1. The standard simple linear regression model assumes an approximate linear relationship between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where the ϵ_i are normal random variables with mean 0 and variance σ^2

(a) (2 points) What else, if anything, do the standard assumptions of linear regression say about the ϵ_i?

> **Solution:** The standard assumptions additionally require that the ϵ_i are *independent*. The variance σ^2 is the same for all i (homoscedasticity).

(b) (6 points) The sum $\epsilon_1 + \cdots + \epsilon_n$ is a random variable. What can be said about the distribution of this random variable? (Give the name of the distribution and the values of any relevant parameters that define it.)

> **Solution:** The sum is normal with mean 0 and variance $n\sigma^2$. The sum of independent normal random variables is normal (Wackerly Theorem 6.3). Using linearity of expectation and independence:
>
> $$E[\epsilon_1 + \cdots + \epsilon_n] = 0, \qquad \mathrm{Var}(\epsilon_1 + \cdots + \epsilon_n) = n\sigma^2.$$
>
> Therefore $\epsilon_1 + \cdots + \epsilon_n \sim N(0,\, n\sigma^2)$.

2. Consider the following regression summary output from a study examining factors affecting the sale price of houses in a specific neighborhood. The variables considered are:

- price, the sale price of the house in *thousands* of dollars,
- size, the square footage of the house in *hundreds* of square feet, and
- age, the age of the house in years.

```
Call:
lm(formula = price ~ size + age, data = housing)

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     120.000      12.000   10.000   2.1e-10 ***
size             15.000       3.000    5.000   1.5e-05 ***
age              -2.000       0.500   -4.000   0.00032 ***
```

(a) (5 points) What is the predicted sale price (in *dollars*) for a house that is 2,000 square feet and is 10 years old? (Hint: note the units of measurement in the regression description above.)

Solution: The model is

$$\widehat{\text{price}} = 120 + 15 \cdot \text{size} - 2 \cdot \text{age},$$

where price is in thousands of dollars, size is in hundreds of square feet, and age is in years.
A house of 2,000 sq ft has size = 20 (hundreds of sq ft), and age = 10.

$$\widehat{\text{price}} = 120 + 15(20) - 2(10) = 120 + 300 - 20 = 400 \text{ (thousands of dollars)}.$$

Converting to dollars: $400{,}000 \times 1000/1000 = \$400{,}000$.
The predicted sale price is \$400,000.

(b) (8 points) Write a sentence interpreting the coefficient of `age` in this regression, comprehensible to a non-statistician.

Solution: Holding house size fixed, each additional year of age is associated with a predicted decrease in sale price of \$2,000 (since the coefficient -2 is in units of thousands of dollars per year).

3. We fit a regression model with two independent variables `X1` and `X2` in `R`, and we have only $n = 6$ rows of data.

```
> lmod1 <- lm(Y ~ X1 + X2, data = df)
> summary(lmod1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0357     3.6093   1.949    0.146
X1            0.3420     0.4562   0.750    0.508
X2            1.6580     1.0599   1.564    0.216
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2557 on 3 degrees of freedom
Multiple R-squared:  0.9965,Adjusted R-squared:  0.9942
F-statistic: 432.3 on 2 and 3 DF,  p-value: 0.0002033
```

(a) (4 points) In terms of n, the number of rows of data, and p, the number of parameters in the regression model, how many numerator and denominator degrees of freedom does the F-statistic have (in general; your answer should be in terms of n and p). Hint: be sure to check that your answer is consistent with the 2 and 3 in the output above.

Solution: The model has p parameters (here $p = 3$: intercept, β_1, β_2). The F-statistic for testing whether all slope coefficients are zero has

$$\text{numerator df} = p - 1, \qquad \text{denominator df} = n - p.$$

Check: $p - 1 = 3 - 1 = 2$ and $n - p = 6 - 3 = 3$, consistent with "2 and 3" in the output.

