1. Simple polynomial regression

a.

# Read data

data <- read.csv("F:/statistics/konstantinos/Housing.csv",header = T)

head(data)

housing <- data

housing$x2 <- housing$lotsize^2

head(housing)

b.



Diagnostic Plots



c)

The highest value of the leverage found is 0.2547.

d.

Leverage values are computed from the standardized predictor variable. The predictor variable is first converted into its standard score. Then, the Leverage value is computed as; (z score + 1)/Number of observations.

Leverage values are measured using the Hat Matrix.

e)

i)

F-test

Hypotheses

H0: The model is not statistically significant.

Ha: The model is statistically significant.

The null hypothesis is rejected in the p value associated with the F-statistic is less than 0.05

Conclusion:

Reject the null hypothesis. There is sufficient evidence to conclude that the model is statistically significant.

T-test

Hypotheses

H­0: The coefficient of lot size is not significant in predicting house price.

Ha: The coefficient of lot size is not significant in predicting house price.

Reject the null hypothesis if P(> |t|) < 0.05

Conclusion:

P < 0.001 therefore, reject the null hypothesis. There is sufficient evidence to conclude that the coefficient of lot size is significantly different from 0.

Hypotheses

H0: The polynomial of lot size is not significant in predicting the prices of the houses.

Ha: The polynomial of lot size is significant in predicting the prices of the houses.’

Reject the null hypothesis if P(|t|) < 0.05

Conclusion:

Reject the null hypothesis because P < 0.001.

f.

Fun plot of the model



Fun polynomial plot of degree 3



g.

It seems the 3rd order degree model is likely to produce better results than the polynomial model. The fit values are not scattered in the 3rd order model like they are in the polynomial model.

2. Multiple regression with a dummy variable

a.

data house;

set housing;

driveway1 = .;

if (driveway = "yes") THEN driveway1 = 1;

if (driveway = "no") THEN driveway1 = 0;

driveway0 = .;

if (driveway = "no") THEN driveway0 = 1;

if (driveway = "yes") THEN driveway0 = 0;

run;

b.

An interaction variable is created between bedrooms and the bathrooms.

data house1;

set house;

bed\_bath = bedrooms \* bathrms;

run;

c.



The presented output is of interest because it has the model fit summary and the significance of the dummy variable and the interaction term.

The initial model with four variables had an Adj R-Sq value of 0.5497. This model with the interaction term and the dummy variable added has an Adjusted R-squared value of 0.5497.

c)

The dummy variable is significant in the model. The interaction between the number of bedrooms and the number of bathrooms is not significant. It therefore seems, it is the dummy variable that increased the fitness of the model based on the significance of the added variables.

d)

The interaction term has no effect on the regression model.