7CCMMS61 – Statistics for Data Science
Coursework: mobile.csv Data Analysis

Katherine Poole
K23074532

Summary

This report presents the findings from the 'mobile.csv' data set, which provides data from a Melbourne, Australia telephone shop about 161 mobile phone model's prices and features. The report commences with a brief introduction, followed by exploratory data analysis (EDA) and fitting a linear regression model. The report concludes with analysis of the factors that most affect mobile phone prices in the data and discusses potential limitations and future research inquiries conducted on this and similar datasets might consider.

Introduction

Mobile phones have a near ubiquitous presence in the modern world. Unfortunately, mobile phones remain quite expensive, which is partially due to their variations in features. The 'mobile.csv' data provides data on 161 mobile phones that was collected from a mobile phone shop in Melbourne, Australia in 2021. The data contains prices in Australian dollars (AUS) for each phones, plus various hardware and software features that might affect price. Figure 1.1 contains the complete set of variables present in the data.

Variable Name	Description	Type
price	Price of phone in Australian dollars (AUS)	Quantitative
weight	Weight of phone in grams	Quantitative
resolution	Number of pixels across the screen	Quantitative
ppi	Number of pixels per inch	Quantitative
cpu_core	Number of mini-CPUs on main CPU	Quantitative
cpu_freq	CPU speed in GHz	Quantitative
ram	Random access memory measured in GB	Quantitative
internal_memory	Amount of storage in GB	Quantitative
RearCam	Mega pixels	Quantitative
Front_Cam	Mega pixels	Quantitative
battery	Battery life measured in mAh	Quantitative
thickness	Thickness of phone in millimeters	Quantitative

Figure 1.1 – Data variables and descriptions for mobile phone data
Source: mobile.csv

The purpose of this report is to analyze potential relationships between phone features and prices. Increasing variability and complication in phone software and hardware make it more difficult for consumers to discern fair prices. This analysis intends to understand what factors most affect the price of mobile phones to inform purchasing decisions. Following a methodology statement with descriptions of data processing and statistical methods, the report will discuss exploratory data analyses completed with the data. Next, the report will detail statistical analyses conducted, followed by analysis of patterns and insight derived from these insights. The report will conclude with a summary of findings, limitations, and further potential research in the area.

Methodology

This report reflects a statistical approach to analyze the relationship between mobile phone features and their prices utilizing the 'mobile.csv' dataset. The study conducts exploratory data analysis and implements linear regression analysis to understand pricing dynamics derivable from the mobile data. EDA involves visualization and assessment of variable distributions with histograms and skewness calculations. Informed by this first step, logarithmic transformations are made on certain variables to normalize data and mitigate skewness. Scatterplots and calculation of correlation coefficients are then calculated between each variable to assess the presence and strength of linear relationships. Following EDA, a linear regression model is fit to

the data. Linear regression is the best fit model for the mobile phone data according to the evidence gathered in EDA that implies clear linear relationships alongside the question guiding this inquiry: what factors affect mobile phone price? Model variable selection employs a backwards elimination testing procedure, systematically removing non-significant predictors based on a 5% significance level. This iterative elimination refines the model fit and ensures inclusion of only significant variables that meaningfully contribute to prediction of price. After deriving the best fit model, necessary diagnostic checks including residuals and qq-plots are conducted to ensure the validity of the model before making inferences.

Exploratory Data Analyses

Our inquiry into the "factors that affect phone price" requires analysis that relates the response variable price (Y) with any number of the explanatory variables phone features (X_i). A linear regression model comparing mobile features X_i against price Y might be the best fit for these requirements. Before proceeding with linear regression, it is imperative to first become familiarized with and validate certain aspects of our data.

Descriptive statistics provide a method to accomplish this. Figure 2.1 displays the mean, median, standard deviation (SD), minimum, and maximum value for each variable in the mobile phone data. The measures of mean, median, and standard deviation provide estimates of the "middle" of our data and insight into individual data point's distance from this middle.[1] The minimum and maximum values provide the data's range. The mean measures of explanatory variables in figure 2.1 are frequently higher than their median measure. Additionally, the standard deviation is often quite large in the context of a variable's range. This indicates that our data might be right skewed.

