Linear Regression on Cellphone Price

By: Andy Hu, Tim Yang, Amanda Gao, Yingqian Shi

April 16, 2025

1. Introduction

1.1 Motivation

By analyzing key smartphone features, we can use linear regression to examine the relationships between these features and the price of a phone. This inferential approach provides valuable insights into how different phone specifications influence its price. For consumers, it allows them to understand the extent to which factors like weight, CPU performance, memory, and screen resolution contribute to the cost of a phone, helping them evaluate whether the price of any phone they have their eyes on is justified. For companies, this model reveals which features have the strongest relationships with price, enabling them to refine their pricing strategies, and make data-driven decisions regarding product offerings and prices.

Cellphone usage has skyrocketed in the past years as technology has evolved, where over 98% of Americans own a phone (Sidoti et al., 2024). The smartphone industry is massive, and there is a lot of competition between major brands like Apple and Samsung; these companies are competing for a share of the more than 1 billion phones sold globally each year (Laricchia, 2024). Our research seeks to understand how these companies determine their pricing strategies, offering consumers a clearer understanding of the factors that influence the costs of their devices.

1.2 Research Question

How are a cell phone's specifications, specifically its resolution (in ppi) and number of cores, associated with its final market price?

1.3 Dataset Description

We obtained our dataset from Kaggle, a data science and machine learning platform. The dataset contains various hardware and software features of mobile phones along with their corresponding selling prices. The data was collected through an observational study by recording existing phone specifications and market prices, rather than through a controlled experiment.

Variable Name	**Description**	**Type**
Price	Price of a phone (dollars)	Continuous
Weight	Weight of a phone (grams)	Continuous
PPI	Phone Pixel Density (pixels per inch)	Continuous
CPU Frequency	CPU Frequency clock speed (GHz)	Continuous
Battery	Capacity of battery (mAh)	Continuous
Thickness	Thickness of the phone (mm)	Continuous
Internal Memory	Memory in the phone (0GB, 4GB, 8GB, 16GB, 32GB, 64GB, 128GB, 256GB)	Categorical
CPU Cores	Number of Cores in CPU (0, 2, 4, 6, 8)	Categorical

1.4 Pre-selection of Variables

We excluded product ID and sales number because they are unique identifiers. We excluded resolution, RAM, rear cam, and front cam because we are unsure about their units. Moreover, some of the covariates seem to overlap with one another, like memory vs. RAM which seem to indicate the same thing, but are recorded in differing units in the dataset. We made educated guesses on units of the covariates above since they are not explicitly stated in the dataset.

