1. [10 marks] A pair of random variables $(X, Y)$ has a joint probability density function given by

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } 0 < y < x < 1 \\ 0 & \text{else.} \end{cases}$$

(a) Determine the marginal probability density function of $X$. [2 marks]

$$f_X(x) = \int_{-} f_{X,Y}(x,y)\, dy$$

$$= \int_0^x 2\, dy = 2x, \qquad x \in (0,1)$$

(b) Compute $\mathbb{E}[Y \mid X = x]$. [3 marks]

$$\mathbb{E}[Y|X=x] = \int y\, f_{Y|X}(y|x)\, dy$$

Assuming $x \in (0,1)$,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}$$

$y \in (0, x)$

$$\mathbb{E}[Y|X=x] = \int_0^x y \cdot \frac{1}{x}\, dy = \frac{1}{x} \times \frac{x^2}{2} = \frac{x}{2}$$

(c) Compute $\text{Cov}(X, Y)$. [3 marks]

$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

$\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y|X]] = \mathbb{E}[X^2/2] = \int_0^1 \frac{1}{2}x^2 \times 2x = \left[\frac{1}{4}x^4\right]_0^1 = \frac{1}{4}$

$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[X/2] = \frac{1}{2}\mathbb{E}[X] = 1/3$

$\mathbb{E}[X] = \int_0^1 x \cdot 2x\, dx = \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}$

$\text{Cov}(X,Y) = 1/4 - \frac{1}{3} \times \frac{2}{3} = \frac{1}{36}$

(d) Are $X$ and $Y$ independent? Justify your answer. [2 marks]

Many possible answers:

(1) As $\text{cov}(X,Y) \neq 0$, $X$ and $Y$ are not independent

(2) $\mathbb{E}[Y|X=x] \neq \mathbb{E}[Y]$, so $X$ and $Y$ are not independent.

(3) (long approach)

$f_Y(y) = \int_y^1 2\, dx = 2(1-y), \quad y \in (0,1)$

$f_{X,Y}(x,y) \neq f_X(x) f_Y(y)$ so $X$ and $Y$ are not independe

2. [6 marks] Let $U \sim \text{Uniform}(0,1)$ and define the random variable

$$X = -\log(\sqrt{1+3U} - 1).$$

(a) Find the probability density function of $X$. [4 marks]

$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(-\log(\sqrt{1+3U} - 1) \leq x\right)$ $\quad x > 0$

$= \mathbb{P}(\sqrt{1+3U} \geq 1 + e^{-x})$

$= \mathbb{P}(U \geq \frac{1}{3}((1+e^{-x})^2 - 1))$

$= 1 - \frac{1}{3}((1+e^{-x})^2 - 1)$

$f_X(x) = \frac{d}{dx} F_X(x) = \frac{2}{3} e^{-x}(1+e^{-x}) = \frac{2}{3}e^{-x} + \frac{2}{3}e^{-2x}$,

(b) The moment generating function of $X$ is

$$M_X(t) = \frac{2}{3(1-t)} + \frac{2}{3(2-t)}, \qquad t < 1.$$

Using the moment generating function, determine the variance of $X$. [2 marks]

$M_X(t) = \frac{2}{3}(1-t)^{-1} + \frac{2}{3}(2-t)^{-1}$

$\mathbb{E}(X) = M_X'(0)$, $M_X'(t) = \frac{2}{3}(1-t)^{-2} + \frac{2}{3}(2-t)^{-2}$

$= \frac{2}{3} + \frac{2}{3} \times \frac{1}{4} = 5/6$

$\mathbb{E}(X^2) = M_X''(0)$, $M_X''(t) = \frac{4}{3}(1-t)^{-3} + \frac{4}{3}(2-t)^{-3}$

