Instructions

There are 9 questions worth a total of 100 marks. You should attempt them all.

QUESTION 1

(a) For the following data,

5, 20, -3, 3, 2

(i) Compute the mean, and standard deviation.

[2 marks]

5.4, 8.68

(ii) What would happen to the mean and standard deviation, if 20 was removed from the calculations?

[2 marks]

Both would get smaller.

(b) If this exam were a game, and you knew that your professor's typical TRUE/FALSE question construction yielded TRUE being the answer 75% of the time, which of the following would be your optimal strategy if you really are guessing the correct answer? (Choose one)

[2 marks]

(i) Flip a coin

(ii) Write down TRUE

(iii) Write down FALSE

(iv) Not answer

(ii)

[Total: 6 marks]

— END OF QUESTION 1 —

QUESTION 2

Here are two sequences of 20 coin flips from a fair coin.

A: H,T,H,T,H,H,T,T,T,H,T,H,H,T,T,H,T,H,T,T

B: H,T,H,H,T,T,H,H,T,T,T,T,T,T,T,H,H,H,T

(a) What does the word "fair" imply?

[2 marks]

The probability of H or T is 0.5

(b) One of the sequences is the result of an actual coin flip, and the other is a human written sequence based on what they think might be seen in an actual sequence. Which one is the actual coin flip? Why do you think this?

[2 marks]

B, because it has a long run of T's. This is likely to happen in the short-term. When people write down a sequence that they think might occur they tend to put a more balanced number of H's and T's, and tend not to write long runs of either.

[Total: 4 marks]

— END OF QUESTION 2 —

QUESTION 3

(a) The _____ hypothesis often represents a claim to be tested. The _____ hypothesis represents an alternative claim under consideration and is often represented by a range of possible parameter values.

[3 marks]

Fill in the blanks using what you think is the best choice among these words: *statistical, null, alternative, permutation, random, p-value, power, type I error, type II error, rejection, fail to reject*

null, alternative, in this order

(b) Explain in your own words what the term p-value means.

[3 marks]

Assuming that the null hypothesis is true, the probability of observing a value of the test statistic as or more extreme than this.

(c) The 2004 National Technology Readiness Survey sponsored by the Smith School of Business at the University of Maryland surveyed 418 randomly sampled Americans, asking them how many spam emails they receive per day in proportion to their total number of emails. The survey was repeated on a new random sample of 499 Americans in 2009.

(i) What are the hypotheses for evaluating if the average spam emails per day has changed from 2004 to 2009.

[3 marks]

$H_o : \mu_{2004} = \mu_{2009}$, $H_a : \mu_{2004} \neq \mu_{2009}$, where μ is the average proportion of spam for each year.

(ii) Write out the steps required to conduct the hypothesis test using permutation.

[4 marks]

1. Consider the data to have two columns: year, proportion of spam
2. Compute the average proportion for each year, of the data, and the absolute value of the difference between the two. This is your test statistic value, call it d.
3. Take one of the variables and randomly sample, to scramble the values
4. Compute the average proportion for each year, of the permuted data, and the absolute value of the difference between the two, call this d_p.
5. Repeat steps 3, 4 many times, lets say 1000. Collect each of the d_p, and save.
6. Count the number of d_p values bigger than d. This is the p-value. Make the decision to reject or fail to reject H_o based on the p-value.

(iii) If the permutation p-value was 0.15, what would your decision be?

[2 marks]

Fail to reject, because it would be considered to be large.

[Total: 15 marks]

— END OF QUESTION 3 —

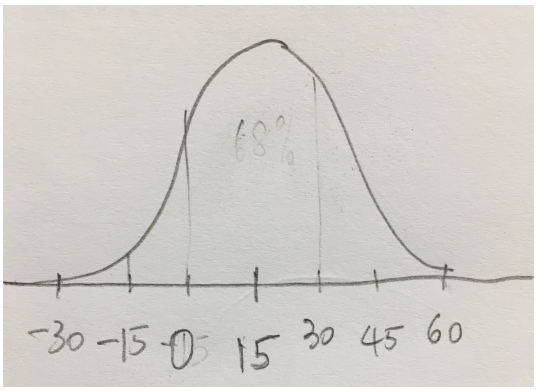
QUESTION 4

This question is about statistical distributions.