(b) (4 points) The p-value `0.216` associated with the coefficient of `X1` is the result of a hypothesis test. What are the null and alternative hypotheses of this test?

Solution: Note: the p-value `0.216` actually appears in the row for `X2` in the output above. The question should read "coefficient of `X2`".

The corresponding test for `X2` is:

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_a : \beta_2 \neq 0,$$

where β_2 is the population coefficient of `X2`, with `X1` already in the model.

(c) (4 points) The coefficient of `X2` is not statistically significant (its p-value is 0.216), yet the overall F-test is highly significant (p-value = 0.0002). A student says: "This is a contradiction — if the model is significant overall, all of its predictors should be significant individually." Is the student correct? Explain why or why not.

Solution: The student is **not** correct. The overall F-test asks whether *at least one* predictor has a nonzero coefficient, while the individual t-tests ask whether each predictor contributes *beyond the others already in the model.*

It is entirely possible for the predictors together to explain a large fraction of the variance (making the F-test significant) while one predictor's marginal contribution—after accounting for the other—is small and not individually significant. This situation is common with multicollinearity: if `X1` and `X2` are correlated, neither may be significant on its own even though together they explain Y well.

(d) (6 points) Consider the following sequence of steps that `R` performs to produce the entry `0.216` in the row for `X2`:

1. Compute the estimate `1.6580`.
2. Compute the standard error `1.0599`.
3. Compute the t-value `1.564`.
4. Compute the p-value `0.216`.
5. Declare the result not statistically significant at level $\alpha = 0.10$.

Which, if any, is the first of these steps to require the assumption that the errors ϵ_i are normally distributed, and how is that assumption used?

Solution: Step 4 is the first step that requires normality of the errors.

Steps 1–3 can be computed directly from the data and the formulas (see Wackerly Chapter 11). Even to establish that these formulas give unbiased estimates does not require normality. The OLS estimate $\hat{\beta}_2$ and its standard error can be computed without normality, and the t-statistic is simply their ratio.

However, in order to assign a *p-value* to the t-statistic, we need to know its sampling distribution under H_0. Under the assumption that the errors are normally distributed, $\hat{\beta}_2/\widehat{\mathrm{SE}}(\hat{\beta}_2)$ follows a t-distribution with $n-p$ degrees of freedom. Without normality, we would only know the distribution of the t-statistic approximately (by the CLT for large n), and the exact p-value from the t_{n-p} distribution would not be justified. Step 5 (declaring significance) follows from the p-value and requires no additional assumptions.

4. Suppose we have a linear regression model with 2 independent variables U, V and n data points, so the defining equation is

$$Y_i = \beta_0 + \beta_1 U_i + \beta_2 V_i + \epsilon_i$$

with the usual assumptions on the errors ϵ_i. The textbook tells us that linear regression is an orthogonal projection, which is a linear map.

(a) (2 points) What is the domain of this linear map?

Solution: The domain is $\mathbb{R}^n$, the space of all possible response vectors $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$.

(b) (4 points) What is the codomain (also sometimes called the target or range) of this linear map?

Solution: This question has an ambiguity: a map need not be onto, so the target of the map need not be the same as the image of the map.

In this case, the codomain, or target is $\mathbb{R}^n$. The map sends $\mathbf{Y} \in \mathbb{R}^n$ to the fitted values $\hat{\mathbf{Y}} \in \mathbb{R}^n$ (specifically, $\hat{\mathbf{Y}}$ lies in the column space of the design matrix, which is a 3-dimensional subspace of $\mathbb{R}^n$). So the range is the column space of of the design matrix, which is 3-dimensional, not n-dimensional.

(c) (6 points) How can we describe the matrix of this linear map? (Hint: note that it is not sufficient to write a formula with X in it, because no X is defined in this problem!)

