Fondamenti di Analisi dei Dati

from data analysis to predictive techniques

Prof. Antonino Furnari (antonino.furnari@unict.it)
Corso di Studi in Informatica
Dip. di Matematica e Informatica
Università di Catania

Università di Catania

Linear Regression

A fundamental statistical technique for modelling relationships between variables and making predictions on unseen data. Linear regression allows us to understand how changes in independent variables influence target variables, making it invaluable for trend analysis, predictive modeling, and hypothesis testing.

The Auto MPG Dataset

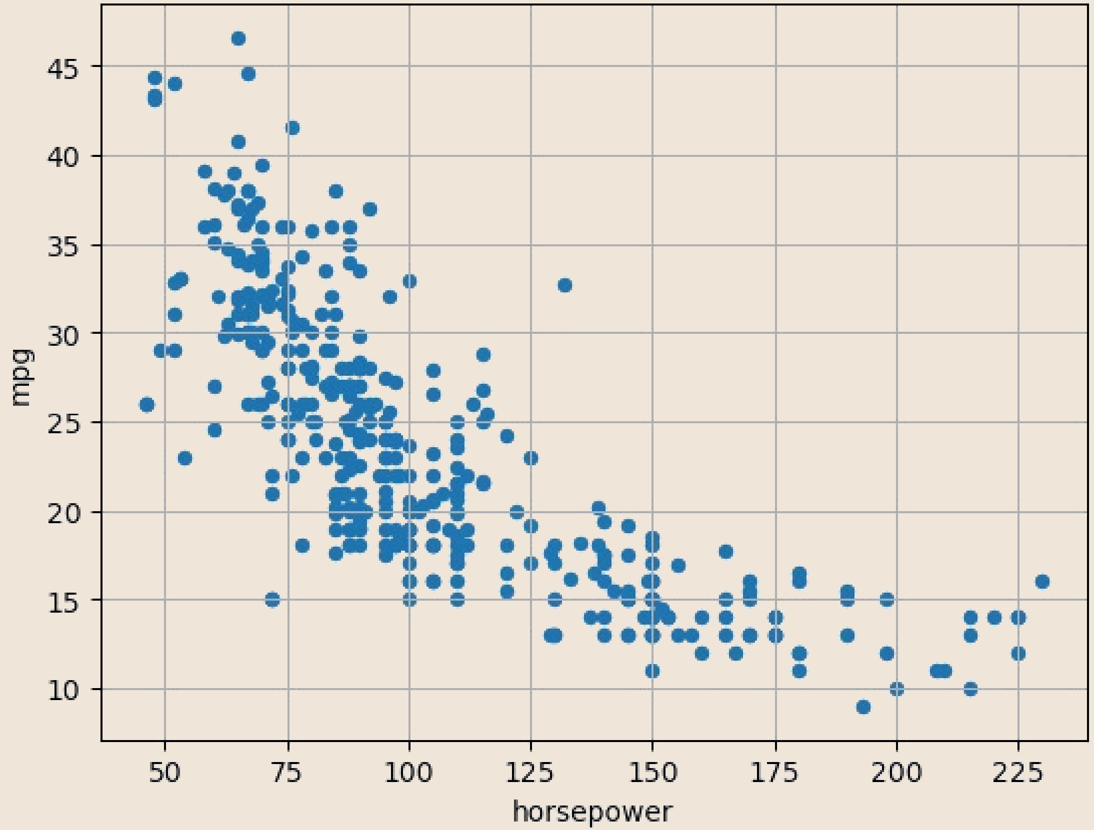
We'll explore linear regression using the Auto MPG dataset, containing 398 measurements across 8 variables that describe car characteristics and fuel efficiency.

