Instructions

There are nine questions worth a total of 100 marks. You should attempt them all.

QUESTION 1

(a) This question is about simple descriptive statistics. Here is a small data set

5, 5, 10, 12, 15, -10

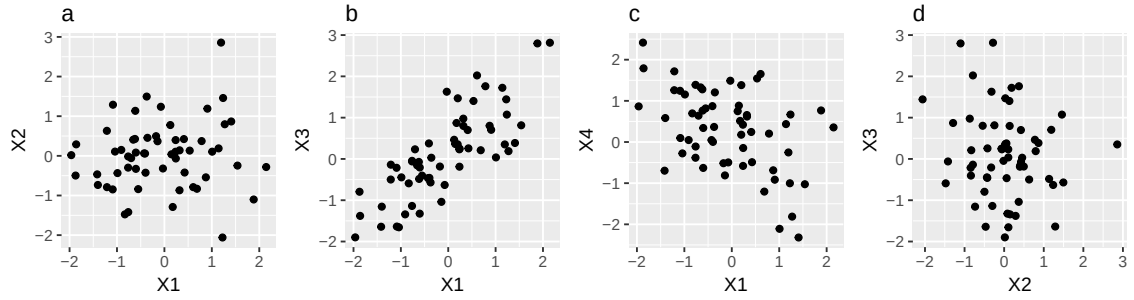
(a) Compute the sample median.
7.50

(b) Could the standard deviation by -5?
No, it cannot be negative

(c) If 50 was added to the numbers what would the **median** be?
10

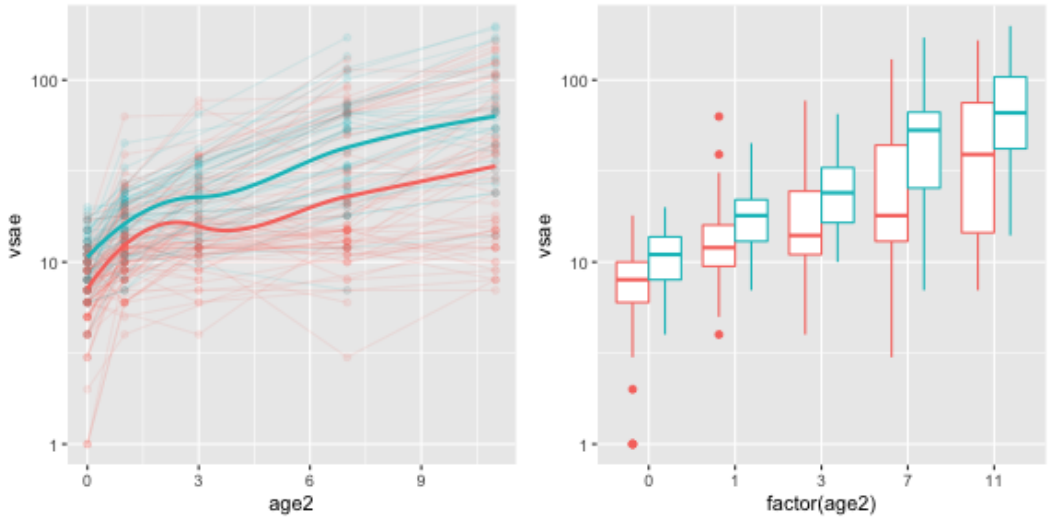
(d) If 5 was added to all the values how would the mean change?
It would increase by 5

(b) Match these correlations (-0.5, 0.8, 0.1, -0.1) to the plot.



a: 0.1, b: 0.8, c: -0.5, d: -0.1

(c) Both of the displays below show the variable vsae (assessing social skills) for children measured at several times, for two groups of children, one diagnosed as autistic and the other not.



- Which plot does this code generate?

```
ggplot(autism, aes(x=factor(age2), y=vsae, colour=bestest2)) +
  geom_boxplot() + scale_y_log10()
```

  The left one, because the right one would have boxplots.
- Why do you think the y axis is put on a log scale?
  The distribution of the original variable is right-skewed, and the log transformation fixes this.
- Which plot is the best design to answer "Does the variation in vsae scores differ between the two groups as they get older?"
  The right one, because you can compare the size of the boxes for both groups.