Solution: The matrix of this linear map is the hat matrix H. Since the design matrix in this problem has columns for the intercept, U, and V, let us call it $\mathbf{D}$:

$$\mathbf{D} = \begin{pmatrix} 1 & U_1 & V_1 \\ 1 & U_2 & V_2 \\ \vdots & \vdots & \vdots \\ 1 & U_n & V_n \end{pmatrix}.$$

Then the hat matrix is

$$H = \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T,$$

which is an $n \times n$ matrix. It is symmetric ($H^T = H$) and idempotent ($H^2 = H$), and it projects any vector in $\mathbb{R}^n$ orthogonally onto the column space of $\mathbf{D}$.

5. (4 points) You obtain a list of 4 samples y_1, y_2, y_3, y_4 from a random variable Y which has either the t-distribution or the F-distribution (degrees of freedom unknown).

   This list turns out to be $0.34, 4.36, -0.93, 1.37$. Is the random variable Y more likely to have the t- or F-distribution? Why?

Solution: The random variable Y is more likely to have the **t-distribution**.

The key observation is that the sample contains a negative value (-0.93). The F-distribution is always nonnegative (it is the ratio of two nonnegative chi-squared random variables), so values below zero are impossible under the F-distribution. The t-distribution, on the other hand, is symmetric about zero and takes both positive and negative values. Since we observed a negative value, Y cannot have an F-distribution; it must have the t-distribution.

6. Suppose we define two matrices in `R` by

```
A <- matrix(-1, nrow = 2, ncol = 2)
B <- matrix(1, nrow = 2, ncol = 2)
```

   (a) (2 points) What is the matrix `A` (that is, the output of `print(A)`)?

Solution:

```
     [,1] [,2]
[1,]   -1   -1
[2,]   -1   -1
```

All entries are -1.

   (b) (2 points) What is the matrix `B` (that is, the output of `print(B)`)?

Solution:

```
     [,1] [,2]
[1,]    1    1
[2,]    1    1
```

All entries are 1.

   (c) (4 points) What will the output of `print( A * B )` be?

Solution: In R, `*` denotes element-wise multiplication. Each entry of `A` is -1 and each entry of `B` is 1, so every entry of `A * B` is $(-1)(1) = -1$:

```
     [,1] [,2]
[1,]   -1   -1
[2,]   -1   -1
```

(d) (4 points) What will the output of `print( A %*% B)` be?

Solution: In R, `%*%` denotes matrix multiplication.

$$AB = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} (-1)(1)+(-1)(1) & (-1)(1)+(-1)(1) \\ (-1)(1)+(-1)(1) & (-1)(1)+(-1)(1) \end{pmatrix} = \begin{pmatrix} -2 & -2 \\ -2 & -2 \end{pmatrix}.$$

```
     [,1] [,2]
[1,]   -2   -2
[2,]   -2   -2
```

7. (4 points) Fill in the blanks in the following definition of level (of significance) and type I error:

Solution: A type I error is made if H_0 is rejected when H_0 is true.

The probability of making a type I error (assuming that H_0 is true) is denoted by α and is called the level of the test.

8. Let X_1, X_2, X_3, X_4 be independent and identically distributed random variables which are all normal with mean 0 and variance 1. Let $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$. Let $W = \sum_{i=1}^{4}(X_i - \bar{X})^2$.

(a) (4 points) What is the distribution of W? (Give the values of any parameters that are relevant.)

Solution: $W = \sum_{i=1}^{4}(X_i - \bar{X})^2$ is the sum of squared deviations from the mean of $n = 4$ i.i.d. $N(0, 1)$ random variables. By Wackerly Theorem 7.3, this follows a chi-squared distribution with $n - 1 = 3$ degrees of freedom:

$$W \sim \chi^2(3).$$

(b) (4 points) Notice that $\bar{X}$ appears in the formula defining W. Are $\bar{X}$ and W independent random variables? How do you know?

Solution: Yes, $\bar{X}$ and W are independent. This is part of theorem 7.3 in Wackerly.