Variable Name	Mean	Median	SD	Min	Max
Price_AUS	2,215.60	2,258.00	768.19	614.00	4,361.00
weight_gr	170.43	153.00	92.89	66.00	753.00
resoloution	5.21	5.15	1.51	1.40	12.20
ppi	335.06	294.00	134.83	121.00	806.00
cpu.core	4.86	4.00	2.44	0.00	8.00
cpu.freq	1.50	1.40	0.60	0.00	2.70
internal.mem	24.50	16.00	28.80	0.00	128.00
ram	2.20	2.00	1.61	0.00	6.00
RearCam	10.38	12.00	6.18	0.00	23.00
Front_Cam	4.50	5.00	4.34	0.00	20.00
battery	2,842.11	2,800.00	1,366.99	800.00	9,500.00
thickness	8.92	8.40	2.19	5.10	18.50

Figure 2.1 – Descriptive statistics
Source: mobile.csv data

[1]MacMillan, Andrew, David Preston, Jessica Wolfe, and Sandy Yu. "13.1: Basic Statistics- Mean, Median, Average, Standard Deviation, Z-Scores, and P-Value." Engineering LibreTexts, March 11, 2023.

Histograms provide a variable's rough distribution and allow visual identification of skewness. According to the histogram output in figure 2.2, some of the mobile data variables exhibit right skewness. Variables '*Weight Gr'*, '*Resolution', 'PPI', 'Internal Mem', 'Front Cam', 'Battery',* and '*Thickness'* appear to have an abundance of low values, but only a handful of very high values that create a waning tail in the positive direction. To capture the precise degree of asymmetry, we can calculate each variable's skewness measure. Skewness measures are shown in figure 2.3 in column '*Skewness (raw)*'.

As suspected, many explanatory X_i variables in the mobile data exhibit varying levels of skewness. The proposed response variable Y 'price' demonstrates low skewness. Variables coded orange in figure 2.3 show medium-high right skewness and red variables show high right skewness. A linear regression model run on untreated skewed data can lead to undesirable parameter estimates and confidence interval calculations. To ameliorate the impact of skewed data on regression estimates, data must be transformed. Viable transformation methods for skewed data include natural log, square root, and inverse transformations.[2] Because all skewed variables demonstrate right skew, at-issue mobile data variables are transformed using a logarithmic method. To account for 0 values in some of the transformed variables, a constant c=1 is added to each value before logarithmic transformation.[3] Skewness measures of the logarithmically transformed variables are shown in figure 2.3 in column '*Skewness (log)*'.

[2] Gonzalez-Blanks, Ana, Jessie M. Bridgewater, and Tuppett M. Yates. "Statistical Approaches for Highly Skewed Data: Evaluating Relations between Maltreatment and Young Adults' Non-Suicidal Self-Injury." Journal of Clinical Child & Adolescent Psychology 49, no. 2 (2020): 147–61. https://doi.org/10.1080/15374416.2020.1724543.
[3] Muldoon, Ariel. "The Log-0 Problem: Analysis Strategies and Options for Choosing C in Log(y + c)." Very statisticious, September 19, 2018. https://aosmith.rbind.io/2018/09/19/the-log-0-problem/#thinking-about-0-values.