[Total: 0 marks]

— END OF QUESTION 1 —

QUESTION 2

Below are the first few rows of data measuring the tuberculosis incidence in Australia over a number of years.

```
# A tibble: 29 x 22
    iso2  year  m_04 m_514 m_014 m_1524 m_2534 m_3544 m_4554
   <chr> <int> <int> <int> <int>  <int>  <int>  <int>  <int>
 1    AU  1980    NA    NA    NA     NA     NA     NA     NA
 2    AU  1981    NA    NA    NA     NA     NA     NA     NA
 3    AU  1982    NA    NA    NA     NA     NA     NA     NA
 4    AU  1983    NA    NA    NA     NA     NA     NA     NA
 5    AU  1984    NA    NA    NA     NA     NA     NA     NA
 6    AU  1985    NA    NA    NA     NA     NA     NA     NA
 7    AU  1986    NA    NA    NA     NA     NA     NA     NA
 8    AU  1987    NA    NA    NA     NA     NA     NA     NA
 9    AU  1988    NA    NA    NA     NA     NA     NA     NA
10    AU  1989    NA    NA    NA     NA     NA     NA     NA
# ... with 19 more rows, and 13 more variables: m_5564 <int>,
#   m_65 <int>, m_u <int>, f_04 <int>, f_514 <int>, f_014 <int>,
#   f_1524 <int>, f_2534 <int>, f_3544 <int>, f_4554 <int>,
#   f_5564 <int>, f_65 <int>, f_u <int>
```

(a) What are the variables?

iso2, year, gender, age

(b) Is the data in tidy format? If no, sketch out what a tidy format of this data would look like.

No

	iso2	year	gender	age	count
1	AU	1980	m	04	NA
2	AU	1981	m	514	NA
3	AU	1982	m	1524	NA

(c) What does "NA" in the data mean?

Missing value

(d) What wrangling verb would have been used to pick the rows corresponding to Australia?

filter

[Total: 0 marks]

— END OF QUESTION 2 —

QUESTION 3

(a) A file with the following name "COMELB.TAB" will typically have what sort of information in it?

(a) temporal data, (b) spatial data, (c) economic data, (d) gambling data, (e) comma separated values

Spatial, typically map polygons

(b) Data in a web page is often provided in what format (choose all that apply):

(a) JSON, (b) html table, (c) comma separated values, (d) sqlite, (e) wav

html table

(c) The PISA data was provided as ".sav" format. What is this format?

SPSS binary file

[Total: 0 marks]

— END OF QUESTION 3 —

QUESTION 4

This question is about tidying and wrangling data. In the french fries data, 10 week sensory experiment, 12 individuals assessed taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do they taste?), fried in one of 3 different oils, replicated twice.

```
> head(french_fries)
   time treatment subject rep potato buttery grassy rancid painty
61    1         1       3   1    2.9     0.0    0.0    0.0    5.5
25    1         1       3   2   14.0     0.0    0.0    1.1    0.0
62    1         1      10   1   11.0     6.4    0.0    0.0    0.0
26    1         1      10   2    9.9     5.9    2.9    2.2    0.0
63    1         1      15   1    1.2     0.1    0.0    1.1    5.1
27    1         1      15   2    8.8     3.0    3.6    1.5    2.3
```

(a) What processing steps do you need to do to examine the replicates against each other?

gather the rating variables, and then spread the rep column into two columns

(b) What processing steps do you need to do to check the completeness of the experiment, that is, whether each taster tasted the chips for each week, for each oil type?

Count the number of values that are not missing, over time and treatment

(c) If I want to create a new variable called "yucky" which is a sum of the ratings on grassy, rancid and painty, what processing steps are needed?

mutate(yucky=grassy+rancid+painty)

(d) To study the temporal trend for the average rancid rating over the weeks of the study, what needs to be done to the data?

Select the variables, time, treatment, subject, rep, rancid. Compute mean of rancid by time, treatment and subject. (You need to take subject and oil into account because there are likely differences between oil type, and subjects perception that would invalidate combining them.