$= \frac{4}{3} + \frac{4}{3} \times \frac{1}{8} = \frac{3}{2}$

$\text{Var}(X) = \frac{3}{2} - \left(\frac{5}{6}\right)^2 = 0.8055\ldots$

3. [10 marks] Maximal oxygen consumption is a way to measure the physical fitness of an individual. It is the amount of oxygen in millilitres a person uses per kilogram of body weight per minute (mL/kg/min). A medical researcher is investigating whether athletes have a greater mean maximal oxygen consumption than non-athletes. A total of 24 students was sampled from an American university. Each student was asked whether they were on an athletic scholarship or not, before having their maximal oxygen consumption level measured. The data are summarised in the table below:

| | $n$ | $\bar{x}$ | $s$ |
|---|---|---|---|
| Athlete | 10 | 57 | 6.9 |
| Non-athlete | 14 | 49 | 6.1 |

(a) Do athletes have higher maximal oxygen consumption than non-athletes? Answer this question by carrying out an appropriate hypothesis test. You should clearly state the null and alternative hypotheses, compute a $p$-value, and write a conclusion in a manner that can be understood by the medical researcher. [5 marks]

Let $\mu_A$ be maximal oxygen consumption for athletes and $\mu_N$ be maximal oxygen consumption for non-athletes.

Test $H_0 : \mu_A = \mu_N$ against $H_1 : \mu_A > \mu_N$

pooled sample variance $s_p^2 = \dfrac{(10-1)\times 6.9^2 + (14-1)\times 6.1^2}{10+14-2}$

$= 41.465$

test statistic $t = \dfrac{(57-49)-0}{\sqrt{41.465}\sqrt{1/10+1/14}} = 3.0006$

p-value $= \mathbb{P}(T_{22} \geq 3.0006)$ is between 0.001 and 0.005,

This is strong evidence against $H_0$, suggesting that athletes have higher maximal oxygen consumption.

(b) Briefly explain the role of the significance level in hypothesis testing. [1 mark]

The significance level controls the rate of type I errors, i.e. if the null hypothesis is true, the significance level is the probability of rejecting the null.

(c) Construct an approximate 95% confidence interval for the true mean maximal oxygen consumption for athletes. [4 marks]

95% CI

$$\bar{x}_A \pm t_{n-1;0.975} \times \text{S.e.}(\bar{x}_A)$$

$$57 \pm 2.262 \times 6.9/\sqrt{10}$$

$$57 \pm 4.936$$

or $(52.064, 61.936)$.

4. [10 marks] Energy drinks have become widely popular among adolescents and are also consumed by athletes, particularly those who have just begun their sporting career. A recent paper presented a study on the consumption of energy drinks by teenagers engaged in sports, including quantity consumed and factors that might be associated with consumption. A total of 707 students, selected randomly from sports classes at various schools, completed a questionnaire on energy drink consumption. The following table shows the crosstabulation of regular energy drink consumption by gender:

| | Energy Drinks | |
|---|---|---|
| Gender | Yes | No |
| Female | 192 | 90 |
| Male | 296 | 129 |

(a) Assuming this sample is representative of all teenagers engaged in sports, give a 95% confidence interval for the true difference in the proportions of females and males who consume energy drinks. What does the interval say about the difference in energy drink consumption between genders? [4 marks]

$\hat{p}_F = \frac{192}{192+90} = 0.6809$ $\hat{p}_M = \frac{296}{296+129} = 0.6965$

95% CI

$$\hat{p}_F - \hat{p}_M \pm z_{0.975} \times \text{s.e.}(\hat{p}_F - \hat{p}_M)$$

$$(0.6809 - 0.6965) \pm 1.96 \times \sqrt{\frac{0.6809 \times (1-0.6809)}{282} + \frac{0.6965 \times (1-0.6965)}{425}}$$

$$-0.0156 \pm 1.96 \times 0.03561$$

$$-0.0156 \pm 0.0698$$

or $(-0.0854, 0.0542)$.

As the interval covers zero, there is no evidence of a difference of proportions at the 5% significance level.