(a) The Capital Asset Pricing Model is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 15% (i.e. an average gain of 15%) with a standard deviation of 15%. A return of 0% means the value of the portfolio doesn?t change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money.

(a) Make a sketch of the distribution.

[3 marks]



(b) What percent of years does this portfolio lose money, i.e. have a return less than 0%?

[3 marks]

0 is one standard deviation below the mean. Between one standard deviation below, and one above contains 68% of values. Thus to the left of 0, will have 16% of returns.

(c) What is the cutoff for the highest 2.5% of annual returns with this portfolio?

[2 marks]

45%

(b) For an exponential distribution, $f(x \mid \lambda) = e^{-\lambda x} \quad x \geq 0$,

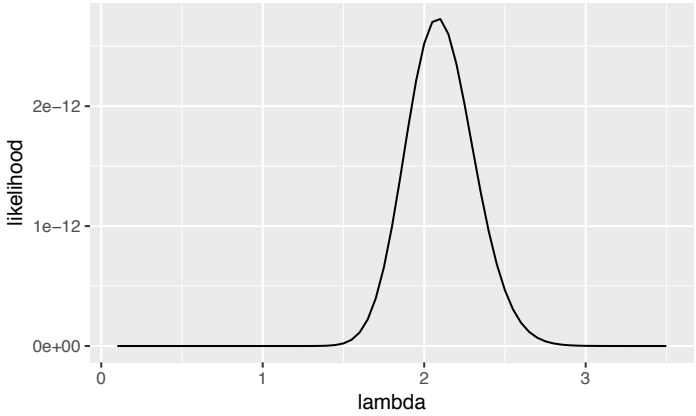
(a) you have a sample of data, $x_1 = 0.97, x_2 = 1.71, x_3 = 0.66$, write down the likelihood function.

[3 marks]

$l(\lambda) = e^{-\lambda 0.97} e^{-\lambda 1.71} e^{-\lambda 0.66}$

(b) Here is a plot of the likelihood function for a much larger sample, $n = 100$. What would be the best estimate for λ?

[2 marks]



About 2.1

(c) Fill in the blank on this statement of the Central Limit Theorem

Let $\{X_1, ..., X_n\}$ *be a random sample of size* n *- that is, a sequence of independent and identically distributed random variables drawn from a distribution mean given by* μ *and finite variance given by* σ*. The sample average is defined* $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$*, then as* n *gets large the distribution of* $\bar{X}$ *approximates a* _____

[2 marks]

$N(\mu, \sigma/\sqrt{n})$

[Total: 15 marks]

— END OF QUESTION 4 —

QUESTION 5

This question is about model building on the OECD PISA 2015 data, using a generalised linear model. A description of the variables used is as follows:

Variable name	Description	Coding
ST004D01T	Gender	1=Female, 2=Male
ANXTEST	Personality: Test Anxiety (WLE)	
PARED	Index highest parental education in years of schooling	
JOYSCIE	Enjoyment of science (WLE)	
WEALTH	Family wealth (WLE)	
ST013Q01TA	How many books are there in your home?	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
ST012Q01TA	How many TVs in your home;	1= None, 2= One, 3= Two, 4= Three or more
SENWT	Weight	Reflects how the student represents other students in Australia based on socioeconomic and demographic characteristics

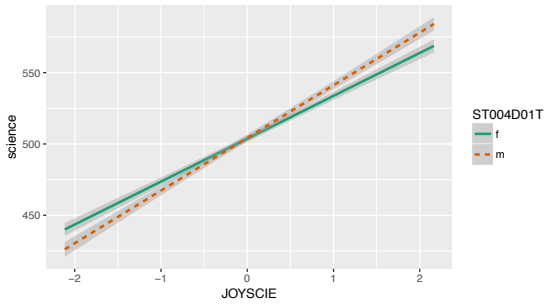
(a) Weighted multiple regression is needed because this is survey data with weights reflecting the relative representation of students in the population. (TRUE or FALSE).

[2 marks]

TRUE

(b) This plot shows science score against enjoyment of science, coloured by gender. Does it suggest an interaction between ST004D01T and JOYSCIE would be helpful in the model? Explain your answer.

