Property Valuation data:

**Interpretation of the Regression Output Table**

The table provides the summary of a **multiple linear regression model** with an interaction term (x2:x3). Here’s a detailed breakdown:

**1. Coefficients & Their Significance**

| **Term** | **Estimate** | **Std. Error** | **t-value** | **p-value** | **Interpretation** |
| --- | --- | --- | --- | --- | --- |
| **(Intercept)** | 48.0895 | 16.1081 | 2.985 | **0.0114**\* | The expected value of the response when x2 = 0 and x3 = 0 is **48.09**. This is statistically significant at the 5% level. |
| **x2** | -1.2561 | 2.2327 | -0.563 | 0.5841 | For a one-unit increase in x2 (holding x3 constant), the response decreases by **1.2561**, but this effect is **not statistically significant** (p > 0.05). |
| **x3** | -0.5461 | 0.3467 | -1.575 | 0.1412 | For a one-unit increase in x3 (holding x2 constant), the response decreases by **0.5461**, but this effect is **not statistically significant** (p > 0.05). |
| **x2:x3 (Interaction)** | 0.0686 | 0.0497 | 1.380 | 0.1927 | The interaction term suggests that the effect of x2 on the response depends on x3 (and vice versa), but this interaction is **not statistically significant** (p > 0.05). |

**Key Takeaways:**

* Only the **intercept** is statistically significant (p < 0.05).
* Neither x2**nor**x3 alone has a significant effect on the response.
* The **interaction term (**x2:x3**)** is also **not significant**, meaning there’s no strong evidence that the effect of one predictor depends on the other.

**2. Model Fit Statistics**

| **Statistic** | **Value** | **Interpretation** |
| --- | --- | --- |
| **Residual Standard Error (RSE)** | 5.06 | The average deviation of observed values from the predicted values is **5.06 units**. Lower values indicate better fit. |
| **Multiple R²** | 0.5458 | **54.58%** of the variation in the response is explained by the model. |
| **Adjusted R²** | 0.4322 | After penalizing for extra predictors, **43.22%** of the variation is explained. |
| **F-statistic (p-value)** | 4.806 (0.02011) | The overall model is **statistically significant** (p < 0.05), meaning at least one predictor has a significant effect. |

**Key Takeaways:**

* The model explains **moderate variation (R² = 54.58%)**, but the **Adjusted R² (43.22%)** is lower, suggesting some predictors may not be useful.
* The **F-test (p = 0.02011)** confirms that the model is better than a null model (with no predictors).

**3. Overall Conclusions**

1. **The intercept is significant**, but **none of the predictors (**x2**,**x3**, or their interaction) are statistically significant** at the 5% level.
2. **The model is significant overall (F-test p = 0.02011)**, but individual predictors do not contribute significantly.
   * This could mean **multicollinearity** (high correlation between predictors) or **weak effects** of x2 and x3.
3. **The interaction term (**x2:x3**) is not significant**, meaning the effect of x2 does not meaningfully change with x3 (and vice versa).

**1. Simple Linear Model (**y ~ x2 + x3**)**

model = lm(y ~ x2 + x3, data=df)

summary(model)

**Key Results:**

| **Term** | **Estimate** | **Std. Error** | **t-value** | **p-value** | **Interpretation** |
| --- | --- | --- | --- | --- | --- |
| Intercept | 27.4899 | 6.2697 | 4.385 | 0.0007\*\*\* | Significant baseline value when x2=0, x3=0. |
| x2 | 1.7167 | 0.6086 | 2.821 | 0.0144\* | Significant positive effect on y. |
| x3 | -0.0848 | 0.0955 | -0.888 | 0.3907 | No significant effect. |

**Model Fit:**

* **R² = 0.4736**: Predictors explain **47.36%** of variance in y.
* **Adjusted R² = 0.3927**: Penalizes for non-significant predictors.
* **F-statistic (p=0.01543)**: Model is **significantly better** than null model.

