for your plot, a smooth trend?

(b) [3] A regression analysis with its output is shown in Figure 22. According to this Figure, is there a significant effect of income on food expenditure, and, from this analysis, is the trend upward, downward, or non-existent? Explain briefly in each case.

(c) [2] A plot of residuals against fitted values is shown in Figure 23, for the regression shown in Figure 22. What is the major problem with this residual plot? Explain briefly (that is, say how you know).

(d) [2] In an attempt to fix the problem, a Box-Cox analysis is run, as shown in Figure 24. What do you conclude from this Figure?

(e) [2] A regression that uses the results of the Box-Cox analysis is fitted, and the residuals vs. fitted values are plotted in Figure 25. The statistician on the project knows that there was one family whose food expenditure was much less than you would expect from their income. Aside from this, do you think the residual plot is satisfactory now? Explain briefly.

6. A debt is money that is owed to another person or company that has not yet been paid. In the past, being in debt was seen as a bad thing, but attitudes towards debt have changed over time, and borrowing money to pay for a large purchase like a house or a car or an education can now be seen as being financially responsible.

   An economist designed a survey in which they tried to find out how attitudes towards debt are influenced by other things such as age, money-management skills, or "locus of control". This last is a concept from psychology. A person can have an "external locus of control", when they feel that most of the things that happen to them are caused by outside events, or an "internal locus of control", where most of the things that happen to them are caused by things inside the person (such as decisions they made).

   A complete list of the survey items is shown in Figure 26. All variables are quantitative (and are treated as such here), even though many of them are more ordinal than quantitative (scores on a scale). Some of the items have a yes/no answer; in those cases, 1 means "yes" and 0 means "no". After removing surveys with missing answers, there were complete surveys from 304 respondents. Some of the data is shown in Figure 27.

   (a) [3] A regression is fitted predicting attitudes to debt from all the other variables. This is shown in Figure 28. Since there were a lot of explanatory variables, I decided to remove the five variables shown in the `update` line at the top of Figure 29; the resulting regression output is shown in the remainder of that Figure. Finally, I ran the test shown in Figure 30.

   Why was it necessary to run the test shown in Figure 30, and what do you conclude from it in the context of the data?

   (b) [2] What other comparison of models `debt.1` and `debt.2` (that is, not a hypothesis test) suggests that the conclusion from your test in the previous part makes sense?

   (c) [3] For the model `debt.2`, Figure 31 shows a plot of residuals against fitted values, and Figure 32 shows a normal quantile plot of residuals for the same model. Comment briefly on each of these two plots, and make an overall comment considering the two plots together.

(d) [2] What other plot or plots would you like to see to confirm your impressions about the model `debt.2`?

(e) [2] Two of the explanatory variables in the model `debt.2` are `agegp` and `ccarduse`. Based on what you know or can guess, does the *sign* of the slope estimate for each variable (that is, whether it is positive or negative) make sense? Explain briefly.

7. The exponential distribution has (continuous) density function $f(x) = \beta e^{-\beta x}$ for $x \geq 0$, where β is a non-negative parameter. The distribution has mean $1/\beta$. Our aim in this problem is to estimate β using Bayesian methods. Stan has an `exponential` distribution with one input (that is β), and also a `uniform` distribution, whose two inputs are the lower and upper limits of that distribution.

(a) [2] Before looking at any data, we think that the mean μ is almost certainly between 2 and 10. What does this tell us about β, before looking at any data?

(b) [3] Write the `model` section of a Stan program to estimate β, assuming a uniform prior between the two limits you found above, and calling your data `x`.

(c) [3] Add appropriate sections to your Stan code to make a complete Stan program. You can assume that there will be 10 observations in your data, and that the observations will be decimal numbers.

(d) [2] Assuming that your Stan code has been saved in a file `expo.stan`, what R code would compile it to C++?

(e) [3] Some data is shown in Figure 33, stored in a variable `w`. What R code will set up this data suitably, and draw random samples from the posterior distribution of β for the data in `w`?

(f) [2] A summary of the simulated posterior distribution of β is shown in Figure 34. What is a 90% posterior interval for β?

(g) [3] Somebody says to you that you have to interpret the interval of the previous part in this way: "in 90% of all possible samples, the procedure will give you an interval that contains the true value of β". Is that what your interval of the previous part says, or not? If not, what does that interval actually say? Explain briefly.

Use the rest of this page if you need more space. Be sure to label any answers here with the question and part they belong to.

