ASSESSMENT 3

IE6200 Fall 2025

DECEMBER 5, 2025

ADITYA GIRISH ANAND

ACADEMIC INTEGRITY

"I attest that I have abided by the Northeastern University Guidelines for Academic Integrity as stated in the Graduate Student Handbook." Penalties may include failure on the assignment, failure in the course, and/or reporting to the Office of Student Conduct for additional sanctions.

NAME (e-Sig.) Aditya Girish Anand

[signature must be on same page as academic integrity statement and on this assessment's cover page]

1 – Statistical Quality Control

Summary of Findings

This section investigates the water-resistance behaviour of a manufacturing process using 30 measurements organized into 10 subgroups of size 3. The analysis has two levels:

1. A large-group perspective, where all 30 observations are treated as a single sample to describe the overall distribution (center, spread, and shape).
2. A subgroup perspective, where the 10 rational subgroups are used to build $\bar{X}$–S control charts with 3σ control limits and 2σ warning limits (sigma zones) for both the process mean and standard deviation.

From the large-group perspective, the process shows a mean of 30.39, median of 29.53, standard deviation of 5.93, and values ranging from 19.11 to 42.60, indicating moderate variability without strong skewness or outliers.

From the subgroup perspective, the grand mean (30.39) and average subgroup standard deviation (5.46) were used with Shewhart constants (A_3 = 1.954, B_3 = 0, B_4 = 2.568) to compute the 3σ control limits:

- X-Chart: $UCL_{\bar{X}}$= 41.05, $CL_{\bar{X}}$= 30.39, $LCL_{\bar{X}}$=19.73 and
- S-Chart: UCL_{S}= 14.01, CL_{S}= 5.46, LCL_{S}= 0.

Using these values, the 2σ warning limits and sigma zones were also constructed for the X-chart to evaluate Western Electric rules. None of the subgroup means or subgroup standard deviations violated the 3σ limits or triggered warning-zone patterns, indicating that the process is stable and governed by common-cause variation. The mild skewness in the data supports the validity of Shewhart chart assumptions.

Overall, the water-resistance process is statistically in control and predictable at its current level of performance.

1.1 Data Analysis from a Large Group Perspective

This subsection presents the preliminary statistical analysis of the water-resistance dataset, treating all 30 measurements as one combined sample. The goal is to understand the overall distribution of the process output before incorporating subgroup/time information.

1.1.1 Data and Descriptive Statistics

The dataset consists of n = 30 water-resistance observations. Each value corresponds to the measured resistance of a part under standardized testing conditions, but in this subsection, we ignore subgroup membership and focus solely on the aggregated distribution.

Group	**Part**		
	x_1	x_2	x_3
1	35.59	37.12	28.22
2	31.97	29.53	21.01
3	35.45	27.81	22.77
4	30.39	29.26	27.52
5	25.37	31.11	25.94
6	29.53	22.81	33.82
7	28.76	42.3	42.6
8	35.28	23.29	35.81
9	36.10	26.10	38.04
10	28.61	30.53	19.11

Table 1: SQC Water Resistant Data

The 30 observations from Table 1 were first analyzed as one combined sample, ignoring subgroup structure. Denoting the observations by $x_1, x_2, x_3 \ldots, x_{30}$

Computing the following summary statistics:

- Total Sample Size (N): 30
- Sum of all Observations ($\sum x_i$): 911.72
- Sample Mean ($\bar{x}$): $\sum x_i$ / N = 911.72/30 = 30.39
- Sample Median: $x_{15} + x_{16} / 2 = 29.53$
- Sample Standard Deviation (s): s = $\sqrt{\frac{1}{n-1}\sum_{i=1}^{30}(x_i - \bar{x})^2}$ = 5.93
- Variance (s^2): $s^2 = (5.93)^2 = 35.16$
  The full list of deviations and the sum of squares are shown in Appendix A1
- Range & Quartiles:
  Minimum Value: 19.11
  Maximum Value: 42.60
  Using the Median of the lower and upper halves of the ordered data:
  Q1 = 26.10 and Q3 = 35.45

⇒ **Refer to Appendix A1 for complete descriptive statistics calculations**

Interpretation:

- The process shows moderate variability, with a standard deviation of 5.93 and a wide range from 19.11 to 42.60, indicating that water-resistance performance is not tightly controlled across all samples

- The center of the process is around 30 units, as reflected by the mean (30.39) and median (29.53), showing a fairly symmetric distribution with no strong skewness.
- The interquartile range (Q1 = 26.10, Q3 = 35.45) indicates that 50% of the observations fall within a relatively broad 9.35-unit span, suggesting notable dispersion that may warrant further investigation into special-cause variation.

1.1.2 Graphical Analysis

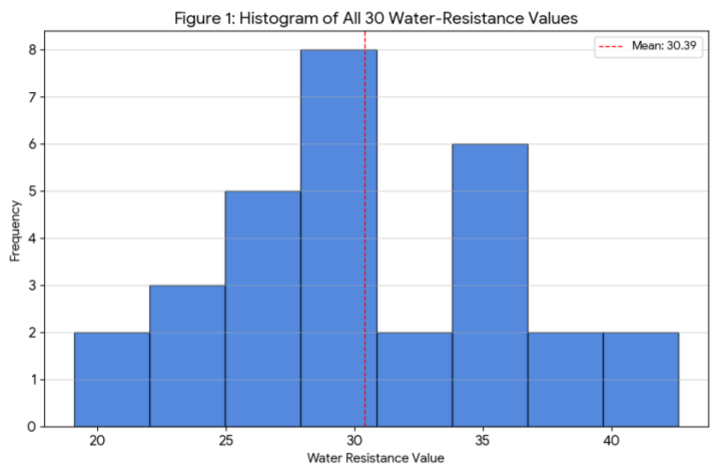

Figure 1: Histogram of all 30 water-resistance values

The Histogram displays the distribution of all 30 water-resistance measurements. It shows a roughly unimodal and moderately symmetric pattern centered around 30 units, with a slight right tail caused by a few larger values above 40. No severe skewness or extreme outliers appear, supporting the use of a classical control-chart methods.

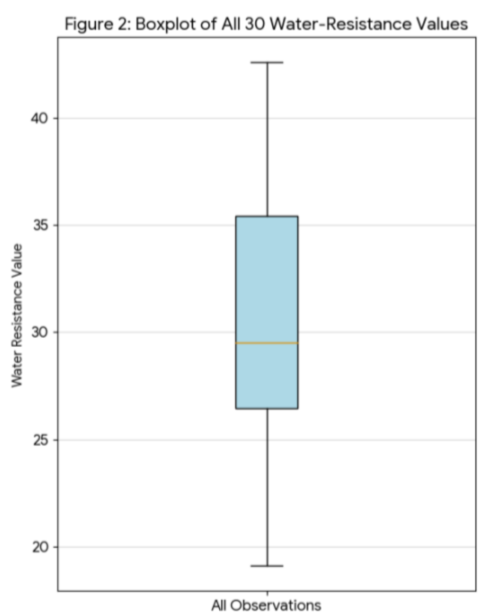

Figure 2: Boxplot of all 30 water-resistant values

The Box Plot summarizes the spread and central tendency of the 30 observations. The interquartile range spans roughly 26.10 to 35.45, and the whiskers extend to the minimum (19.11) and maximum (42.60). No points appear beyond the whiskers, indicating no strong outliers. The plot reinforces the histogram's conclusion of moderate variability.