	displacement	cylinders	horsepower	weight	acceleration	model_year	origin	mpg
0	307.0	8	130.0	3504	12.0	70	1	18.0
1	350.0	8	165.0	3693	11.5	70	1	15.0
2	318.0	8	150.0	3436	11.0	70	1	18.0
3	304.0	8	150.0	3433	12.0	70	1	16.0
4	302.0	8	140.0	3449	10.5	70	1	17.0
...	...	...	...	...	...	...	...	...
393	140.0	4	86.0	2790	15.6	82	1	27.0
394	97.0	4	52.0	2130	24.6	82	2	44.0
395	135.0	4	84.0	2295	11.6	82	1	32.0
396	120.0	4	79.0	2625	18.6	82	1	28.0
397	119.0	4	82.0	2720	19.4	82	1	31.0

Visualizing the Relationship

Before building models, we explore the relationship between **horsepower** and **MPG** through visualization. The scatterplot reveals a clear negative correlation: as horsepower increases, fuel efficiency decreases.

This inverse relationship makes intuitive sense—**more powerful engines typically consume more fuel**. Linear regression will help us quantify this relationship precisely.



The Regression Model Framework

Regression models study relationships between variables by defining a mathematical model where

$$Y = f(X) + \epsilon$$

Here, f is a deterministic function predicting Y from X, while ϵ captures everything not explained by f:

Model Imperfection

Our estimated function may not perfectly capture the true underlying process, leading to prediction errors.

Missing Variables

Y may depend on variables beyond X that we haven't observed or included in our model.

Inherent Randomness

Some processes contain stochastic elements that cannot be fully captured by deterministic functions.

Having the analytical form of f is powerful—it allows us to deeply understand the connection between variables, not just make predictions.

Simple Linear Regression

Simple linear regression models the linear relationship between two variables X and Y.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

In our example, we model MPG from horsepower using the equation:

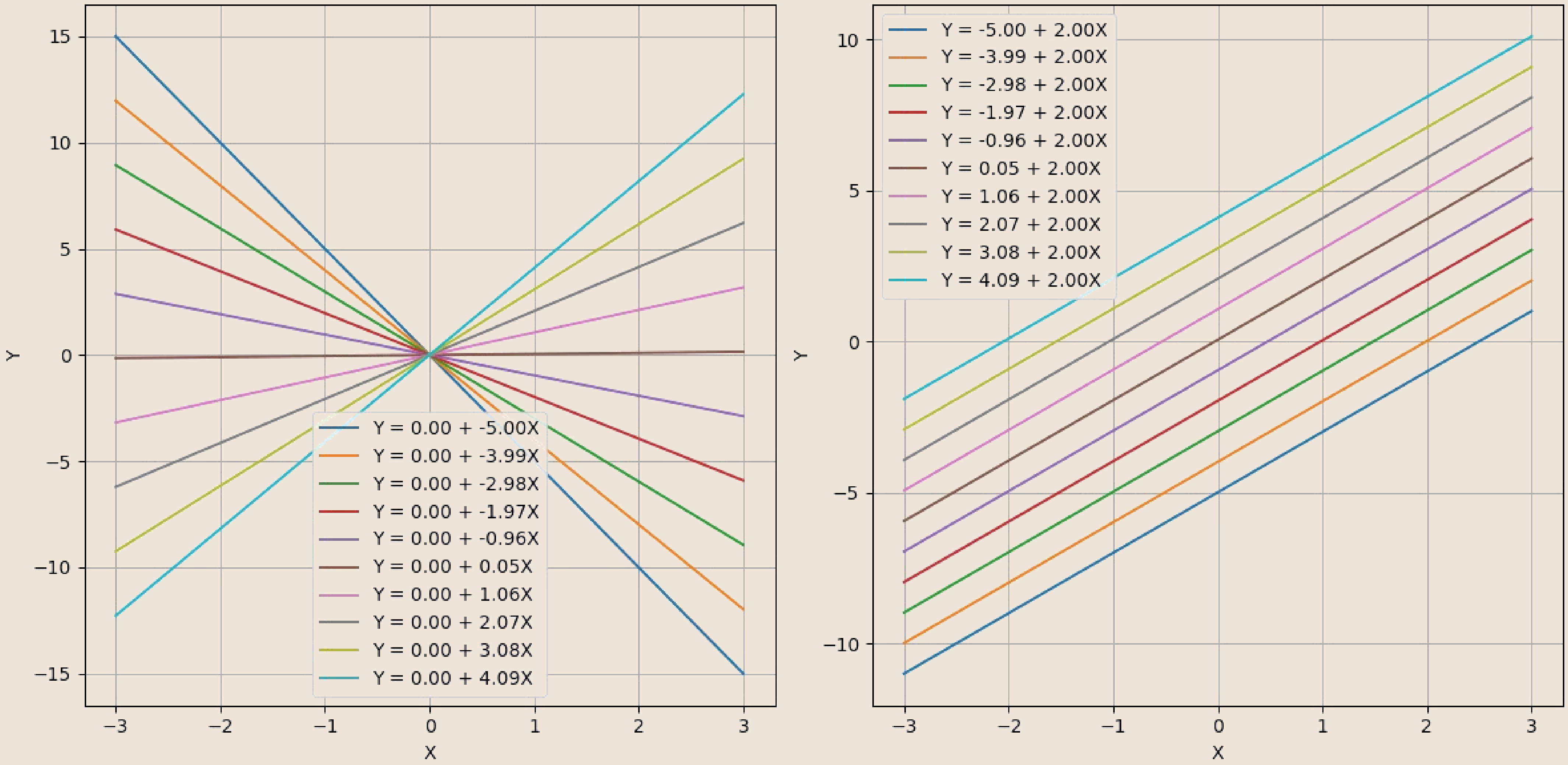
$$mpg = \beta_0 + \beta_1 \cdot horsepower + \epsilon$$