(b) Another factor recorded on the questionnaire was the frequency of practising sports. The following table gives the summary of results for the 681 students who indicated they practised at least once a week:

| | Energy Drinks | | |
|---|---|---|---|
| Practising Sports | Yes | No | |
| Daily | 328 | 146 | 474 |
| 2-3 times per week | 28 | 13 | 41 |
| Once per week | 114 | 52 | 166 |
| | 470 | 211 | |

Based on this table, is there evidence of an association between energy drink consumption and frequency of practising sports? [6 marks]

Test $H_0$: 'Energy drinks' and 'Practising sports' are independent.
against $H_1$: Some association between 'Energy drinks' and 'practising sports'.

expected counts:

$E_1 = \frac{470 \times 474}{681} = 327.14$

$E_2 = 474 - 327.14 = 146.86$

$E_3 = \frac{470 \times 41}{681} = 28.30$

$E_4 = 41 - 28.30 = 12.7$

$E_5 = \frac{470 \times 166}{681} = 114.57$

$E_6 = 166 - 114.57 = 51.43$

$$X^2 = \sum \frac{(e_i - o_i)^2}{e_i}$$

$$= \frac{(327.14 - 328)^2}{327.14} + \frac{(146.86 - 146)^2}{146.86}$$

$$+ \frac{(28.30 - 28)^2}{28.30} + \frac{(12.7 - 13)^2}{12.7}$$

$$+ \frac{(114.57 - 114)^2}{114.57} + \frac{(52 - 51.43)^2}{51.43}$$

$$= 0.0264$$

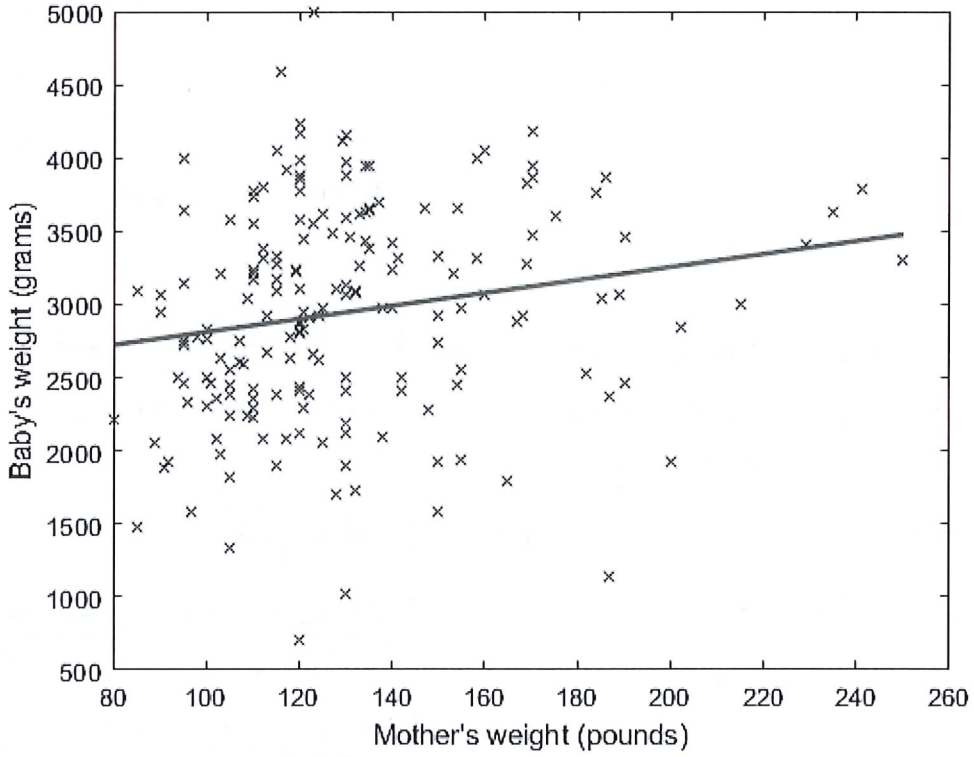
p-value $= \mathbb{P}(\chi^2_2 \geq 0.0264) \geq 0.975$

degrees of freedom $= (3-1) \times (2-1) = 2$

There is no evidence against $H_0$, suggesting 'Energy drinks' and 'Practising sport' are independent.