9. A study was conducted to investigate factors affecting the daily number of steps walked by adults. Thirty individuals were tracked over a 2-week period using a fitness monitor, and the average number of steps walked per day, Y, was recorded for each. Four additional variables were recorded for each individual:

x_1 = age, x_2 = body mass index (BMI), x_3 = hours of sedentary work per day, and x_4 = self-reported motivation score (on a scale of 1–10). Consider the three models given below:

$$\begin{aligned} \text{Model I:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \\ \text{Model II:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \\ \text{Model III:} \quad & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3 x_4 + \epsilon \end{aligned}$$

Indicate whether each of the following statements is true or false, and give a short reason for each answer.

(a) (4 points) Models I and II can be compared using an F-test.

Solution: True. Model II is a special case of Model I obtained by setting $\beta_4 = 0$, so Model II is nested within Model I. Nested models can always be compared using a partial F-test.

(b) (4 points) After fitting Models I and II and computing their sums of squared errors, $\text{SSE}_\text{I} \geq \text{SSE}_\text{II}$.

Solution: False. Adding a predictor (going from Model II to Model I) can only decrease or maintain the SSE, never increase it, because the larger model includes all fits of the smaller model as special cases. Therefore $\text{SSE}_\text{I} \leq \text{SSE}_\text{II}$, which is the opposite of the stated inequality.

(c) (4 points) Models II and III can be compared using an F-test.

Solution: True. Model II is a special case of Model III, obtained by setting $\beta_4 = 0$ in Model III.

(d) (4 points) After fitting Models II and III and computing their R^2 values, $R^2_\text{III} \geq R^2_\text{II}$.

Solution: True. The key point is that R^2 increases (or stays the same) whenever a model includes more terms that allow a better fit.

10. Suppose X has the uniform distribution on $[-1, 1]$ and $Y = X^2$.

(a) (1 point) Find $E[X]$

Solution: By symmetry of the uniform distribution on $[-1, 1]$, $E[X] = 0$.

(b) (2 points) Find $E[Y]$

Solution:

$$E[Y] = E[X^2] = \int_{-1}^{1} x^2 \cdot \frac{1}{2}\, dx = \frac{1}{2} \cdot \frac{x^3}{3}\Big|_{-1}^{1} = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

(c) (2 points) Are X and Y independent?

Solution: **No.** $Y = X^2$ is a deterministic function of X, so knowing Y gives information about X. Specifically, knowing Y tells us $|X|$, which reduces our uncertainty about X.

(d) (4 points) Find $\text{Cov}(X, Y)$.

Solution:

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

We have $E[X] = 0$ and $E[Y] = 1/3$. Also,

$$E[XY] = E[X \cdot X^2] = E[X^3] = \int_{-1}^{1} x^3 \cdot \frac{1}{2}\, dx = 0,$$

since x^3 is an odd function integrated over a symmetric interval.
Therefore $\text{Cov}(X, Y) = 0 - (0)(1/3) = 0$.

(e) (6 points) Suppose we fit a simple regression $Y = \beta_0 + \beta_1 X + \epsilon$ using $n = 9999$ i.i.d. draws from this joint distribution. What approximate values do you expect for $\hat{\beta}_0$ and $\hat{\beta}_1$?

Solution: The OLS estimators converge (by the law of large numbers) to the population regression coefficients. The population slope is

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{0}{\text{Var}(X)} = 0,$$

and the intercept is

$$\beta_0 = E[Y] - \beta_1 E[X] = \frac{1}{3} - 0 \cdot 0 = \frac{1}{3}.$$

So we expect $\hat{\beta}_1 \approx 0$ and $\hat{\beta}_0 \approx 1/3$.

(f) (4 points) What approximate value would you expect for R^2 in this regression?

Solution: Since $\hat{\beta}_1 \approx 0$, the regression line is essentially the horizontal line $\hat{Y} = 1/3$ (the sample mean of Y). The fitted values $\hat{Y}_i \approx \bar{Y}$, so the regression explains essentially none of the variation in Y. Therefore $R^2 \approx 0$.
This makes sense: although X and $Y = X^2$ are not independent, they are *uncorrelated*, and the linear regression can only capture linear relationships. The (linear) regression of Y on X has no predictive power here.

