Name: Daljeet Maken (dmm6393)

1. (**4x4 = 16 points**) Data were gathered for *y* = hours of sleep the previous day and *x* = hours of studying the previous day for *n* = 116 college students. The estimated regression equation is found to be *ŷ* = 7.56 – 0.269*x*.

(a) What is the estimated slope of the regression line? Write a sentence that interprets the slope in the context of the variables for this problem. That is, explain exactly what the slope indicates about the relationship between hours of sleep and hours of study.

The estimated slope is -0.269

Interpretation: The slope indicates that the response variable (hours of sleep the previous day) **decreases** by 0.269 hours for every unit (1 hour) **increase** in the predictor variable (hours of studying).

(b) What is the estimated intercept of the regression line? In the context of these variables, what is the interpretation of the intercept?

The estimated intercept of the regression line is 7.56. The interpretation of the intercept is that students who studied 0 hours the previous day, the estimated value of average hours of sleep the previous day is 7.56.

(c) For students who studied 2 hours the previous day, what is the estimated value of average hours of sleep the previous day?

The estimated value of average hours of sleep the previous day is:

ŷ = 7.56 – 0.269 \* 2 = 7.022

(d) Suppose that a student studied 2 hours the previous day and slept 6 hours that day. What is the value of the residual for that person? (A residual is defined as the difference between the observed and predicted values of *y* for an individual.)

We know predicted ŷ = 7.56 – 0.269 \* 2 = 7.022

Observed value from the question = 6

Residual = ei = yi − y^i = Observed – Predicted = 6 - 7.022 = -1.022

2. (**2x4 = 8 points**) Two of the following statements about a population model for a simple regression are correct and two are incorrect. Which two statements are correct? Explain your answer.

(a) *Yi* = 𝛽0 + 𝛽1*Xi* + 𝜀𝑖

Correct. That is, any response Yi will be a combination of the linear trend (β0 + β1 \* xi) plus some error 𝜀𝑖

(b) *Yi* = 𝛽0 + 𝛽1*Xi*

Incorrect. The error term 𝜀𝑖 is missing.

(c) E(*Yi* ) = 𝛽0 + 𝛽1*Xi* + 𝜀𝑖

Incorrect. The linearity condition is that the mean of the error, E(𝜀𝑖), at each value of the predictor, xi, is zero. Therefore E(*Yi* ) = 𝛽0 + 𝛽1*Xi*

(d) E(*Yi* ) = 𝛽0 + 𝛽1*Xi*

Correct. The linearity condition is that the mean of the error, E(𝜀𝑖), at each value of the predictor, xi, is zero. We know that *Yi* = 𝛽0 + 𝛽1*Xi* + , therefore

E(Yi) = E(𝛽0 + 𝛽1*Xi* + 𝜀𝑖) = 𝛽0 + 𝛽1*Xi* since E(𝜀𝑖) = 0

3. (**4x4 = 16 points**) The fitted line plot below gives results for a straight-line regression between *y* = left forearm length (cm) and *x* = left foot length (cm). The data are from *n* = 55 college students



(a) Write a sentence that gives the value of the slope and interprets it in the context of this situation.

The estimated slope is 0.5963

Interpretation: The slope indicates that the response variable (left forearm length) **increases** by 0.5963 cm for every unit (1 cm) **increase** in the predictor variable (left foot length).

(b) Write a sentence that gives the value of *R*2 and interprets it in the context of this situation.

The r-squared value is 45.6%. It means that 45.6% of the variation in the left forearm length is reduced by taking into account left foot length. Or, we can say that 45.6% of the variation in left forearm length is 'due to' or is 'explained by' left foot length.

(c) Use the equation to estimate the average left forearm length of college students with a left foot length of 25 cm.

The estimated value of average left forearm length is:

ŷ = 10.16+0.5963 \* 25 = 25.0675

(d) The Minitab output includes the information that “**S**=1.65245”. Explain what is measured by this statistic.

The quantity S = 1.65245, is the square root of MSE. That is, in general, S=sqrt(MSE) – therefore MSE = 1.65245^2 = 2.730591003

The mean square error estimates σ2, the common variance of the many subpopulations.

4. (**10x4 = 40 points**) Infant weights in pounds have an upward linear trend with age in months. Data from a sample of 5 babies in a local community, including one newborn and four others who are 1 month, 2 months, 3 months and 5 months old, were used to obtain an estimated regression equation based on the least squares criterion with a slope of 0.2838 pounds per month. Some of the information is given in the following table.



(a) What is the equation of the population regression line in this setting? *[Hint: There should be no numbers in this equation, just* 𝛽*’s.]*

The equation of the population regression line is:

population regression line

(b) What is the estimated regression equation? *[Hint: There should be numbers in this equation. Use the information in the question and in the table, particularly the 3-month old baby.]*

Since we know the predicted value for the 3 months old. We also have estimated regression equation: y^i=b0+b1xi

Therefore: 8.8271 = b0+0.2838\*3, gives b0 = 8.8271-0.2838\*3 = 7.9757

Therefore the estimated regression equation for this particular case is:

y^i=7.9757 + 0.2838 \* xi

(c) Based on the estimated regression equation, what is the predicted birth weight of a newborn in this community?