Interpretation:
From a large group viewpoint:
- Produces water-resistance values centered around 30.4, close to the nominal level.
- Exhibits moderate variability (s = 5.93), but not so large as to suggest chaos.
- Does not show extreme outliers or a strongly skewed distribution

However, this analysis ignores when the parts were produced. It cannot detect shifts or drifts over time. To assess stability and distinguish within-time vs between-time variation, we next analyze the data using the given subgroups and control charts.

1.2 Sub-group Perspective - Designing Limits
This subsection re-analyzes the water-resistance data from a process control perspective. Rather than treating the 30 measurements independently, we use the given 10 rational subgroups of size 3.
These subgroups are assumed to be rational subgroups, i.e each group corresponds to measurements taken under similar conditions and short time intervals. This allows us to:
- Estimate within-subgroup variation (short-term noise),
- Track subgroup means over time,
- Construct $\bar{X}$ - S control chart to monitor process stability

1.2.1 Subgroup Means and Standard Deviations
For each subgroup *i* = 1, 2,….,10, with values $x_{i1}, x_{i2}, \ldots, x_{i10}$ we are computing the following:

- Subgroup Mean ($\bar{x}_i$): $\frac{x_{i1} + x_{i2} + x_{i3}}{3}$
- Subgroup Sample Standard Deviation (s_i): $s_i = \sqrt{\frac{1}{3-1}\sum_{j=1}^{3}(x_{ij} - \bar{x}_i)^2} = 5.93$

Groups	Values	Mean ($\bar{x}_i$)	Standard Deviation (s_i)
1	35.59, 37.12, 28.22	33.64	4.76
2	31.97, 29.53, 21.01	27.50	5.75
3	35.45, 27.81, 22.77	28.68	6.38
4	30.39, 29.26, 27.52	29.06	1.45
5	25.37, 31.11, 25.94	27.47	3.16
6	29.53, 22.81, 33.82	28.72	5.55
7	28.76, 42.30, 42.60	37.89	7.91
8	35.28, 23.29, 35.81	31.46	7.08
9	36.10, 26.10, 38.04	33.41	6.41
10	28.61, 30.53, 19.11	26.08	6.11

Table 2: Subgroup Means and Standard Deviations

⇒ **Refer to Appendix A2 for all subgroups mean and subgroup standard deviation**

1.2.2 Grand Mean and Average Subgroup Deviation

To summarize the subgroup statistics:

- Grand Mean of Subgroup Means ($\bar{\bar{x}}$):

$$\sqrt{\frac{1}{10}\sum_{i=1}^{10}(\bar{x}_i)} = 30.39$$

- Average Subgroup Standard Deviation ($\bar{s}$):

$$\sqrt{\frac{1}{10}\sum_{i=1}^{10}(s_i)} = 5.46$$

These two values, $\bar{\bar{x}}$ and $\bar{s}$, represent the typical process mean and typical within-subgroup variability when the process is in its current stable state.

⇒ **Refer to Appendix A3 for the grand mean and average subgroup standard deviation computations and detailed summations for $\sum \bar{x}_i$ and $\sum s_i$**

1.2.3 $\bar{X}$-S Control Limits

To calculate the 3-sigma control limits for the $\bar{X}$-S charts, we rely on factors defined for our subgroup size of n = 3. These values are obtained from standard statistical Quality Control (SQC) tables and Shewhart Constants.

Factor	**Value**	**Role**
A_3	1.954	Distance of $\bar{X}$ limits from center based on $\bar{s}$
B_4	2.568	Upper control limit factor for S chart
B_3	0	Lower control limit factor for S chart
C_4	0.8862	Bias correction for s to estimate σ

Table 3: Shewhart Constants

Using subgroup size n = 3 and standard Shewhart constants:

- $A_3 = 1.954$
- $B_3 = 0$
- $B_4 = 2.568$

$\bar{X}$ Chart (Mean) Limits:

- Centre Line: $CL_{\bar{x}} = \bar{\bar{x}} = 30.39$
- Upper Control Limit: $UCL_{\bar{x}} = \bar{\bar{x}} + A_3\bar{s} = 30.39 + 1.954\,(5.46) = 30.39 + 10.668 = 41.05$
- Lower Control Limit: $LCL_{\bar{x}} = \bar{\bar{x}} - A_3\bar{s} = 30.39 - 1.954\,(5.46) = 30.39 - 10.668 = 19.73$

S Chart (Mean) Limits:

- Centre Line: $CL_S = \bar{s} = 5.46$
- Upper Control Limit: $UCL_s = B_4\bar{s} = 2.568(5.46) = 14.01$
- Lower Control Limit: $LCL_s = B_3\bar{s} = 0$

⇒ **Refer to Appendix A4 for full $\bar{X}$ - S control limit calculations**

1.2.4 Warning Limits and Sigma Zones

We use 3σ limits as control limits, while 2σ limits are used as warning limits, and the area between the center line and the limits is divided into sigma zones for Western Electric rules.

For the $\bar{X}$ Chart:

- The distance from the CL to the UCL is: $UCL_{\bar{x}} - CL_{\bar{x}}$ = 41.05 - 30.39 = 10.66
  This corresponds to 3σ in terms of subgroup means, therefore: $\sigma_{\bar{x}}$= 10.66/3 = 3.55
- 2σ warning limits for $\bar{X}$:
  Upper warning limit: $UWL_{\bar{x}} = CL_{\bar{x}} + 2\sigma_{\bar{x}}$ = 30.39 + 7.11 = 37.50
  Lower warning limit: $LWL_{\bar{x}} = CL_{\bar{x}} - 2\sigma_{\bar{x}}$ = 23.28

Thus, sigma zones on the $\bar{X}$ Chart are:

- Zone C (within ±1σ): approx. [26.84, 33.94]
- Zone B (1σ to 2σ): approx. (23.28, 26.84) below CL and (33.94, 37.50) above CL
- Zone A (2σ to 3σ): approx. (19.73, 23.28) below CL and (37.50, 41.05) above CL