11. Suppose we have 3 variables `Y`, `X`, and `Z` defined in `R`. We consider 3 linear models defined by

```
lmod1 <- lm(Y ~ X + Z)
lmod2 <- lm(Y ~ -1 + I(X+Z))
lmod3 <- lm(Y ~ X + offset(2*Z))
```

The equation that defines the model `lmod1` is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

(a) (4 points) What equation defines `lmod2`?

Solution: The `-1` removes the intercept, and `I(X+Z)` creates a single predictor equal to $X + Z$. So the model has one parameter β_1 and no intercept:

$$Y = \beta_1(X + Z) + \epsilon.$$

(b) (4 points) What equation defines `lmod3`?

Solution: The `offset(2*Z)` term fixes the coefficient of $2Z$ at 1 (i.e., it is not estimated but included as a known offset). So the model is:

$$Y = \beta_0 + \beta_1 X + 2Z + \epsilon.$$

Here only β_0 and β_1 are estimated; the coefficient of Z is fixed at 2.

(c) (6 points) If we execute the command `anova(lmod1, lmod2)` we see the output:

```
> anova(lmod1,lmod2)
Analysis of Variance Table

Model 1: Y ~ X + Z
Model 2: Y ~ -1 + I(X + Z)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1      7 13.719
2      9 17.967 -2   -4.2486 1.084  0.389
```

We see a p-value (`0.389`) in this output, which comes from a statistical test. What are the null hypothesis and the alternative hypothesis of this statistical test?

Solution: The F-test in `anova(lmod1, lmod2)` compares Model 1 (larger) to Model 2 (smaller/restricted). Model 2 is obtained from Model 1 by imposing two constraints: (1) $\beta_0 = 0$ (no intercept), and (2) $\beta_1 = \beta_2$ (the coefficients of X and Z are equal, since both are absorbed into the single parameter multiplying $X + Z$).

$$H_0 : \beta_0 = 0 \text{ and } \beta_1 = \beta_2 \quad \text{(Model 2 is adequate)}$$

$$H_a : \text{at least one of these constraints fails (Model 1 is needed)}$$

The large p-value (0.389) means we fail to reject H_0; the data do not provide evidence against the restricted model.

Multiple Choice: Circle the letter corresponding to the best response.

12. (4 points) Which of the following best describes the Hill criterion of *strength of association*?

(A) The association between the exposure and outcome must be statistically significant at the $\alpha = .05$ level.

(B) A large effect size makes it less likely that the association is due to confounding or bias alone.

(C) The association must be replicated in at least two independent studies before it can be considered causal.

(D) The exposure must precede the outcome in time.

Solution: (B). The Hill criterion of strength of association refers to the magnitude of the effect: a large effect size is harder to explain away by confounding or bias alone.

13. (4 points) A study finds that individuals who smoke more cigarettes per day have a higher risk of lung cancer, and that those who quit smoking see their risk decrease over time. Which Hill criterion does this most directly address?

(A) Consistency

(B) Gradient

(C) Specificity

(D) Coherence

Solution: (B) Gradient (also called biological gradient or dose-response relationship). The observation that more cigarettes $\rightarrow$ higher risk, and quitting $\rightarrow$ decreasing risk, directly reflects a dose-response gradient.

14. (4 points) A student argues: "Our regression model shows a highly significant association between X and Y with a very small p-value, so by the Hill criterion of strength of association, we have strong evidence of a causal relationship." What is the most important flaw in this argument?

(A) A small p-value indicates a large effect size, which is not the same as a causal relationship.

(B) Statistical significance reflects sample size as much as effect size, and in any case no single Hill criterion is sufficient to establish causation on its own.

(C) The student should have used a confidence interval rather than a p-value to assess strength of association.