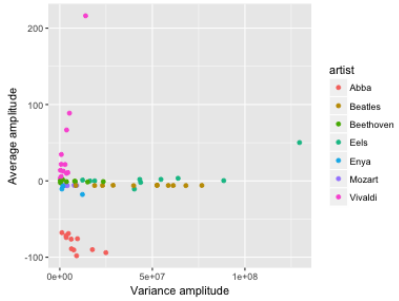
[Total: 0 marks]

— END OF QUESTION 4 —

QUESTION 5

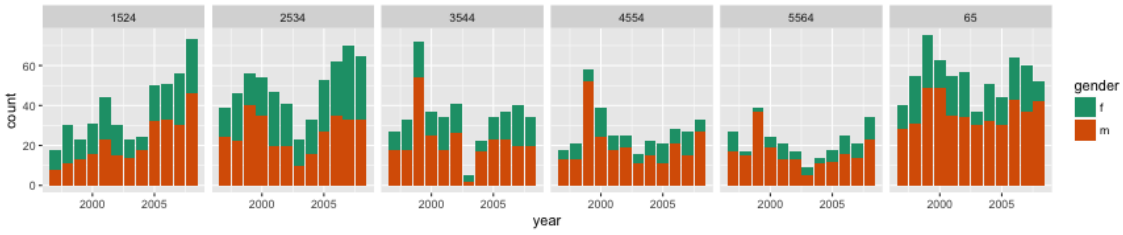
The grammar of graphics provides a mapping from variables in the tidy data to visual elements of a plot. For each of the following plots, specify the grammar that created it, all seven components. (The R code creating the plots is provided to help you.)

(a)
```
ggplot(music, aes(x=lvar, y=lave, colour=artist)) + geom_point() +
  xlab("Variance amplitude") + ylab("Average amplitude") +
  theme(aspect.ratio=1)
```



DATA: music
AESTHETICS/MAPPINGS: x=lvar, y=lave, colour=artist
GEOM: point
STAT: identity
POSITION: identity
COORDINATE: cartesian
FACET: none

(b)
```
ggplot(tb, aes(x = year, y = count, fill = gender)) +
  geom_bar(stat = "identity") +
  facet_grid(~ age) +
  scale_fill_brewer(palette="Dark2")
```

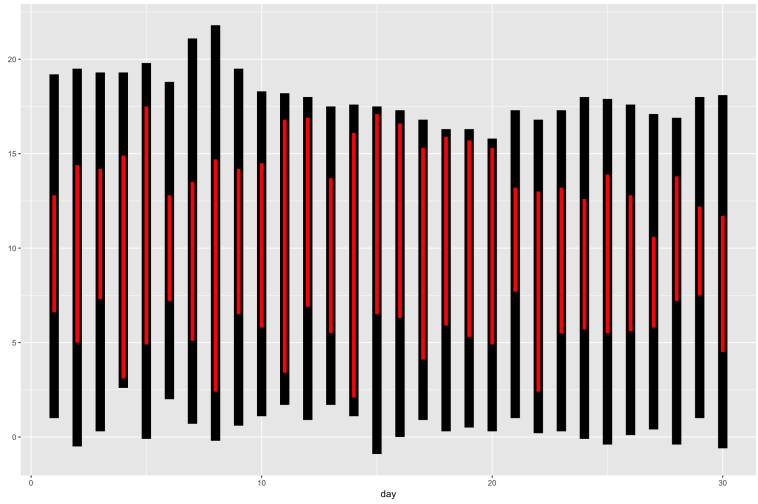


DATA: tb
AESTHETICS/MAPPINGS: x=year, y=count, fill=gender
GEOM: bar
STAT: identity
POSITION: identity
COORDINATE: cartesian
FACET: none

(c)
```
melbtemp_global <- melbtemp %>%
  filter(year < 2017, month==6, day<31) %>%
  group_by(day) %>%
  summarize(TMIN=min(TMIN, na.rm=T), TMAX=max(TMAX, na.rm=T))

july <- melbtemp %>% filter(year==2017, month==6, day<31)
ggplot() +
```

```
geom_linerange(data=melbtempglobal,
               aes(x=day, ymin=TMIN, ymax=TMAX), size=5) +
geom_linerange(data=july,
               aes(x=day, ymin=TMIN, ymax=TMAX), size=2, colour="red")
```



