**What is a treatment effect?**

A treatment effect measures the difference in outcomes between a group that received a treatment and a control group that did not. It quantifies the impact of the treatment on the group, helping to understand its effectiveness.

**How does lasso regression work?**

Lasso regression is a type of linear regression that adds a penalty for including too many predictors in the model. This penalty term leads to some coefficients being exactly zero, effectively selecting a simpler model that includes only the most important predictors. Lasso helps prevent overfitting and improves model interpretability by eliminating irrelevant predictors.

**Assume the significance level for a hypothesis test is = 0.05. Consider the following statement: “If the p-value is 0.001, the conclusion is to reject the null hypothesis.” Is this statement correct? Explain.**

Yes, the statement is correct. The p-value of 0.001 is less than the significance level of 0.05, indicating that the observed data are highly unlikely under the null hypothesis. Therefore, we reject the null hypothesis, suggesting that the observed effect is statistically significant.

**Based on 1988 census data for the 50 States in the United States, the correlation between the number of churches per State and the number of violent crimes per State was 0.85. Consider the following statement: “We can conclude that there is a causal relationship between the number of churches and the number of violent crimes committed in a city.” Is this statement correct? Explain.**

No, the statement is not correct. A high correlation of 0.85 indicates a strong association between the number of churches and violent crimes, but correlation does not imply causation. Without further experimental or observational evidence that controls for other variables, we cannot conclude that there is a causal relationship between the two.

**What is “overfitting”?**

Overfitting happens when your model learns the details and noise in the training data to the extent that it performs poorly on new data. It's like memorizing the answers to a test without understanding the concepts, so when faced with new questions, the performance drops. The model becomes too complex, capturing random fluctuations rather than the underlying trend.

**Consider the following statement: “In a simple linear regression analysis we assume that the errors are independent and normally distributed with variance 0 and constant mean.” Is this statement correct? Explain.**

This statement is partially correct but has a crucial error. We do assume that the errors are independent and normally distributed with a constant mean (usually zero), but the variance should not be 0; instead, it should be constant (homoscedasticity). A variance of 0 would imply no variation in errors, which is unrealistic in real-world data.

**Why does it often make sense to look at the Log of a variable instead of the variable itself?**

Looking at the Log of a variable can help stabilize the variance, make the data more normal distribution-like, and make relationships more linear, especially when dealing with skewed data. It's useful in dealing with data that spans several orders of magnitude, making patterns more discernible and relationships easier to model. Plus, it can convert multiplicative relationships into additive ones, simplifying the analysis.

**How does elasticity relate to the following regression model: log 𝑦 = 𝛼 + 𝛽 log 𝑥 + 𝜀? Explain.**

In this regression model, elasticity represents the percentage change in 𝑦 for a one percent change in 𝑥, captured by the coefficient 𝛽. Because both 𝑦 and 𝑥 are logged, the coefficient 𝛽 directly measures the elasticity between the two variables. This model is particularly useful in economics to understand how responsive an outcome variable (like demand) is to changes in a predictor variable (like price).

**Explain cross-validation.**

Cross-validation is a technique used to assess how well a predictive model will perform on unseen data. By dividing the dataset into a number of smaller, separate parts or "folds," the model is trained on some of these folds while being tested on the remaining ones. This process is repeated multiple times, with different folds used as the training and testing sets each time, allowing the model's performance to be evaluated across different subsets of the data. It helps in identifying how well the model generalizes to new data, reducing the likelihood of overfitting by not relying on a single train-test split. Ultimately, cross-validation provides a more robust estimate of the model's predictive accuracy.

**Are AIC and CV related concepts?**

AIC (Akaike Information Criterion) and cross-validation (CV) are related in the sense that both are methods used for model selection and to evaluate model performance, but they approach the problem from different angles. AIC is a metric that comes from information theory, providing a measure of the relative quality of a statistical model for a given set of data. It balances the model's goodness of fit with the complexity of the model, penalizing overcomplicated models to prevent overfitting. CV, on the other hand, directly assesses a model's predictive performance by using multiple data splits. While both aim to select the best model and avoid overfitting, AIC does so through a formulaic approach based on likelihood and model complexity, whereas CV uses a more empirical, data-driven method by evaluating how the model performs across different subsets of the data.