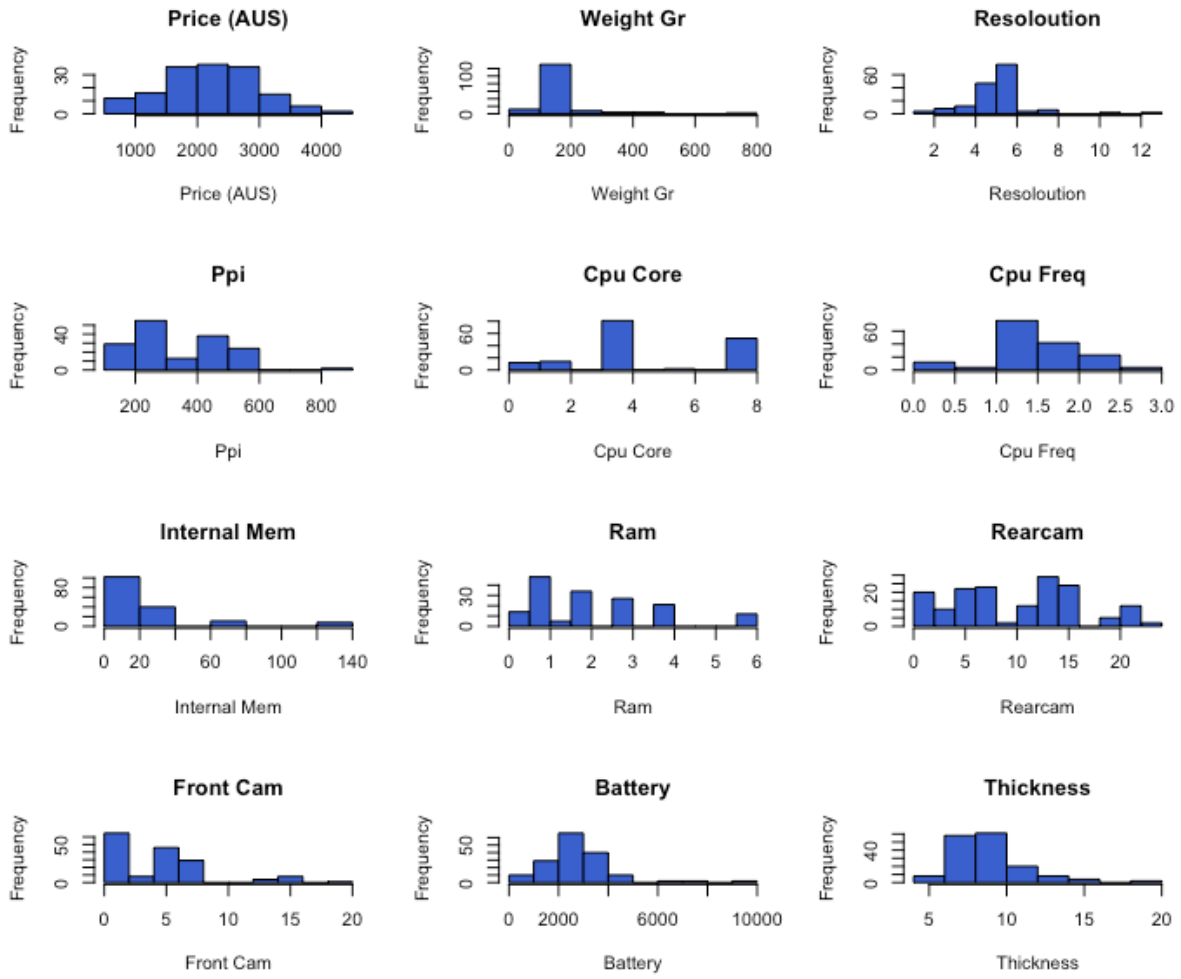

Figure 2.2 - Histograms for all variables in mobile phone data
Source: mobile.csv data

Variable Name	Skewness (raw)	Skewness (log)
price	0.051375415	N/A
weight	3.907869070	1.298251831
resolution	1.154285770	-1.048368920
ppi	0.591322536	N/A
cpu_core	-0.008908678	N/A
cpu_freq	-0.503149909	N/A
ram	0.777988400	N/A
internal_memory	2.345338722	-0.610357731
RearCam	0.104945073	N/A
Front_Cam	1.148043221	-0.310868851
battery	2.053827649	-0.354557236
thickness	1.558218312	0.580646428

Figure 2.3 – Skewness measure for all variables in mobile phone data
Source: mobile.csv data

Logarithmic transformation of each skewed variable results in a smaller absolute value skewness factor. Interestingly, transformation of the variable *'resolution'* only generates a marginally smaller factor. The log transformed histogram, however, appears to demonstrate a more normal distribution than its raw form.

In addition to these initial data inquiries, establishment of relationships between explanatory variables X_i and response variable price is required prior to proceeding with regression. Scatterplots are a method to visualize potential relationships and identify possible outliers. To validate use of logarithmic transformations of some variables X_i, scatterplots for both raw and logarithmic versions of each variable are analyzed.

According to the scatterplots in figure 2.4, many of the phone features X_i appear to have a relationship with price Y. The exact nature and extent of these relationships differ, however. For example, X variables PPI, CPU freq, internal memory, RAM, and front/rear cameras seem to share a positive linear relationship with Y. As their relative values increase, so does price. X variable thickness shares a negative relationship with Y – as a phone's thickness decreases, price increases.