Layer 1:
DATA: melbtempglobal
AESTHETICS/MAPPINGS: x=day, ymin=TMIN, ymax=TMAX
GEOM: linerange
STAT: identity
POSITION: identity COORDINATE: cartesian
FACET: none
Layer 2:
DATA: july
AESTHETICS/MAPPINGS: x=day, ymin=TMIN, ymax=TMAX
GEOM: linerange
STAT: identity
POSITION: identity
COORDINATE: cartesian
FACET: none

(d) The purpose of using a grammar of graphics to make data plots is:

(i) Be able to compare different designs, (ii) to map variables specifically into elements of the plot, (iii) better connect data plots to statistics, (iv) all of these

(iv) all of these

[Total: 0 marks]

— END OF QUESTION 5 —

QUESTION 6

A small manufacturing firm specializes in making five types of automobile parts. Each part is first cast from iron in the casting shop and then sent to the finishing shop for detailing. The required work hours for producing each part and the profits associated with each part are given below. The total labour hours for the casting and finishing shops over the next month are 700 and 1000 hours respectively. Use this information to answer the following questions

Part	one	two	three	four	five
Casting Hours	2	1	2	3	1
Finishing Hours	3	2	2	1	2
Profit	30	20	40	25	10

(A) State the objective function. (2pt) Maximize profit.

(B) What are the decision variables? (2pt) The number of each part, $x_1, ..., x_5$.

(C) Formulate the mathematical optimization problem. (5pt) See below

$$
\begin{aligned}
(2pt)\quad & \max\ 30 * x_1 + 20 * x_2 + 40 * x_3 + 25 * x_4 + 10 * x_5 \\
& 0 \leq x_1, ..., x_5 \\
(1pt)\quad & x_1, ..., x_5 \in \mathbb{N} \\
(1pt)\quad & 2 * x_1 + 1 * x_2 + 2 * x_3 + 3 * x_4 + x_5 \leq 700 \\
(1pt)\quad & 3 * x_1 + 2 * x_2 + 2 * x_3 + 1 * x_4 + 2 * x_5 \leq 1000
\end{aligned}
$$

(D) Solve the mathematical optimization problem. (5pt) (4pt) For correct answer: $x^* = (0, 0, 350, 0, 0)$. There is another feasible solution with the exact some profit: $x^* = (0, 300, 200, 0, 0)$. Obviously, both solutions are correct. (1pt) If they note that more profit is obtained by producing part three than others at a lower relative cost.

(E) The company currently has orders for 100 units of part one, 200 units of part two and 100 units of part five.

(E.1) Restate the mathematical optimization problem using this additional information. (4pt)

$$
\begin{aligned}
(1pt)\quad & \max\ 30 * x_1 + 20 * x_2 + 40 * x_3 + 25 * x_4 + 10 * x_5 \\
& 0 \leq x_1, ..., x_5 : \qquad\qquad x_1, ..., x_5 \in \mathbb{N} \\
(1pt)\quad & 2 * x_1 + 1 * x_2 + 2 * x_3 + 3 * x_4 + x_5 \leq 700 \\
(1pt)\quad & 3 * x_1 + 2 * x_2 + 2 * x_3 + 1 * x_4 + 2 * x_5 \leq 1000 \\
(1pt)\quad & x_1 \geq 100,\ x_2 \geq 200,\ x_5 \geq 100
\end{aligned}
$$

(E.2) How will these additional constraints affect the optimal solution from question (D)? (4pt)

(0) This is one of the tougher exam questions. The students should realize that the solution to the previous question will change and that the required inputs for x_1, x_2, x_5 will bind.

(2pt) For noting that the required constraints on x_1, x_2, x_5 will bind.

(2pt) For noting that this will lead to a decrease in the number of units for x_3.

(+2EC) if they can work out the actual decrease in the optimum of x_3: $x^{**} = (100, 200, 25, 50, 100)$. Solution:

$$\begin{aligned} x_1^{**} = 100 &\implies (C_l, F_l) = (200, 300) \\ x_2^{**} = 200 &\implies (C_l, F_l) = (200, 400) \\ x_5^{**} = 100 &\implies (C_l, F_l) = (100, 200) \\ &\implies (C_l, F_l) = (500, 900) \\ &\implies x_3^{**} = 2 * C_l + 2 * F_l \\ &\implies x_4^{**} = 3 * C_l + 1 * F_l \end{aligned}$$

and so we only have left 200 units of C_l and 100 units of F_l. This means we can only produce at most 50 units of x_3. However, it may be that we can do better by making some of x_3 and some of x_4 with the remaining units. If you make 50 units of x_4 you still have 50 units with which to make x_3, which would yield $x_3^{**} = 25$ and $x_4^{**} = 50$. You can easily show that this solution is feasible and that producing with $x_3^{**} = 50, x_4^{**} = 0$ yields less profit than $x_3^{**} = 25, x_4^{**} = 50$.