Key Components

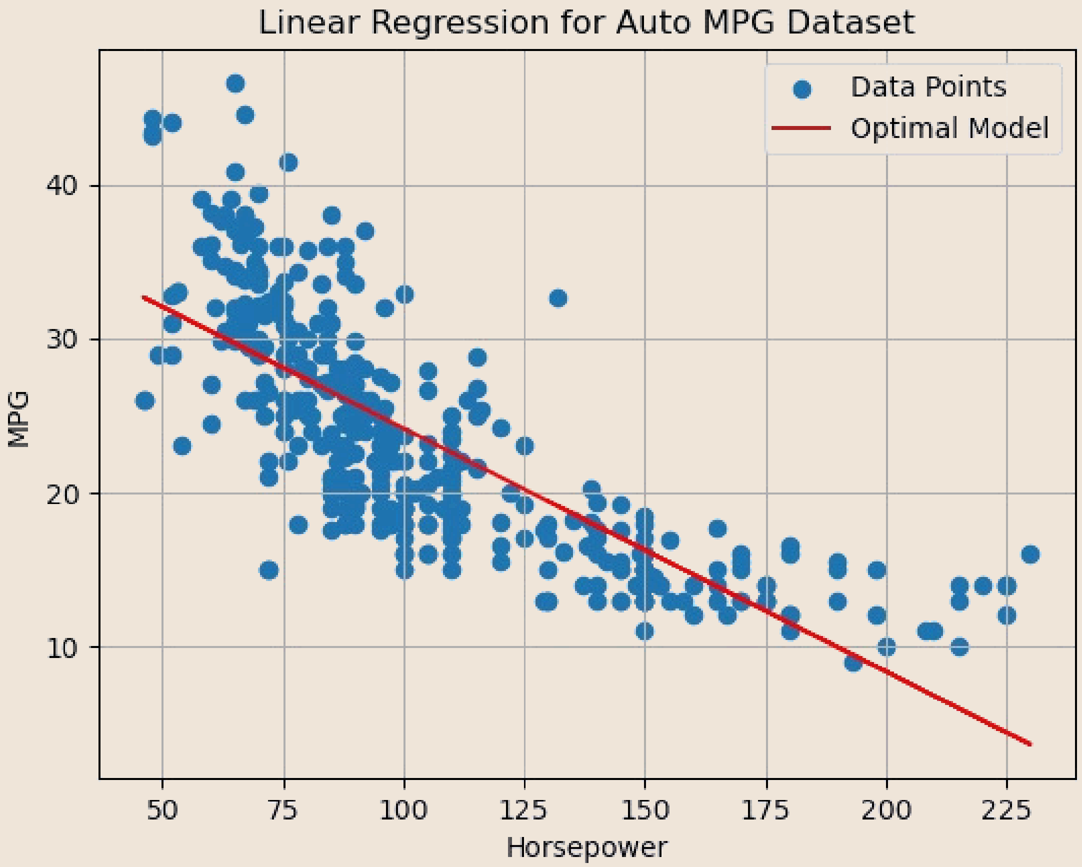
- **X**: Independent variable (regressor)
- **Y**: Dependent variable (regressed variable)
- **β_0, β_1**: Coefficients or parameters
- **ε**: Error term (expected to be small and random)