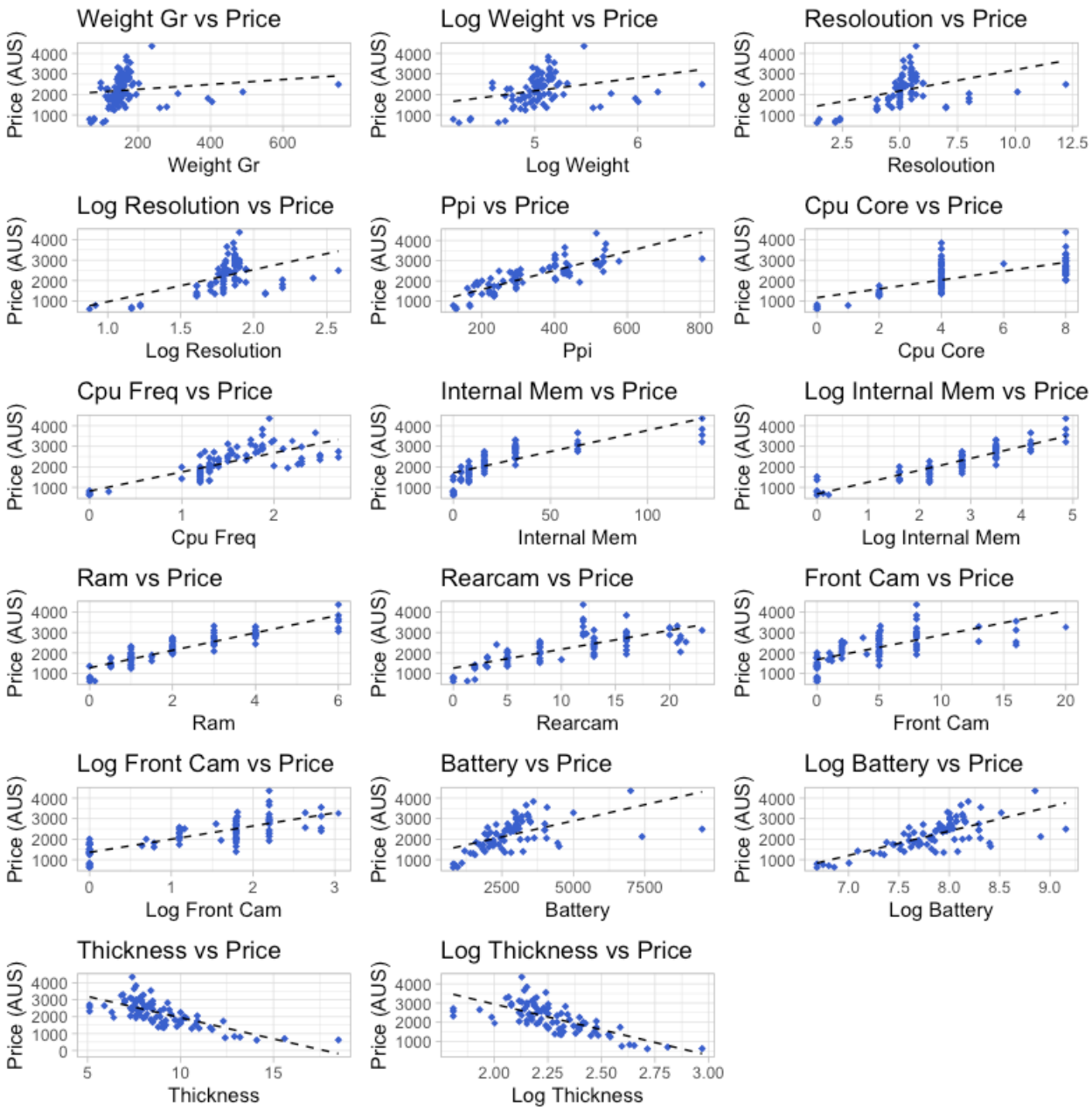

Figure 2.4 – Scatterplots relating all explanatory variables X_i to response variable Y 'price' from mobile phone data
Source: mobile.csv data

Because so many explanatory variables appear to share linear relationships with response variable price, explanatory variables must be tested against each other for collinearity. Collinearity in a regression model can lead to artificially high standard errors for regression estimates, which can result in false rejections of alternative hypotheses (type II errors). A scatterplot matrix relating all X_i variables to each other allows visualization of possible explanatory variable relationships. According to plots in figure 2.5, some explanatory variables share a linear relationship. For example, variable pairs {weight, resolution}, {resolution, battery}, and {internal memory, ram}.