Figures

```
library(MASS)
library(tidyverse)
library(smmr)
library(cmdstanr)
```

Figure 1: Packages

```
cultivar_name,alcohol,malic_acid,ash,mg
grignolino,12.87,4.61,2.48,86
grignolino,13.4,4.6,2.86,112
barolo,13.9,1.68,2.12,101
barbera,11.84,0.89,2.58,94
barolo,13.51,1.8,2.65,110
barolo,13.05,1.73,2.04,92
barolo,14.38,1.87,2.38,102
barolo,13.5,1.81,2.61,96
barbera,11.87,4.31,2.39,82
barbera,12,0.92,2,86
barbera,12.21,1.19,1.75,151
barbera,12.52,2.43,2.17,88
```

Figure 2: Wine data (some)

```
wine
```

```
## # A tibble: 178 x 5
##    cultivar_name alcohol malic_acid   ash    mg
##    <chr>           <dbl>      <dbl> <dbl> <dbl>
##  1 grignolino       12.9       4.61  2.48    86
##  2 grignolino       13.4       4.6   2.86   112
##  3 barolo           13.9       1.68  2.12   101
##  4 barbera          11.8       0.89  2.58    94
##  5 barolo           13.5       1.8   2.65   110
##  6 barolo           13.0       1.73  2.04    92
##  7 barolo           14.4       1.87  2.38   102
##  8 barolo           13.5       1.81  2.61    96
##  9 barbera          11.9       4.31  2.39    82
## 10 barbera          12         0.92  2       86
## # i 168 more rows
```

Figure 3: Wine data after being read in (some)

```
##
##      One-sample t test power calculation
##
##               n = 25.38969
##           delta = 10
##              sd = 20
##       sig.level = 0.05
##           power = 0.677
##     alternative = two.sided
```

Figure 4: Power analysis

```
wine %>% group_by(cultivar_name) %>%
  summarize(n = n())
```

```
## # A tibble: 3 x 2
##   cultivar_name     n
##   <chr>         <int>
## 1 barbera          71
## 2 barolo           59
## 3 grignolino       48
```

