**Question 1. [15 marks] Basics**

Consider a dataset containing the weights (in kg) of adults in a certain region. The mean weight is 70 kg, and the standard deviation is 8 kg.

(a) [8 marks] (i) Explain the concept of standardization and its purpose. [4 marks] (ii) Using the given mean and standard deviation, standardize the weight of an individual who weighs 75 kg. Show your calculations. [4 marks]

(b) [7 marks] (i) Interpret the 95% confidence interval for the mean weight if it is given as (67.24, 72.76). [3 marks] (ii) Explain the effect on the confidence interval when the confidence level is increased from 95% to 99%. [4 marks]

**Question 2. [30 marks] SLR, MLR**

Given the following dataset of car prices and their corresponding ages:

| **Car Age (years)** | **Price (in $1000s)** |
| --- | --- |
| 2 | 22 |
| 3 | 20 |
| 4 | 18 |
| 5 | 16 |
| 6 | 14 |
| 7 | 12 |

The summary output of the linear model is:

Call:

lm(formula = car\_prices ~ car\_age)

Residuals:

Min 1Q Median 3Q Max

-1.080 -0.540 0.000 0.540 1.080

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.0000 0.9574 25.078 0.00001 \*\*\*

car\_age -2.0000 0.2000 -10.000 0.0002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8729 on 4 degrees of freedom

Multiple R-squared: 0.9615, Adjusted R-squared: 0.9519

F-statistic: 100 on 1 and 4 DF, p-value: 0.0002

(a) [5 marks] Interpret the coefficient for car age. What does it imply about the impact of the car's age on its price?

(b) [5 marks] Discuss the significance of both the intercept and the coefficient.

(c) [5 marks] What is the R-squared value, and how does it assist in evaluating the model fit?

(d) [5 marks] Calculate the 99% confidence interval for the coefficient of car age.

(e) [5 marks] Explain the difference between the prediction interval and the confidence interval of a mean response.

(f) [5 marks] Compare and contrast the F-test and t-test in the context of this model.

**Question 3. [20 marks] GLM**

A hospital wants to predict the likelihood of readmission of patients based on various factors such as age, severity of illness, and length of stay. The dataset includes:

* Age: The age of the patient.
* Severity: A score representing the severity of the illness (1 to 10).
* Length of Stay: The number of days the patient stayed in the hospital.
* Readmission: A binary variable indicating whether the patient was readmitted (Yes=1, No=0).

(a) [8 marks] Use a Generalized Linear Model (GLM) framework to formulate this problem.

(b) [6 marks] Describe the components of the GLM you would use.

(c) [6 marks] Explain how you would make inferences using this GLM.

**Question 4. [15 marks] Bayesian Linear Regression**

Consider a dataset describing house prices (in $1000s) and their corresponding square footage. Assume a Gaussian prior distribution for the regression coefficients.

(a) [9 marks] Given the likelihood function for house prices, derive the posterior distribution for the regression coefficients.

(b) [6 marks] Explain the role of the prior distribution in Bayesian Linear Regression and how it affects the posterior distribution.

**Question 5. [20 marks] Model Selection and Evaluation**

A marketing company wants to predict customer churn based on several features, including customer age, monthly spending, and contract length. The goal is to evaluate different modeling approaches: linear regression, generalized linear models (GLM), and Bayesian linear models to determine the best method for predicting customer churn.

(a) [7 marks] Model Selection: Explain how you would choose among the three modeling approaches (linear regression, GLM, or Bayesian linear models).

(b) [7 marks] Feature Engineering and Selection: Describe how you would prepare the variables for modeling.

(c) [6 marks] Model Evaluation: Discuss the methods you would use to assess the performance of your chosen model.

下面是模拟测试卷的答案，包含中英文对照：

---

### \*\*Question 1. [15 marks] Basics\*\*

Consider a dataset containing the weights (in kg) of adults in a certain region. The mean weight is 70 kg, and the standard deviation is 8 kg.

\*\*Question 1. [15 分] 基础知识\*\*

考虑一个包含某地区成年人体重（单位：千克）的数据集。平均体重为70千克，标准差为8千克。

---

\*\*(a) [8 marks]\*\*

(i) Explain the concept of standardization and its purpose. [4 marks]

解释标准化的概念及其目的。 [4 分]

Standardization transforms data to have a mean of zero and a standard deviation of one, making different datasets comparable.

标准化将数据转换为均值为零，标准差为一，使得不同的数据集具有可比性。

(ii) Using the given mean and standard deviation, standardize the weight of an individual who weighs 75 kg. Show your calculations. [4 marks]

使用给定的均值和标准差，将体重为75千克的个体标准化。展示你的计算过程。 [4 分]

\[

z = \frac{x - \mu}{\sigma} = \frac{75 - 70}{8} = \frac{5}{8} = 0.625

\]

---

\*\*(b) [7 marks]\*\*

(i) Interpret the 95% confidence interval for the mean weight if it is given as (67.24, 72.76). [3 marks]

如果95%的置信区间为(67.24, 72.76)，解释其意义。 [3 分]

This interval means we are 95% confident that the true mean weight of the population lies between 67.24 kg and 72.76 kg.