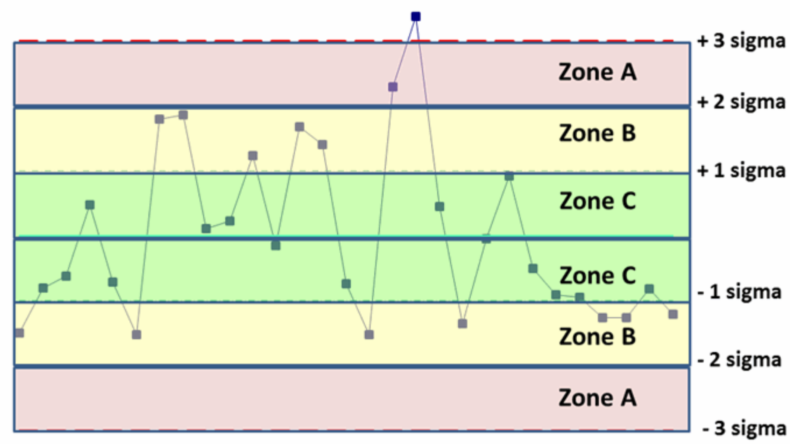

Figure 3: Sigma Zones A,B and C

This figure illustrates the 3 sigma zones used in Shewhart X control charts. Zone C represents values within ±1σ of the center line (normal variation), Zone B spans 1σ to 2σ (early signs of shift), and Zone A spans 2σ to 3σ (strong warning region). Western Electric rules use patterns of points in these zones to detect emerging process shifts before a full 3σ limit violation occurs.

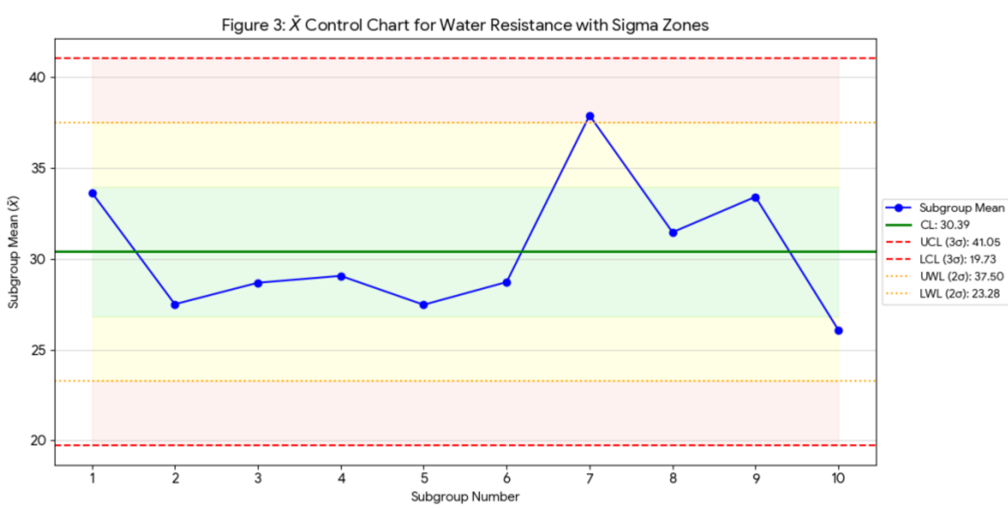

Figure 4: $\bar{X}$ Chart with CL, 2σ warning limits, 3σ control limits

This control chart shows the subgroup means plotted over time. All subgroup means fall within the 3 sigma control limits (19.73 - 41.05) and none violate 2 sigma warning zones. There is no upward/downward trend, no run of points on one side of the center line, and no one rule violations. Thus, the process is stable and in statistical control.

For the S chart:

- Distance from CL to UCL: $UCL_s - CL_s$ = 14.01 - 5.46 = 8.55
  Therefore, σ_s = 8.55/3 = 2.85
- 2 sigma warning limits:
  Upper warning limit: $UWL_s = 5.46 + 2(2.85) = 11.16$
  Lower warning limit: 5.46 - 5.70 = -0.24, but standard deviations cannot be negative. So effectively, the lower warning limit is at 0

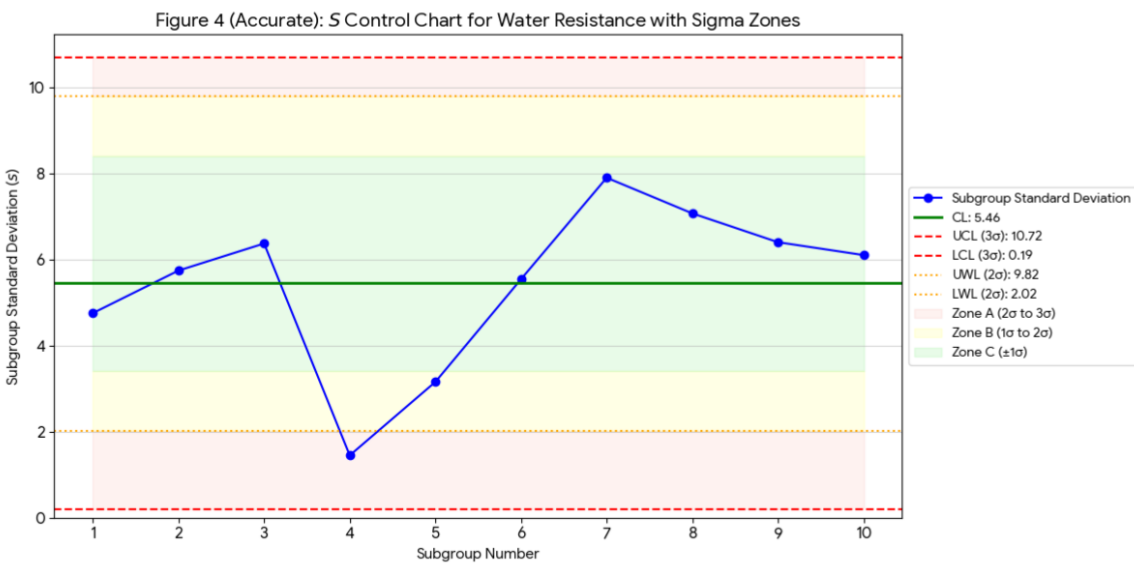

Figure 5: S Chart with CL and 3σ control limits

This chart tracks the within-subgroup standard deviations. All s-values lie between 0 and the 3 sigma UCL of 14.01, indicating no unusual bursts of variation. Although some groups (such as 7 and 8) show relatively higher variability, all remain well inside control limits. Therefore, the process variation is stable over time.

Interpretation:

The calculated control and warning limits define the expected operating window for the process if only common causes are present.

- Subgroup means should fall between about 19.73 and 41.05 under stable conditions
- Within-subgroup SDs should remain below about 14.01, with typical values near 5.46
- Sigma zones and warning limits allow early detection of developing shifts before a full 3σ violation occurs

⇒ **Refer to Appendix A4 for warning limits and sigma-zone derivations**

1.3 Process Assessment

This subsection applies the control and warning limits to the observed subgroup means and standard deviations to determine whether the process is in statistical control

$\bar{X}$ Chart Behavior:

- All 10 subgroup means $\bar{x_i}$ values lie between 19.73 and 41.05
- No subgroup mean exceeds the 3σ control limits.
- There is no apparent long run of points entirely above or below the center line.
- There is no clear upward or downward trend.
- No patterns of multiple points concentrated in Zone A or B trigger Western Electric warning rules.