[Total: 0 marks]

— END OF QUESTION 6 —

QUESTION 7

The management of a large tropical resort would like to improve their guests overall satisfaction by building at least one new entertainment attraction. The following attractions are currently under consideration: swimming pool, laser-tag, outdoor basketball court and movie theater. Resort management aims to provide the facilities that will be used by the largest number of guests at the resort. The resort faces the following budget and land restrictions: the total budget is $400,000 and the resort can use no more than 14 acres of land.

The swimming pool and laser-tag facilities both require lockers. However, due to health and safety regulations, these facilities can only be built in the laser-tag arena. Therefore, if the swimming pool is to be built, laser-tag must also be built. However, laser-tag can be built independent of the swimming pool. The resorts current outdoor space limitations ensure that there is only enough land to build either the basketball court or the movie theater, but not both.

Predicted daily usage and facilities costs (in $1,000) are shown below:

From	Usage (number of guests)	Cost ($1,000)	Land (acres)
Swimming Pool	325	$130	4
Laser-Tag	350	$180	4
Movie Theater	270	$150	5
Basketball Court	210	$80	8

(A) State the objective function. (1pt) maximize overall guest facilities usage.

(B) What are the decision variables in part (A)? (2pt) $X_i = 1$ if facility i is built, zero else. If students do not put the values this variable takes, deduct one mark.

(C) Formulate the mathematical optimization problem based on the above description. (9pt) See the following breakdown.

$(2pt)$ $\max 325 * X_1 + 350 * X_2 + 270 * X_3 + 210 * X_4$

$(1pt)$ $X_i \in \{0, 1\}$

$(2pt)$ $130 * X_1 + 180 * X_2 + 150 * X_3 + 80 * X_4 \leq 400$

$(1pt)$ $4 * X_1 + 4 * X_2 + 5 * X_3 + 8 * X_4 \leq 14$

$(1pt)$ $X_1 \leq X_2,$ or $X_1 - X_2 \leq 0$

$(1pt) X_3 + X_4 \leq 1$(accept $=$ as well)

$(1pt) x_1 + x_2 + x_3 + x_4 \geq 1$

[Total: 0 marks]

— END OF QUESTION 7 —

QUESTION 8

We observe data $\{y_i, x_{1i}\}$, $i = 1, ..., n$ from the following linear regression model:

$$Y = \beta_0 + X_1\beta_1 + \epsilon.$$

Use this regression model to answer the following questions.

(A) State the unconstrained optimization problem associated with minimizing the sum of squared errors associated with the above regression model, i.e., the usual least square objective function. Is this a linear or nonlinear optimization problem in the unknown parameters? (1pt) (.5pt) for stating that it is nonlinear or quadratic. Two, (.5pt) for the following problem

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2.$$

(B) Assume we know that β_1 satisfies $\beta_1 = a + h$, for some known constant h and unknown parameter a. State the constrained optimization problem associated with minimizing the sum of squared errors. (1pt)

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \;\; s.t.\; \beta_1 = a + h.$$

(C) Substitute the constraint in problem (B) into the least squares objective function and state the resulting optimization problem.(1pt)

$$\min_{\beta_0,a} \sum_{i=1}^{n} (y_i - \beta_0 - hx_{i1} - ax_{i1})^2.$$

(D) Solve, in the unknown parameters, the optimization problem in part (C) and give explicit formula for all unknown parameters. Hint: the calculations will simplify if you define a new outcome variable $y_i^* = y_i - x_i h$. (8pt)

$$\min_{\beta_0,a} \sum_{i=1}^{n} (y_i^* - \beta_0 - ax_{i1})^2$$

$$0 = \sum_{i=1}^{n} (y_i^* - \hat{b} - \hat{a}x_{i1})$$

$$0 = \sum_{i=1}^{n} x_{i1}(y_i^* - \hat{b} - \hat{a}x_{i1})$$

$$\Longrightarrow$$

$$\begin{pmatrix} \hat{b} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} \bar{y}^* - \hat{a}\bar{x} \\ \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x})(y_i^*-\bar{y}^*)}{\sum_{i=1}^{n}(x_{i1}-\bar{x})^2} \end{pmatrix}$$

[Total: 0 marks]

— END OF QUESTION 8 —

QUESTION 9

A linear model of number of tvs and books in the household is fitted to math scores. These are the results.

