Brian Tenneson

ADS 534

07/22/2024

Statement regarding academic integrity: Large Language Models (LLMs) were utilized alongside human oversight to perform data analysis and generate insights across various contexts and datasets.

1. SOLUTION: To formulate the regression models for the given toxicology study, we'll use the data on the effects of air pollution (filtered air vs. concentrated air particles, CAPs) and preexisting pulmonary disease (exposure to SO2) on pulmonary inflammation as measured by neutrophil numerical density (Nn). The categorical variables are CAPs and SO2, and the dependent variable is Nn.

1. Main Effect Model

The main effect model includes only the primary effects of the CAPs and SO2, without considering the interaction between them. Here's how this model can be expressed:

$$\text{Nn}_i = \beta_0 + \beta_1 \times \text{CAPs}_i + \beta_2 \times \text{SO2}_i + \epsilon_i$$

Where:

Nn_i is the neutrophil numerical density for rat i.

β_0 is the intercept (baseline level of Nn when both CAPs and SO2 are at their reference levels).

β_1 is the coefficient for the effect of exposure to concentrated air particles (CAPs). This is the expected change in Nn due to exposure to CAPs, assuming no exposure to SO2.

β_2 is the coefficient for the effect of exposure to sulfur dioxide (SO2). This is the expected change in Nn due to exposure to SO2, assuming no exposure to CAPs.

CAPs_i is a binary indicator (0 or 1), where 1 indicates exposure to CAPs.

SO2_i is a binary indicator (0 or 1), where 1 indicates exposure to SO2.

ϵ_i is the error term.

2. Interaction Model

The interaction model includes both the main effects of CAPs and SO2 and their interaction. This model can help assess if the effect of one variable depends on the level of the other variable.

$$\text{Nn}_i = \beta_0 + \beta_1 \times \text{CAPs}_i + \beta_2 \times \text{SO2}_i + \beta_3 \times (\text{CAPs}_i \times \text{SO2}_i) + \epsilon_i$$

Where:

β_3 is the interaction term coefficient.

Interpretation:

β_3: The expected change in Nn due to the interaction between CAPs and SO2. It measures how the combined effect of exposure to both CAPs and SO2 differs from the sum of their individual effects.

Example Interpretation Using the Models:

- If β_1 is positive, it suggests that exposure to CAPs increases the level of Nn, indicating more pulmonary inflammation.

- If β_3 is significantly different from zero, it indicates that the effect of CAPs on Nn is modified by the presence of SO2 exposure, and vice versa.

These models will allow researchers to quantitatively assess the independent and combined effects of air pollution and preexisting pulmonary disease on inflammation in rats.

(b) (10 points) Consider a test of whether there is a difference in the health effects of air pollution inhalation for healthy animals and that for chronic bronchitic animals (i.e. those who received SO2) under the interaction model. What is the null hypothesis corresponding to this test, in terms of the regression coefficients under the interaction model?

SOLUTION: In the interaction model described earlier:

$$\text{Nn}_i = \beta_0 + \beta_1 \times \text{CAPs}_i + \beta_2 \times \text{SO2}_i + \beta_3 \times (\text{CAPs}_i \times \text{SO2}_i) + \epsilon_i$$

The interaction term $\beta_3 \times (\text{CAPs}_i \times \text{SO2}_i)$ represents how the effect of CAPs on the neutrophil numerical

density (Nn) changes depending on whether the rat was exposed to SO2 or not. Essentially, this coefficient measures whether the impact of air pollution is different for animals with chronic bronchitis compared to healthy animals.

Null Hypothesis

The null hypothesis for testing whether the health effects of air pollution differ between healthy animals and those with chronic bronchitis (as related to the interaction between CAPs and SO2) is:

$$H_0 : \beta_3 = 0$$

This hypothesis states that there is no interaction effect between CAPs and SO2 on the neutrophil numerical density, meaning the effect of air pollution on Nn is the same regardless of the SO2 exposure status. If β_3 is not significantly different from zero, it suggests that the impact of CAPs on Nn is not modified by whether the animal has chronic bronchitis due to SO2 exposure.