Geometric Interpretation

Specific values of β_0 and β_1 identify a line in the 2D plane. Our goal is to find the line that best represents our data.



Finding the Best Fit Line



We estimate coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ from data to create a model that represents our observations well. The resulting regression line captures the linear relationship between horsepower and MPG.

Ordinary Least Squares (OLS)

To estimate optimal coefficients, we define a good model as one that predicts Y well from X. Given the i^{th} realization of X, we define:

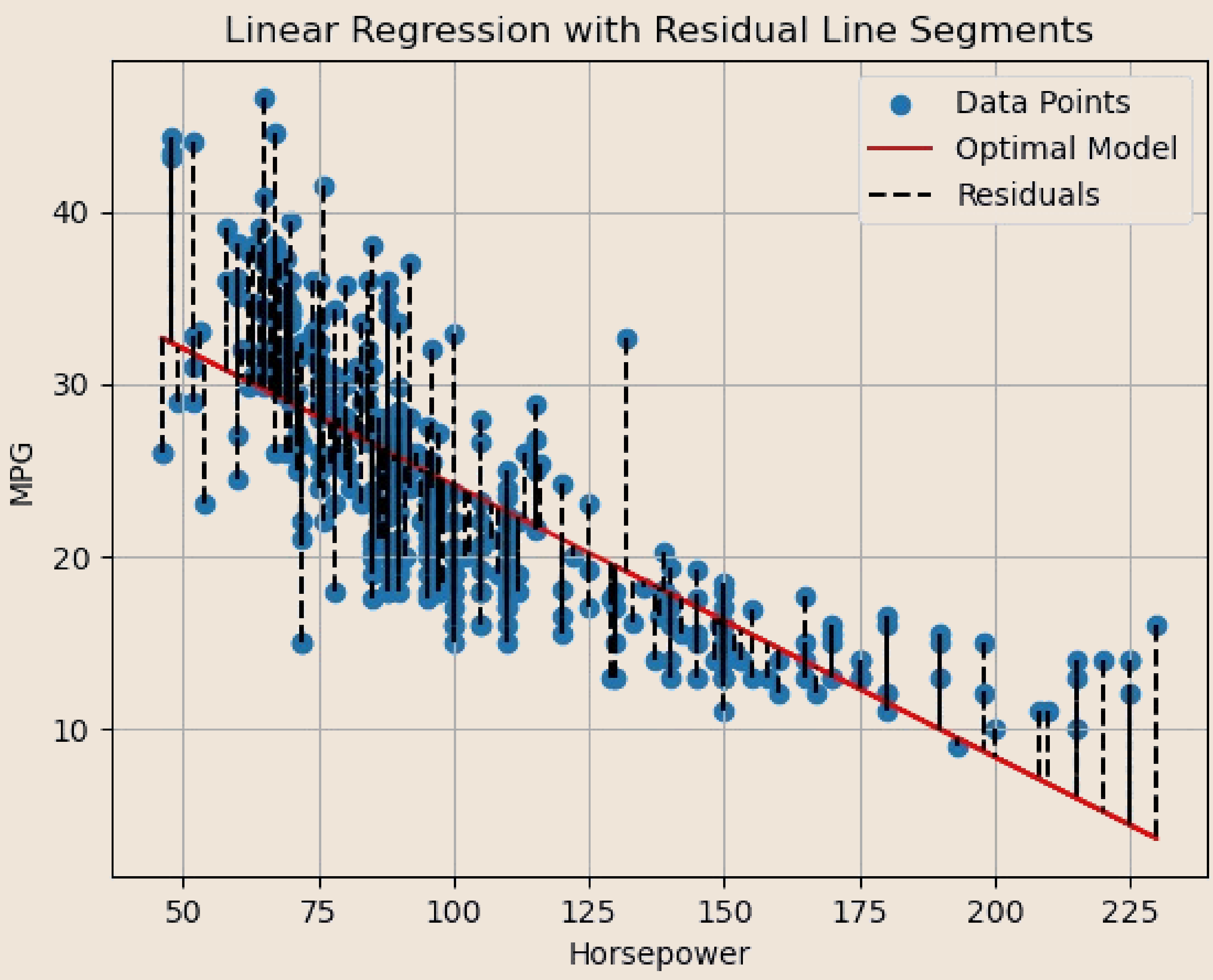
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