**ANOVA Comparison (vs. Null Model):**

anova(null\_model, model)

* **p=0.01543**: Confirms the full model is statistically significant.

**2. Interaction Model (**y ~ x2 + x3 + x2:x3**)**

model\_int = lm(y ~ x2 + x3 + x2:x3, data=df)

summary(model\_int)

**Key Results:**

| **Term** | **Estimate** | **Std. Error** | **t-value** | **p-value** | **Interpretation** |
| --- | --- | --- | --- | --- | --- |
| Intercept | 48.0895 | 16.1081 | 2.985 | 0.0114\* | Significant baseline. |
| x2 | -1.2561 | 2.2327 | -0.563 | 0.5841 | No standalone effect. |
| x3 | -0.5461 | 0.3467 | -1.575 | 0.1412 | No standalone effect. |
| x2:x3 | 0.0686 | 0.0497 | 1.380 | 0.1927 | Interaction is **not significant**. |

**Model Fit:**

* **R² = 0.5458**: Slightly better fit than simple model.
* **Adjusted R² = 0.4322**: Lower due to extra non-significant term.
* **F-statistic (p=0.02011)**: Model is still significant overall.

**3. Prediction for New Data (**x2=10, x3=50**)**

predict(model, new.data, interval="confidence", level=0.90)

* **Fit = 40.415**: Predicted y value.
* **90% CI = [34.847, 45.983]**: Confidence interval for the **mean response**.

predict(model, new.data, interval="prediction", level=0.90)

* **90% PI = [29.604, 51.227]**: Wider interval for **individual predictions**.

**4. Standardized Coefficients**

beta\_x2 = coef(model)["x2"] \* (sd(df$x2)/sd(df$y)) # 0.78

beta\_x3 = coef(model)["x3"] \* (sd(df$x3)/sd(df$y)) # -0.29

* x2**has a stronger effect** (0.78 SD change in y per SD change in x2).
* x3**has a weak negative effect** (-0.29 SD).

**Key Takeaways & Recommendations**

1. **Simple Model (**y ~ x2 + x3**)**:
   * x2**is significant**, but x3 is not.
   * **Model is useful** (F-test p=0.015).
2. **Interaction Model**:
   * **Interaction term (**x2:x3**) is not significant** (p=0.19).
   * **Avoid including it** (Adjusted R² drops to 0.4322).
3. **Predictions**:
   * At x2=10, x3=50, expect y ≈ 40.4 (90% CI: 34.8–46.0).
4. **Next Steps**:
   * **Drop**x3 (non-significant) and refit: y ~ x2.
   * **Check for outliers** (residual analysis).
   * **Validate assumptions** (normality, homoscedasticity).

**Final Decision**

* **Best Model**: y ~ x2 (simpler, more interpretable).
* **Avoid Interaction**: No evidence it improves the model.

Logistic Regression:

1. (Intercept): -10.5

This is the baseline log-odds of loan approval when all predictor variables are 0.

It’s statistically significant (p = 0.043), but it's not meaningful alone because 0 values for variables like income or credit score are unrealistic.

2. Income: 0.0015

Interpretation: For every 1-unit increase in income, the log-odds of approval increase by 0.0015.

Odds Ratio = exp(0.0015) ≈ 1.0015

So, a higher income slightly increases the chance of approval.

Not statistically significant (p = 0.133)

3. CreditScore: 0.0200

Interpretation: A 1-point increase in credit score increases the log-odds of approval by 0.02.

Odds Ratio = exp(0.02) ≈ 1.0202

Each extra credit score point increases approval odds by about 2%.

Statistically significant (p = 0.045)

This is an important predictor.

4. LoanAmount: -0.0002

Interpretation: Each additional 1 unit in loan amount slightly decreases the log-odds of approval.

Odds Ratio = exp(-0.0002) ≈ 0.9998

So, bigger loans reduce approval chances slightly.