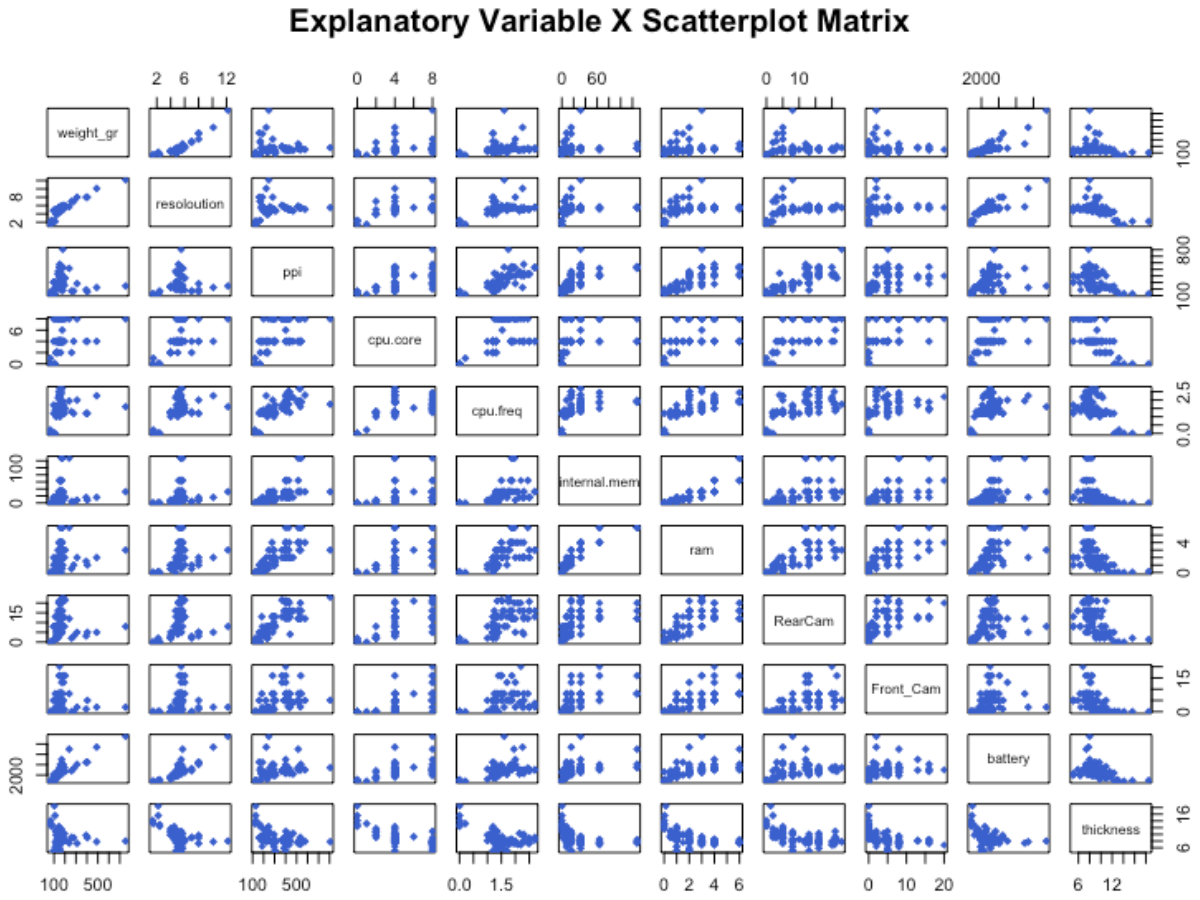

Figure 2.5 – Scatterplots relating all explanatory variables X_i to each other from mobile phone data
Source: mobile.csv data

To further test suspected collinearity, we can calculate correlation coefficients between each variable. Correlation coefficients offer insight into the strength of linear relationships between variables. The matrix in Figure 2.6 displays each variables correlation against all others. Response variable price appears to share a strong linear relationship with a number of variables, including variables 'PPI', 'internal memory', and 'rear camera' among others. Sets of explanatory variables, such as {resolution, weight}, {battery, weight}, and {internal memory, ram}, also appear to share quite strong linear relationships with each other. Variables sharing a correlation coefficient of 0.75 or greater are highlighted in green. After fitting the final linear regression model, it will be imperative to check the variable inflation factors (VIF) associated with each explanatory variable.

