1. The data consists of one independent variable *X*, which has 5 levels (i.e. *X* = 1*,* 2*,* 4*,* 5*,* 6) and a numeric dependent variable *Y* . Variable *X*2 and *X*3 are the square (*X*2 = *X*2) and cube (*X*3 = *X*3) of *X*, respectively. Below are part of the SAS code and outputs of fitting a SLR, quadratic regression

and cubic regression models of *Y* on *X*.

PROC REG DATA=data;

MODEL Y=X; title ’Simple linear regression model’; RUN; PROC REG DATA=data;

MODEL Y=X X2; title ’Quadratic regression model’; RUN; PROC REG DATA=data;

MODEL Y=X X2 X3; title ’Cubic regression model’; RUN; PROC glm; class X;

model Y=X; title ’one-way ANOVA’; RUN;

Simple linear regression model Analysis of Variance

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Source | DF | | Sum of  Squares | Mean  Square | | F Value | Pr > F |
| Model | 1 | | 114.11434 | 114.11434 | | 11.26 | 0.0047 |
| Error | 14 | | 141.90316 | 10.13594 | |  |  |
| Corrected Total | 15 | | 256.01750 |  | |  |  |
| Parameter Estimates | | | | | | | |
|  |  | Parameter | | Standard |  |  |  |
| Variable | DF | Estimate | | Error | t Value | Pr > | |t| |
| Intercept | 1 | -7.50304 | | 1.76625 | -4.25 | 0.0008 | |
| X | 1 | 1.43472 | | 0.42759 | 3.36 | 0.0047 | |

Quadratic regression model Analysis of Variance

Sum of Mean

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Source | DF | Squares | Square | F | Value | Pr > F |
| Model | 2 | 209.01170 | 104.50585 |  | 28.90 | <.0001 |
| Error | 13 | 47.00580 | 3.61583 |  |  |  |
| Corrected Total | 15 | 256.01750 |  |  |  |  |

Parameter Estimates Parameter Standard

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | DF | Estimate Error | | t Value | Pr > |t| | |
| Intercept | 1 | 1.04319 1.97378 | | 0.53 | 0.6060 | |
| X | 1 | -5.32626 1.34422 | | -3.96 | 0.0016 | |
| X2 | 1 | 0.96029 0.18745 | | 5.12 | 0.0002 | |
|  |  | Cubic regression model | |  |  | |
|  |  | Analysis of Variance | |  |  | |
|  |  | Sum of | | Mean |  | |
| Source |  | DF Squares | | Square | F Value Pr > F | |
| Model |  | 3 215.40920 71.80307 | | | 21.22 | <.0001 |
| Error |  | 12 40.60830 3.38402 | | |  |  |
| Corrected Total |  | 15 256.01750 | | |  |  |
|  |  | Parameter Estimates | | |  |  |
|  |  | Parameter Standard | | |  |  |
| Variable | DF | Estimate Error t Value | | | Pr > | |t| |
| Intercept | 1 | -4.35868 | 4.36821 | -1.00 | 0.3381 | |
| X | 1 | 1.45448 | 5.10018 | 0.29 | 0.7804 | |
| X2 | 1 | -1.21839 | 1.59489 | -0.76 | 0.4597 | |
| X3 | 1 | 0.20125 | 0.14637 | 1.37 | 0.1943 | |

one-way ANOVA The GLM Procedure

Sum of

Source DF Squares Mean Square F Value Pr > F

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 4 | 216.9241667 | 54.2310417 | 15.26 | 0.0002 |
| Error | 11 | 39.0933333 | 3.5539394 |  |  |
| Corrected Total | 15 | 256.0175000 |  |  |  |

1. Use the SAS outputs above to find out a reasonable regression model of *Y* on *X*. Provide the

**appropriate test statistics, df and p-value to justify your choice** and explain briefly.

1. Suppose we are ONLY interested three pairwise comparisons among X=1,5 and 6. That is 5-1 (which means *µx*=5 *−µx*=1), 6-1 and 6-5. Below are the Fisher’s LSD output for these three pairwise comparisons. Use the Bonferroni approach to test all three pairwise comparisons (i.e. 5-1, 6-1 and 6-5). Provide the Bonferroni confidence intervals **numerically**. Set the familywise error rate at 0.05.

t Tests (LSD) for Y

NOTE: This test controls the Type I comparisonwise error rate, not the familywise error rate.