[3 marks]



Because the lines cross, and the standard error bars are separated, it suggests an interaction would be useful.

(c) Below is a summary of the coefficients in the fitted model. Which variable might be considered the least important for the model? Should it be dropped and the model re-fitted?

[3 marks]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	403.05	7.20	56.01	0.00
ST004D01Tm	1.03	1.59	0.65	0.52
JOYSCIE	23.49	0.91	25.95	0.00
ANXTEST	-6.22	1.09	-5.71	0.00
PARED	6.59	0.41	16.06	0.00
WEALTH	4.90	1.02	4.82	0.00
ST013Q01TA	18.35	0.56	32.52	0.00
ST012Q01TA	-13.37	1.20	-11.18	0.00
ST004D01Tm:JOYSCIE	6.24	1.28	4.88	0.00
ST004D01Tm:ANXTEST	-4.91	1.59	-3.10	0.00

ST004D01T is not significant. However it is important for predicting science when included as an interaction with both JOYSCIE and ANXTEST.

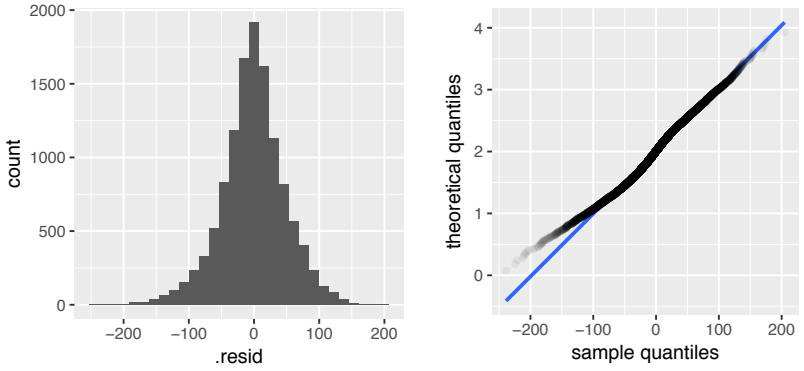
(d) Explain how the number of TVs in the household tends to affect the average science score.

[3 marks]

The coefficient is negative. It is significantly different from 0. This suggests that each increase of one TV in the household reduces average science score by about 13 points.

(e) Below are plots of the residuals from the fitted model. Discuss the normaility assumption, and whether it is satisfied here.

[2 marks]



The histogram shows a very clear bell-shape. The normal probability plot shows most of the residuals match the theoretical quantiles. There is a small deviation from normality, in the lower values - there are a few much lower than expected residuals.

(f) This is a summary of the goodness of fit of the model for the model fitted with interaction terms (above) and a simpler model with no interaction terms. Which is more parsimonious (simple and accurate)?

[3 marks]

Model A: (as fitted above, with interactions)

```
glm(formula = science ~ ST004D01T * JOYSCIE + ST004D01T * ANXTEST +
    PARED + WEALTH + ST013Q01TA + ST012Q01TA, data = aus_nomiss,
    weights = SENWT)
...
      Null deviance: 40898612  on 12117  degrees of freedom
Residual deviance: 29085855  on 12108  degrees of freedom
AIC: 145211
```

Model B: (no interactions)

```
 glm(formula = science ~ ST004D01T + JOYSCIE + ANXTEST + PARED +
    WEALTH + ST013Q01TA + ST012Q01TA, data = aus_nomiss, weights = SENWT)
 ...
    Null deviance: 40898612  on 12117  degrees of freedom
Residual deviance: 29170010  on 12110  degrees of freedom
AIC: 145242
```

There is a slight improvement on the explanation of variation in science scores with model A, as seen by comparing the residual deviance which drops by 84155. The additional terms only add 2 df. It is a very slight improvement, though, so there is only a very small gain in predictive accuracy. It could be argued either way - for model B preference, that the interaction doesn't substantially help get a better explanation of average science scores.

[Total: 16 marks]

— END OF QUESTION 5 —

QUESTION 6

(a) The pedestrian sensor data collected in Melbourne has counts of passersby for each hour of each day over several years. To best model the data, to predict pedestrian counts at one sensor location, which of the following choices of generalised linear model would be most appropriate. Explain your choice.