(An unusually large p-value but did appear in STAT1201, 2017, sem 2)

5. [14 marks] In a study of factors thought to be associated with birth weight, data from 189 births was collected at the Baystate Medical Center, Springfield, Massachusetts during 1986. Two of the variables recorded were the baby's weight (grams) at birth and the mother's weight (pounds) at last menstrual period. The data are plotted below with the least squares regression line




The output on the next page show the results of a linear regression fit in MATLAB for the relationship between the mother's weight (`Mother`) and the baby's weight (`Baby`).

```
Linear regression model:
    Baby ~ 1 + Mother

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    _______    _______    __________

    (Intercept)      2369.6      228.49     10.371    3.3968e-20
    Mother           4.4291      1.7135     2.5848      0.010504


Number of observations: 189, Error degrees of freedom: 187
Root Mean Squared Error: 718
R-squared: 0.0345,  Adjusted R-Squared 0.0293
F-statistic vs. constant model: 6.68, p-value = 0.0105
```

(a) Briefly interpret the value 4.4291 in the regression output. [1 mark]

A one pound increase in mother's weight is associated with an increase in 4.4291 grams of the baby's weight on average.

(b) Does the mother's weight explain much of the variation in the baby's weight at birth? Justify your answer. [1 mark]

Only 3.45% of the variation in baby's weight is explained by the mother's weight. ($R^2$ from MATLAB output).

(c) Give a 90% confidence interval for the intercept term in the linear relationship between the mother's weight and the baby's weight at birth. [2 marks]

90% CI  estimate $\pm$ $t_{187;0.95}$ $\times$ s.e.(estimate)

$2369.6 \pm 1.645 \times 228.49$  using df = $\infty$ in table

$2369.6 \pm 375.8$ (g)

or (1993.7, 2745.5) grams

(d) Does the data provide evidence of an association between the mother's weight and the baby's weight at birth? State the null and alternative hypotheses and report the appropriate test statistic and $P$-value from the output. What do you conclude? [3 marks]

Let $\beta_1$ be the true slope in the linear relation between mother's weight and baby's weight.

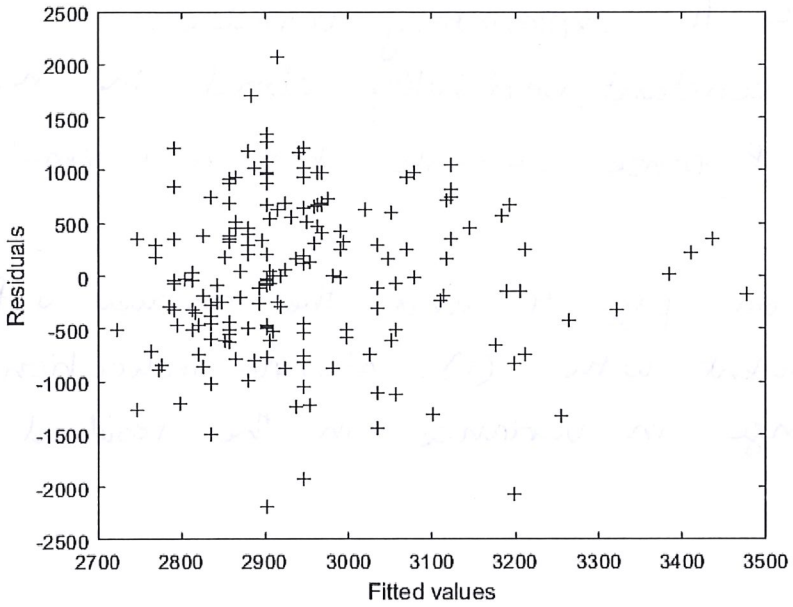
Test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$