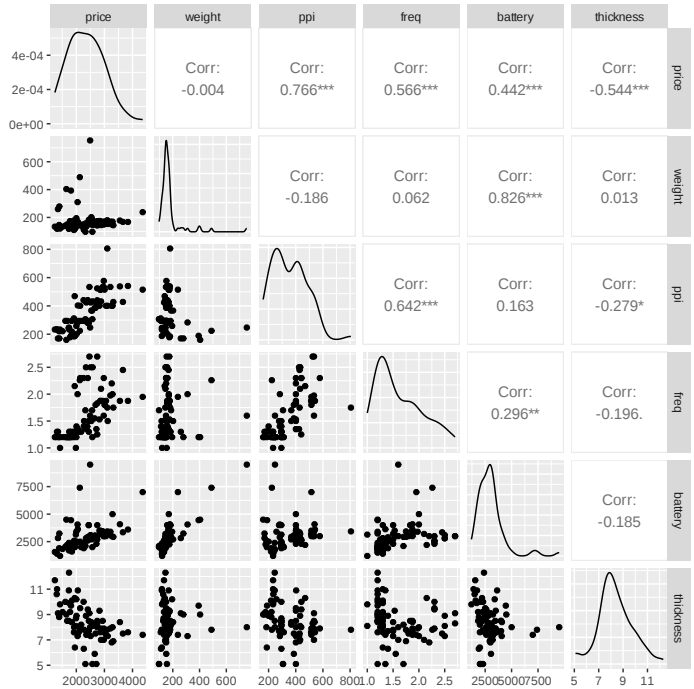
2. Analysis

2.1 Initial Data Cleaning

We first drop the columns specified in 1.1 Pre-selection of Variables as they will be irrelevant to our analysis. After exploring the primary statistics about the data and finding unique values for each of the columns, we discovered that there are no NA values. However, the memory column had unique values of 0.004, 0.128, 0.256, 0, 4, 8, 16, 32, 64, 128, which are obviously not all on the same scale. For memory values below 1, we multiply it by 1000 so that all memory measurements are on the same scale of gigabytes. We then discovered that some phones had 0 cores or 0 GB of memory, which would be outliers in the dataset and don't make sense in terms of cellphone characteristics, so we filter out rows with those properties. Finally, we delete duplicate entries in the dataset, bringing us to a total of 75 observations to work with and do our inference on. We are left with 7 covariates and 1 response (price).

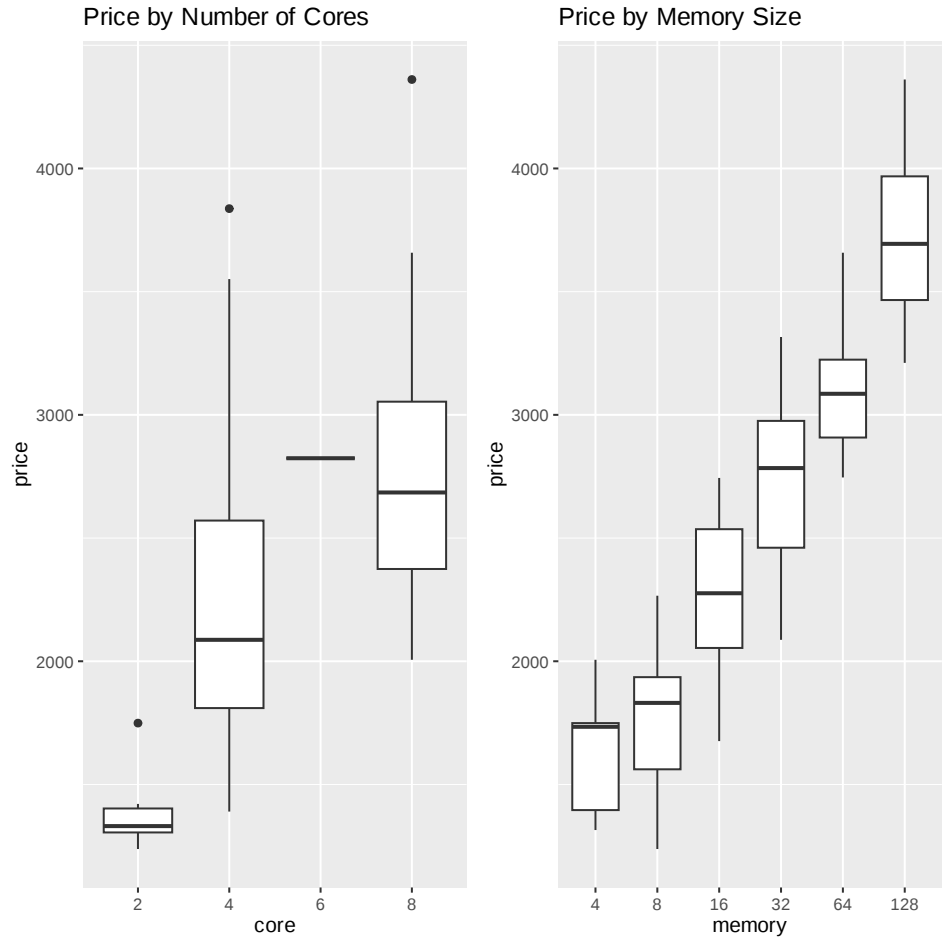
2.2 Exploratory Data Analysis

For all of the numerical columns in our data, we can plot their relationship in a pairwise manner. The diagonals below simply show the histogram distribution of each variable. The lower triangle shows scatter plots comparing each pair of variables, while the upper triangle shows the correlation coefficients, indicating the strength and direction of the relationship between the two variables.



