Sample Midterm

Problem 1

Background

Consider the simple linear regression model:

$$y_i = \beta_1 x_i + \varepsilon_i \quad \text{for} \quad i = 1, \dots, n$$

where the intercept is set to zero. We are tasked with deriving the least squares estimator of β_1.

The least squares method minimizes the sum of squared errors (SSE), given by:

$$S(\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$$

The error (or residual) for each observation is:

$$\varepsilon_i = y_i - \beta_1 x_i$$

Solution

To find the estimator $\hat{\beta}_1$, we minimize $S(\beta_1)$ with respect to β_1. First, expand the SSE:

$$S(\beta_1) = \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_1 x_i + \beta_1^2 x_i^2)$$

Now, differentiate $S(\beta_1)$ with respect to β_1:

$$\frac{dS(\beta_1)}{d\beta_1} = \sum_{i=1}^{n} (-2y_i x_i + 2\beta_1 x_i^2)$$

Set the derivative equal to zero to find the minimum:

$$0 = \sum_{i=1}^{n} (-2y_i x_i + 2\beta_1 x_i^2)$$

Simplify:

$$0 = -2\sum_{i=1}^{n} y_i x_i + 2\beta_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i = \beta_1 \sum_{i=1}^{n} x_i^2$$

Solving for β_1 gives:

$$\beta_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Thus, the least squares estimator for β_1 without an intercept is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Problem 2

Part A

First, we calculate calculate $\hat{\beta}_1$ (the slope):

$$\begin{aligned}\hat{\beta}_1 &= r\left(\frac{S_y}{S_x}\right) \\ &= 0.21 \times \frac{0.91}{0.50} = 0.3822\end{aligned}$$

Where r is the correlation, S_y is the standard deviation of Y, and S_x is the standard deviation of X.

Now, we calculate $\hat{\beta}_0$ (the intercept):

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X} \\ &= -0.04 - (0.3822 \times 0.50) = -0.2311\end{aligned}$$

Where $\bar{Y}$ is the mean of Y and $\bar{X}$ is the mean of X.

For the standard deviations, we need to calculate:

$$\begin{aligned}SE(\hat{\beta}_1) &= \sqrt{\frac{1-r^2}{n-2}} \times \frac{S_y}{S_x} \\ SE(\hat{\beta}_0) &= SE(\hat{\beta}_1) \times \sqrt{\frac{\sum x^2}{n}}\end{aligned}$$

And so:

$$\begin{aligned}SE(\hat{\beta}_1) &= \sqrt{\frac{1-r^2}{n-2}} \times \frac{S_y}{S_x} \\ &= \sqrt{\frac{1-0.21^2}{100-2}} \times \frac{0.91}{0.50} = 0.1796\end{aligned}$$

For $SE(\hat{\beta}_0)$, we approximate $\sum x^2$ using the variance:

$$\begin{aligned}Var(X) &= \frac{\sum(x-\bar{x})^2}{n} = SD^2 = 0.50^2 = 0.25 \\ \sum(x-\bar{x})^2 &= n \times Var(X) = 100 \times 0.25 = 25 \\ \sum x^2 &= \sum(x-\bar{x})^2 + n\bar{x}^2 = 25 + 100 \times 0.50^2 = 50\end{aligned}$$

And so:

$$\begin{aligned}SE(\hat{\beta}_0) &= SE(\hat{\beta}_1) \times \sqrt{\frac{\sum x^2}{n}} \\ &= 0.1796 \times \sqrt{\frac{50}{100}} = 0.1270\end{aligned}$$

And we can fill in our least squares table as:

	Estimate	Standard Deviation
$\hat{\beta}_0$	-0.2311	0.1270
$\hat{\beta}_1$	0.3822	0.1796

Now for ANOVA::

$$SSR = \hat{\beta}_1^2 \times \sum(x - \bar{x})^2 = 0.3822^2 \times 25 = 3.6494$$
$$SSE = (n-1)S_y^2 - SSR = 99 \times 0.91^2 - 3.6494 = 78.2506$$
$$SST = SSR + SSE = 3.6494 + 78.2506 = 81.9000$$

And so we can fill in our ANOVA table as:

	Sum of squares	d.f.	Mean squares
Regression	3.6494	1	3.6494
Sum of squares of residuals	78.2506	98	0.7985
Total	81.9000	99	

Part B

For the two groups:

$$\text{Group 1 (X = 0):} \quad n_1 = 50, \bar{Y}_1 = -0.04 - 0.3822 \times 0 = -0.04$$
$$\text{Group 2 (X = 1):} \quad n_2 = 50, \bar{Y}_2 = -0.04 + 0.3822 \times 1 = 0.3422$$

The pooled standard deviation is:

$$s_p^2 = \frac{SSE}{n-2} = \frac{78.2506}{98} = 0.7985$$

And the t-statistic is:

$$\begin{aligned} t &= \frac{\bar{Y}_2 - \bar{Y}_1}{s_p\sqrt{\frac{2}{n}}} \\ &= \frac{0.3422 - (-0.04)}{\sqrt{0.7985} \times \sqrt{\frac{2}{100}}} \\ &= \frac{0.3822}{0.8936 \times 0.1414} \\ &= 3.0233 \end{aligned}$$

Finally, the degrees of freedom for this test is $n - 2 = 98$.