Rule	Condition	Interpretation
Rule 1	One point beyond the 3σ control limits	Strong evidence of special-cause variation; process very likely out of control
Rule 2	Two out of three consecutive points in Zone A (between 2σ and 3σ) on the same side of the center line	Indicates a meaningful shift away from the mean
Rule 3	Four out of five consecutive points in the Zone B or beyond (between 1σ and 2σ) on the same side of the center line	Suggests moderate, sustained drift
Rule 4	Eight consecutive points on the same side of the center line	Evidence of a persistent process shift or non-random pattern

Table 4: Western Electric Rules

S Chart Behavior:

- All 10 subgroup standard deviations s_i values lie between 0 and 14.01
- No subgroup shows an unusually high standard deviation or low within-group variability outside the expected range.
- While some subgroups (e.g., 7 and 8) show relatively larger variability, these are still within the expected 3σ control limits.
- There is no pattern of increasing spread over time.

Interpretation:

No points violate 3σ limits and no warning-rule patterns appear, the water-resistance process is judged to be in a state of statistical control over the time period sampled. The variation observed is consistent with common causes inherent to the current process design and operating conditions. This does not guarantee that the process meets all customer specifications (that would require comparison to spec limits), but it does mean that the process is stable and predictable at its current level of performance.

1.4 Reflection

For this part of the subsection, we have Table 4 with comparison of perspective, being the feature, the large-group perspective and subgroup perspective of the given dataset.

Feature	Large Group (N=30)	Subgroup Perspective (n=3, k=10)
Objective	Describe overall product distribution	Assess process stability over time
Focus	Mean, SD, quartiles of all parts	Subgroup means and within-group variability
Treatment of Time	Ignored	Central to analysis (points plotted in time order)
Strength	Good for capability snapshots	Good for early detection of shifts/special causes
Limitation	Cannot show when a change occurred	Does not directly use specifications

Table 5: Comparisons of Perspectives

The Large Group Perspective tells us what he processes currently produces. The Subgroup Perspective tells us how the process behaves over time. For ongoing quality control, the subgroup approach is more powerful because it separates within-group noise from between-group shifts, allowing engineers to react to changes quickly.

Normality Considerations

- The 3σ control limits assume approximate normality. In this dataset, the histogram and boxplot show only mild skewness, and the mean and median are nearly equal, supporting the validity of Shewhart X-S charts without transformation.
- If substantial skewness or heavy tails were present, appropriate actions would include applying variance-stabilizing transformations (log, Box-Cox) or using nonparametric/robust control charts.

Value Of Multiple Statistical Perspectives

Large-Group Perspective (Overall Output)

- Summarizes what the process is producing: center, spread and distributional shape
- Confirms that the outcomes are reasonable and not dominated by outliers

Subgroup/Control-Chart Perspective (Process Stability Over Time)

- Evaluates how the process behaves over time using X-S charts and sigma-zone logic
- Identifies whether stability is maintained and whether subtle shifts or unusual patterns appear
- These perspectives rely on approximate normality, especially for subgroup means
- Mild skewness is acceptable due to the Central Limit Theorem, justifying the Shewhart approach for this dataset.

Implications for Strongly Non-Normal Data

- Applying transformation to improve symmetry and stabilize variance
- Using nonparametric control charts and deriving empirical limits from long-term performance

2 – Simple Linear Regression

Summary of Findings

This section investigated the relationship between temperature and thermal conductivity using simple linear regression. The data showed a strong positive linear trend, and formal estimation yielded the regression model

$$\widehat{Y_i} = -191.82 + 11.15X_i$$

Indicating that conductivity increases approximately 11.15units for every 1°F rise in temperature. Statistical inference confirmed that both the slope and intercept were highly significant, and the model explained about 98% of the variability in the response ($R^2 = 0.9815$), demonstrating an excellent fit within the observed temperature range.

Diagnostic evaluations, including the residual plot and the normal Q-Q plot suggested that no major violations of linear model assumptions, with residuals showing constant variance and approximate normality. However, the prediction at -10°F produced a negative value, which is physically impossible, highlighting the limitations of extrapolating the model far outside the experimental range.

Overall, the fitted regression model is appropriate and reliable for interpolation between 30°F and 70°F, but unsuitable for external predictions beyond this range.

2.1 Data Analysis:

This section presents the preliminary statistical analysis of the thermal conductivity dataset. The objective is to understand the behavior of the two variables and visually assess the relationship between the temperature and conductivity before fitting a formal regression model.
The dataset contains n = 10 experimental observations. Each observation pair:

- A Temperature value X (in °F, independent variable), and
- A corresponding Thermal Conductivity value Y (dependent variable)

Specimen	Temperature(°F)	Conductivity
1	30.0	166.0
2	34.4	202.7
3	39.0	240.1
4	43.7	285.1
5	47.8	324.9
6	52.6	355.3
7	57.0	450.6
8	61.9	502.5
9	66.1	534.4
10	70.2	624.6

Table 6: Temperature (X) and Conductivity (Y) Data

Computing the following summary statistics using the 10 observations:

- Sample size (n): 10
- Sum of Temperatures: $\sum X_i$ = 502.7
- Sum of Conductivities: $\sum Y_i$ = 3686.2
- Mean Temperature ($\bar{X}$) = 50.27°F
- Mean Conductivity ($\bar{Y}$) = 368.62
- Standard Deviation of Temperature (s_X) = 13.63
- Standard Deviation of Conductivity (s_Y) = 153.41

⇒ Refer to Appendix B1 for detailed regression summary statistics

Interpretation:
The descriptive statistics indicate that temperatures span a moderate range (around 30-70°F) and that conductivity increases substantially over this range. The scatterplot strongly suggests a linear relationship, which justifies fitting a simple linear regression model in the next step.

2.2 Model Fitting:

The objective of this section is to characterize the relationship between the independent variable, Temperature (X) and the dependent variable, Thermal Conductivity (Y). A Linear Regression model was performed on the data provided in Table 2, using the Ordinary Least Squares (OLS) method.