这个区间意味着我们有95%的信心认为总体的真实平均体重大约在67.24千克和72.76千克之间。

(ii) Explain the effect on the confidence interval when the confidence level is increased from 95% to 99%. [4 marks]

解释当置信水平从95%提高到99%时，对置信区间的影响。 [4 分]

Increasing the confidence level to 99% will make the confidence interval wider. This means the interval will cover a broader range to ensure a higher confidence level.

将置信水平提高到99%会使置信区间变宽。这意味着为了确保更高的置信水平，区间将覆盖更广的范围。

---

### \*\*Question 2. [30 marks] SLR, MLR\*\*

Given the following dataset of car prices and their corresponding ages:

| Car Age (years) | Price (in $1000s) |

|-----------------|-------------------|

| 2 | 22 |

| 3 | 20 |

| 4 | 18 |

| 5 | 16 |

| 6 | 14 |

| 7 | 12 |

\*\*Question 2. [30 分] 简单线性回归，多重线性回归\*\*

给定以下车龄和相应价格的数据集：

The summary output of the linear model is:

```

Call:

lm(formula = car\_prices ~ car\_age)

Residuals:

Min 1Q Median 3Q Max

-1.080 -0.540 0.000 0.540 1.080

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.0000 0.9574 25.078 0.00001 \*\*\*

car\_age -2.0000 0.2000 -10.000 0.0002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8729 on 4 degrees of freedom

Multiple R-squared: 0.9615, Adjusted R-squared: 0.9519

F-statistic: 100 on 1 and 4 DF, p-value: 0.0002

```

---

\*\*(a) [5 marks] Interpret the coefficient for car age. What does it imply about the impact of the car's age on its price?\*\*

解释车龄的系数。它对车价的影响是什么？ [5 分]

The coefficient for car age is -2.0000, implying that for each additional year of age, the car's price decreases by $2000.

车龄的系数是-2.0000，意味着车龄每增加一年，车价减少2000美元。

---

\*\*(b) [5 marks] Discuss the significance of both the intercept and the coefficient.\*\*

讨论截距和系数的显著性。 [5 分]

The intercept is 24.0000, indicating the expected price of a new car (age 0) is $24,000. Both the intercept and the coefficient are highly significant (p-values < 0.01), suggesting a strong relationship between car age and price.

截距是24.0000，表示新车（车龄为0）的预期价格为24,000美元。截距和系数都非常显著（p值 < 0.01），表明车龄和价格之间有很强的关系。

---

\*\*(c) [5 marks] What is the R-squared value, and how does it assist in evaluating the model fit?\*\*

什么是R平方值，它如何帮助评估模型拟合？ [5 分]

The R-squared value is 0.9615, indicating that approximately 96.15% of the variability in car prices can be explained by car age. A higher R-squared value suggests a better fit of the model.

R平方值是0.9615，表明车价的约96.15%的变异性可以通过车龄来解释。较高的R平方值表明模型拟合较好。

---

\*\*(d) [5 marks] Calculate the 99% confidence interval for the coefficient of car age.\*\*

计算车龄系数的99%置信区间。 [5 分]

---

\*\*(e) [5 marks] Explain the difference between the prediction interval and the confidence interval of a mean response.\*\*

解释预测区间和均值响应置信区间的区别。 [5 分]

A prediction interval estimates the range for a single new observation, while a confidence interval estimates the range for the mean response of all observations. Prediction intervals are typically wider than confidence intervals due to the additional uncertainty.

预测区间估计单个新观测值的范围，而置信区间估计所有观测值的均值响应的范围。由于额外的不确定性，预测区间通常比置信区间宽。

---

\*\*(f) [5 marks] Compare and contrast the F-test and t-test in the context of this model.\*\*

在该模型的背景下比较和对比F检验和t检验。 [5 分]

The t-test assesses the significance of individual regression coefficients, while the F-test assesses the overall significance of the model. In this context, the t-test evaluates the significance of the car age coefficient, and the F-test evaluates if the model as a whole significantly explains the variability in car prices.

t检验评估单个回归系数的显著性，而F检验评估模型的整体显著性。在此背景下，t检验评估车龄系数的显著性，F检验评估整个模型是否显著解释了车价的变异性。

---

### \*\*Question 3. [20 marks] GLM\*\*

A hospital wants to predict the likelihood of readmission of patients based on various factors such as age, severity of illness, and length of stay.