Houses:

```
> summary(pisalm1)

Call:
glm(formula = math ~ nbooks + ntvs, data = pisa_au_nomiss, weights = W_FSTUWT)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1434.47   -202.04    -21.14    182.60   1401.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 472.5491     4.0099  117.85   <2e-16 ***
nbooks       20.7757     0.5065   41.02   <2e-16 ***
ntvs        -12.9661     0.9983  -12.99   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for gaussian family taken to be 112872.8)

    Null deviance: 1581870839  on 12117  degrees of freedom
Residual deviance: 1367453601  on 12115  degrees of freedom
AIC: 144148

Number of Fisher Scoring iterations: 2
```

(a) What variables contribute significantly to the model?

nbooks, ntvs

(b) Write down the linear model equation.

$\hat{y} = 472.5 + 20.8nbooks - 13.0ntvs$

(c) What does looking at the null deviance and residual deviance tell you about the model fit?

The model only explains a very small amount in the variation in math scores, because the difference between these two numbers is not very big.

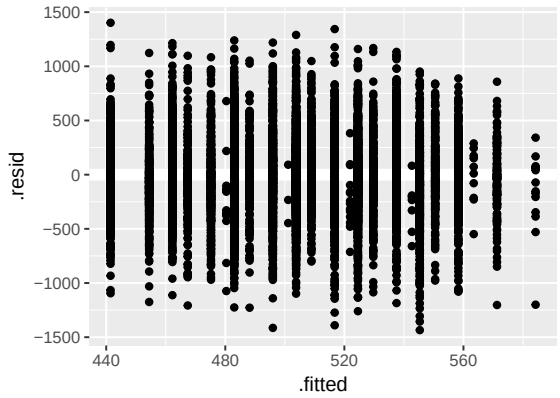
(d) What criteria was (algebraicly) optimised to yield the fitted model?

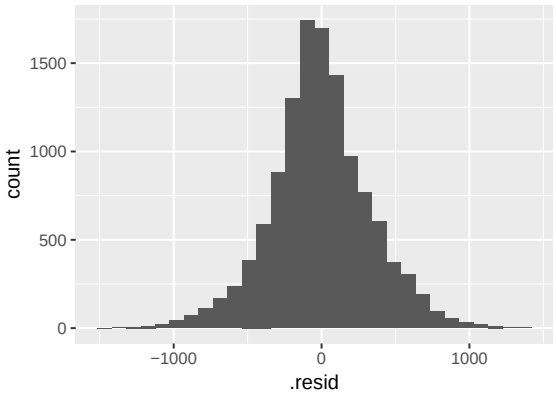
$\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$

(e) Write down the interpretation of the effect on nbooks on the model, assuming that ntvs is fixed. (nbooks has these categories 1=0-10 books, 2=11-25 books, 3=26-100 books, 4=101-200 books, 5=201-500 books, 6=More than 500 books.)

With each increase in a category of books math score increases by 21 points, on average. The more books in the household the higher the average student's score on the math test.

(f) Below are some plots of the residuals from the model fit. Explain what you learn about the assumption that the error has a normal distribution by looking at these.





There is nothing of concern about the error assumption. The residuals are mostly spread evenly above and below 0, in the residuals vs fitted plot. The histogram is unimodal and symmetric.

(g) The full model is fit to the data using additional variables gender, ANXTEST, PARED, JOYSCIE, WEALTH. And the new AIC=142044. Discuss whether this is a better model.

Technically yes, because AIC drops by about 2000. But the predictive power of the model is still weak. These variables explain very little of the variability in math scores.

(h) True or false. "A statistically significant regression model means that the explanatory variable CAUSES the response variable."

