**Problem 1: Statistical Inference**

1. Summary of the model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coefficient | Standard Error | Test statistic | P-value |
| Intercept | 7.8243876 | 1.1298792 | 6.925 | 1.81e-11 |
| CompPrice | 0.0942545 | 0.0078637 | 11.986 | < 2e-16 |
| Income | 0.0130501 | 0.0034908 | 3.738 | 0.000213 |
| Advertising | 0.1369399 | 0.0210394 | 6.509 | 2.33e-10 |
| Population | -0.0002007 | 0.0007010 | -0.286 | 0.774780 |
| Price | -0.0924395 | 0.0050621 | -18.261 | < 2e-16 |
| Age | -0.0447919 | 0.0060226 | -7.437 | 6.59e-13 |
| Education | -0.0423034 | 0.0373738 | -1.132 | 0.258371 |
| UrbanYes | -0.1559036 | 0.2133519 | -0.731 | 0.465380 |
| USYes | -0.1062926 | 0.2832192 | -0.375 | 0.707640 |

Hypothesis test for “Price” coefficient:

H­0: the coefficient of Price is zero

H1: the coefficient of Price is not zero

Test statistic = -18.261

Null distribution: distribution of the test statistic for Price under the assumption that Price is 0

P-value = <2e-16

Conclusion: There’s evidence that Price is significantly different from 0 at a significance level 0.05

1. σ2 = 3.733413. This is an estimate for the variance of the residuals in the multiple linear regression model. It is a measure of how well the multiple linear regression model explains the variation in Sales.
2. The estimated regression coefficient for Advertising (0.1369399) represents the expected change in the response variable (Sales, in thousands) for a one-unit increase in the predictor Advertising budget (in dollars), while holding all other predictors constant.
3. RSS for the full model: 145.031

RSS for the reduced model: 3182.275

1. H0: B1 = B2 = B3 = B4 = B5 = B6 = B7 = B8 = B9

H1: at least one Bj is non-zero

Test-statistic = 51.3752

Null distribution: When εi ~ N(0, σ2) and we assume H0 is true, F\* has a null distribution of F9, 390

p-value = < 2.2e-16

We reject H0 we find evidence of a relationship between Y and at least one of the predictors, at significance level 0.05

1. The estimated sales (f(X)): 15.80707

The confidence interval (95%): [14.91953, 16.69461]

1. The prediction for Y: 15.80707

The prediction Interval (95%): [11.90593, 19.70821]

1. Prediction for Y: -21.16873

The confidence interval (95%): [-24.49992, -17.83755]

When we set the car seat price to $450 while keeping other predictors the same, our model predicts a negative sales value. This result indicates a limitation of our model – it struggles to provide accurate predictions when prices are set far beyond the range of observed data. Linear regression assumes a linear relationship within the data range it was trained on, making extrapolation unreliable. So the negative prediction at $450 for car seats highlights the need for caution when using the model outside its data range and suggests the importance of considering alternative approaches for extreme values.

**Problem 2: The Challenge of Multiple Testing**

1. m \* α type 1 errors expected
2. To calculate the probability P(V ≥ 1), which represents the family-wise error rate, we need to consider the complement of the probability that no type 1 errors occur. P(V ≥ 1) = 1 - P(V = 0).

If we’re controlling the Type 1 error for each null hypothesis at level α, the probability of a Type 1 error for a single test is α. So the probability of no Type 1 errors for a single test is (1-α). Since we are conducting m hypothesis tests, and each test is assumed to be independent, the probability of no Type 1 errors for all m tests is (1 - α)m.

The probability of at least one Type 1 error (P(V ≥ 1)) is:

P(V ≥ 1) = 1 - P(V = 0) = 1 - (1 - α)m

1. An application where the multiple testing problem could be problematic if not addressed is in genomics research. In genomics, scientists often perform thousands of statistical tests to identify genes or genetic variants associated with specific traits or diseases. Without appropriate correction methods for multiple testing, the volume of tests dramatically increases the risk of obtaining false-positive results. This could lead to the erroneous identification of genes as significant contributors to a trait or disease when in fact, they are not, potentially resulting in misleading conclusions and wasted resources. Properly addressing the multiple testing problem is important to ensure validity and reliability of findings in genomics research.
2. Yes, this approach makes statistical sense. With the colleague’s approach, (1) the probability of making a Type 1 error for a single test is α/m because we are dividing the overall significance level by the number of tests and (2) The probability of not making a Type 1 error for a single test is (1 - α/m).

To calculate the probability of not making a Type 1 error for all m tests:

(1 - α/m) \* (1 - α/m) \* ... \* (1 - α/m) = (1 - α/m)m

This expression represents the probability that no Type 1 error occurs for all m tests. To calculate the probability of making at least one Type 1 error, we take the complement:

1 - (1 - α/m)m

To show that FWER ≤ α: 1 - (1 - α/m)m ≤ α  
We can use the limit as m approaches infinity to prove this inequality:

lim (m -> ∞) [1 - (1 - α/m)m]

Using calculus, it can be shown that this limit approaches 1 - e^(-α)

1. The code is in the R script. There were 6 significant predictors using the cutoff.
2. One drawback could be that it can be overly conservative when the number of tests (m) is very large. It can make it difficult to identify truly significant results because the adjusted significance level becomes very small and can lead to a higher likelihood of false negatives.

**Problem 3: Diagnostics for MLR**

1. Relationship between Y and X = (X1, X2, …, Xp) is approximately linear. E(ε) = 0. Var(ε) = σ2. ε's are uncorrelated.
2. False. The key assumption in linear regression is that the errors have a mean of zero and constant variance, but there is no requirement for a specific distributional assumption for εi. However, distributional assumptions can be useful for making inferences and conducting hypothesis tests about the model parameters, but they are not necessary for estimating the coefficients themselves through least squares estimation.
3. True. The lm() function is used for fitting linear regression models, provides p-values for individual hypothesis tests of the form H0: Bj = 0 vs H1: Bj ≠ 0 for each predictor Xj. These hypothesis tests assume that the random error terms are normally distributed. This assumption is important for valid inference using the t-distribution to compute p-values and confidence intervals.
4. Checking every pairwise scatterplot between the response and each predictor can be time-consuming and not sufficient to assess the linearity assumption comprehensively. While examining individual scatterplots can provide insights into the relationship between individual predictors and the response, it may miss interactions and dependencies that exist when multiple predictors are considered together. In complex models, the linearity assumption may hold for combinations of predictors even if it doesn’t hold for individual variables. Visual diagnostics like residual plots or partial residual plots offer a more holistic view by considering the joint effect of all predictors, allowing for a more accurate assessment of linearity and the detection of non-linear patterns and interactions that might be missed when examining scatterplots one by one.
5. In model m2, include both horsepower and its quadratic term as predictor variables. The quadratic terms allow for a quadratic relationship between horsepower and mpg which can help accommodate non-linearity. The Residuals vs Fitted plot has a relatively flat red line indicating that the linearity assumption is met. The quadratic polynomial model m2 is better.

