

# UTS 37010

**Assessment Task 1: Assignment Part A**  
**Autumn 2025**

**Coversheet**

**Student Acknowledgement**

I acknowledge that I am aware of the University rules regarding plagiarism and academic misconduct. I confirm that this assignment has not been previously submitted for assessment at UTS or any other institution.

**Name**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   **Student Number**: \_\_\_\_\_\_\_\_\_\_\_\_   
**Signature**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Question 1: Some Maths [12 marks]

## (a) Chi-squared and Confidence Interval [3 marks]

Given for , and is the sample variance, the test statistic

To construct a one-sided confidence interval to test versus , use

So the lower bound for is

## (b) Covariance of Least Squares Estimators [3 marks]

Consider the model

with . Then the least squares estimators are

Using ANOVA theory under the constraint , we have

## (c) Least Squares Estimator in Log-linear Model [3 marks]

Given the model

Take logs

Then the transformed model is

By least squares

Exponentiate both sides

Which simplifies to

## (d) Bias of Estimator [3 marks]

We examine if is unbiased

Recall

So

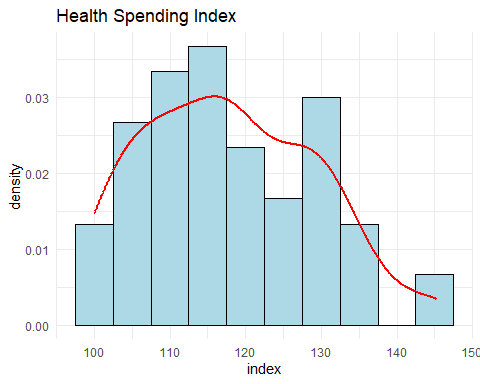
Since , use the MGF

Hence

So is a **biased** estimator of .

# Question 2: One and two-sample analysis [12 marks]

## (a) Histogram and normal curve [3 marks]



The histogram is moderately symmetric but with some slight right skew. The red density curve does not perfectly match a bell shape, suggesting mild deviations from normality, but the sample appears roughly normal for the purposes of t-tests.

## (b) One-sample t-test (Is mean ?) [3 marks]

##   
## One Sample t-test  
##   
## data: health\_data$index  
## t = -4.8314, df = 59, p-value = 1.003e-05  
## alternative hypothesis: true mean is not equal to 125  
## 95 percent confidence interval:  
## 115.0513 120.8787  
## sample estimates:  
## mean of x   
## 117.965

**Hypotheses**

**Decision**

The p-value = 1.003e-05 is less than 0.05, so we reject .

**Conclusion**

There is strong evidence that the mean health spending index is significantly different from 125.

## (c) Two-sample t-test between time periods [3 marks]

##   
## Welch Two Sample t-test  
##   
## data: july2021\_dec2023 and jan2019\_june2021  
## t = 0.94188, df = 57.574, p-value = 0.1751  
## alternative hypothesis: true difference in means is greater than 15  
## 95 percent confidence interval:  
## 13.58709 Inf  
## sample estimates:  
## mean of x mean of y   
## 126.3767 109.5533

**Hypotheses**

**Decision**

The p-value = 0.1751 is greater than 0.05, so we do not reject .

**Conclusion**

There is insufficient evidence that the health spending index increased by more than 15 units between the two periods.

## (d) F-test for variance (Food vs Alcohol & Tobacco) [3 marks]

##   
## F test to compare two variances  
##   
## data: food\_index and alco\_index  
## F = 0.53075, num df = 59, denom df = 59, p-value = 0.008134  
## alternative hypothesis: true ratio of variances is less than 1  
## 95 percent confidence interval:  
## 0.0000000 0.8173311  
## sample estimates:  
## ratio of variances   
## 0.5307495

**Hypotheses**  
 -

**Decision**

The p-value = 0.008134 is less than 0.05, so we reject .

**Conclusion**

There is strong evidence that the variance in food spending is less than the variance in alcohol and tobacco spending.

# Question 3: Four-sample analysis [12 marks]

## (a) Residuals for category 3 [3 marks]

From the ANOVA model

To find residuals without fitting the model: 1. Compute the overall mean 2. Compute the group mean for category 3 3. Estimate treatment effect 4. Residuals:

These are deviations from category 3’s own group mean.

## [1] 277.0151

Sample variance of residuals = **277.02**, reflecting variability within category 3.

## (b) One-way ANOVA test for any mean difference [3 marks]

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(category) 3 9522 3174 15.83 2.06e-09 \*\*\*  
## Residuals 236 47306 200   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**ANOVA Output** F = 15.83, p-value = 2.06e-09

**Hypotheses**

: All group means are equal  
: At least one group mean is different

**Decision**

Reject (p < 0.05)

**Conclusion**

Significant evidence that at least one group’s spending index mean is different.