[3 marks]

- binomial(link = "logit")
- gaussian(link = "identity")
- poisson(link="log")
- Gamma(link = "inverse")

poisson(link="log"), because the response variable is counts and the poisson distribution is the best match for this type of variable.

(b) In the OECD PISA data, students from different countries are given standardised tests for reading, math and science. Within each country the data is collected by first generating a random sample of schools, and then a random sample of 15 year olds within schools.

(i) What are the main assumptions made for fitting a multiple regression model?

[3 marks]

Linear relationship between response and predictors, independent and identically distributed errors, homoskedasticity, normality of error distribution

(ii) Explain the hierarchical dependencies that violate the multiple linear regression model assumptions, that would lead to using a multilevel model for predicting math scores based on other variables collected on each student.

[3 marks]

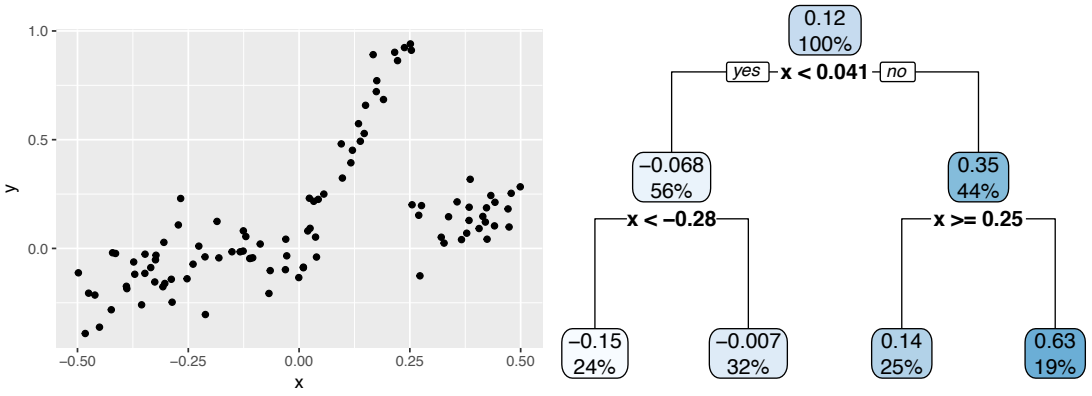
Students are sampled within schools, which possibly breaks the independent and identically distributed errors. Because we only have a sample of schools, and we would like to infer the results to other schools, we need to consider these as random effects.

[Total: 9 marks]

— END OF QUESTION 6 —

QUESTION 7

(a) Regression (decision) trees are fit to data, by recursively partitioning it into subsets. Below is a data set (x, y) and the resulting fitted regression tree.



(i) What value of x defines the first partition.

[2 marks]

0.041

(ii) How many terminal nodes in the tree?

[2 marks]

4

(iii) Write down the decisions that would need to be followed to obtain fitted values for the model.

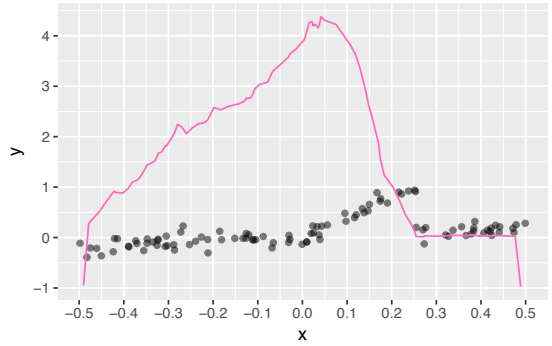
[4 marks]

If x< 0.041 then check
... if x< -0.28 then predict y=-0.150
... else, then predict y=-0.007
else check
... if x >= 0.25 then predict y=0.14
... else, then predict y= 0.63

(iv) Partitions are decided by optimising the criteria,

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

and SS_L, SS_R are the equivalent sum of squares for the left and right partition. This is a plot of the function, showing the partitions that were evaluated in order to decide on the best.



Which value of x corresponds to the optimal value of the function?