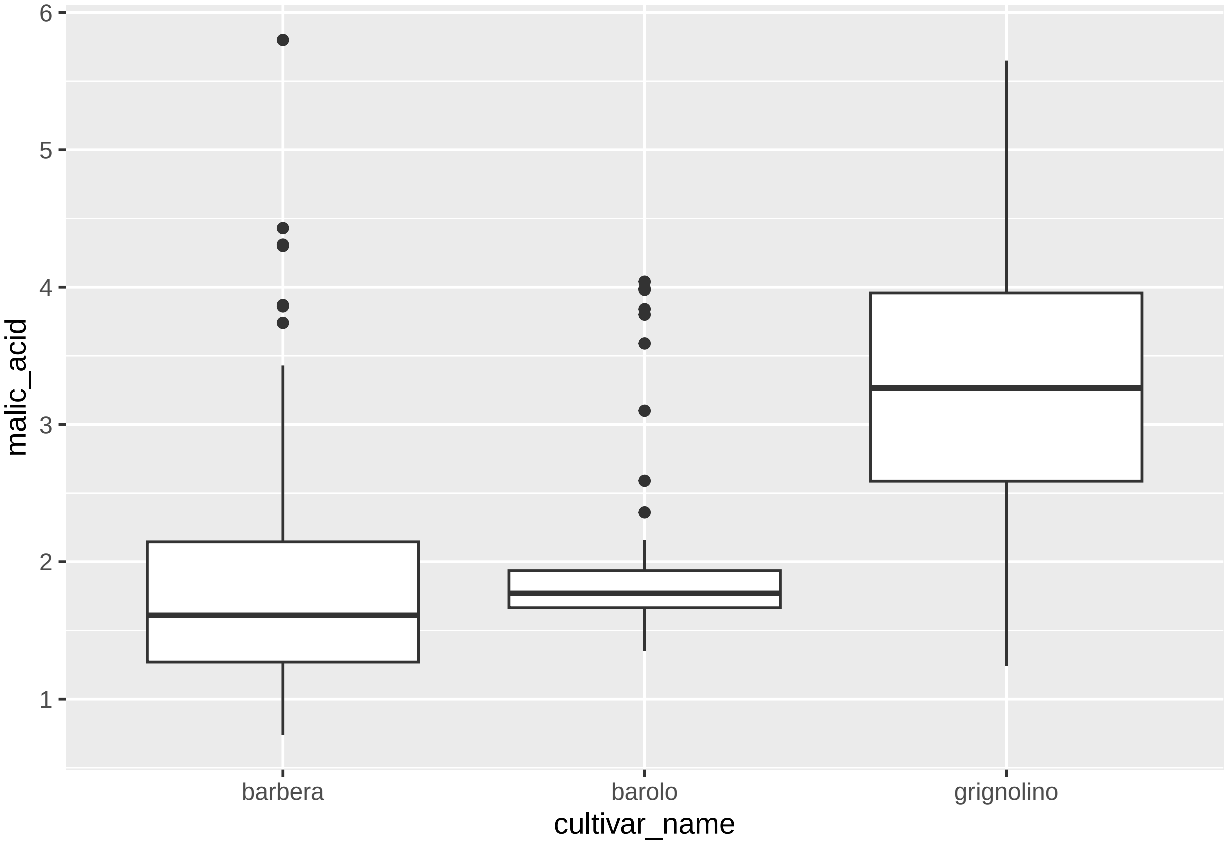

Figure 5: Wine data summary and plot

```
median_test(wine, malic_acid, cultivar_name)
```

```
## $grand_median
## [1] 1.865
##
## $table
##             above
## group         above below
##   barbera        25    46
##   barolo         21    38
##   grignolino     43     5
##
## $test
##        what        value
## 1 statistic 4.119291e+01
## 2        df 2.000000e+00
## 3   P-value 1.135205e-09
```

Figure 6: Wine data Mood Median Test

```
pairwise_median_test(wine, malic_acid, cultivar_name)
```

```
## # A tibble: 3 x 4
##   g1      g2           p_value adj_p_value
##   <chr>   <chr>          <dbl>       <dbl>
## 1 barbera barolo     3.97e- 2    1.19e- 1
## 2 barbera grignolino 1.52e-11    4.55e-11
## 3 barolo  grignolino 2.15e-12    6.44e-12
```

Figure 7: Wine data pairwise median tests

```
wine %>% filter(cultivar_name == "barolo") -> barolo
tibble(sim = 1:10000) %>%
  rowwise() %>%
  mutate(my_sample = list(sample(barolo$malic_acid, replace = TRUE))) %>%
  mutate(my_mean = mean(my_sample)) %>%
  ggplot(aes(sample = my_mean)) + stat_qq() + stat_qq_line()
```

Figure 8: Wine data mystery code

```
d1
```

```
## # A tibble: 2 x 3
##      id     a     b
##   <dbl> <dbl> <dbl>
## 1     1    10    11
## 2     2     8     9
```

Figure 9: Dataframe d1

```
d1 %>% pivot_longer(-id, names_to = "name", values_to = "value")
```

Figure 10: Code for dataframe `d1`

```
d2
```

```
## # A tibble: 2 x 5
##     row  m_ht  f_ht  m_wt  f_wt
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     7   180   150    80    60
## 2     8   185   160    90    55
```

Figure 11: Dataframe `d2`

```
## # A tibble: 8 x 4
##     row gender what  measure
##   <dbl> <chr>  <chr>   <dbl>
## 1     7 m      ht        180
## 2     7 f      ht        150
## 3     7 m      wt         80
## 4     7 f      wt         60
## 5     8 m      ht        185
## 6     8 f      ht        160
## 7     8 m      wt         90
## 8     8 f      wt         55
```

Figure 12: Dataframe `d2` output

```
d3
```

```
## # A tibble: 2 x 3
##      id   x_1   x_2
##   <dbl> <dbl> <dbl>
## 1     4    10    11
## 2     6     8     9
```

Figure 13: Dataframe `d3`

```
d3 %>% pivot_longer(-id, names_to = c(".value", "col"), names_sep = "_")
```

Figure 14: Code for dataframe `d3`

```
d4
```

```
## # A tibble: 4 x 3
##     row x     measure
##   <dbl> <chr>   <dbl>
## 1     7 m_ht      180
## 2     7 f_ht      150
## 3     7 m_wt       80
## 4     7 f_wt       60
```

Figure 15: Dataframe d4

```
d4 %>% separate(x, into = c("gender", "what"), sep = "_")
```

Figure 16: Code for dataframe d4

```
d5
```

```
## # A tibble: 4 x 3
##     row group     x
##   <dbl> <chr> <dbl>
## 1     1 a        14
## 2     1 b        15
## 3     2 a        16
## 4     2 b        17
```

Figure 17: Dataframe d5

```
## # A tibble: 2 x 3
##     row     a     b
##   <dbl> <dbl> <dbl>
## 1     1    14    15
## 2     2    16    17
```

Figure 18: Dataframe d5 output

```
d6
```

```
## # A tibble: 4 x 3
##   x         y z
##   <chr> <dbl> <chr>
## 1 c        16 low
## 2 b        18 high
## 3 a        20 medium
## 4 b        22 low
```

Figure 19: Dataframe d6

```
d6 %>% pivot_wider(names_from = z, values_from = y)
```