For each observation, we calculate the residual:

$$e_i = y_i - \hat{y}_i$$

The Residual Sum of Squares (RSS) serves as our global error indicator:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$



By minimizing RSS, we find the line that best fits the data. This optimization yields formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize prediction errors:

First, we write the RSS as a function of the parameters, keeping the data fixed:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We aim to find:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} RSS(\beta_0, \beta_1)$$

The minimum can be found setting:

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = 0 \qquad \frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Doing the math, it can be shown that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Empirical Risk Minimization

Empirical Risk Minimization (ERM) offers an alternative perspective on finding optimal model coefficients. It's a fundamental principle in machine learning, particularly for supervised learning, where the goal is to learn a function that performs well on unseen data.

If we use residuals as a **cost function** (or loss function) to quantify the error for a single data point:

$$L(h(x_i), y_i) = (h(x_i) - y_i)^2 = e_i^2$$

The **Empirical Risk** represents the average loss over our observed training data:

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} (h(x_i) - y_i)^2 = \frac{1}{N} RSS$$

Minimizing this empirical risk leads directly back to our previous approach:

$$\hat{h} = \underset{h \in \mathcal{H}}{\arg\min}\, R_{emp}(h) = \underset{h \in \mathcal{H}}{\arg\min}\, RSS$$

From a learning perspective, solving this optimization problem means finding the optimal set of parameters β that minimize the average prediction error, precisely what Ordinary Least Squares (OLS) achieves by minimizing the Residual Sum of Squares.

Interpreting the Coefficients

For our Auto MPG example, OLS produces these estimates:

$$y = 39.94 - 0.15x$$

39.94

Intercept (β_0)

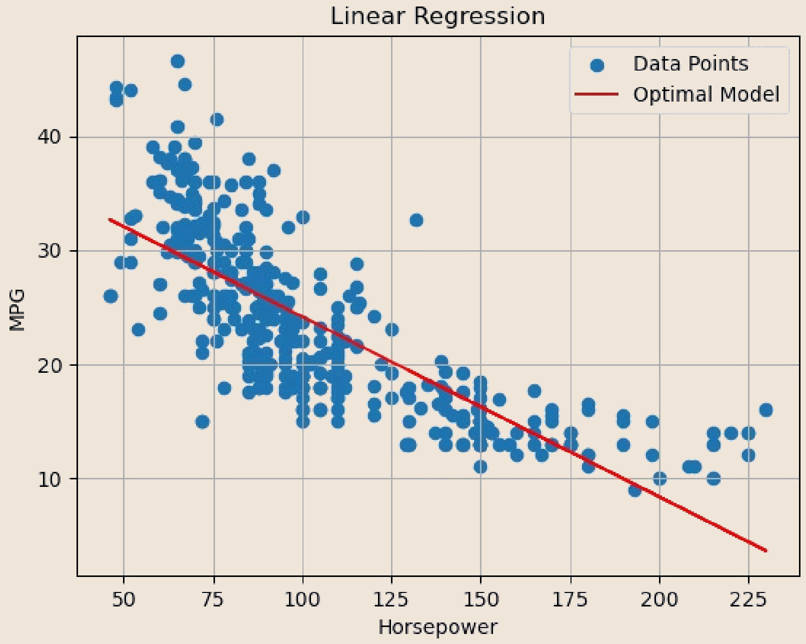
Predicted MPG when horsepower equals zero

-0.16

Slope (β_1)

Change in MPG per unit increase in horsepower

The negative slope reveals that each additional unit of horsepower is associated with a 0.16 decrease in MPG. This quantifies the fuel efficiency trade-off for more powerful engines.