The column on the far left shows pairwise plots between our response, the price, with all the other covariates. We notice that `weight` and `thickness` are negatively correlated with price, while `ppi`, `freq`, and `battery` are positively correlated with the price. Looking at the relationship between each covariate with the response, we see that the relationships are all linear. Also, we are wary that there may be signs of multicollinearity because many of the pairwise plots between covariates (between two variables that are not the price column) seem to have a relationship, such as the one between `battery` and `weight` (correlation of 0.826), or `freq` and `ppi` (correlation of 0.642).

The pairwise plots above only graphed our numerical variables, but we can also visualize the effects of our categorical covariates, `core` and `memory`, to the response of price.



We notice that for both `core` and `memory`, a higher level corresponds to a higher selling price. This makes sense as more cores and more memory lead to higher performance, which can make the price of a phone higher. We see that there is very little data for `core = 6` and that it is completely surrounded within the same bounds as `core = 8`, so we can combine them together in our data as well (merge `core = 6` into the `core = 8` level). This is done to simplify our model for inference.

2.3 Methodology and Assumptions

The objective is to regress on price, given a collection of different covariates describing a phone's characteristics. Since we are modelling a numerical response, then it follows that a linear regression model may be suitable for the task. We will incorporate an additive linear model because it is easier to interpret and simpler to isolate effects of different variables. Some assumptions we make in this linear model are listed below:

1. Linearity: The relationship between the price and the phone characteristics is linear.

2. Independence: Each phone example comes from an independent and identically distributed dataset. This means that the price of one phone is marketed independent of the price of another.

3. Homoscedasticity: The residuals exhibit constant variance. This means that the variability of the difference between the actual phone price and predicted phone price should stay roughly the same across different phone characteristics.

4. Normality: The residuals comes from a normal distribution. This means that the difference between the actual and predicted phone prices should follow a bell curve shape. This helps to ensure that the regression model works as expected and provides reasonable estimates for hypothesis testing and confidence intervals.

We will revisit some of these assumptions after fitting our model to verify model diagnostics and the suitability of a linear model. Some ways we can do this is to test homoscedasticity through a residual vs. fitted value plot, and we can also test normality through a QQ plot.

2.4 Model Selection

Recalling back to the exploratory data analysis we conducted, we should be wary of multicollinearity because many of the pairwise plots between covariates showed a strong relationship. This is something to be wary of because when two covariates are highly correlated, it can lead to unstable model coefficients that cancel each other out, and these coefficients will not be meaningful in terms of doing inference or drawing reasonable relationships between variables. One way to check for high multicollinearity is to look at the Variance Inflation Factor (VIF) scores. We begin by checking the VIF scores for a full model including all covariates.

	GVIF	Df	GVIF^(1/(2*Df))
weight	6.136215	1	2.477139
ppi	3.500995	1	1.871095
core	2.380969	2	1.242191
freq	2.041126	1	1.428680
memory	6.296669	5	1.202018
battery	8.236277	1	2.869891
thickness	1.569519	1	1.252804

We notice that `battery` has the highest standardized GVIF score of nearly 3, indicating high levels of multicollinearity. As such, we will remove it from the model and recheck the GVIF scores to see if there are any others to consider.

	GVIF	Df	GVIF^(1/(2*Df))
weight	1.281456	1	1.132014
ppi	3.471809	1	1.863279
core	2.153667	2	1.211420
freq	2.009113	1	1.417432
memory	4.086851	5	1.151168
thickness	1.564339	1	1.250735

We observe that the highest standardized GVIF score is 1.87 for the covariate `ppi`, which is acceptable and indicates low multicollinearity. With this confidence, we can now proceed with model selection.

We will run a stepwise model selection method which starts with a model that includes all covariates. It then goes through a series of steps, where it considers all models with one more variable or one less variable, and checks how well each new model fits the data. This continues until the model finds the model with the best score and terminates when no further improvements in the score can be made.