Alternative Hypothesis

The alternative hypothesis would be:

$$H_A : \beta_3 \neq 0$$

This states that there is an interaction effect, implying that the health effect of air pollution (CAPs) differs between animals based on their chronic bronchitis status (SO2 exposure).

(c) (10 points) Perform an $\alpha = 0.05$ level test of the null hypothesis in (b). What

do you conclude?

SOLUTION:

This is the analysis that came from running python code:

Parameter	Coefficient	Std Error	t-value	P-value	95% CI Lower	95% CI Upper
Intercept	0.2381	0.0067	35.326	1.98E-103	0.2249	0.2514
CAPs	0.2377	0.0092	25.7623	7.95E-75	0.2195	0.2559
SO2	0.0977	0.009	10.8219	6.38E-23	0.08	0.1155
CAPs*SO2	0.1263	0.0129	9.827	1.08E-19	0.101	0.1516

The coefficient of CAPs*SO2 has a 95% confidence interval of 0.101 to 0.1516 and the p-value is $1.08E-19 < \alpha$ so we reject the null hypothesis (with enthusiasm).

Can you perform this test under the main effect model? If yes,

carry out the test. If not, explain why.

SOLUTION: No because β_3 is absent in the main effect model so there is no chance it isn't zero.

(d) (10 points) Suppose the investigator neglected to mention that half of the animals received SO2, and you fit a regression model for Nn using CAPs exposure only. That is, you ignore whether the animal is chronic bronchitic or not. Write down the corresponding regression model.

SOLUTION:

$$\text{Nn}_i = \beta_0 + \beta_1 \times \text{CAPs}_i + \epsilon_i$$

Fit both this model and the main effect model that includes CAPs and SO2, but not the interaction, to the data.

Model	Intercept	CAPs	SO2	R_squared	F_statistic	P_value_CAPs	P_value_SO2
Simple Linear Regression	0.293	0.289		0.674	564.809	1.97E-68	
Multiple Linear Regression	0.203	0.303	0.160	0.879	985.770	2.74E-117	2.45E-60

Based on the results of these two models, do you think that SO2 confounds the association between CAPs

and Nn? Explain why?

SOLUTION: The coefficient for CAPs remains relatively stable (from 0.2895 to 0.3028) when SO2 is included in the model. This indicates that the relationship between CAPs and Nn does not change substantially when controlling for SO2.

R-squared Increase: The significant increase in R-squared value (from 0.674 to 0.879) suggests that SO2 explains additional variability in Nn that is not captured by CAPs alone.

Based on these observations, while SO2 is an important variable that explains additional variation in Nn, it does not appear to strongly confound the association between CAPs and Nn. The association between CAPs and Nn remains consistent even when accounting for SO2. Thus, SO2 does not significantly confound the relationship between CAPs and Nn, although it is an important predictor in its own right.

(e) (10 points) Another way to justify whether SO2 confounds the association between CAPs and Nn is by looking at the pairwise association among these three variables directly. Choose appropriate measures and/or tests to investigate these pairwise associations. Based on the results you get, do you think that SO2 confounds the association between CAPs and Nn? Explain why? SOLUTION:

Here, I have calculated some Pearson r-coefficients among rat ID, CAPs, SO2, and Nn.

r-value	rat	caps	so2	nn
rat	1.00	0.02	-0.03	-0.01
caps	0.02	1.00	-0.08	0.82
so2	-0.03	-0.08	1.00	0.38
nn	-0.01	0.82	0.38	1.00

We can see that CAPs and Nn have a significant correlation (r=0.82) and SO2 and Nn have an appreciable correlation (r=0.38). No other variables have a significant correlation.

p-value	rat	caps	so2	nn
rat	N/A	0.70	0.58	0.91
caps	0.70	N/A	0.17	1.97E-68
so2	0.58	0.17	N/A	5.16E-11
nn	0.91	1.97E-68	5.16E-11	N/A

These are the p-values from the above analysis. Notice that CAPs and Nn have a significant p-value (p=1.97E-68), SO2 and Nn has a significant p-value (p=5.16E-11), and other pairs do not have a significant p-value (<0.05). Based on the results, SO2 does not confound the association between CAPs and Nn. Although SO2 is significantly associated with Nn, it is not significantly associated with CAPs. The association between CAPs and Nn remains significant and strong even when considering SO2, indicating no confounding effect.