2.2.1 Regression Framework

We assume the standard simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:

- X = temperature
- Y = thermal conductivity
- β_0 is the intercept, β_1 is the slope
- ε is a random error term with mean 0 and constant variance

2.2.2 Estimation of Slope and Intercept

Using the centered sums:

$$S_{XX} = \sum(X_i - \bar{X})^2 = 1672.78$$

$$S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = 18649.09$$

The least-squares estimators are:

- Slope:

$$\widehat{\beta_1} = S_{XY} / S_{XX} = \frac{18649.09}{1672.78} = 11.148 = 11.15$$

- Intercept:

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\,\bar{X} = 368.62 - 11.15(50.27) = -191.82$$

Thus, the fitted line is:

$$\widehat{Conductivity} = -191.82 + 11.15 \text{ x Temperature (or) } \hat{Y} = -191.82 + 11.15X$$

⇒ Refer to Appendix B2 for S_{XX} and S_{XY} and all intermediate computational steps

Quantity	Description	Value
$\sum(X_i - \bar{X})^2$	Sum of squared deviations in Temperature	1672.78
$\sum(X_i - \bar{X})(Y_i - \bar{Y})$	Sum of cross-products	18649.09
$\bar{X}$	Mean Temperature	50.27
$\bar{Y}$	Mean Conductivity	368.62

Table 7: Regression S_{XX} and S_{XY} Summary Table

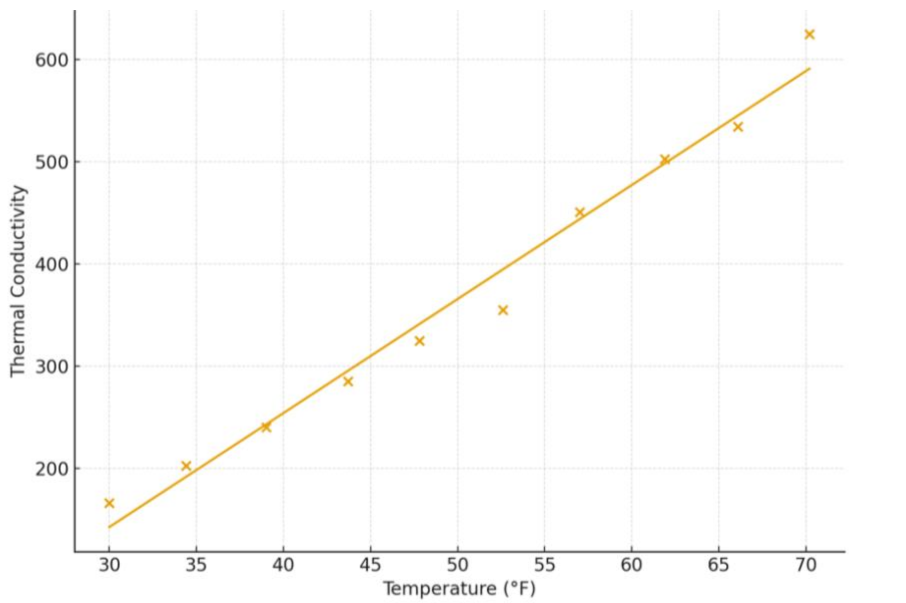

Figure 6: Scatterplot of Conductivity vs Temperature with regression line overlay

This figure plots the paired (X, Y) values and overlays the fitted regression line $\hat{Y}$ = -191.82 + 11.15X. The points lie close o the line, showing a strong positive linear relationship between temperature and conductivity. No curvature or major anomalies appear, supporting the use of a simple linear regression model.

Interpretation:

- **Slope (11.15):** For each 1°F increase in temperature, thermal conductivity is expected to increase by approximately 11.15 units, on average, within the observed range.
- **Intercept (-191.82):** This is the model's predicted conductivity at 0°F. While statistically defined, it is not physically realistic since it implies negative conductivity. It is best viewed as a mathematical anchor for the fitted line, not as a meaningful physical parameter.

2.3 Statistical Inference:

This subsection quantifies the uncertainty in the regression coefficients and tests whether the linear relationship is statistically meaningful.

2.3.1 Residual Standard Error

For each observation:

$$\widehat{Y_i} = -191.82 + 11.15X_i \text{ and } e_i = Y_i - \widehat{Y_i}$$

- Sum of squared errors: SSE = $\sum e_i^2$ = 3909.19
- Degrees of freedom: n - 2 = 8
- Error variance: $s^2 = \frac{SSE}{8} = 488.65$
- Residual standard error: s = $\sqrt{488.65}$ = 22.11

⇒ Refer to Appendix B3 for SSE, residuals, and error variance calculations

2.3.2 Standard Errors and t-tests

- SE $(\widehat{\beta_1})$ = 0.54
- SE $(\widehat{\beta_0})$ = 28.05

Test for the slope: (i) H_0: $\beta_1 = 0$ vs. (ii) H_a: $\beta_1 \neq 0$

$t_{\beta 1} = \widehat{\beta_1} / SE(\widehat{\beta_1}) = \frac{11.15}{0.54} = 20.63$

Test for intercept: (i) H_0: $\beta_0 = 0$ vs. (ii) H_a: $\beta_0 \neq 0$

$t_{\beta 0} = \frac{-191.82}{28.05} = -6.84$

With 8 degrees of freedom, both p-values are far below 0.001, so both coefficients are statistically significant.

⇒ Refer to Appendix B3 for the derivation of standard errors of the coefficients

2.3.3 Coefficient of Determination (R^2)

- Total sum of squares:

SST = $\sum(Y_i - \bar{Y})^2$ = 211819.50

$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{3909.19}{211819.50} = 0.9815$

⇒ Refer to Appendix B3 for the SST calculations and R^2 determination

Interpretation:

- The slope is highly significant, confirming a strong linear effect of temperature on conductivity.
- $R^2 = 0.9815$ means about 98.15% of the variability in conductivity is explained by temperature, indicating an excellent fit within the observed range.

2.4 Model Evaluation and Assumptions:

2.4.1 Diagnostic Plots

To evaluate the assumptions:

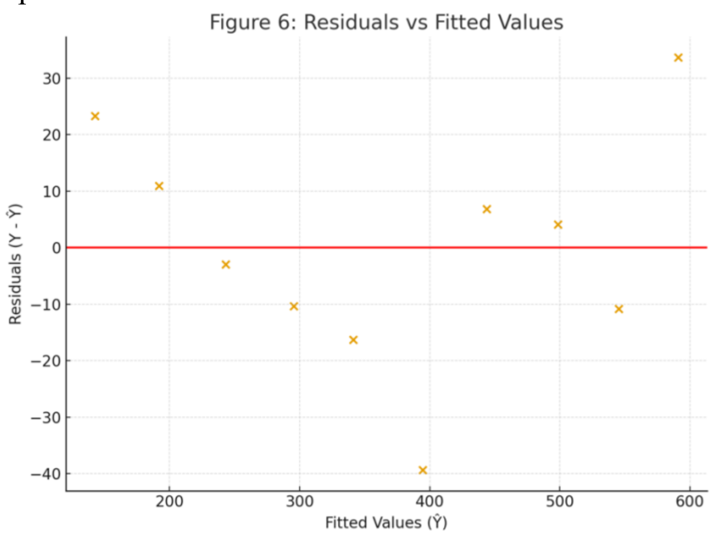

Figure 7: Residuals vs Fitted Values Plot

This diagnostic plot assesses assumptions of linearity and constant variance. Residuals are scattered randomly around zero with no clear trend, pattern, or funnel shape, suggesting that:

- The linear relationship is appropriate and the variance of the residuals is approximately constant (no heteroscedasticity). The plot validates the use of standard least-squares inference.