Figure 20: Code for dataframe d6

```
engel
```

```
## # A tibble: 234 x 2
##    income foodexp
##     <dbl>   <dbl>
##  1   420.    256.
##  2   541.    311.
##  3   901.    486.
##  4   639.    403.
##  5   751.    496.
##  6   946.    634.
##  7   829.    631.
##  8   979.    700.
##  9  1310.    831.
## 10  1492.    815.
## # i 224 more rows
```

Figure 21: Food expenditure data (some)

```
engel.1 <- lm(foodexp ~ income, data  = engel)
summary(engel.1)
```

```
##
## Call:
## lm(formula = foodexp ~ income, data = engel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -622.00  -54.02    3.22   52.87  398.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.33302   15.52094   5.885 1.39e-08 ***
## income       0.54654    0.01458  37.497  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100.2 on 232 degrees of freedom
## Multiple R-squared:  0.8584, Adjusted R-squared:  0.8578
## F-statistic:  1406 on 1 and 232 DF,  p-value: < 2.2e-16
```

Figure 22: Food expenditure: regression analysis

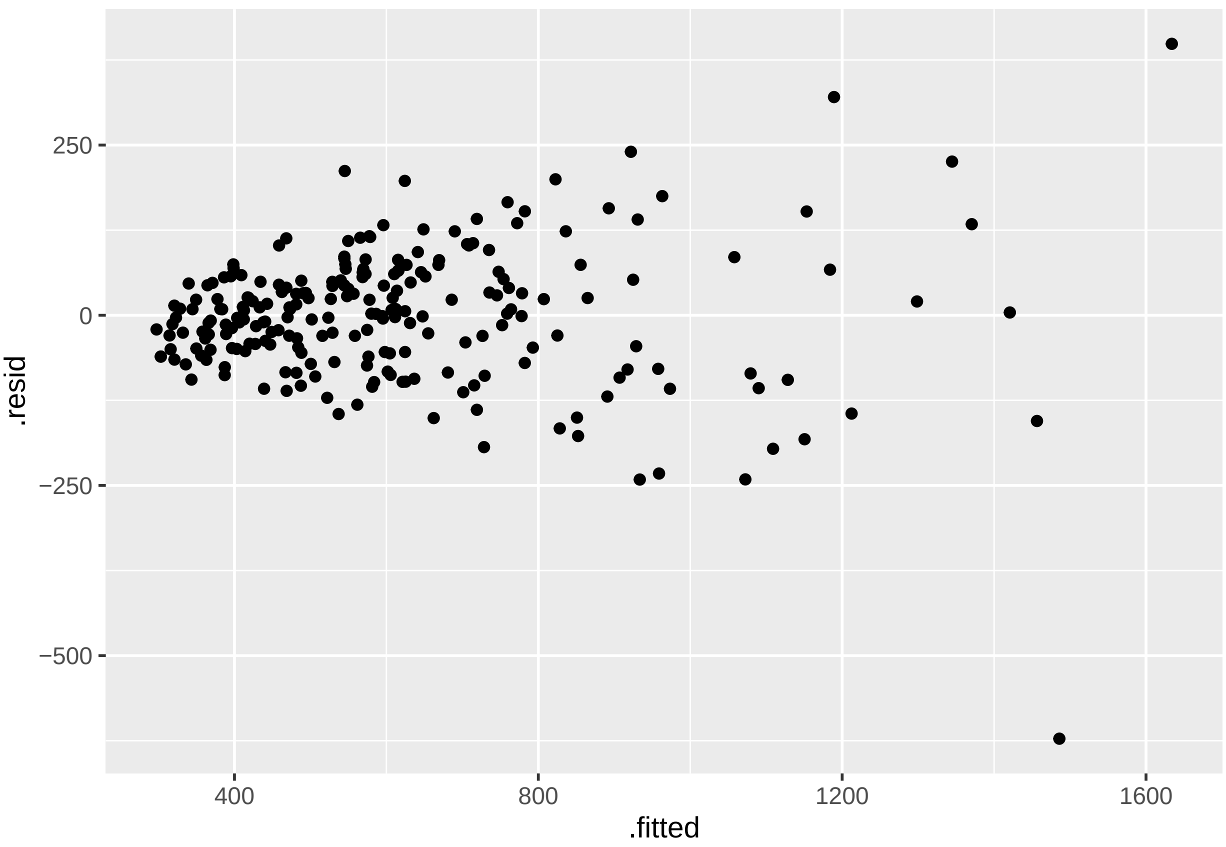

Figure 23: Food expenditure: residual plot 1

```
boxcox(foodexp ~ income, data = engel)
```