A table with numbers and text

Description automatically generated A graph with numbers and symbols

Description automatically generated with medium confidence

A diagram of a disease

Description automatically generated A diagram of a graph

Description automatically generated with medium confidence

**a) Based on the scatterplot of Disease versus Cigarette, does there appear to be a linear relationship between cigarette consumption and heart disease? If so, does the relationship appear to be negative or positive?**

The scatterplot shows a pattern that suggests a linear relationship between cigarette consumption and heart disease. The relationship is positive, meaning that as cigarette consumption increases, the number of deaths from heart disease also appears to increase. The data points tend to rise upwards as we move from left to right across the plot.

**(b) What patterns or problems, if any, do you see in the residuals versus fits plot? Would you feel reasonably comfortable in fitting a simple linear regression model to this data set?**

The residuals versus fits plot shows that the residuals do not appear to be randomly scattered around the horizontal axis; there seems to be a pattern where residuals for lower fitted values are mostly positive, and for higher fitted values, they are mostly negative. This could indicate that a simple linear model might not be the best fit for the data, as it does not capture all the variability in the response variable. While some caution is warranted, a simple linear regression could still provide a reasonable approximation for an initial analysis.

**(c) State the size of the fitted slope, 𝛽መ, and give an interpretation for it.**

The fitted slope, 𝛽, is 0.05568. This means that for each additional cigarette consumed per capita, there is an associated increase of approximately 0.05568 in the death rate from heart disease per 100,000 people aged 35-64. It quantifies the rate of increase in heart disease deaths with respect to per capita cigarette consumption.

**(d) Do you think that natural variability alone could account for such a large value of 𝛽መ as found here? Explain.**

Considering the p-value associated with 𝛽 is 0.000, it suggests that the likelihood of observing such a large slope due to natural variability alone is extremely low. This p-value indicates that the relationship between cigarette consumption and heart disease deaths is statistically significant. The low p-value means that the observed association is unlikely to be due to chance, and there is likely a real underlying relationship. Therefore, natural variability alone is not likely to account for the large value of 𝛽 observed here.

**(e) Using the regression output to determine whether sufficient statistical evidence exists to conclude that there is a positive linear relationship between Cigarette and Disease at the 1% level of significance.**

The regression output shows that the slope coefficient (𝛽) for Cigarette has a p-value of 0.000, which is less than the 1% level of significance (0.01). This very low p-value indicates that the positive relationship between Cigarette consumption and Disease is statistically significant at the 1% level. Therefore, we have sufficient statistical evidence to conclude that an increase in cigarette consumption is associated with an increase in the rate of heart disease deaths. This conclusion is based on the data from the 1960s for the 21 countries examined.

**(f) Based on R2, assess the strength of the linear relationship between Cigarette and Disease.**

The R-squared value of the regression is 49.6%, which indicates that approximately half of the variation in the heart disease death rates can be explained by the per capita cigarette consumption. While this is a substantial proportion, it also suggests that there are other factors not included in the model that account for the remaining variability in heart disease rates. An R-squared value of 49.6% represents a moderate-to-strong relationship, but not a perfect one. It shows that cigarette consumption is an important factor, but not the only factor affecting heart disease death rates.

**(g) Do the p-value for 𝛽 and the value of R2 provide contradictory evidence on the strength of the linear relationship between smoking and heart disease? Explain.**

The p-value for 𝛽 and the R2 value do not provide contradictory evidence; instead, they complement each other. The p-value tells us that the relationship between smoking and heart disease is statistically significant, while the R2 value gives us information about how much of the variance in heart disease is explained by smoking. A significant p-value alongside a moderate-to-strong R2 suggests that while cigarette consumption is a significant predictor of heart disease, it is not the only factor, and other variables may also play a significant role. Together, these statistics support the conclusion that there is a significant and substantive relationship between smoking and heart disease, but the relationship is not all-encompassing.