	log_ weight	log_resol ution	ppi	cpu. core	cpu. freq	log_inter nal_mem	ram	Rear Cam	log_front _cam	log_ battery	log_ thickness
Price_ AUS	0.309	0.520	0.818	0.687	0.727	0.901	0.897	0.740	0.777	0.710	-0.704
log_ weight		0.876	0.108	0.328	0.405	0.378	0.290	0.137	0.193	0.824	-0.252
log_ resolution			0.307	0.556	0.625	0.585	0.404	0.378	0.417	0.881	-0.596
ppi				0.488	0.713	0.759	0.749	0.774	0.635	0.473	-0.476
cpu.core					0.492	0.591	0.483	0.611	0.659	0.597	-0.699
cpu.freq						0.732	0.634	0.625	0.511	0.660	-0.558
log_inter nal_mem							0.879	0.685	0.768	0.719	-0.661
ram								0.648	0.711	0.645	-0.502
RearCam									0.726	0.490	-0.535
log_ front_cam										0.555	-0.647
log_ battery											-0.546
log_ thickness											

Figure 2.6 – Correlation coefficients for all regression variables in mobile phone data
Source: mobile.csv data

Statistical Analysis

The guiding inquiry of this report – the factors that affect mobile phone prices – and the salient linear relationships revealed in exploratory data analysis indicate that a linear regression model might be a good fit for the mobile phone data. A linear regression model relates some explanatory variables X_i against a response variable Y. The model informs us of (1) the linear estimates of each explanatory variable (i.e. the average rate of change for variable X_i given a 1-unit change in response Y), (2) the standard error of these estimates, (3) the estimate's associated T-value, and (4) the estimate's associated P-value. The basic structure of a multiple linear regression model is shown in figure 3.1.[4]

[4] Kalli, Maria. "Statistics for Data Analysis week 11 reading, multiple linear regression." (lecture reading, King's College London, London, England. 2023).

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} \epsilon_i$$

- β_j *for j = 0, 1, 2, ..., p are the true (but unknown) model parameters*
- ϵ *is the true (but unknown) random variation that cannot be explained by the regression model, i.e. error terms*

Figure 3.1 – Multiple linear regression model

Exploratory data analysis, particularly the scatterplots and correlation coefficient calculations, revealed that many explanatory variables in the mobile phone dataset share a linear relationship with the response variable price. To avoid model saturation and overfitting, it is imperative to proceed with a test based variable selection. For this analysis, the selected testing procedure is backwards elimination. Backwards elimination follows an iterative procedure[5]:

(1) Start with all predictor variables in the regression model (e.g. the full model)
(2) Remove the predictor variable X_i with the largest P-value
(3) Fit the new model and remove the next least significant predictor
(4) Continue this process until all "non-significant" predictors are removed

The result of backwards elimination is the final "best fit" regression model for a data set. For the mobile phone data set, the model resulting from backwards elimination fits the equation found in figure 3.2. The fitted model's results are summarized in figure 3.3.[i]

The tests for ß$_j$ coefficients are significant at a 0.05% level. Each explanatory variable in the fitted model has a p-value less than this, which implies a linear relationship. The smaller a variable's p-value is the stronger a relationship it shares with response variable Y. R-squared values indicate the *overall* model's linear fit, i.e. with respect to all variables X_i. Multiple R-squared (0.9429) and adjusted R-squared (0.9403) estimates are both quite close to 1. *This suggests that the selected linear regression model is a good fit for the mobile phone data because over 94% of variation in mobile phone price is explained by the phone features X_i included in the fitted model*. Additionally, we can be certain that the model is adequately fine-tuned via backwards elimination because the multiple R-squared and adjusted R-squared estimates are only marginally different from each other.

$$\begin{aligned}\mathbf{Price}_i = \beta_0 + \beta_1 \mathbf{log\ resolution}_i + \beta_2 \mathbf{ppi}_i + \beta_3 \mathbf{cpu\ core}_i + \\ \beta_4 \mathbf{log\ internal\ mem.}_i + \beta_5 \mathbf{ram}_i + \beta_6 \mathbf{log\ battery}_i + \\ \beta_7 \mathbf{log\ thickness}_i + \varepsilon_i\end{aligned}$$

Figure 3.2 – Fitted regression model for mobile phones data
Source: mobile.csv data

[5] Kalli, Maria. "Statistics for Data Analysis week 11 reading, multiple linear regression."