Alpha 0.05

Error Degrees of Freedom 11

Error Mean Square (MSE) 3.553939

Critical Value of t 2.201

Comparisons significant at the 0.05 level are indicated by \*\*\*.

Difference

|  |  |  |  |
| --- | --- | --- | --- |
| X  Comparison | Between  Means | 95% Confidence  Limits | |
| 6 - 5 | 7.125 | 3.532 | 10.718 \*\*\* |
| 6 - 1 | 7.908 | 4.739 | 11.077 \*\*\* |
| 5 - 1 | 0.783 | -3.004 | 4.571 |

1. We can also apply Tukey’s HSD approach as well as Scheff´e approach to test all three pairwise comparisons in **part b** (i.e. 5-1, 6-1 and 6-5). Compare the three approaches (i.e. Bonferroni, Tukey and Scheff´e), and find out which one is the best? Explain briefly. Set the familywise error rate at 0.05.
2. An automobile manufacturer want to compare the gasoline consumption rate (miles per gallon) on five particular brands of cars. Five cars were used, one for each brand. Four drivers were randomly selected in this study. Each driver drove each car twice over a 25-mile test course and the miles per gallon (mpg) were recorded. Therefore, 5 *×* 4 *×* 2 = 40 mpg values were recorded. A crude SAS analysis output is the following. Set *α* = 0*.*05.

|  |  |  |
| --- | --- | --- |
| Source | DF | Sum of Squares |
| Driver |  | 280.2847500 |
| Car |  | 94.7135000 |
| Driver\*Car |  | 2.4465000 |
| Rep(Driver\*Car) |  | 3.5150000 |
| Corrected Total |  | 380.9597500 |

1. Provide the appropriate model for this study. Clearly label and define all subscripts. Your answer for the model should be something like: *yij* = *µ* + *τi* + *Eij* for *i* = 1*, . . . ,* 8 (make sure to clearly define the *yij*, *τi*, etc. in your model.
2. Fill in the degrees of freedom in the above ANOVA source table. Test the Car effect. Provide the **test statistic with its df**, the **range of p-value** (using the table from the course package) and conclusion.
3. Can you improve the power of the test for the Car effect **without increasing the sample size**? If so, explain briefly how to increase the power. **No calculation is needed.**
4. The table below describes the features of five car brands used in the study. The investigator want to compare the MPG (1) between the Japanese and US cars; (2) between Sedan and SUV. Write out the two contrasts for the comparisons. You answer should be something like: *γ*1 = *µ*1 *−* 1 *µ*2 *−* 1 *µ*4.

2

2



3. A study was conducted to investigate the age effect on the severity of a certain disease. The independent variable *X* is the age of the patient. The response *Y* is a binary variable with 0 for the mild (less severe) symptoms, and 1 for the severe symptoms. Below are part of the SAS outputs for the logistic regression of *Y* on *X*.

Analysis of Maximum Likelihood Estimates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | DF | Estimate | Standard  Error | Wald  Chi-Square | Pr > ChiSq |
| Intercept | 1 | -25.2738 | 2.6552 | 90.6038 | <.0001 |
| X | 1 | 0.5129 | 0.0517 | 98.4202 | <.0001 |

1. Find out the age group (of the patients) that will have 50% or higher chance to get severe symptoms.
2. What is the assumption for the logistic regression model? Explain in the context of this study.
3. Interpret the slope (i.e. 0.5129) within the context.

4. Adam is tracking the total case trend of flu for two countries (say country A and B). The response variable is the cumulative number of flu cases. The explanatory variable is the number of days since 100th case in that country. Each country has 30 observations. From the scatter plot, John found out that two countries seem to have different slopes. This is probably because that the majority (minority) of people in Country A (B) wear masks when they are out in public. Help John to set up a linear regression model so that he can test whether the slopes of country A (say *βA*) & B (say *βB*) are the same or not. **Write out the model formula and explain how to test** *H*0: *βA* = *βB* vs. *Ha*: *βA /*= *βB*. **Define the variable(s) if necessary**. Briefly describe how to find out the test statistic. Provide the **degrees of freedom** for the test statistic in **number(s)**.