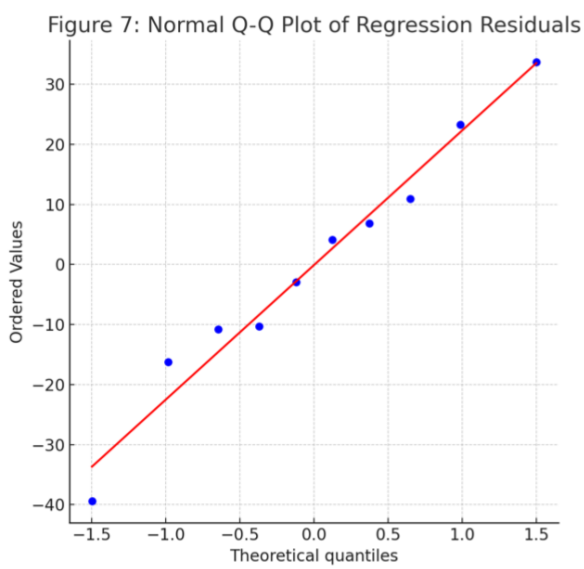

Figure 8: Normal Q-Q Plot of Residuals

This plot compares the regression residuals to a theoretical normal distribution. The points fall reasonably close to the straight reference line, with only mild deviations at the extreme, typical for small samples (n = 10). This indicates that the residual are approximately normally distributed, supporting the validity of the t-tests, confidence intervals, and prediction intervals.

Observations:

- Residuals appear randomly scattered around zero with no obvious pattern, supporting linearity and constant variance.
- The Q-Q plot shows residuals close to the straight line, suggesting approximate normality.

Interpretation: (within the range (30 - 70°F)**)**

- The linear model is appropriate. There is no evidence of strong curvature or heteroscedasticity.
- The usual OLS-based confidence intervals are valid for interpolation within this range.

2.5 Prediction and Confidence:

2.5.1 Point Prediction at -10°F

Using the fitted equation:

$$\hat{Y}(-10) = -191.82 + 11.15(-10) = -303.30$$

The model predicts negative thermal conductivity, which is physically impossible, immediately suggesting that extrapolation is unsafe.

⇒ **Refer to Appendix B4 for the full computation of the point prediction at -10°F**

2.5.2 Confidence Interval for Mean Response at -10°F

Standard error for the mean at $x_0 = -10$:

$$SE_{mean}(x_0) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}}} = 33.32 \text{ with } t_{0.975,8} = 2.306$$

The margin of error is:

$ME_{mean} = 2.306 \times 33.32 = 76.83$

So the 95% CI for the mean conductivity at -10°F is: [-380.13, -226.48] → Entirely Negative

⇒ **Refer to Appendix B4 for the full computation of Confidence Interval at -10°F**

2.5.2 95% Prediction Interval for a New Observation at -10°F

Standard error for a single new observation:

$$SE_{pred}(x_0) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}}} = 39.98$$

The margin of error is:

$ME_{pred} = 2.306 \times 39.98 = 92.20$

95% prediction interval: [-395.50, -211.10] → Again, the Entire Interval is Negative

⇒ **Refer to Appendix B4 for the full computation of Prediction Interval at -10°F**

Interpretation:
Both the point prediction and the intervals at -10°F are physically impossible, proving that:

- Although the model fits very well between 30-70°F, it should not be used to extrapolate far beyond that range, especially into regions where physics dictate that conductivity must be non-negative.

Closer to the sample mean (around 50°F), the squared distance term $(x_0 - \bar{X})^2/S_{XX}$ would be small. Intervals would be much narrower and would produce realistic, positive predictions.

2.6 Open-Ended Reflection:

2.6.1 Outliers and Influential Points

In small-sample regression, outliers and high-leverage points can strongly affect the slope and intercept.

- Outliers: Observations with large residuals (far from the line vertically)
- Influential points: Observations whose removal would substantially change the fitted line

For this dataset, the residual analysis did not flag any extreme outliers. However, in a full analysis one would compute:

- Leverage values
- Studentized residuals and
- Cook's distance

To objectively assess influence. If problematic points exist, options include verifying measurement accuracy or using robust regression methods.

2.6.2 Extrapolation and Physical Constraints

The -10°F prediction demonstrates that we must collect more data near the low-temperature region of interest. It also suggests fitting a local model just around that region or fit a nonlinear model (power or exponential relationship) that respects the physical constraint that conductivity remains non-negative and may approach zero as temperature decreases.

Appendix – Sample Calculations

Appendix A

A1-Large Group Descriptive Statistics (Water Resistance)

A1.1-Raw Data and Mean

Values: 35.59, 37.12, 28.22, 31.97, 29.53, 21.01, 35.45, 27.81, 22.77, 30.39, 29.26, 27.52, 25.37, 31.11, 25.94, 29.53, 22.81, 33.82, 28.76, 42.30, 42.60, 35.28, 23.29, 35.81, 36.10, 26.10, 38.04, 28.61, 30.53, 19.11
Sum: 911.75
Mean: 911.75/30 = 30.3917 = 30.39

A1.2-Deviations and Squared Deviations

For each observation x_i, compute $d_i = x_i - \bar{x}$ and d_i^2

i	xi	di = xi – 30.3917	di²
1	35.59	5.1983	27.022323
2	37.12	6.7283	45.270021
3	28.22	-2.1717	4.716281
4	31.97	1.5783	2.491031
5	29.53	-0.8617	0.742527
6	21.01	-9.3817	88.016295
7	35.45	5.0583	25.586399
8	27.81	-2.5817	6.665175
9	22.77	-7.6217	58.090311
10	30.39	-0.0017	0.000003
11	29.26	-1.1317	1.280745
12	27.52	-2.8717	8.246661
13	25.37	-5.0217	25.217471
14	31.11	0.7183	0.515955
15	25.94	-4.4517	19.817633
16	29.53	-0.8617	0.742527
17	22.81	-7.5817	57.482175
18	33.82	3.4283	11.753241
19	28.76	-1.6317	2.662445
20	42.30	11.9083	141.807609
21	42.60	12.2083	149.042589
22	35.28	4.8883	23.895477
23	23.29	-7.1017	50.434143
24	35.81	5.4183	29.357975
25	36.10	5.7083	32.584689
26	26.10	-4.2917	18.418689
27	38.04	7.6483	58.496493
28	28.61	-1.7817	3.174455
29	30.53	0.1383	0.019127
30	19.11	-11.2817	127.276755

Sum of squared deviations: $\sum_{i=1}^{30}(x_i - \bar{x})^2$= 1020.83
Sample variance: s^2 = 1020.83/30-1 = 35.17
Standard deviation: s = $\sqrt{s}$ = $\sqrt{35.17}$ = 5.93

A1.3-Median and Quartiles

Ordered Data (from smallest to largest): 19.11, 21.01, 22.77, 22.81, 23.29, 25.37, 25.94, 26.10, 27.52, 27.81, 28.22, 28.61, 28.76, 29.26, 29.53, 29.53, 30.39, 30.53, 31.11, 31.97, 33.82, 35.28, 35.45, 35.59, 35.81, 36.10, 37.12, 38.04, 42.30, 42.60