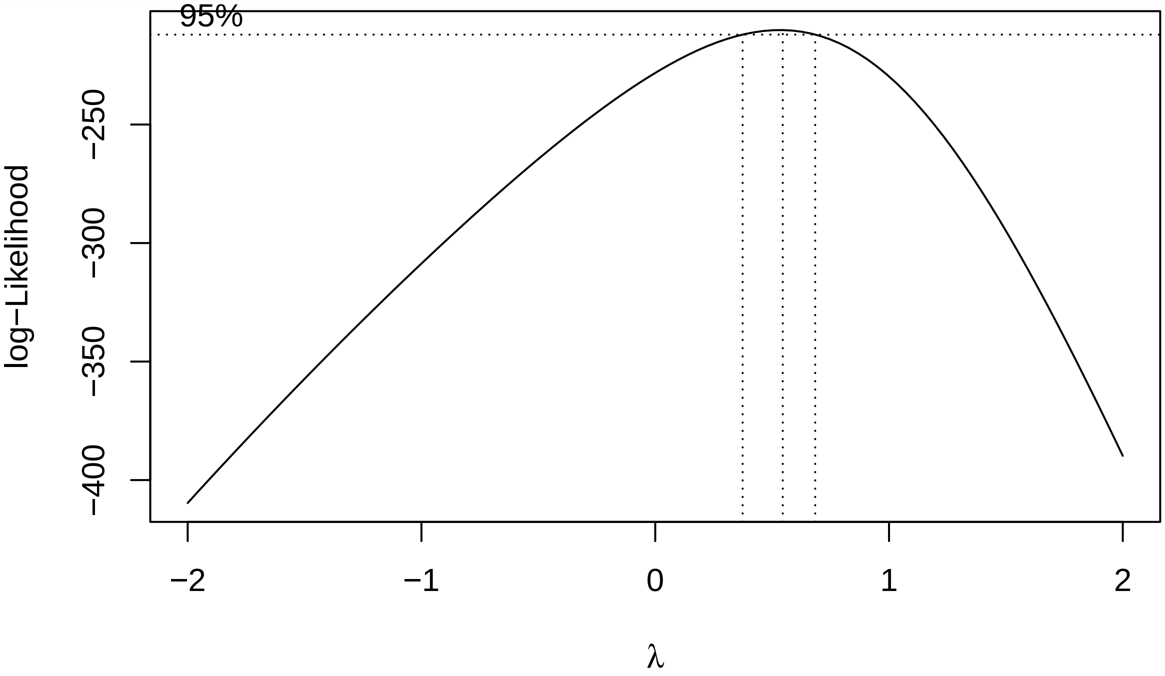

Figure 24: Food expenditure: Box-Cox

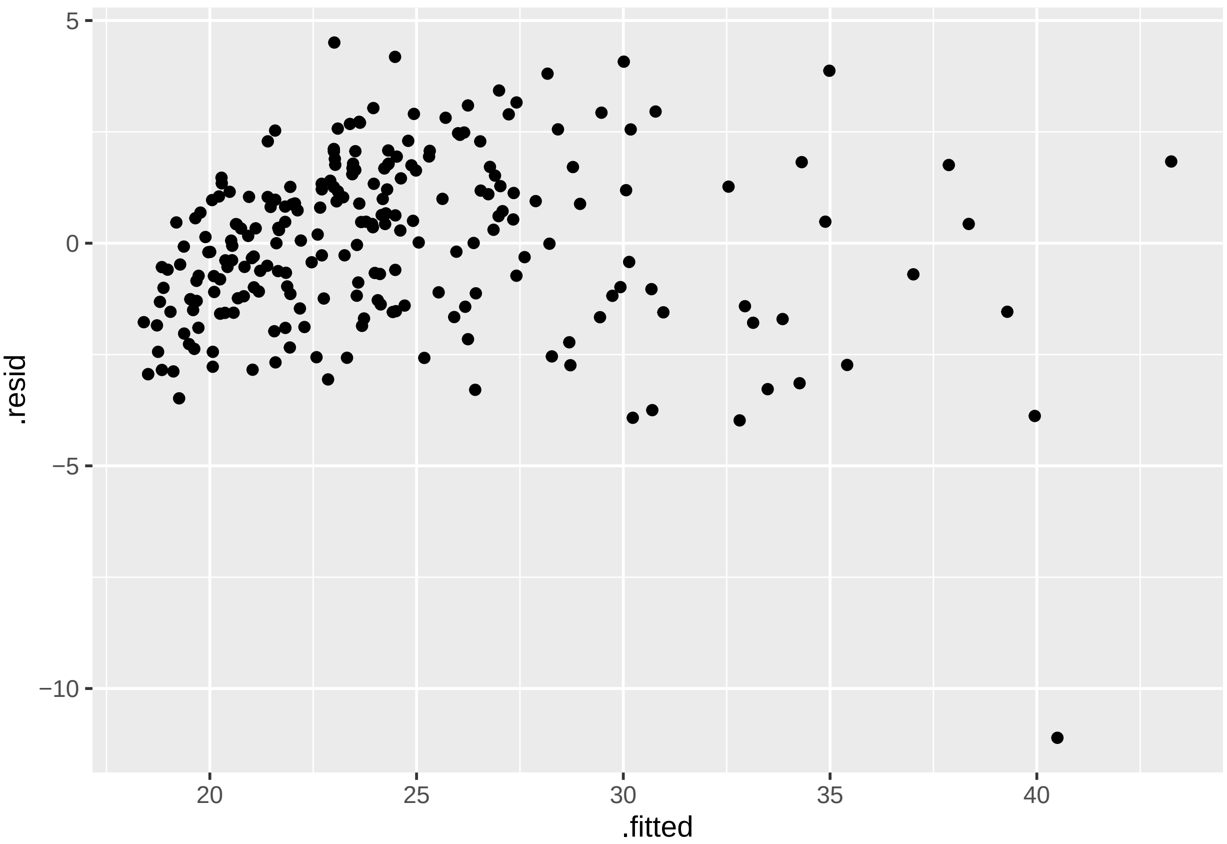

Figure 25: Food expenditure: residual plot 2

- `incomegp` income group (1=lowest, 5=highest)
- `house` security of housing tenure (1=rent, 2=mortgage, 3=owned outright)
- `children` number of children in household
- `singpar` is the respondent a single parent?
- `agegp` age group (1=youngest)
- `bankacc` does the respondent have a bank account?
- `bsocacc` does the respondent have a building society (credit union) account?
- `manage` self-rating of money management skill (high values=high skill)
- `ccarduse` how often did s/he use credit cards (1=never... 3=regularly)
- `cigbuy` does s/he buy cigarettes?
- `xmasbuy` does s/he buy Christmas presents for children?
- `locintrn` score on a locus of control scale (high values=internal)
- `prodebt` score on a scale of attitudes to debt (high values=favourable to debt (response variable)

Figure 26: Debt survey items

```
debt
```

```
## # A tibble: 304 x 13
##    incomegp house children singpar agegp bankacc bsocacc manage ccarduse cigbuy
##       <dbl> <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>  <dbl>    <dbl>  <dbl>
##  1        3     3        0       0     4       1       0      5        2      0
##  2        5     2        2       0     2       1       0      5        3      0
##  3        3     3        0       0     4       1       0      4        2      0
##  4        4     2        0       0     2       1       0      5        2      0
##  5        4     2        0       0     2       1       0      4        2      0
##  6        2     1        1       0     4       1       0      4        