This t-statistic can be used to test the null hypothesis $H_0 : \mu_0 = \mu_1$ against the alternative hypothesis $H_A : \mu_0 \neq \mu_1$.

Problem 3

Background

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

Let $z_i = a + bx_i$, and consider the transformed model:

$$y_i = \gamma_0 + \gamma_1 z_i + \delta_i \tag{2}$$

Solution

We aim to show that:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i$$

The least squares estimators for model (1) are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Now we turn to model (2). Given $z_i = a + bx_i$, we have:

$$z_i - \bar{z} = b(x_i - \bar{x})$$

since

$$\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i = \frac{1}{n}\sum_{i=1}^n (a + bx_i) = a + b\bar{x}$$

We now compute:

$$S_{zz} = \sum_{i=1}^n (z_i - \bar{z})^2 = \sum_{i=1}^n [b(x_i - \bar{x})]^2 = b^2 \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 S_{xx}$$
$$S_{zy} = \sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y}) = \sum_{i=1}^n [b(x_i - \bar{x})](y_i - \bar{y}) = b\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = bS_{xy}$$

The least squares estimators for model (2) are:

$$\hat{\gamma}_1 = \frac{S_{zy}}{S_{zz}} = \frac{bS_{xy}}{b^2 S_{xx}} = \frac{\hat{\beta}_1}{b}$$
$$\hat{\gamma}_0 = \bar{y} - \hat{\gamma}_1 \bar{z} = \bar{y} - \left(\frac{\hat{\beta}_1}{b}\right)(a + b\bar{x}) = \bar{y} - \frac{\hat{\beta}_1 a}{b} - \hat{\beta}_1 \bar{x}$$
$$= (\bar{y} - \hat{\beta}_1 \bar{x}) - \frac{\hat{\beta}_1 a}{b} = \hat{\beta}_0 - \frac{\hat{\beta}_1 a}{b}$$

The predicted values from model (2) are:

$$\hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i$$
$$= \left(\hat{\beta}_0 - \frac{\hat{\beta}_1 a}{b}\right) + \left(\frac{\hat{\beta}_1}{b}\right)(a + bx_i)$$
$$= \hat{\beta}_0 - \frac{\hat{\beta}_1 a}{b} + \frac{\hat{\beta}_1 a}{b} + \hat{\beta}_1 x_i$$
$$= \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Therefore, the predicted values from both models are identical for all x_i and z_i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i$$

Problem 4

The p-value is defined as the probability of obtaining a test statistic at least as extreme as the observed one, assuming the null hypothesis is true.

For a two-tailed t-test:

$$p\text{-value} = 2 \cdot \mathbb{P}(T > |t|)$$

where T follows a t-distribution and t is the observed t-statistic. We define P as the random variable representing the p-value, and consider its CDF:

$$F_P(x) = \mathbb{P}(P \leq x), \quad \text{for } 0 \leq x \leq 1$$

Under the null hypothesis:

$$P = 2 \cdot \mathbb{P}(T > |t|)$$

Therefore:

$$\begin{aligned} F_P(x) &= \mathbb{P}(2 \cdot \mathbb{P}(T > |t|) \leq x) \\ &= \mathbb{P}(\mathbb{P}(T > |t|) \leq x/2) \\ &= \mathbb{P}(|t| \geq T^{-1}(1 - x/2)) \end{aligned}$$

Where T^{-1} is the inverse of the t-distribution's CDF.

Now, for a uniform distribution on $[0, 1]$, the CDF should be $F(x) = x$ for $0 \leq x \leq 1$. Under the null hypothesis, t follows a t-distribution. Therefore:

$$\begin{aligned} \mathbb{P}(|t| \geq T^{-1}(1 - x/2)) &= 2 \cdot (1 - (1 - x/2)) \\ &= x \end{aligned}$$

This shows that $F_P(x) = x$ for $0 \leq x \leq 1$, which is the CDF of a uniform distribution on $[0, 1]$.

As an aside, this result is known as the probability integral transform and holds not just for t-tests, but for all continuous test statistics under their null hypothesis.