Median: Average of 15th and 16th values: $\tilde{x} = \frac{29.53+29.52}{2}$ = 29.53

Q1: Lower Half: median of the first 15 values = 26.10

Q3: Upper Half: median of the last 15 values = 35.45

A2-Subgroup Means and Standard Deviations

Data organized into Subgroups (n = 3 observations each):

1. (35.59, 37.12, 28.22)
2. (31.97, 29.53, 21.01)
3. (35.45, 27.81, 22.77)
4. (30.39, 29.26, 27.52)
5. (25.37, 31.11, 25.94)
6. (29.53, 22.81, 33.82)
7. (28.76, 42.30, 42.60)
8. (35.28, 23.29, 35.81)
9. (36.10, 26.10, 38.04)
10. (28.61, 30.53, 19.11)

For each subgroup: $\bar{x_i} = \frac{x_{i1} + x_{i2} + x_{i3}}{3}$, $s_i = \sqrt{\frac{1}{3-1} \sum_{j=1}^{3} (x_{ij} - \bar{x_i})^2}$

Calculation for Subgroup 1: (35.59, 37.12, 28.22)

- Subgroup Mean ($\bar{x_1}$): $\frac{35.59+37.12+28.22}{3} = \frac{100.93}{3}$ = 33.64
- Deviation:
  d_{11} = 35.59 - 33.64 = 1.9467
  d_{12} = 37.12 - 33.64 = 3.4767
  d_{13} = 28.22 - 33.64 = -5.4233
- Squared Deviation:
  d_{11}^2= (1.9467)^2 = 3.79
  d_{12}^2= (3.4767)^2 = 12.09
  d_{13}^2= (-5.4233)^2 = 29.41
- Sum of Squared Deviation: 3.79 + 12.09 + 29.41 = 45.29
- Sample SD (n = 3): $s_1 = \sqrt{\frac{45.29}{3-1}} = \sqrt{22.645}$ = 4.76

Using the same steps for all other subgroups, we get the following table.

Groups	Values	Mean ($\bar{x_i}$)	Standard Deviation (s_i)
1	35.59, 37.12, 28.22	33.64	4.76
2	31.97, 29.53, 21.01	27.50	5.75
3	35.45, 27.81, 22.77	28.68	6.38
4	30.39, 29.26, 27.52	29.06	1.45
5	25.37, 31.11, 25.94	27.47	3.16

6	29.53, 22.81, 33.82	28.72	5.55
7	28.76, 42.30, 42.60	37.89	7.91
8	35.28, 23.29, 35.81	31.46	7.08
9	36.10, 26.10, 38.04	33.41	6.41
10	28.61, 30.53, 19.11	26.08	6.11

A3-Grand Mean and Average Subgroup SD

Sum of subgroup means:
$\sum_{i=1}^{10} \overline{x_i}$ = 33.64 + 27.50 + 28.68 + 29.06 + 27.47 + 28.72 + 37.89 + 31.46 + 33.41 + 26.08 = 303.92

Grand mean: $\bar{\bar{x}} = \frac{303.92}{10}$ = 30.392 = 30.39

Sum of subgroup SDs:
$\sum_{i=1}^{10} s_i$ = 4.76 + 5.75 + 6.38 + 1.45 + 3.16 + 5.55 + 7.91 + 7.08 + 6.41 + 6.11 = 54.56

Average SD: $\bar{s} = \frac{54.56}{10}$ = 5.456 = 5.46

A4- $\overline{X}$-S Control and Warning Limits

A4.1- $\overline{X}$-Chart Limits

Using n = 3 and constants A_3 = 1.954:
Centre: $CL_{\bar{x}} = \bar{x} = 30.39$
Upper Control Limit: $UCL_{\bar{x}} = \bar{x} + A_3\bar{s}$ = 30.39 + 1.954(5.46) = 30.39 + 10.668 = 41.05
Lower Control Limit: $LCL_{\bar{x}} = \bar{x} - A_3\bar{s}$ = 30.39 - 1.954(5.46) = 30.39 - 10.668 = 19.73

A4.2- S-Chart Limits

Using n = 3 and constants B_3 = 0, B_4 = 2.569:
Centre: $CL_s = \bar{s}$ = 5.46
Upper Control Limit: $UCL_s = B_4\bar{s}$ = 2.568(5.46) = 14.01
Lower Control Limit: $LCL_s = B_3\bar{s}$ = 0

A3.4-Warning Limits and Sigma Zones

$\overline{X}$-Chart:
3 sigma: CL to UCL is: $UCL_{\bar{x}} - CL_{\bar{x}}$ = 41.05 - 30.39 = 10.66
Sigma: $\frac{10.66}{3}$ = 3.554
2 sigma warning limits:
Upper Limit: $UWL_{\bar{x}} = CL_{\bar{x}} + 2\sigma$ = 30.39 + 7.11 = 37.50
Lower Limit: $LWL_{\bar{x}} = CL_{\bar{x}} - 2\sigma$ = 23.28

S-Chart:
3 sigma: CL to UCL: $UCL_s - CL_s$ = 14.01 - 5.46 = 8.55
Sigma: $\frac{8.55}{3}$ = 2.85

2 sigma warning limits:
Upper Limit: $UWL_s = CL_s + 2\ \sigma$ = 5.46 + 2(2.85) = 11.16
Lower Limit: $LWL_s = CL_s - 2\ \sigma$ = 5.46 = 2(2.85) = -0.248, thus → 0 (truncated to 0)

Appendix B

B1-Regression Data and Summaries

- Sample Size (n): 10
- Sum of Temperature ($\sum x$): 30.0 + 34.4 + 39.0 + 43.7 + 47.8 + 52.6 + 57.0 + 61.9 + 66.1 + 70.2 = 502.7
- Sum of Conductivity ($\sum y$): 166.0 + 202.7 + 240.1 + 285.1 + 324.9 + 355.3 + 450.6 + 502.5 + 534.4 + 624.6 = 3,686.2
- Mean Temperature ($\bar{X}$): $\bar{X} = \frac{\sum X}{n}$ = 502.7 /10 = 50.27 °F
- Mean Conductivity ($\bar{Y}$): $\bar{Y} = \frac{\sum Y}{n}$ = 3.686.2 /10 = 386.62 units

B2-Sums of Squares and Cross-Products

Define: $X_i = X_i - \bar{X}$ and $Y_i = Y_i - \bar{Y}$
Then compute $(X_i)^2$ and X_iY_i

X_i	Y_i	$X_i = X_i - 50.27$	$Y_i = Y_i - 368.62$	$(X_i)^2$	X_iY_i
30.0	166.0	-20.27	-202.62	410.8729	4107.1074
34.4	202.7	-15.87	-165.92	251.8569	2633.1504
39.0	240.1	-11.27	-128.52	126.8809	1448.2880
43.7	285.1	-6.57	-83.52	43.1649	548.7062
47.8	324.9	-2.47	-43.72	6.1009	108.0084
52.6	355.3	2.33	-13.32	5.4289	-31.0356
57.0	450.6	6.73	81.98	45.2929	551.7154
61.9	502.5	11.63	133.88	135.2569	1557.0244
66.1	534.4	15.83	165.78	250.5889	2624.2974
70.2	624.6	19.93	255.98	397.2049	5101.6814