The score we will be using is the AIC score, where a lower AIC score indicates a better model fit. We will be using this because it not only describes how well the model fits the data, but also penalizes a model for being too complex; this is important because simpler models are preferred for inference unless the added complexity can be justified to significantly improve the model fit. Choosing the lowest (best) AIC score indicates a model that is both accurate and simple.

We will also keep track of R^2 and adjusted R^2 values, although this is not the criterion upon which we are evaluating different models. R^2 measures how well the model explains the variation in our data, while adjusted R^2 is the same but penalizes for more complex models similar to the AIC score. The table below shows the stepwise selection procedure:

Model	AIC Score	R^2	Adjusted R^2
Full model - battery	810.89	0.9146	0.8997
Full model - battery - weight	809.07	0.9144	0.901

From stepwise selection based on the AIC criterion, we reach our final reduced model which removed weight (from the already removed battery due to its high standardized GVIF).

We see that not only did the AIC score improve, but the adjusted R^2 value went up. This means that removing the `weight` covariate did not significantly reduce the model's ability to explain the data, making its removal justifiable.

When we went to the final model with less covariates, the R^2 value went down; this is expected because this metric does not penalize complexity, so a more complex model will always have an R^2 value greater than or equal to one of its nested model. After all, more covariates offers more explanatory power, even if not by much.

We can also calculate Mallow's C_p value, which compares a model's fit to the perfect model while also accounting for complexity. We summarize the main model statistics in the two tables below:

Metric	**Value**
R^2	0.9144
Adjusted R^2	0.901

Metric	**Value**	**Expected**
Mallow's C_p	10.153	11

2.5 The Final Model

Our final model regressed price on 5 covariates: `ppi`, `core`, `freq`, `memory`, and `thickness` (`battery` and `weight` were removed from the original 7 covariates).

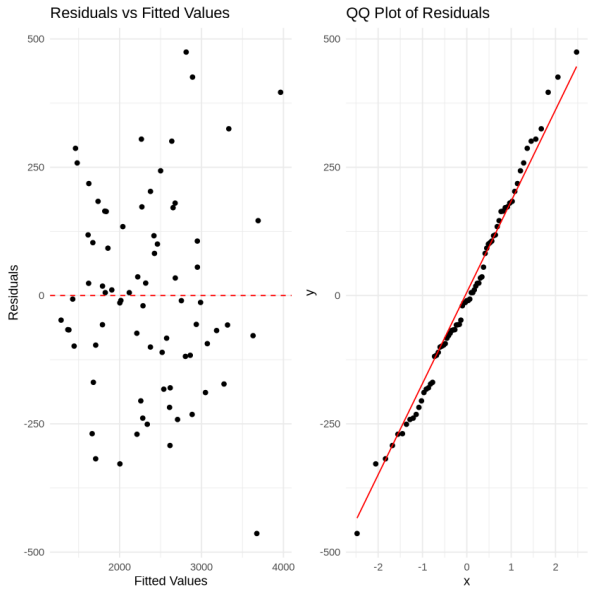
Coefficient	Estimate	Significant at 5% Level?
(Intercept)	1727.1527	Yes
ppi	1.1896	Yes
core4	226.1831	Yes
core8	506.0274	Yes
freq	138.2568	No
memory8	-31.1117	No
memory16	154.8842	No
memory32	409.2588	Yes
memory64	798.9093	Yes
memory128	1389.1719	Yes
thickness	-72.9221	Yes

The coefficient column is the covariate name. The estimate column gives the estimated effect that the covariate has on the price. The significance at a 5% level indicates if a covariate has has a real and significant effect on the price, and the relationship is not just random chance.

2.6 Model Diagnostics

	GVIF	Df	GVIF^(1/(2*Df))
ppi	2.786128	1	1.669170
core	2.101039	2	1.203950
freq	1.907578	1	1.381151
memory	3.664452	5	1.138678
thickness	1.540315	1	1.241094

We do not have high standardized GVIF values, letting us disregard high multicollinearity issues. We now verify the linear model assumptions about homoscedasticity and normality of residuals.



On the left, we plotted residuals vs. fitted value, and the plot shows no trends or noticeable relationship; this random scatter of points around 0 means that the variability of the residuals is consistent across different examples and phone characteristics, allowing us to conclude homoscedasticity of the residuals (constant variance).