## (c) Multiple comparisons to identify differing groups [3 marks]

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = index ~ factor(category), data = data)  
##   
## $`factor(category)`  
## diff lwr upr p adj  
## 2-1 -7.233333 -13.921567 -0.5450996 0.0282732  
## 3-1 10.190000 3.501766 16.8782337 0.0006128  
## 4-1 -1.710000 -8.398234 4.9782337 0.9113847  
## 3-2 17.423333 10.735100 24.1115670 0.0000000  
## 4-2 5.523333 -1.164900 12.2115670 0.1445481  
## 4-3 -11.900000 -18.588234 -5.2117663 0.0000400

**Significant Differences**

3 vs 1, 3 vs 2, and 4 vs 3 (p < 0.001) - 2 vs 1 (p = 0.028) is marginally significant

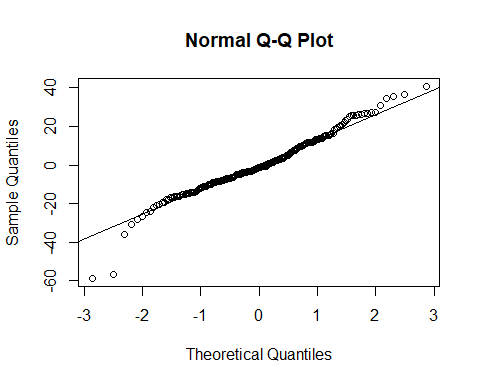
4 vs 1 and 4 vs 2 are **not** significant

**Conclusion**

Category 3 is significantly different from others, particularly categories 1 and 2.

## (d) Normality of residuals [3 marks]

##   
## Shapiro-Wilk normality test  
##   
## data: resid\_vals  
## W = 0.96977, p-value = 5.494e-05



**Shapiro-Wilk Test Result** W = 0.96977, p-value = 5.494e-05

**Hypotheses**

: Residuals are normally distributed  
: Residuals are not normally distributed

**Decision**

Reject (p < 0.05)

**Conclusion**

Residuals are **not normally distributed**, supported by both a low p-value and visible deviations from the normal line in the Q-Q plot.

# Question 4: Linear regression [18 marks]

## (a) Variance inflation factor (VIF) for multicollinearity [3 marks]

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## index2 index3   
## 1.023089 1.023089

**Interpretation**: VIF values were

index2: **1.023**

index3: **1.023**

Both are well below 5, suggesting no multicollinearity concern.

## (b) Fit the model and write regression equation [3 marks]

##   
## Call:  
## lm(formula = index1 ~ index2 + index3, data = wide\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.0317 -4.6274 0.4186 3.0628 18.8968   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.40676 8.83378 1.404 0.166   
## index2 0.48704 0.06582 7.400 6.86e-10 \*\*\*  
## index3 0.40285 0.04795 8.401 1.49e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.061 on 57 degrees of freedom  
## Multiple R-squared: 0.7211, Adjusted R-squared: 0.7113   
## F-statistic: 73.67 on 2 and 57 DF, p-value: < 2.2e-16

**Estimated equation**

Model fit

Adjusted R² = **0.711**, F(2,57) = **73.67**, p < 0.001.

## (c) Hypothesis test

## Does alcohol index decrease health? [3 marks]

## 2.5 % 97.5 %  
## (Intercept) -5.2825810 30.0960935  
## index2 0.3552412 0.6188393  
## index3 0.3068328 0.4988707

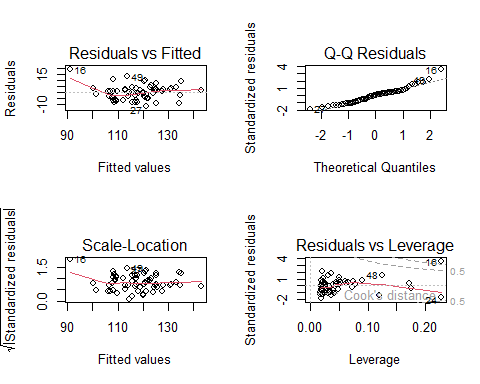
**Hypotheses**

95% CI for : (0.355, 0.619). Since the entire interval is > -0.3, we **reject** .

**Conclusion**

There is strong evidence that an increase in alcohol index is associated with more than a 0.3 unit **increase**, not decrease, in health index.

## (d) Diagnostic plots to assess regression assumptions [3 marks]



**Assessment**

**Linearity**

Some curve in Residuals vs Fitted → mild non-linearity

**Normality**

Q-Q plot shows slight deviation at tails

**Homoscedasticity**

Scale-Location plot is fairly level

**Influence**

Observations 16, 48, 24 show high leverage (see Cook’s D)

Assumptions mostly hold but should monitor influential points.

## (e) Test for independence of residuals [3 marks]

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Durbin-Watson test  
##   
## data: fit  
## DW = 0.96084, p-value = 3.419e-06  
## alternative hypothesis: true autocorrelation is greater than 0

**Durbin-Watson test**: DW = 0.961, p = 3.42e-06 → **Reject**

**Conclusion**: Residuals are **autocorrelated**, violating independence assumption. Inference from regression may be biased.

## (f) Influence analysis and refitting [3 marks]

##   
## Call:  
## lm(formula = index1 ~ index2 + index3, data = wide\_data[-influential\_points,   
## ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2697 -3.9651 0.4319 2.7681 13.9270   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.85307 8.54867 -0.334 0.74   
## index2 0.53646 0.06498 8.256 3.35e-11 \*\*\*  
## index3 0.47743 0.04799 9.949 6.71e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.316 on 55 degrees of freedom  
## Multiple R-squared: 0.7897, Adjusted R-squared: 0.782   
## F-statistic: 103.3 on 2 and 55 DF, p-value: < 2.2e-16

**Model improvement**

Adjusted R² increased from **0.711** → **0.782** - Standard error reduced from 6.06 → **5.32**

**Conclusion**

Excluding influential points improved model fit and increased the estimated impact of both predictors.

## References

* Lecture Notes: *37010 Statistics Chapter 3* (Scott Alexander, UTS)
* R Documentation: aov(), TukeyHSD(), shapiro.test()