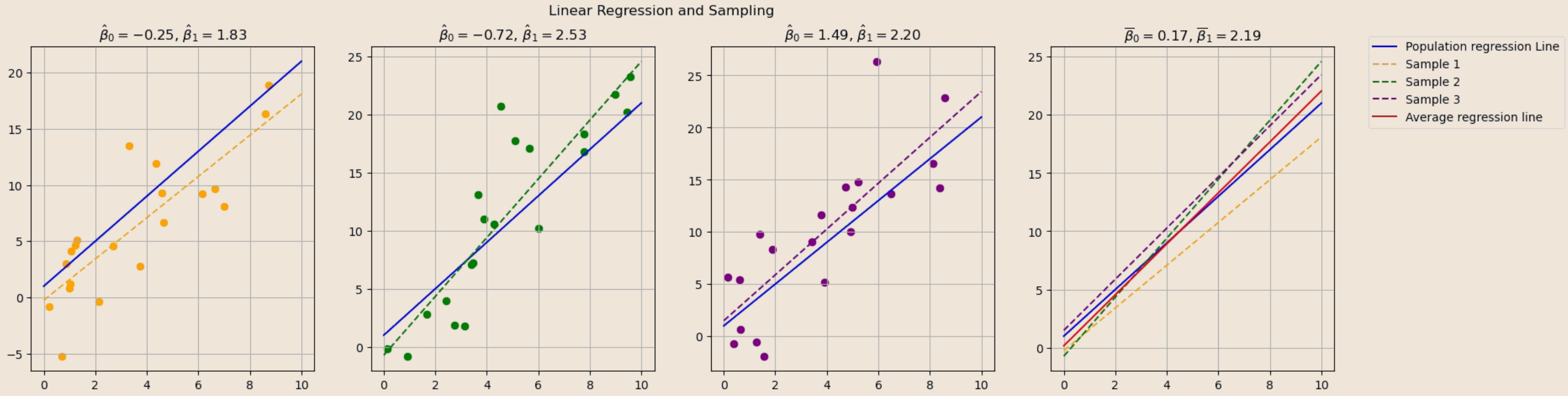


Coefficient Accuracy and Confidence

Our coefficient estimates are derived from sample data and have inherent variability. Different samples from the same population yield different regression lines, though the average line approaches the true population relationship. Let us consider an ideal population for which:

$$Y = 2x + 1$$

Ideally, given a sample from the population, we expect to obtain $\hat{\beta}_0 \approx 1$ and $\hat{\beta}_1 \approx 2$. In practice, different samples may lead to different estimates and hence different regression lines, as shown in the plot below:



Confidence Intervals and Standard Errors of Coefficients

In practice, we can see formulas used to estimate coefficients as **estimators**. Hence, we can compute standard errors and confidence intervals.