1      0
##  7        2     3        0       0     4       1       0      5        1      0
##  8        2     3        0       0     4       1       0      5        1      0
##  9        2     3        2       0     4       0       1      4        2      0
## 10        2     2        2       1     3       1       0      4        1      1
## # i 294 more rows
## # i 3 more variables: xmasbuy <dbl>, locintrn <dbl>, prodebt <dbl>
```

Figure 27: Debt data (some)

```r
debt.1 <- lm(prodebt ~ ., data = debt)
summary(debt.1)
```

```
##
## Call:
## lm(formula = prodebt ~ ., data = debt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95085 -0.46986 -0.01442  0.40263  1.87677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.04642    0.31682  12.772  < 2e-16 ***
## incomegp     0.06463    0.03373   1.916 0.056336 .
## house       -0.05331    0.06751  -0.790 0.430378
## children     0.03813    0.03898   0.978 0.328749
## singpar      0.02054    0.17372   0.118 0.905984
## agegp       -0.10206    0.04761  -2.144 0.032899 *
## bankacc      0.06248    0.12123   0.515 0.606641
## bsocacc     -0.11198    0.08344  -1.342 0.180628
## manage      -0.12820    0.04556  -2.814 0.005231 **
## ccarduse     0.18779    0.05258   3.571 0.000415 ***
## cigbuy      -0.15448    0.08731  -1.769 0.077894 .
## xmasbuy      0.20147    0.11928   1.689 0.092298 .
## locintrn    -0.13942    0.04371  -3.190 0.001579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6562 on 291 degrees of freedom
## Multiple R-squared:  0.2043, Adjusted R-squared:  0.1715
## F-statistic: 6.226 on 12 and 291 DF,  p-value: 8.916e-10
```

Using a dot on the right side of a model formula means “all the other variables”.

