AP Statistics: Linear Regression

From Simple Studies, https://simplestudies.edublogs.org & @simplestudiesinc on Instagram

General Notes

- Equation: $\hat{y} = a + bx$
  - b = Slope = $\Delta y / \Delta x$
- Line of Best Fit: Minimal residual (vertical distance between line and points)
- r: Shows correlation (not causation)
  - Sign of r = Sign of b (Slope)
  - The closer r is to ±1, the stronger r is
- r^2: Correlation of determination (convert to a percentage)
- LSLR: Least squares regression line
- Extrapolation: Extending the LSLR to values outside the range of data used to generate it

Interpretations

- Slope (b)
  - For each [*x*], the [*y*] increases/decreases on average by [*b*]
- Correlation Coefficient (r)
  - There is a [*strong/weak*], [*positive/negative*], and [*linear/non-linear*] relationship between [*x*] and [*y*]
  - As [*x*] increases, it appears that [*y*] [*increases/decreases*].
- Correlation of determination (r^2)
  - r^2% variation in [*y*] can be explained by a linear relationship with [*x*]
- Put all interpretations in the context of the problem!

Residuals

- $y - \hat{y}$
- If the graph of the residuals shows a pattern, a non-linear model should be used

- “Since the residual graph [*does/does not*] show a pattern, a [*linear/nonlinear*] model should be used”
- Positive residual = Underestimate; Negative residual = Overestimate

Predictor	Coef	SE Coef	T	P
Constant	0.137 [1]	0.126 [2]	1.09 [3]	0.289 [4]
Wind velocity	0.240 [5]	0.019 [6]	12.63 [7]	0.000 [8]

S = 0.237 [9] R-Sq = 0.873 [10] R-Sq (adj) = 0.868 [11]

From HYPERLINK "https://secure-media.collegeboard.org/apc/ap11_frq_statistics.pdf"https://secure-media.collegeboard.org/apc/ap11_frq_statistics.pdf, p. 10 (modified)

Table

- When presented with a table (like the one seen in the image), you DO NOT need: *2, 3, 4, 9, 11*
- Line of best fit: $\hat{y} = 1 + 5x$
- 6 is the standard deviation of the slope (SE_b)
- 7 is the t-score of the slope
- 8 is the p-value of 7
- 9 is the standard deviation of the residuals

Confidence Intervals

- Step 1: Conditions
  - “The scatterplot shows a somewhat linear trend”
  - “A residual plot shows no pattern”
  - “All points are independent and random”
  - If the problem states that the conditions are satisfied, you can write: “The stem of the problem states the conditions are satisfied.”
- Step 2: Formula
  - List the formula, substitution, unrounded answer, and degrees of freedom
  - Formula: $b \pm t^*(SE_b)$
  - d.f. = n - 2
- Step 3: Interpretation
  - “Based on these samples, we are _% confident that for every [*x*], [*y*] will [*increase/decrease*] between [*lower value*] and [*upper value*]”

Hypothesis Testing

- Step 1: Hypothesis
    - H_0: $\beta = 0$ (There is no linear relationship between [*x*] and [*y*])
    - H_a: $\beta \neq 0$ (There is linear relationship between [*x*] and [*y*])
- Step 2: Conditions
    - "The scatterplot shows a somewhat linear trend"
    - "A residual plot shows no pattern"
    - "All points are independent and random"
    - If the problem states that the conditions are satisfied, you can write: "The stem of the problem states the conditions are satisfied."
- Step 3: Formula
    - List the formula, substitution, unrounded answer, and degrees of freedom
    - Formula: $t = \frac{b - \beta}{SE_b}$
    - d.f. = n - 2
- Step 4: P Value
    - 2P(t > or < _) = _
- Step 5: Conclusion
    - "Assuming H_0 is true, since the p value [*p value*] is [*greater/less*] than α = _, we [*fail to reject/reject*] H_0"
    - "We [*have/do not have*] sufficient evidence to suggest H_a, that there is a linear relationship between [*x*] and [*y*]"