ST231 Lab Report 1

Deadline: 3 February 2026, 1 pm

Data Description

The data in the file `dia1.csv` is an adapted subset from the `diamonds` dataset in the `ggplot2` package. It consists of information on 800 round diamonds. The variables considered for this lab report are:

- `price`: the sale price of the diamond in US Dollars.
- `weight`: the weight of the diamond in carat.

Instructions

Carefully read the Lab Report Guidance on moodle before you start working on this lab report!

In the following you will explore the relationship between the weight of a diamond and its price.

1. [**2 marks**] Produce a scatterplot that illustrates the relationship between the weight of a diamond and its price. Fit a quadratic and a cubic regression model to the data with `price` as the response variable and `weight` as the explanatory variable. Add the fitted curves of the two polynomial models to the scatterplot.
2. [**3 marks**] Based on the evidence from the plot in Question 1, critically evaluate and compare the fit of the two models to the data.
3. [**3 marks**] For each of the two fitted models, produce a residual plot, that is a plot of residuals against fitted values. Critically evaluate and compare the two plots.

[**2 marks**] for style and quality of presentation.

Feedback:

1. R markdown syntax and cheat sheet.
2. Hide code blocks, `echo=FALSE`.
3. `fig.cap`
4. Question 2 and 3

Solutions:

1. The scatterplot below shows the data and the fitted curves for the quadratic regression model (blue solid line) and the cubic regression model (dashed red line) which both predict the price of a diamond from its weight.

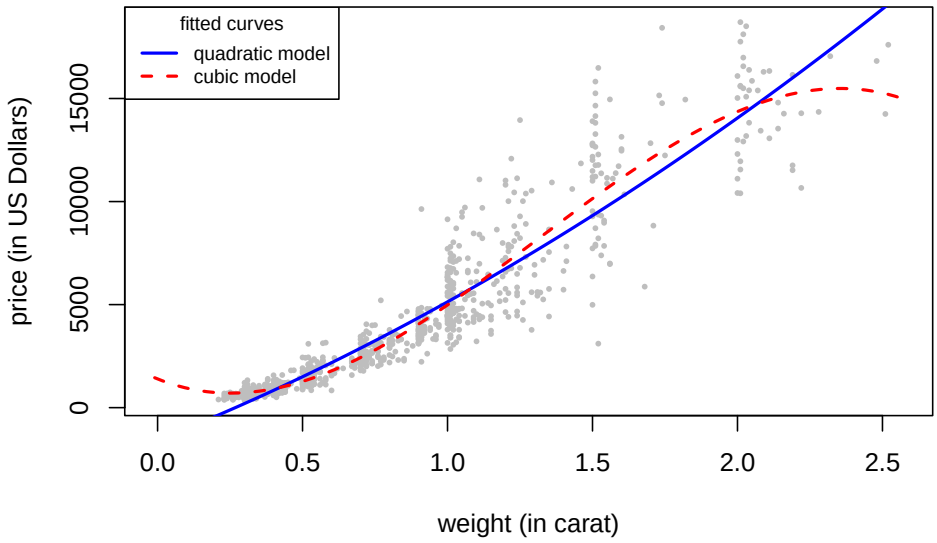

Figure 1: A scatterplot of diamond prices in Dollars against the weight of the diamond in carat. The fitted curves for a quadratic regression model and a cubic regression model have been added to the plot.

2. Both models take account of the fact that the relationship between the weight of diamonds and their price is curved rather than linear. However the quadratic model systematically underestimates prices for diamonds that have a small weight, even predicting negative prices. In contrast the cubic model appears to fit the data more closely. In particular, it avoids predicting negative prices for lighter diamonds. However, it shows some artifacts that come from the properties of a cubic function. The fitted cubic curve attains a maximum at around 2.25 carat and then decreases which

does not fit with the general observation that diamonds tend to be more expensive the heavier they are. (We observe a similar artifact weights close to zero where the curve attains a minimum.)

Note: In this specific example the artifacts of the cubic model occur at the boundary of the range of the data where there are few observations to infer the shape of the relationship between weight and price. Here we would always be cautious.

4. The residual plot for the quadratic regression model indicates some mild non-linearity as the smoother first decreases and then increases. The non-linearity is less pronounced in the residual plot for the cubic regression model as the smoother is initially flat and only increases for larger fitted values. In both plots, the observations form a right-opening megaphone pattern, which indicates hetero-scedasticity with a variance that is increasing with fitted value.

   Note: When we fit polynomial models, we always include any lower order terms. For example, in the cubic modoel we included a cubic term `I(weight^3)`, but also a quadratic term `I(weight^2)` and a linear term `weight`.

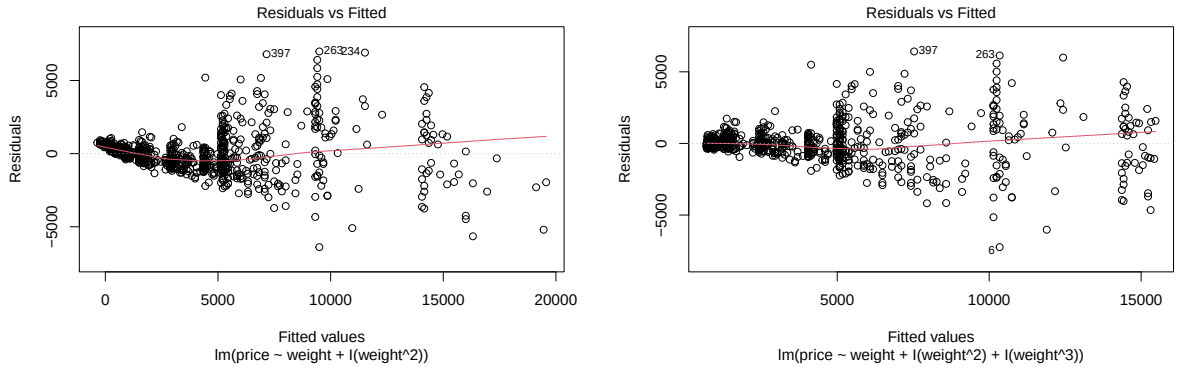

Figure 2: Residual versus fitted values plots. Left: residual plot for the quadratic regression model. Right: residual plot for the cubic regression model.