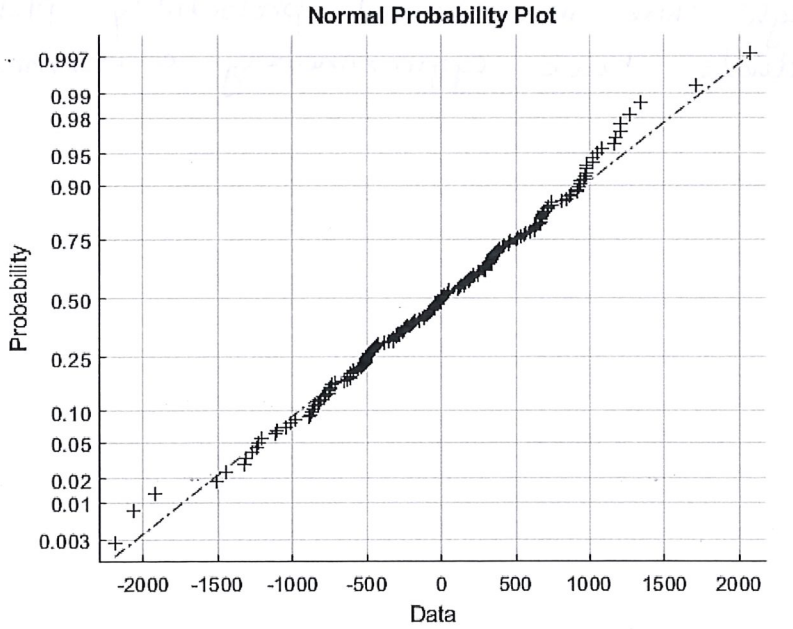
test statistic = 2.5848

p-value = 0.0105

There is strong evidence against $H_0$, suggesting an association between mother's weight and baby's weight.

(e) The following figures were generated in MATLAB to help check the assumptions underlying the linear regression model. State these assumptions and comment on their validity for this data with reference to these figures and the figure on page 9. [3 marks]



(additional space for answer to part (e) — not all this space is needed)

Assumptions

(1) Mean of the response is a linear function of the explanatory variable

(2) constant variability about the mean

(3) Response variable has a normal distribution

- Plot on page 9 and the residual v fitted plot is consistent with (1). Also no indication of a change in variance in the residual v fitted plot.

- Straight-line in normal probability plot indicates residuals have approximately a normal distribution

(f) A second linear model was fitted with the additional explanatory variable `Smoke` which is a dummy variable that takes the value 1 when the mother is a smoker and 0 otherwise. The edited output is given below.

```
>> birthlm2 = fitlm(birth,'Baby~Mother+Smoke')

birthlm2 =


Linear regression model:
    Baby ~ 1 + Mother + Smoke

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    _______    _______    __________

    (Intercept)      2501.1     230.84     10.835    1.6091e-21
    Mother           4.2367     1.6899     2.5071      0.013028
    Smoke           -272.08                           0.010749


Number of observations: 189, Error degrees of freedom:
Root Mean Squared Error: 708
R-squared:       ,  Adjusted R-Squared 0.0578
F-statistic vs. constant model: 6.76, p-value = 0.00146
>> birthlm2.CoefficientCovariance

ans =

   1.0e+04 *

    5.3285   -0.0374   -0.5389
   -0.0374    0.0003    0.0008
   -0.5389    0.0008    1.1150

>> var(birth.Baby)

ans =

   5.3175e+05
```

From this output determine (i) the standard error for the estimate of the coefficient of `Smoke`, (ii) the error degrees of freedom and (ii) the R-squared value. [4 marks]

(space for answer to part (f) — not all this space is needed)

(1) s.e. for smoke

$\sqrt{1.1150 \times 10^4} = 105.5936$

(2) degrees of freedom

$= 189 - 3 = 186$

(3) $S^2 = 708^2$

$R^2 = 1 - \dfrac{186 \times 708^2}{(189-1) \times 5.3175 \times 10^5} = 0.0674$

END OF EXAMINATION