Variable	ß Estimate	Std. Error	T-value	P-value
Intercept	554.6318	537.1144	1.033	0.303414
log_resolution	-595.0041	159.4554	-3.731	0.000268
ppi	1.2874	0.1819	7.078	4.95E-11
cpu.core	42.3511	9.2945	4.557	1.06E-05
log_internal_mem	71.9766	34.8744	2.064	0.040718
ram	174.8146	23.6195	7.401	8.38E-12
log_battery	433.7596	97.5826	4.445	1.68E-05
log_thickness	-831.5841	123.3923	-6.739	3.05E-10

Residual Std. Error:	*187.8 on 153 degrees of freedom*
Multiple R-squared:	*0.9429*
Adjusted R-squared:	*0.9403*
P-value:	*< 2.2e-16*

Figure 3.3 – Summary of estimates for fitted linear regression model
Source: mobile.csv data

Before accepting the validity of and making inferences from the fitted model, diagnostic checks must be completed. Reasonable diagnostic checks for a linear regression model include residuals plots to check for random variation of the errors and qq-plots to confirm that errors are normally distributed.

The residuals plot in figure 3.4 shows the model's fitted values plotted against its residual values. The residuals appear to have near constant symmetrical variation and no discernable pattern in distribution. This implies constant variation and validates the results of the fitted regression model.

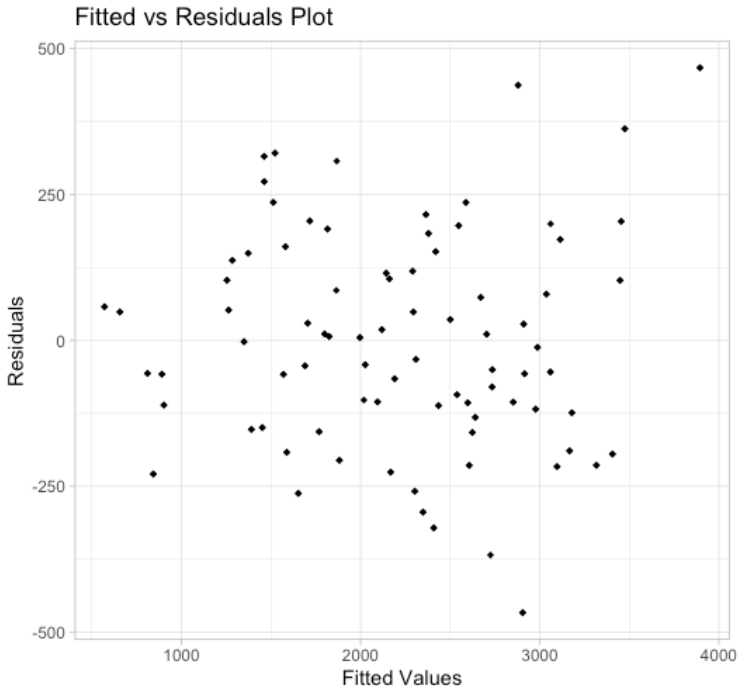

Figure 3.4 – Residuals plot for fitted regression model
Source: mobile.csv data

Figure 3.5 illustrates the fitted model's normal qq-plot. The qq-plot provides enough evidence that the residual errors follow a normal distribution. This outcome validates the results of the fitted regression model.