[2 marks]

0.041

(b) Random forests are an ensemble model built by combining predictions from models built on bootstrap samples, and by also random sampling the available variables. This question is about fitting a forest model to Melbourne property auction prices.

(i) Explain what a bootstrap sample is.

[2 marks]

Bootstrap takes a sample of the data, with replacement, with the same number of observations.

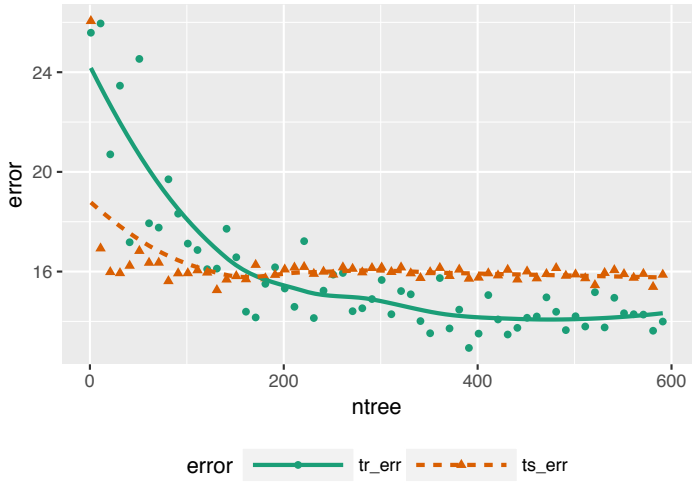
(ii) Error in the model is computed using the "out-of-bag" cases for each tree. What does "out-of-bag" mean?

[2 marks]

For each tree, some of the original observations are left out by the bootstrap sampling. These cases are therefore not used to construct the tree - they are the "out-of-bag" cases. Computing the predictive error on these cases gives a more accurate estimate of the model performance on new data.

(iii) The plot below shows the training (solid line) and test error (dashed) for forests built on differing numbers of trees. Explain why the test error is higher than training error as the number of trees increases. What is the recommended size of the forest?

[4 marks]



With increasing complexity, the model better fits the training data, but may fail to perform on a new sample. This is called overfitting. The best model balances complexity with predictive accuracy, which is where the training and test errors are closest. For this data, the balance occurs a little under 200 trees.

(c) Choose all of the following that are TRUE. (Feel free to explain your choices if needed, because partial credit might be awarded.)

[3 marks]

(i) Boosting as a method iteratively builds multiple models by re-weighting the cases, and combines the predictions.

(ii) Boosting generally leads to higher predictive accuracy.

(iii) Cross-validation is not generally needed with boosting because it doesn't typically over-fit the training data.

(iv) Boosting and bagging are both ways of re-sampling the data, and tend to give similar predictive accuracy.

(v) None of these are true.

(i) and (ii) are true.

[Total: 21 marks]

— END OF QUESTION 7 —

QUESTION 8

The game rock, paper, scissors is fun to play with a friend when you've got time to spare. Suppose you are player A, and your friend is player B. You are going to bet \$1 each time you play. Each time you play you have three choices, rock, paper or scissors, and your opponent has the same. If you play rock, and your opponent plays scissors, you win, but if they play paper you lose. Its a tie if you both make the same choice. Rock beats scissors, scissors beats paper, and paper beats rock.

(a) Write down the payoff matrix.

[3 marks]

	rock	paper	scissors
rock	0	-1	1
paper	1	0	-1
scissors	-1	1	0

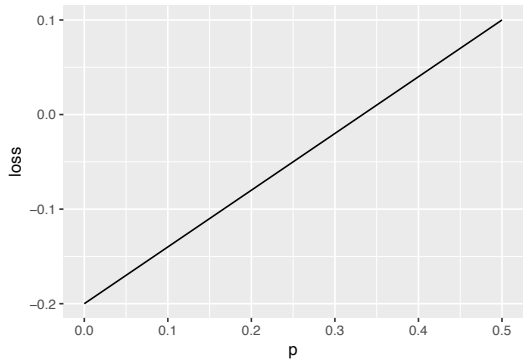
(b) Suppose that your friend chooses rock, paper, scissors with these probabilities: 0.5, 0.3, 0.2, write down your loss function if your strategy is to choose with these probabilities p, p, $1-2p$, where $p = 0 - 0.5$.