Figure 28: Debt data regression 1

```
debt.2 <- update(debt.1, .~. - singpar - bankacc - house - children - bsocacc)
summary(debt.2)
```

```
##
## Call:
## lm(formula = prodebt ~ incomegp + agegp + manage + ccarduse +
##     cigbuy + xmasbuy + locintrn, data = debt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99736 -0.43552  0.00559  0.40031  1.81132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.08091    0.29233  13.960  < 2e-16 ***
## incomegp     0.06025    0.03063   1.967 0.050125 .
## agegp       -0.13047    0.04143  -3.149 0.001805 **
## manage      -0.14141    0.04389  -3.222 0.001416 **
## ccarduse     0.18775    0.05149   3.647 0.000314 ***
## cigbuy      -0.13220    0.08560  -1.544 0.123579
## xmasbuy      0.22305    0.11479   1.943 0.052963 .
## locintrn    -0.14165    0.04330  -3.271 0.001198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6554 on 296 degrees of freedom
## Multiple R-squared:  0.1926, Adjusted R-squared:  0.1735
## F-statistic: 10.09 on 7 and 296 DF,  p-value: 2.546e-11
```

`update` requires a model to update, and then how to update it. This one means “leave everything the same except take out the five explanatory variables listed.”

Figure 29: Debt data regression 2

```
anova(debt.2, debt.1)
```

```
## Analysis of Variance Table
##
## Model 1: prodebt ~ incomegp + agegp + manage + ccarduse + cigbuy + xmasbuy +
##     locintrn
## Model 2: prodebt ~ incomegp + house + children + singpar + agegp + bankacc +
##     bsocacc + manage + ccarduse + cigbuy + xmasbuy + locintrn
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    296 127.14
## 2    291 125.31  5     1.836 0.8528 0.5134
```

Figure 30: Debt data: a test

```r
ggplot(debt.2, aes(x = .fitted, y = .resid)) + geom_point()
```

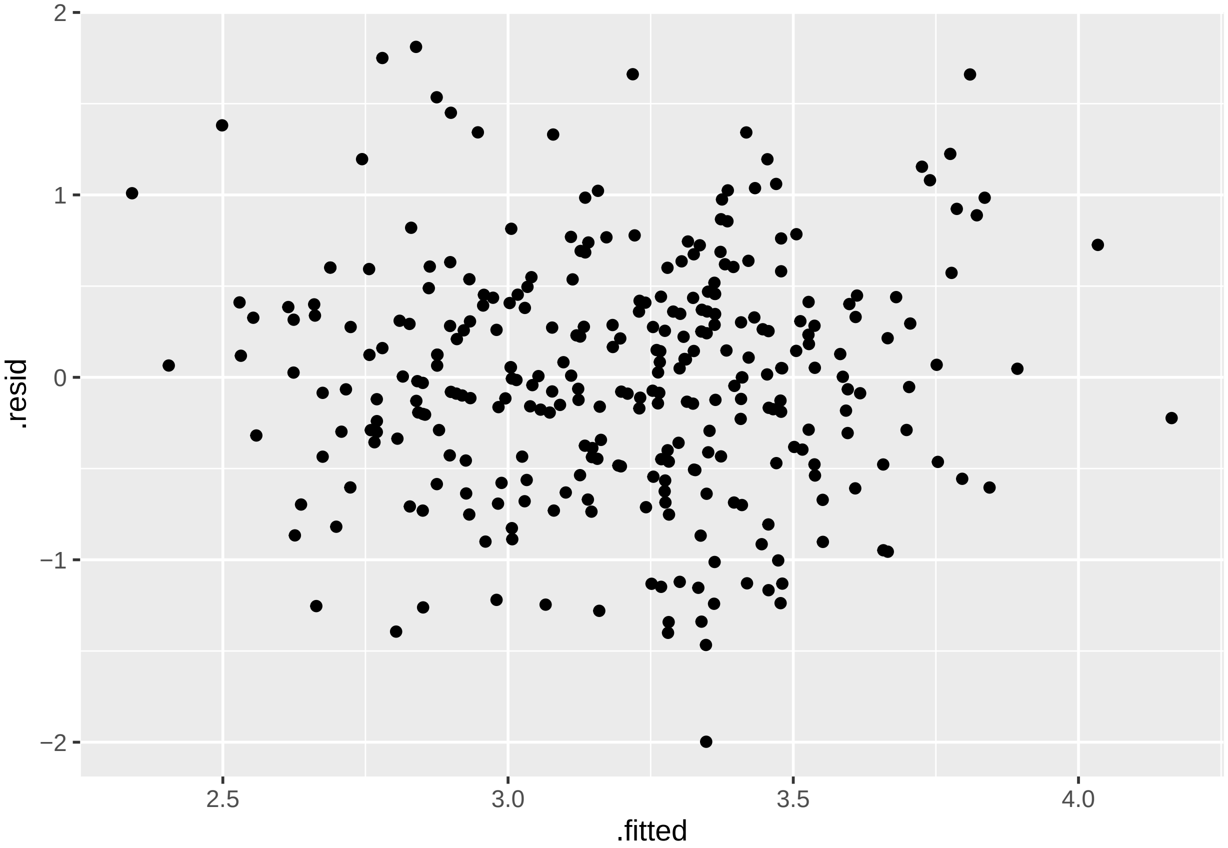

Figure 31: Debt data: residuals vs. fitted values from model `debt.2`

```
ggplot(debt.2, aes(sample = .resid)) +
  stat_qq() + stat_qq_line()
```