Statistically significant (p = 0.046)

5. Married\_dummy: 1.5

Interpretation: Being married increases the log-odds of approval by 1.5 units compared to unmarried.

Odds Ratio = exp(1.5) ≈ 4.48

Married applicants are over 4 times more likely to be approved.

Marginally significant (p = 0.060) — just above the 0.05 threshold

**1. Deviance Analysis**

**Null Deviance (13.87, df=9)**

* Measures how well the response variable (Failure) is predicted by a model with **only an intercept** (no predictors).
* Higher null deviance indicates poorer fit of the null model.
* **Degrees of freedom (df) = 9**: Number of observations minus 1 (n - 1).

**Residual Deviance (5.42, df=5)**

* Measures how well the response is predicted by the **current model** (with predictors like Temperature).
* **Significant drop** from null deviance (13.87 → 5.42) suggests the model explains much of the variability.
* **df = 5**: Number of observations minus the number of parameters estimated (n - p - 1, where p = predictors).

**Key Takeaway**:

* The **reduction in deviance (13.87 - 5.42 = 8.45)** indicates the model improves prediction significantly compared to the null model.

**2. AIC (Akaike Information Criterion) = 15.42**

* Balances **model fit** (lower deviance) and **complexity** (number of parameters).
* **Lower AIC = Better model**. For small datasets, AIC < 20 is often reasonable.
* Useful for comparing models (e.g., if you add/remove predictors).

**Rule of Thumb**:

* Compare with the null model’s AIC (not provided here). If the current AIC is lower, the model is preferable.

**3. Statistical Significance of Deviance Change**

To formally test if the model improvement is significant:

p\_value <- 1 - pchisq(null\_deviance - residual\_deviance, df=9-5)

* If p\_value < 0.05, the model is significantly better than the null.

✅ Conclusion:

Significant predictors: CreditScore, LoanAmount, (Intercept)

Marginal predictor: Married\_dummy

Non-significant: Income

Model appears to fit the small dataset quite well, but more data would improve reliability.

**Water flow data:**

**Comprehensive Interpretation of the Linear Regression Analysis**

**1. Model Summary**

summary(model)

**Output**:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 23.0819 2.4827 9.297 4.63e-06 \*\*\*

x -0.0376 0.0818 -0.460 0.655

---

Significance codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.631 on 11 degrees of freedom

Multiple R-squared: 0.01876, Adjusted R-squared: -0.07043

F-statistic: 0.2104 on 1 and 11 DF, p-value: 0.6549

**Interpretation**:

* **Relationship**: The slope for x (−0.0376) suggests a negligible negative trend, but it is **not statistically significant** (p = 0.655).
* **Model Fit**:
  + **R² = 0.01876**: Only 1.8% of the variance in y is explained by x.
  + **Adjusted R² = −0.07043**: Penalizes for useless predictors; negative value indicates the model is worse than using the mean.
  + **F-statistic (p = 0.6549)**: The model is **not significantly better** than an intercept-only model.

**Conclusion**: No meaningful linear relationship exists between x and y.

**2. Residual Diagnostics (Q-Q Plots)**

**Plots Generated**:

1. **Raw Residuals**: Slight deviation at the upper tail (potential outlier).
2. **Studentized Residuals**: Clearly flags **observation 4** (x = 77.6) as an outlier (point far from the red line).

**Interpretation**:

* **Normality**: Residuals are approximately normal except for observation 4.
* **Outlier**: Observation 4 has an extreme residual value.

**3. Outlier Detection**

outlier = which(abs(stu.res) > 2)

cat("Outliers at:", toString(outlier))

**Output**:

Outliers at: 4

**Interpretation**:

* **Observation 4** (x = 77.6, y = 15.7) is an outlier (|studentized residual| > 2).