For instance, for our model

$$horsepower \approx \beta_0 + mpg \cdot \beta_1$$

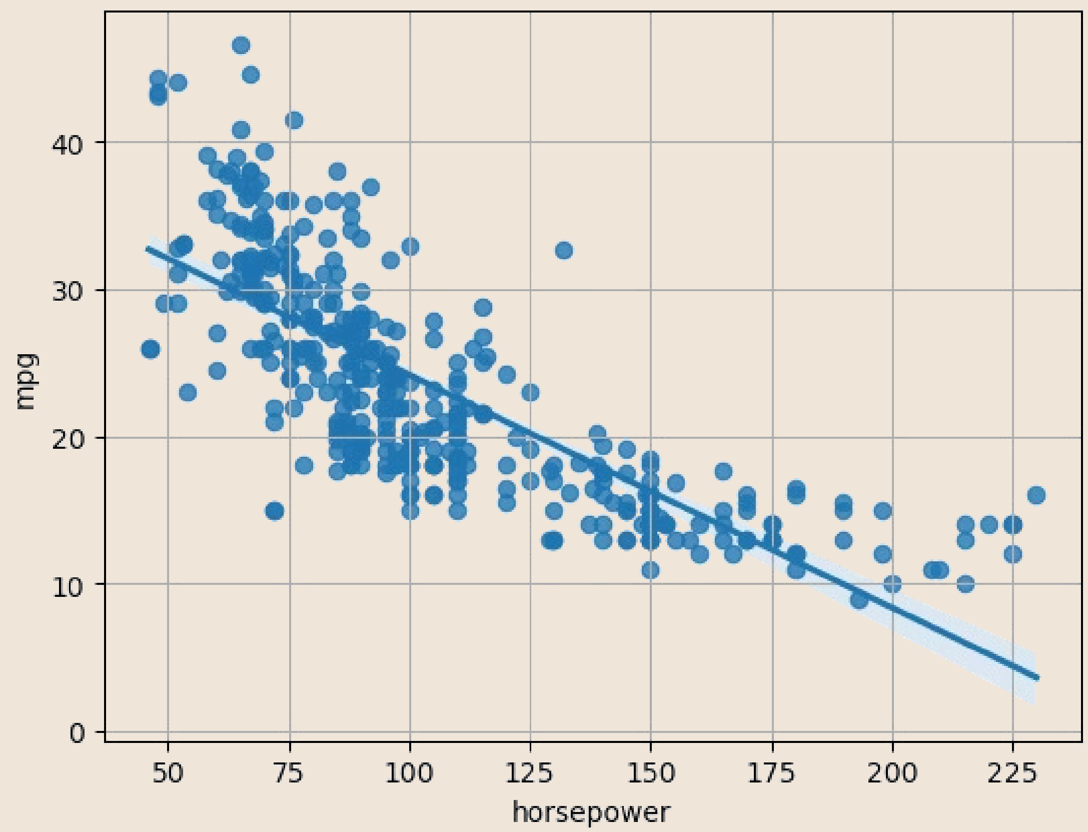
We compute confidence intervals to quantify this uncertainty:

Coefficient	Estimate	Std Error	95% CI
β_0	39.94	0.717	[38.53, 41.35]
β_1	-0.1578	0.006	[-0.17, -0.15]

From the table above, we can say that:

- The value of `mpg` for `horsepower=0` lies somewhere between 38.53 and 41.35;
- An increase of `horsepower` by one unit is associated to a decrease of `mpg` between -0.17 and -0.15.

It is also common to see plots like the following one:



In the plot , the shading around the line illustrates the variability induced by the confidence intervals estimated for the coefficients.

Statistical Significance Testing

We perform hypothesis tests to determine if coefficients are significantly different from zero. For β_1, the null hypothesis states there's no association between X and Y. Indeed, we would have $Y = \beta_0$.

1

Null Hypothesis

H_0: $\beta_1 = 0$ (no relationship exists)

H_0: $\beta_1 \neq 0$ (there is association between X and Y)

2

Calculate t-statistic

Compare estimate to standard error

3

Compute p-value

Probability of observing this result if H_0 is true

4

Decision

Reject H_0 if $p < 0.05$

Example for $horsepower \approx \beta_0 + mpg \cdot \beta_1$:

Coefficient	Estimate	t-stat	p-value	95% CI
β_0	39.94	55.66	0.000	[38.53, 41.35]
β_1	-0.1578	-24.49	0.000	[-0.17, -0.15]

Both coefficients are highly significant ($p \approx 0$), confirming a real relationship between horsepower and MPG.

Assessing Model Accuracy

Statistical tests tell us *if* a relationship exists, but not how well the model fits or predicts. We use two categories of metrics:

Goodness-of-Fit Metrics

Measure how well the model explains training data (statistical approach)

- Residual Standard Error (RSE)
- R^2 (Coefficient of Determination)

Predictive Accuracy Metrics

Measure performance on new, unseen data (machine learning approach)

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

All metrics build upon residuals—the differences between observed and predicted values.