(D) Regression models cannot be used to assess the Hill criteria because they are designed for prediction, not causal inference.

Solution: (B). Statistical significance (small p-value) is driven by both effect size *and* sample size. A tiny effect can be highly significant with a large sample. Moreover, even if there were a large effect, no single Hill criterion is sufficient to establish causality; all or most criteria must be considered together.

15. (4 points) In a least-squares regression model with an intercept, expressed in matrix notation as $Y = X\beta + \epsilon$, the residual vector e is always orthogonal to which of the following?

    (A) The response vector Y

    (B) The fitted values vector $\hat{Y}$

    (C) The column space of the design matrix X

    (D) The null space of $X^T X$

Solution: (C) The column space of the design matrix X. By the normal equations $X^T e = 0$, the residuals are orthogonal to every column of X, i.e., orthogonal to the entire column space. (Since the model includes an intercept, the residuals also sum to zero, which is one specific consequence of this orthogonality. Note that (B) $\hat{Y}$ lies in the column space of X, so $e \perp \hat{Y}$ as well—but (C) is the more general and fundamental statement.)

16. (4 points) The hat matrix H in a regression is defined as

$$H = X(X^T X)^{-1} X^T.$$

    What are the properties of H?

    (A) It is symmetric and idempotent.

    (B) It is diagonal.

    (C) It is always invertible.

    (D) It has eigenvalues strictly greater than zero.

Solution: (A) It is symmetric and idempotent.

- Symmetric: $H^T = (X(X^T X)^{-1} X^T)^T = X((X^T X)^{-1})^T X^T = X(X^T X)^{-1} X^T = H$.
- Idempotent: $H^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H$.

(B) is false in general; (C) is false since H has rank $p < n$ and is therefore singular; (D) is false since eigenvalues of an idempotent matrix are 0 or 1.

17. (4 points) Which of the following best explains why a prediction interval is wider than a confidence interval?

(A) A prediction interval accounts for both the uncertainty in the estimated regression function and the natural variability in future observations.

(B) A prediction interval uses a smaller significance level than a confidence interval.

(C) A confidence interval assumes that future observations will be close to the mean response.

(D) A prediction interval is based on a different test statistic than a confidence interval.

Solution: (A). A confidence interval for the mean response $E[Y \mid X = x^*]$ captures only the uncertainty in estimating $\hat{Y}$. A prediction interval for a *new individual observation* must additionally account for the irreducible random variation $\epsilon \sim N(0, \sigma^2)$ around the mean. This extra term makes the prediction interval wider.

18. (4 points) What is the primary source of additional uncertainty in a prediction interval compared to a confidence interval?

(A) Sampling variability in estimating $\hat{\beta}$

(B) The residual variance, which accounts for random variation in individual observations

(C) The presence of collinearity in the predictors

(D) The need to estimate the variance of the independent variables

Solution: (B). The variance of the prediction error for a new observation Y_{new} at x^* is

$$\text{Var}(Y_{\text{new}} - \hat{Y}^*) = \sigma^2 + \sigma^2 x^{*T}(X^T X)^{-1} x^*,$$

where the first σ^2 term is the residual variance from the new observation itself. This term is absent in the confidence interval variance, making the prediction interval wider.

19. (4 points) In a linear regression analysis with n rows of data, with the usual assumptions, which of the following quantities can never be zero?

A) The estimated intercept, $\hat{\beta}_0$

B) The first residual e_1

C) The last fitted value, $\hat{Y}_n$

D) All of the above quantities can be zero.

Solution: D) All of the above quantities can be zero. There is no reason that any of $\hat{\beta}_0$, e_1, or $\hat{Y}_n$ must be nonzero. Each can be zero for appropriate data. For instance, if the true intercept is zero, $\hat{\beta}_0$ will be close to (or exactly) zero; $e_1 = 0$ if the first observation happens to lie exactly on the fitted regression line; $\hat{Y}_n = 0$ if the predicted value for the last observation is zero.