\*\*Question 3. [20 分] 广义线性模型\*\*

一家医院希望根据各种因素（如年龄、疾病严重程度和住院时间）预测患者再入院的可能性。

---

\*\*(a) [8 marks] Use a Generalized Linear Model (GLM) framework to formulate this problem.\*\*

使用广义线性模型（GLM）框架来表述这个问题。 [8 分]

Let \( Y \) be the binary response variable indicating readmission (1 for Yes, 0 for No). The predictors are age (\( X\_1 \)), severity of illness (\( X\_2 \)), and length of stay (\( X\_3 \)). The model can be written as:

---

\*\*(b) [6 marks] Describe the components of the

GLM you would use.\*\*

描述你将使用的GLM组件。 [6 分]

The components of the GLM are:

- \*\*Link function:\*\* Logit function,

- \*\*Random component:\*\* Binomial distribution for the response variable

- \*\*Systematic component:\*\* Linear predictor

GLM的组成部分是：

- \*\*链接函数:\*\* 对数几率函数，

- \*\*随机部分:\*\* 响应变量 的二项分布

- \*\*系统部分:\*\* 线性预测变量

---

\*\*(c) [6 marks] Explain how you would make inferences using this GLM.\*\*

解释如何使用这个GLM进行推断。 [6 分]

To make inferences using this GLM, we:

- Estimate parameters \( \beta \) using maximum likelihood estimation (MLE)

- Test hypotheses about the parameters using Wald tests or likelihood ratio tests

- Calculate confidence intervals for the parameters

- Assess the goodness-of-fit using deviance or Pearson residuals

使用这个GLM进行推断时，我们：

- 使用最大似然估计（MLE）估计参数 \( \beta \)

- 使用Wald检验或似然比检验进行参数假设检验

- 计算参数的置信区间

- 使用偏差或Pearson残差评估拟合优度

---

### \*\*Question 4. [15 marks] Bayesian Linear Regression\*\*

Consider a dataset describing house prices (in $1000s) and their corresponding square footage.

\*\*Question 4. [15 分] 贝叶斯线性回归\*\*

考虑一个描述房价（单位：千美元）及其对应平方英尺的数据集。

---

\*\*(a) [9 marks] Given the likelihood function for house prices, derive the posterior distribution for the regression coefficients.\*\*

给定房价的似然函数，推导回归系数的后验分布。 [9 分]

Let the likelihood be).

假设似然函数为 。

Assume a Gaussian prior

假设高斯先验。

The posterior distribution \( p(\beta | y) \) is also Gaussian, with:

后验分布 \( p(\beta | y) \) 也是高斯分布，且：

---

\*\*(b) [6 marks] Explain the role of the prior distribution in Bayesian Linear Regression and how it affects the posterior distribution.\*\*

解释先验分布在贝叶斯线性回归中的作用及其对后验分布的影响。 [6 分]

The prior distribution represents our beliefs about the parameters before observing the data. It combines with the likelihood to form the posterior distribution, which updates our beliefs after observing the data. A strong prior can significantly influence the posterior, while a weak (non-informative) prior allows the data to dominate.

先验分布表示我们在观察数据之前对参数的信念。它与似然结合形成后验分布，更新我们在观察数据后的信念。强先验可以显著影响后验，而弱（非信息）先验允许数据占主导地位。

---

### \*\*Question 5. [20 marks] Model Selection and Evaluation\*\*

A marketing company wants to predict customer churn based on several features, including customer age, monthly spending, and contract length.

\*\*Question 5. [20 分] 模型选择和评估\*\*

一家营销公司希望根据多个特征（包括客户年龄、月支出和合同长度）预测客户流失。

---

\*\*(a) [7 marks] Model Selection: Explain how you would choose among the three modeling approaches (linear regression, GLM, or Bayesian linear models).\*\*

模型选择：解释如何在三种建模方法（线性回归、GLM或贝叶斯线性模型）中进行选择。 [7 分]

Model selection involves considering the nature of the dependent variable (binary or continuous), assumptions of the model, uncertainty estimation, and model complexity. For binary outcomes like churn, GLMs (logistic regression) are appropriate. Bayesian models can be used if prior information is available and uncertainty needs to be quantified.

模型选择涉及考虑因变量的性质（二元或连续）、模型的假设、不确定性估计和模型复杂性。对于流失这样的二元结果，GLM（逻辑回归）是合适的。如果有先验信息并且需要量化不确定性，可以使用贝叶斯模型。

---

\*\*(b) [7 marks] Feature Engineering and Selection: Describe how you would prepare the variables for modeling.\*\*

特征工程和选择：描述如何为建模准备变量。 [7 分]

Feature engineering may include scaling numeric features, encoding categorical variables, and creating interaction terms. Feature selection techniques like forward selection, backward elimination, or regularization methods (LASSO) can be used to choose the most relevant variables.

特征工程可能包括缩放数值特征、编码分类变量和创建交互项。特征选择技术如前向选择、后向消除或正则化方法（LASSO）可以用来选择最相关的变量。

---

\*\*(c) [6 marks] Model Evaluation: Discuss the methods you would use to assess the performance of your chosen model.\*\*

模型评估：讨论用于评估所选模型性能的方法。 [6 分]

Model evaluation methods include cross-validation, ROC curves, precision-recall curves, and metrics like accuracy, precision, recall, F1-score, and AUC for classification models. For regression models, metrics like RMSE, MAE, and R-squared are used.

模型评估方法包括交叉验证、ROC曲线、精确召回曲线以及分类模型的准确性、精度、召回率、F1分数和AUC等指标。对于回归模型，使用RMSE、MAE和R平方等指标。

---