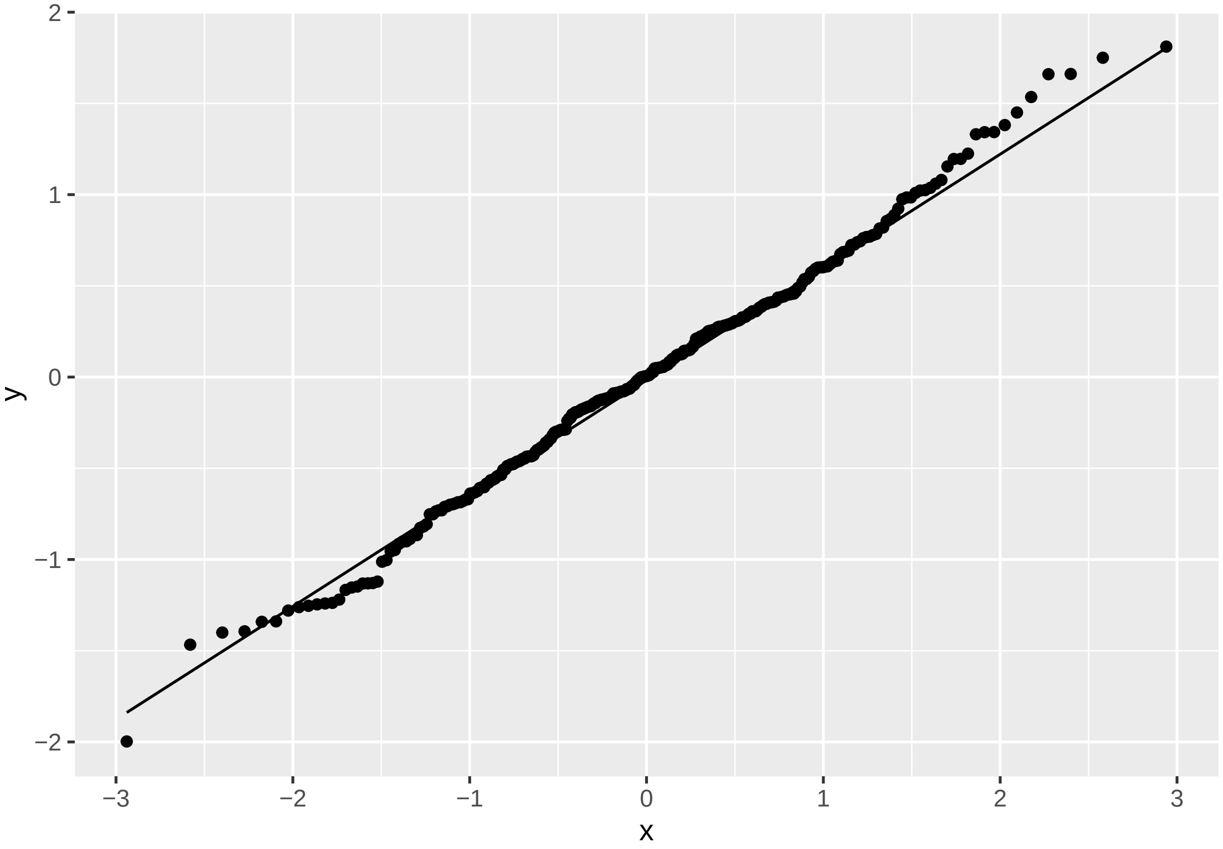

Figure 32: Debt data: normal quantile plot of residuals from model `debt.2`

```
w <- c(0.5, 5.4, 3.7, 13.8, 12.9, 4.0, 17.3, 6.6, 4.8, 2.5)
```

Figure 33: Observed data for estimating β by Bayesian methods

```
expo_fit
```

```
##   variable   mean median   sd  mad     q5    q95 rhat ess_bulk ess_tail
##       lp__ -31.94 -31.76 0.53 0.22 -32.80 -31.59 1.00     1379     1212
##       beta   0.16   0.16 0.04 0.04   0.11   0.24 1.00     1004     1154
```

Figure 34: Summary of posterior distribution of β