By analyzing the QQ plot on the right, we observe that the model's errors match a normal distribution fairly well. In a QQ plot, the points represent the quantiles of our model's prediction errors, and the diagonal qqline shows the expected quantiles if the errors were normal. Since our points closely follow the line and there are no noticeable patterns, we conclude that the errors are normally distributed.

3. Conclusion

Summary

We first analyzed our data by pre-selecting variables and disregarding irrelevant ones such as ID or sales number fields. Then, we did initial data pre-processing alongside exploraty data analysis, which included dropping duplicates, transforming data to the same scale, and encoding categorical data as factors in R. We then checked for multicollinearity issues and used the AIC criterion to do model selection, resulting in a model that regresses on `ppi`, `core`, `freq`, `memory`, and `thickness`. We then underwent model diagnostics to verify our initial assumptions about homoscedasticity of residuals and if they come from a normal distribution. A summary of the main statistics related to the final model is given below:

Metric	**Value**
R^2	0.9144
Adjusted R^2	0.901

Metric	**Value**	**Expected**
Mallow's C_p	10.153	11

Being able to explain 90% (adjusted R^2) of the variability in price after being penalized is impressive, and our C_p value is nearly as expected, indicating a relatively good model fit.

Findings

From our final model, we discover that the covariates which contribute positively to the price are `ppi` (screen resolution), `core` (both 4 and 8 cores), `freq`, `memory` (all levels but 8GB). On the other hand, `thickness` and `memory8` are negatively associated with the price. It is worth noting that `freq`, `memory8`, and `memory16` are all statistically insignificant at the 5% level. This means that despite their positive or negative contribution to the price predictions of a cell phone, we could also choose to ignore their effects.

In addressing our original research question, which focused on the effects of screen resolution (ppi) and the number of cores, we find that `ppi`, `core4`, and `core8` remain significant even after model selection. Their p-values are as follows: 0.00036 for `ppi`, 0.040 for `core4`, and 0.00011 for `core8`. An interpretation of the coefficients is that each additional more pixel per inch in screen resolution is associated with a \$1.19 increase in the price. Relative to a baseline level of two cores, 4 cores is associated with an increase in price by \$226 while 8 cores is associated with an increase in price by \$506. These findings make sense given the nature of the variables: more cores enhance the

phone's performance and require more advanced hardware, while a higher screen resolution results in a sharper display capable of handling high-quality images.

Business Implications

From the customer's perspective, our model helps customers understand how different features influence a phone's price. Knowing this, buyers can evaluate whether a phone is priced unusually high or low without the strong features to support it. From the company's perspective, manufacturers can use these insights to fine-tune their pricing and product design strategies. Features like higher screen resolution and CPU cores have a strong positive association with price, which supports using them as key differentiators in premium models in segmenting product lines.

Limitations

As the data was collected from an observational study, there is a risk of confounding variables not included in our analysis, such as brand (Apple vs Android), age of the phone (newer phones are more expensive than older phones in general). Since we used an additive regression model, we overlooked potential interaction effects between variables, which could affect the model's explanatory power. Also, as technology evolves over time, prices of phones based on current tech can change. In fact, phone specifications may evolve such that the different levels of `core` and `memory` in our model may become outdated in the future.

Future Research Questions

Future research could explore confounding variables such as including brand or camera quality, and could also explore interaction effects to reveal relationships between covariates. As prices related to certain technological specs could evolve over time, a study could also consider temporal effects and time series data to predict prices of phones in the future.

4. References

Laricchia, F. (2024, October 15). *Number of smartphones sold to end users worldwide from 2007 to 2023*. Statista. https://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since-2007/

Sidoti, O., Dawson, W., Gelles-Watnick, R., Faverio, M., Atske, S., Radde, K., & Park, E. (2024, November 13). *Mobile Fact Sheet*. Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/mobile/

Younesi, F. (2022, August 21). *Mobile Price Prediction*. Kaggle. https://www.kaggle.com/datasets/mohannapd/mobile-price-prediction/data