The predicted birth weight is y^i=7.9757 + 0.2838 \* xi

=> y^i=7.9757 + 0.2838 \* 0 = 7.9757

(d) What is the actual birth weight of the newborn in the sample?

Since we know: Residual = 0.0243 = ei = yi − y^i = Observed – Predicted = Observed - 7.9757 = 0.0243

Therefore Observed (actual birth weight) = 0.0243 + 7.9757 = 8

(e) Complete the remaining entries in the table above.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Age | (xi) | 0 | 1 | 2 | 3 | 5 |
| Weight | (yi) | 8 | 8.1 | 8.4 | 9.3 | 9.2003 |
| Predicted weight | (𝑦̂𝑖) | 7.9757 | 8.2595 | 8.5433 | 8.8271 | 9.3947 |
| Residual error | (ei) | 0.0243 | -0.1595 | -0.1433 | 0.4729 | -0.1944 |

Computed using the fact that sum of residuals is 0

(f) Comment on the validity of using the estimated regression equation to predict the weight for a one year old.

The regression equation is not valid for 1 year olds since this will require the model to be "extrapolated" beyond the "scope of the model" (the range of the x values)

(g) Calculate SSE, the sum of residual error squares.

Squaring each residual we get,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0.00059 | +0.02544 | +0.020535 | +0.223634 | +0.037791 |

= 0.3079914

(h) Calculate the sample estimate of the variance, 𝜎2, for the regression model.

MSE is an unbiased estimator of variance for regression model

E{MSE} = σ2

The s2 = MSE = SSE / n-2 => 0.3079914 / (5-2) = 0.1026638

(i) Calculate the value that would be given in Minitab for “**S=**". Write a sentence that interprets this value.

Minitab will show S = 0.320412

This values comes from the sqrt(s2) = sqrt(MSE) = sqrt(0.1026638) = 0.320411922’

This represents the point estimate of σ, the standard deviation of the probability distribution of weights for any age. Stated differently this is the variation in weight from set to set of 5 infants.

(j) Calculate the value of *R*2. To start, you will have to calculate the value of SSTO. Write a sentence that interprets the value of *R*2.

We have yBar = 8.60006

SSTO =  =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0.360072 | +0.25006 | +0.040024 | +0.489916 | +0.360288 |

= 1.50036

SSE = 0.3079914

R2 = 1-SSE/SSTO = 1-0.3079914/1.50036=0.794722 or 79.47%

This is interpreted as: 79.47 percent of the variation in weight of infants is 'explained by' the variation in predictor age of the infants.

5. (**4x5 = 20 points**)

(a) Briefly describe the four assumptions (or conditions) that underlie the simple linear regression model.

The four conditions that comprise the simple linear regression model:

1. The mean of the response, E(Yi), at each value of the predictor, xi, is a Linear function of the xi.
2. The errors, ei, are Independent.
3. The errors, ei, at each value of the predictor, xi, are Normally distributed.
4. The errors, ei, at each value of the predictor, xi, have Equal variances (denoted σ2).

The mnemonic LINE is used to remember.

(b) The scatter plot below shows sample data for *y* = selling price of a house and *x* = square foot area of the house.



(i) Name one condition that may be satisfied by the selling price vs square foot area data and justify your choice.

The conditions satisfied above are:

1. The mean of the response, E(Yi), at each value of the predictor, xi, is a Linear function of the xi.

Since the average selling price at each value of square foot area is linearly increasing as square foot area is increasing

1. ~~The errors, ei, are Independent.~~

~~The residual errors at each values of square foot area are independent of the errors at other square foot areas. Even though the residuals are increasing as square foot area is increasing, they are however independent. (~~ scatter plots don’t indicate independence: independence of errors are not apparent from the scatterplot)

(ii) Name one condition that may not be satisfied by the data and justify your choice.

The condition that is not met is:

1. The errors, ei, at each value of the predictor, xi, have Equal variances (denoted σ2).

The scatter plot clearly shows that the scatter is increasing as square foot area is increasing. This indicates that the variance at each value of the predictor, xi is not equal.

(iii) Would you expect the magnitude of the sample correlation coefficient to be near 0, closer to +1, or closer to –1? Justify your choice.

We expect the sample correlation coefficient to be closer to +1

The scatter plot clearly shows that there is a positive correlation between Selling Price and Square foot area. Such relationships lead to a positive value of the sample correlation coefficient. In the case of perfect +ve correlation the sample correlation coefficient approaches +1.

(iv) Based on the estimated regression equation, 𝑃𝑟𝑖𝑐𝑒̂ = 5049 + 64.87 SQFT, what is the y-intercept estimate? Is this value meaningful? Why or why not?

The value of y-intercept estimate b0 is 5049. This value is not meaningful.

The intercept is the value of the estimated regression equation at x = 0. Here, it tells us that the selling price of a house with 0 square foot area is predicted to be 5049. Clearly, this is meaningless. This happened because we "extrapolated" beyond the "scope of the model". It is not meaningful to have a square foot area of 0, that is, the scope of the model does not include x = 0. So, here the intercept b0 is not meaningful.