Now sum the last two columns:

- $S_{XX} = \sum X_i^2$ = 410.8729 + 251.8569 + 126.8809 + 43.1649 + 6.1009 + 5.4289 + 45.2929 + 135.2569 250.5889 + 397.2049 = 1672.781
- $S_{XY} = \sum X_i\, Y_i$ = 4107.1074 + 2633.1504 + 1448.2880 + 548.7062 + 108.0084 - 31.0356 + 551.7154 + 1557.0244 + 2624.2974 + 5101.6814 = 18649.086

B3-Regression Coefficients, SSE and Standard Errors

B3.1-Regression Coefficients

- Slope: $\widehat{\beta_1}$ = SXX / SXY = $\frac{18649.09}{1672.78}$ = 11.148 = 11.15
- Intercept: $\widehat{\beta_0}$ = $\bar{Y}$ - $\widehat{\beta_1}\bar{X}$ = 368.62 - 11.15(50.27) = 368.62 - 560.4377 = =191.8177 = -191.82
- Fitted Line: $\hat{Y}$ = -191.82 + 11.15X

B3.2-Fitted Values, Residuals, and SSE

For each i: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}X_i$ and $e_i = Y_i - \widehat{Y_i}$

X_i	Y_i	$\widehat{Y_i}$	$e_i = Y_i - \widehat{Y_i}$	e_i^2
30.0	166.0	142.64	23.36	545.74
34.4	202.7	191.69	11.01	121.17
39.0	240.1	242.98	-2.88	8.27
43.7	285.1	295.37	-10.27	105.56
47.8	324.9	341.08	-16.18	261.89
52.6	355.3	394.60	-39.30	1544.19
57.0	450.6	443.65	6.95	48.31
61.9	502.5	498.28	4.22	17.83
66.1	534.4	545.10	-10.70	114.52
70.2	624.6	590.81	33.79	1141.72

Sum of squared residuals: SSE = $\sum e_i^2$ = 545.74 + 121.17 + 8.27 + 105.56 + 261.89 + 1544.19 + 48.31 + 17.83 + 114.52 + 1141.72 = 3909.1906

B3.3-Error Variance, Residual Standard Error

- Degrees of freedom: n - 2 = 8
- $s^2 = \frac{SSE}{n-2} = \frac{3909.1906}{8}$ = 488.6488
- s = $\sqrt{488.6488}$ = 22.1054

B3.4-Standard Errors of β_0 and β_1

- Slope: SE ($\widehat{\beta_1}$) = $\frac{s}{\sqrt{S_{XX}}} = \frac{22.1504}{\sqrt{1672.781}} = \frac{22.1504}{40.894}$ = 0.5405 = 0.54
- Intercept: SE ($\widehat{\beta_0}$) = s$\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$

  Compute inside the square root:

  $\frac{1}{n}$ = 0.1, $\frac{\bar{X}^2}{S_{XX}} = \frac{50.27^2}{1672.781} = \frac{2527.27}{1672.781}$ = 1.5087 → 0.1 + 1.5087 = 1.6087

  SE ($\widehat{\beta_0}$) = 22.1054$\sqrt{1.6087}$ = 22.1054(1.297) = 28.0547 = 28.05

B3.5-T-statistics and R^2

- Slope t-statistic: $t_{\beta 1} = \widehat{\beta_1} / \text{SE}(\widehat{\beta_1}) = \frac{11.15}{0.54} = 20.63$
- Intercept t-statistic: $t_{\beta 0} = \widehat{\beta_0} / \text{SE}(\widehat{\beta_0}) = \frac{-191.8177}{28.05} = -6.84$
- Total sum of squares: $\text{SST} = \sum(Y_i - \bar{Y})^2 = 211819.496$
- Coefficient of determination: $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{3909.1906}{211819.496} = 0.9815$

B4-Confidence and Prediction Intervals at -10°F

Take $x_0 = -10$

B4.1-Point Prediction

$\widehat{Y}(x_0) = \widehat{\beta_0} + \widehat{\beta_1}$

$x_0 = -191.8177 + 11.15(-10) = -181.8177 - 111.50 = -303.30$

B4.2-Standard Error of Mean Response

$\text{SE}_{mean}(x_0) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}}}$

Compute:

$x_0 - \bar{X} = -10 - 50.27 = -60.27$

$(x_0 - \bar{X})^2 = 60.27^2 = 3632.44$

$\frac{(x_0 - \bar{X})^2}{S_{XX}} = \frac{3632.44}{1672.781} = 2.1715 = 2.1715$

$\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} = 0.1 + 2.1715 = 2.2715$

Thus, $\text{SE}_{mean} = 22.1054\sqrt{2.2715} = 22.1054(1.507) = 33.3163 = 33.32$

B4.3-95% Confidence Interval for Mean

Using $t_{0.975,8} = 2.306$

$\text{ME}_{mean} = t_{0.975,8}(\text{SE}_{mean}) = 2.306 \times 33.3163 = 76.8273 = 76.83$

So, $\widehat{Y}(x_0) \pm \text{ME}_{mean} = -303.3037 \pm 76.8273$

Lower Bound: $-303.3037 - 76.8273 = -380.1310$

Upper Bound: $-303.3037 + 76.8273 = -226.4764$

Confidence Interval: [-380.1310, -226.4764]

B4.4-95% Standard Error and 95% Prediction Interval

For a new observation at x_0:

$SE_{pred}(x_0) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}}}$

We already have $\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}} = 2.2715 \rightarrow 1 + 2.2715 = 3.2715$

So, $SE_{pred} = 22.1054\sqrt{3.2715} = 22.1054(1.808) = 39.9828 = 39.98$

- Margin of error for prediction:
  $ME_{pred} = t_{0.975,8}\,(SE_{pred}) = 2.306 \times 39.9828 = 92.2003 = 92.20$
- $\hat{Y}\,(x_0) \pm ME_{pred} = -303.3037 \pm 92.2003$
  Lower Bound: -303.3037 - 92.2003 = -395.5040
  Upper Bound: -303.3037 + 92.2003 = -211.1034
  Prediction Interval: [-395.50, -211.10]

B5-Residual Summary Table

X	Y	Ŷ	Residual (Y−Ŷ)
30.0	166.0	142.68	**+23.32**
34.4	202.7	191.74	**+10.96**
39.0	240.1	243.03	**−2.93**
43.7	285.1	295.44	**−10.34**
47.8	324.9	341.15	**−16.25**
52.6	355.3	394.67	**−39.37**
57.0	450.6	443.73	**+6.87**
61.9	502.5	498.37	**+4.14**
66.1	534.4	545.20	**−10.80**
70.2	624.6	590.91	**+33.69**