**4. Leverage Analysis**

leverage = hatvalues(model)

high.lev = which(leverage > 2 \* mean(leverage))

cat("High Leverage at:", toString(high.lev))

**Output**:

High Leverage at: 4

**Interpretation**:

* **Observation 4** has **high leverage** (unusually large x value), meaning it disproportionately influences the regression line.

**5. Influence Metrics**

**Cook’s Distance**:

inf.cooks = which(cooks.distance(model) > 4/nrow(df))

cat("Influential (Cook's) at:", toString(inf.cooks))

**Output**:

Influential (Cook's) at: 4

**DFFITS**:

inf.dffits = which(abs(dffits(model)) > 2 \* sqrt(mean(leverage)))

cat("Influential (DFFITS) at:", toString(inf.dffits))

**Output**:

Influential (DFFITS) at: 4

**Interpretation**:

* **Observation 4** is **highly influential** by both metrics. Removing it would significantly change the regression results.

**6. Model After Outlier Removal**

df.clean = df[-outlier, ]

model.clean = lm(y ~ x, data = df.clean)

summary(model.clean)

**Output**:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 22.8926 2.1795 10.504 1.24e-06 \*\*\*

x 0.0174 0.0716 0.243 0.812

---

R-squared: 0.005918

**Interpretation**:

* **No Improvement**: Even after removing the outlier:
  + Slope for x remains **insignificant** (p = 0.812).
  + **R² worsened** (0.0059 vs. 0.01876), indicating the outlier was not the sole issue.

**Key Takeaways**

1. **No Linear Relationship**: x does not predict y effectively (p > 0.05, R² ≈ 0).
2. **Observation 4 Issues**:
   * **Outlier**: Extreme residual value.
   * **High Leverage**: Extreme x value.
   * **Influential**: Drives regression results.
3. **Model Unreliable**: Even after outlier removal, the model fails to explain meaningful variance.

**Recommendations**

1. **Investigate Observation 4**: Verify if it’s a data entry error or genuine extreme value.
2. **Explore Alternatives**:
   * **Nonlinear Models** (e.g., polynomial regression).
   * **Additional Predictors** (if available).
3. **Conclusion**: Do not use this linear model for prediction.

**Final Note**: The analysis suggests the dataset may not suit linear regression. Consider exploratory analysis (e.g., scatterplots) to identify better modeling approaches.

**Normality Test:**

**1. Kolmogorov-Smirnov (KS) Test**

**Purpose**:

Tests if a sample follows a **specified distribution** (e.g., normal) by comparing empirical and theoretical CDFs.

**Hypotheses**:

* **H₀**: Data is normally distributed.
* **H₁**: Data is **not** normally distributed.

**R Code**:

r

Copy

Download

ks\_result <- ks.test(data, "pnorm", mean = mean(data), sd = sd(data))

print(ks\_result)

**Interpretation**:

* **Test Statistic (D)**: Maximum vertical distance between empirical and theoretical CDFs.
  + **Smaller D** → Closer to normal.
* **p-value**:
  + **p < 0.05**: Reject H₀ (non-normal).
  + **p ≥ 0.05**: Fail to reject H₀ (normal).

**Limitations**:

* Less powerful than Shapiro-Wilk for small samples.
* Sensitive to outliers.

**2. Shapiro-Wilk Test**

**Purpose**:

Specifically designed for **normality testing** (more reliable than KS for small samples).

**Hypotheses**:

* **H₀**: Data is normally distributed.
* **H₁**: Data is **not** normally distributed.

**R Code**:

r

Copy

Download

shapiro\_result <- shapiro.test(data)

print(shapiro\_result)

**Interpretation**:

* **Test Statistic (W)**: Closer to **1** indicates normality.
* **p-value**:
  + **p < 0.05**: Reject H₀ (non-normal).
  + **p ≥ 0.05**: Fail to reject H₀ (normal).

**Strengths**:

* More accurate for small samples (n < 50).
* Robust against non-normality in tails.