[3 marks]

$L(p) = 0 \times 0.5p + -1 \times 0.3p + 1 \times 0.2p + 1 \times 0.5p + 0 \times 0.3p + -1 \times 0.2p + -1 \times 0.5(1-2p) + 1 \times 0.3(1-2p) + 0 \times 0.2(1-2p) = 0.6p - 0.2$

(c) Plot the loss function. Which is the optimal strategy?

[2 marks]



Choose p=0.5

[Total: 8 marks]

— END OF QUESTION 8 —

QUESTION 9

(a) Bayes theorem states that given two events A, B, where $P(B) \neq 0$,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A), P(B)$ are the probabilities of observing A and B without regard to each other, $P(A|B)$ is the conditional probability, is the probability of observing event A given that B occurred, and $P(B|A)$ is the probability of observing event B given that A occurred.

This forms the basis of the Naive Bayes spam filter. Email messages are scrutinized for the appearance of key words in the subject, such as "Get a Date with the Hottest", that might indicate the message is spam. From previous experience we know that 60% of emails are spam, 1% of spam email have "Get a Date with the Hottest" in the subject, and 0.1% of non-spam emails have this sentence in the subject.

Consider event A to be the message is spam, and B to be the event that "Get a Date with the Hottest" appears in the subject line. Use Bayes theorem to find $P(A|B)$. (Hint: We also have the basic probability property, $P(B) = P(B|A)P(A) + P(B|not\ A)P(not\ A)$.) SHOW YOUR WORKING.

[4 marks]

$P(A) = 0.6, P(B|A) = 0.01, P(B) = P(B|A)P(A) + P(B|not\ A)P(not\ A) = 0.01 \times 0.6 + 0.001 \times 0.4 = 0.0064$

$P(A|B) = \frac{0.01 \times 0.6}{0.0064} = 0.9375$

(b) Given the probability you have calculated for $P(A|B)$, would you be advised to send the email to the spam folder? (If you couldn't find an answer to the previous question use 0.95.)

[2 marks]

Yes, because it is almost certain to be spam if it has that phrase in subject.

[Total: 6 marks]

— END OF QUESTION 9 —

Formula sheet

Hypothesis testing

Statistic	Population	Sample
Mean	μ	$\bar{X}$
Proportion	π	p
Correlation	ρ	r

- one sample test: $H_0 : \mu = \mu_0$ and $H_a : \mu > (or <)\mu_0$, $H_a : \mu \neq \mu_0$
- two sample: $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 > (or <)\mu_2$, $H_a : \mu_1 \neq \mu_2$
- α: Probability of Type I Error
- β: Probability of Type II Error
- One sample test statistic: $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

Statistical distributions

- Uniform, e.g. $P(X = x) = f(x) = 1/10, \quad x \in \{0, ..., 9\}$
- Normal: $N(\mu, \sigma)$, $f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$
- Exponential: $Exp(\lambda)$, $f(x \mid \lambda) = e^{-\lambda x} \quad x \geq 0$
- Poisson: $P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x \in \{0, 1, 2, ...\}$
- Binomial: $P(X = x \mid n, p) = \left(\begin{array}{c} n \\ p \end{array} \right) p^x (1-p)^{n-x} \quad x \in \{0, 1, 2, ..., n\}$
- Pareto: $f(x \mid \alpha, \lambda) = \frac{\alpha \lambda^\alpha}{(\lambda+x)^{\alpha+1}} \quad x > 0, \alpha > 0, \lambda > 0$
- Weibull: $f(x \mid \lambda, k) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{(-x/\lambda)^k}, \quad x \geq 0$
- Gamma: $f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}, \quad x \geq 0 \quad \alpha, \beta > 0$

Likelihood function:

$L(X_1, \ldots, X_n|\theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \theta) = f(x_1|\theta) f(x_2|\theta) \cdots f(x_n|\theta) = \prod_{i=1}^n f(x_i; \theta)$

Regression 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: $b_1 = r\frac{s_y}{s_x}$, $b_0 = \bar{Y} - b_1 \bar{X}$
- $R^2 = 1 - \frac{\sum e^2}{\sum Y^2}$