False

[Total: 0 marks]

— END OF QUESTION 9 —

QUESTION 10

Regression (decision) trees are fit to data, by recursively partitioning it into subsets. Below is a summary of the fit for the PISA data.

```
> pisarp
n= 12118

node), split, n, deviance, yval
      * denotes terminal node

1) root 12118 1581871000 500.3926
  2) nbooks< 2.5 3065  304715000 453.0841 *
  3) nbooks>=2.5 9053 1133123000 514.0852
    6) nbooks< 4.5 6005  684041300 502.2830 *
    7) nbooks>=4.5 3048  402388400 537.0283 *
```

(a) How many observations are in the data?

12118

(b) What variable and value is the first split made on?

nbooks, 2.5

(c) How many terminal nodes in the tree?

3

(d) Write down the decision rules corresponding to the model.

If nbook >= 2.5 then check
... if nbook < 4.5 then predict math= 502.2830
... else, then predict math= 537.0283
else predict math= 453.0841

(e) Partitions are decided by optimising what criteria,

$$SS_T - (SS_L + SS_R) \text{ where } SS_T = \sum_{i=1}^{\#before\ split} (y_i - \bar{y})^2,$$

(f) How many observations are thee in the terminal node where $nbooks < 2.5$?

3065

[Total: 0 marks]

— END OF QUESTION 10 —

Formula sheet

Summary statistics

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}, \qquad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Types of variables: categorical, quantitative, logical, date.

Descriptive words for univariate distributions:

- unimodal, bimodal, multimodal
- symmetric, right-skewed, left-skewed, uniform
- outliers

Descriptive words for bivariate distributions:

- shape: linear, non-linear, no relationship
- strength: weak, moderate, strong
- form: positive, negative

Tidy data

Verbs: gather, spread, nest/unnest, separate/unite

Wrangling data

Verbs: filter, arrange, select, mutate, summarise, group/ungroup

Grammar of graphics

There are seven components of the grammar that define a data plot: DATA, AESTHETICS/MAPPINGS, GEOM, STAT, POSITION, COORDINATE, FACET.

Colour palettes: sequential, diverging, qualitative

Optimization

One variable

For a single variable x and $f(x)$ a continuously differentiable function on $[a, b]$, recall that the conditions for a local optima are as follows:

$$\begin{aligned} f'(x) = 0 &\quad \text{First-order condition,} \\ f''(x) < 0 &\quad \text{Second-order condition: Max,} \\ f''(x) > 0 &\quad \text{Second-order condition: Min.} \end{aligned}$$

Two variables

For two variables x, y and $f(x, y)$ a continuously differentiable function on $[a, b] \times [a, b]$, recall that the conditions for a local optima are as follows:

$$\begin{pmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{First-order condition,}$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} < 0,\ \frac{\partial^2 f(x,y)}{\partial y^2} < 0,\ \left\{ \left(\frac{\partial^2 f(x,y)}{\partial x^2} \right) \left(\frac{\partial^2 f(x,y)}{\partial y} \right) - \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\} > 0 \quad \text{Second-order condition: Max,}$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} > 0,\ \frac{\partial^2 f(x,y)}{\partial y^2} > 0,\ \left\{ \left(\frac{\partial^2 f(x,y)}{\partial x^2} \right) \left(\frac{\partial^2 f(x,y)}{\partial y} \right) - \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\} > 0 \quad \text{Second-order condition: Min.}$$

Models

Simple linear:

$Y = \beta_0 + \beta_1 X + \varepsilon$

- $\varepsilon \sim N(\mu, \sigma)$
- Fitted values: $\hat{Y} = b_0 + b_1 X$
- Residual: $e = Y - \hat{Y}$
- Estimates: $b_1 = r \frac{s_y}{s_x}$, $b_0 = \bar{Y} - b_1 \bar{X}$
- $R^2 = 1 - \frac{\sum e^2}{\sum Y^2}$
- $MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-2)}$
- $RMSE = \sqrt{MSE}$
- $MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{(n-2)}$

Decision trees:

ANOVA criterion: $SS_T - (SS_L + SS_R)$, $SS_T = \sum (y_i - \bar{y})^2$, and SS_L, SS_R are the equivalent values for the two subsets created by partitioning.