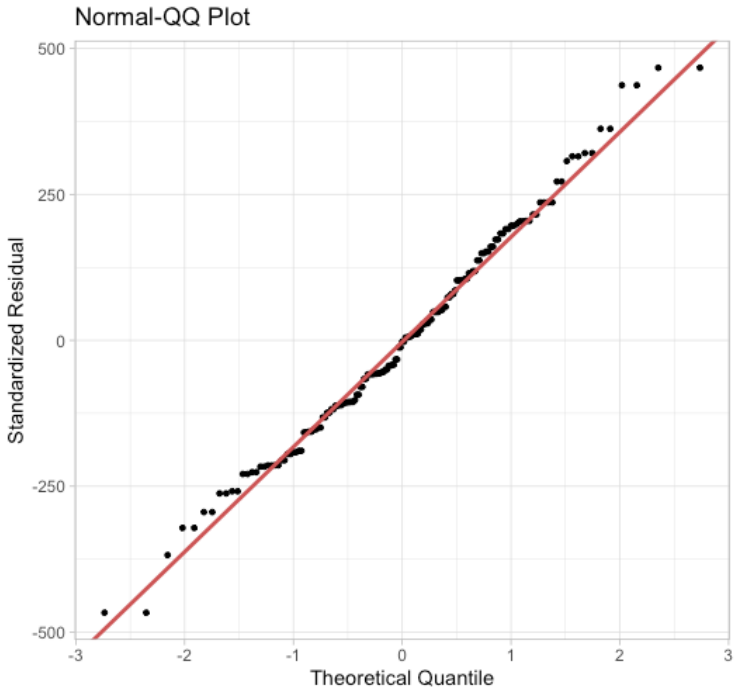

Figure 3.5 – qq-plot for fitted regression model
Source: mobile.csv data

The final test to check the fitted model is calculating the variable inflation factors. This is needed because the mobile phone dataset exhibits signs of collinearity between explanatory variables in exploratory data analysis. The variable inflation factor (VIF) provides a precise measure of regression estimate's standard error inflation caused by collinearity. A VIF measure larger than 10 is cause for concern. Due to depreciated R packages, VIF calculations were not completed on this model. However, based on the linear relationships indicated by multiple explanatory variables in exploratory data analysis scatterplots and correlation coefficients, it would not be shocking to discover that one or more variables in the fitted model have a VIF greater than 10.

Conclusion & Recommendations

This report aims to identify mobile phone features that most affect price based on the mobile phone dataset collected from a shop in Melbourne, Australia. Information discovered in exploratory data analysis along with the nature of the initial inquiry into the dataset "factors that affect phone prices" indicate a strong use case for a linear regression model. According to the fitted linear regression model, in order of magnitude, a mobile phone's ram, ppi, thickness, cpu core, battery, resolution, and internal memory linearly impact its price. When purchasing a new phone, it's recommended that you pay attention to these factors when assessing the fairness of a mobile phone's price.

Although the fitted model's diagnostic residuals and qq-plot imply a good fit, it is important to consider the potential impact collinearity might have. Possible collinearity was identified in exploratory data analysis's linear scatterplots and strong correlation coefficients between explanatory variables X_i. Unfortunately, due to depreciated R packages, variable inflation measures were not able to be assessed for the fitted model. In future modeling using this data, I recommend that VIF is calculated on a fitted model to provide definitive evidence of collinearity. If collinearity is proven, then regularization or normalization methods could help increase data interpretability.

Further, it is important to keep the limited scope of the analyzed data set in mind. Data was collected from a single store in Melbourne, Australia, and the data is only comprised of 161 observations. This is quite a small and specific sample population to make valid sweeping inferences from. In future analyses, layering in data from other areas of Australia and different countries could broaden and strengthen the outcomes of this research.

References & Endnotes

Gonzalez-Blanks, Ana, Jessie M. Bridgewater, and Tuppett M. Yates. “Statistical Approaches for Highly Skewed Data: Evaluating Relations between Maltreatment and Young Adults’ Non-Suicidal Self-Injury.” *Journal of Clinical Child & Adolescent Psychology* 49, no. 2 (2020): 147–61. https://doi.org/10.1080/15374416.2020.1724543.

Kalli, Maria. “Statistics for Data Analysis week 11 reading, multiple linear regression.” (lecture reading, King’s College London, London, England. 2023).

MacMillan, Andrew, David Preston, Jessica Wolfe, and Sandy Yu. “13.1: Basic Statistics- Mean, Median, Average, Standard Deviation, Z-Scores, and P-Value.” Engineering LibreTexts, March 11, 2023. https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Chemical_Process_Dynamics_and_Controls_(Woolf)/13%3A_Statistics_and_Probability_Background/13.01%3A_Basic_statistics-_mean_median_average_standard_deviation_z-scores_and_p-value#:~:text=The%20mean%2C%20median%20and%20mode,actual%20data%20and%20the%20mean.

Muldoon, Ariel. “The Log-0 Problem: Analysis Strategies and Options for Choosing C in Log(y + c).” Very statisticious, September 19, 2018. https://aosmith.rbind.io/2018/09/19/the-log-0-problem/#thinking-about-0-values.

[i] See associated coursework R script for complete backwards elimination procedure.