(f) (5 points) Under each of the two models in (a) (the main effect model and the interaction model), test whether there is a health effect of air pollution inhalation for healthy animals, and also test whether there is a health effect of air pollution inhalation for chronic bronchitic animals (i.e. those who received SO2).

SOLUTION:

Model	Animal Type	Coefficient	P-value
Main Effect	Healthy	0.30	2.74E-117
Main Effect	Chronic Bronchitic	0.46	2.45E-60
Interaction	Healthy	0.24	7.95E-75
Interaction	Chronic Bronchitic	0.36	1.08E-19

Main Effect Model: Both healthy and chronic bronchitic animals show a significant health effect of air pollution inhalation. The effect is stronger for chronic bronchitic animals (0.46) compared to healthy animals (0.30). **Interaction Model:** Both healthy and chronic bronchitic animals show a significant health effect of air pollution inhalation. The effect is stronger for chronic bronchitic animals (0.36) compared to healthy animals (0.24). **In summary**, the results from both models indicate that air pollution inhalation has a significant health effect on both healthy and chronic bronchitic animals. The effect is more pronounced

in chronic bronchitic animals in both the main effect model and the interaction model.

2. SOLUTION:

We obtain the following table using python:

	coef	std err	t	P>|t|	[0.025	0.975]
const	162.8759	25.776	6.319	0.0	108.927	216.825
X1	-1.2103	0.301	-4.015	0.001	-1.841	-0.579
X2	-0.6659	0.821	-0.811	0.427	-2.384	1.052
X3	-8.613	12.241	-0.704	0.49	-34.234	17.008

Note that β_2 is possibly anywhere in the confidence interval -2.384 to 1.052; so severity of illness increase may lead to a decrease in patient satisfaction. However, the opposite might be true (albeit a little less likely) that severity of illness increase may lead to an increase in patient satisfaction.

(b) Test whether there is a regression relationship here; that is, if the regression as a whole explains variability in the response. Using significance level $\alpha = 0.05$, state your null and alternative hypotheses, and your conclusions. What does your test imply about 1, 2, and 3?

Let the null hypothesis be $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$. When we apply an F-test, we get F-statistic: 13.01 and p-value for the F-test: 7.48e-05. The F-statistic is a measure of the ratio of the model's explained variance to the unexplained variance. An F-statistic of 13.01 indicates that the model explains a significant amount of variance when compared to a model with no predictors (just an intercept). Rejecting the null hypothesis implies that there is significant evidence that at least one of the predictors (β_1, β_2, or β_3) contributes to explaining the variability in the response variable. Therefore, the regression model as a whole does significantly better at explaining the response variable compared to the model with no predictors (just the intercept).

(c) Test the null hypothesis that β_1 is equal to 0 at the 0.05 level of significance. What do you conclude? The t-statistic of -4.015 and a p-value of 0.001 suggest that β_1 is significantly different from 0 at the 0.05 level of significance.

(d) Obtain 95% confidence interval estimates of β_1, β_2, and β_3. Interpret your results.

β_1: [−1.841,−0.579]

β_2: [−2.384,1.052]

β_3 : [−34.234,17.008]

We have some assurance that β_1 is negative so that when it increases, overall satisfaction will decrease. For β_2 and β_3, 0 is in the confidence interval, meaning that we can't be sure if β_2 and β_3 are positive or negative.

(e) Obtain a 95% confidence interval estimate of mean satisfaction when $X_1 = 35$, $X_2 = 45$, and $X_3 = 2.2$. Interpret your confidence interval.

This confidence interval means that you can be 95% confident that the true mean satisfaction score, given $X_1 = 35$, $X_2 = 45$, and $X_3 = 2.2$, lies between 62.30 and 80.90. This range captures the uncertainty about the mean satisfaction score based on the current model and the variability of the data used to fit the model.