Key Performance Metrics

Residual Standard Error

RSE estimates the standard deviation of the error term ε, measuring typical prediction error in Y's units.

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

We divide by the "degrees of freedom" ($n-2$) because we estimated *two* parameters (β_0 and β_1). This makes RSE an unbiased estimate of σ (the standard deviation of ϵ).

> **Interpretation:** RSE is an **absolute measure** of the model's "lack of fit," expressed in the **same units as** Y. A smaller RSE means the model fits the data better.

For our MPG model: RSE = 4.91, about 20% of the mean MPG value (23.52)

R^2 Statistic

R^2 measures the proportion of variance explained by the model, ranging from 0 to 1.

$$R^2 = 1 - \frac{RSS}{TSS}$$

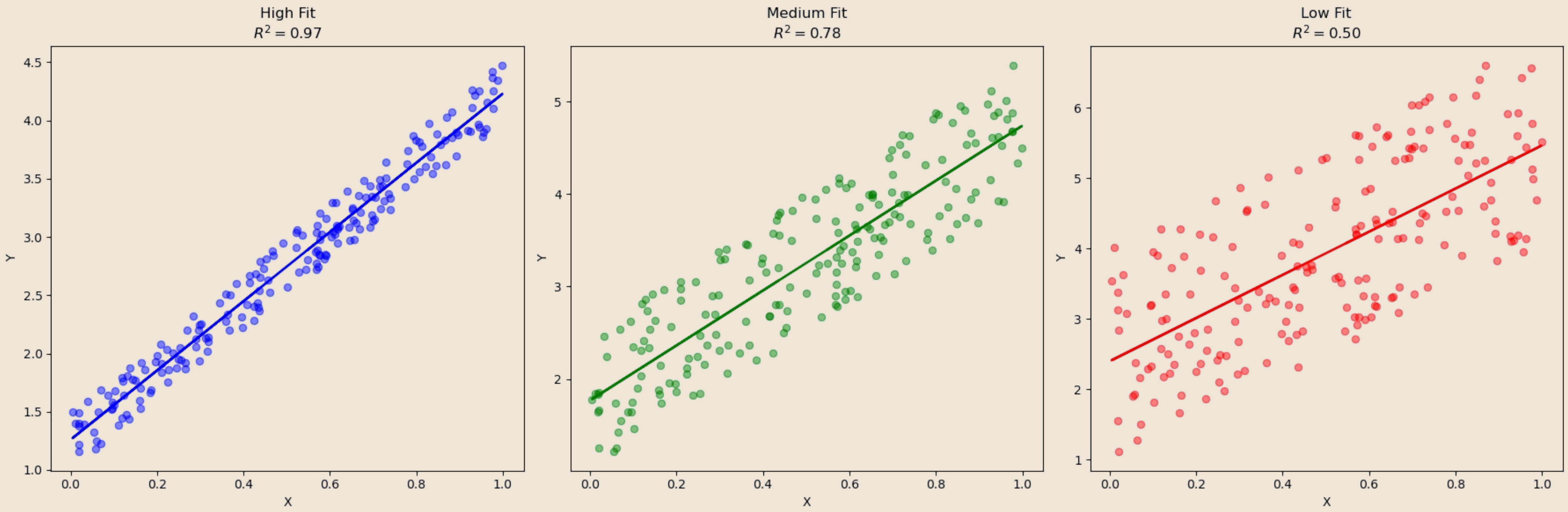
For our model: $R^2 = 0.61$, meaning horsepower explains 61% of MPG variance. For simple linear regression (one X), R^2 is also equal to the square of the Pearson correlation coefficient: $R^2 = \rho^2$.

The total sum of square TTS is the performance of a baseline model predicting the average model $\overline{y}$:

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Visual Examples of Model Fit

Different models achieve varying levels of fit quality. The plots below illustrate how RSS, RSE, and R^2 relate to visual fit:

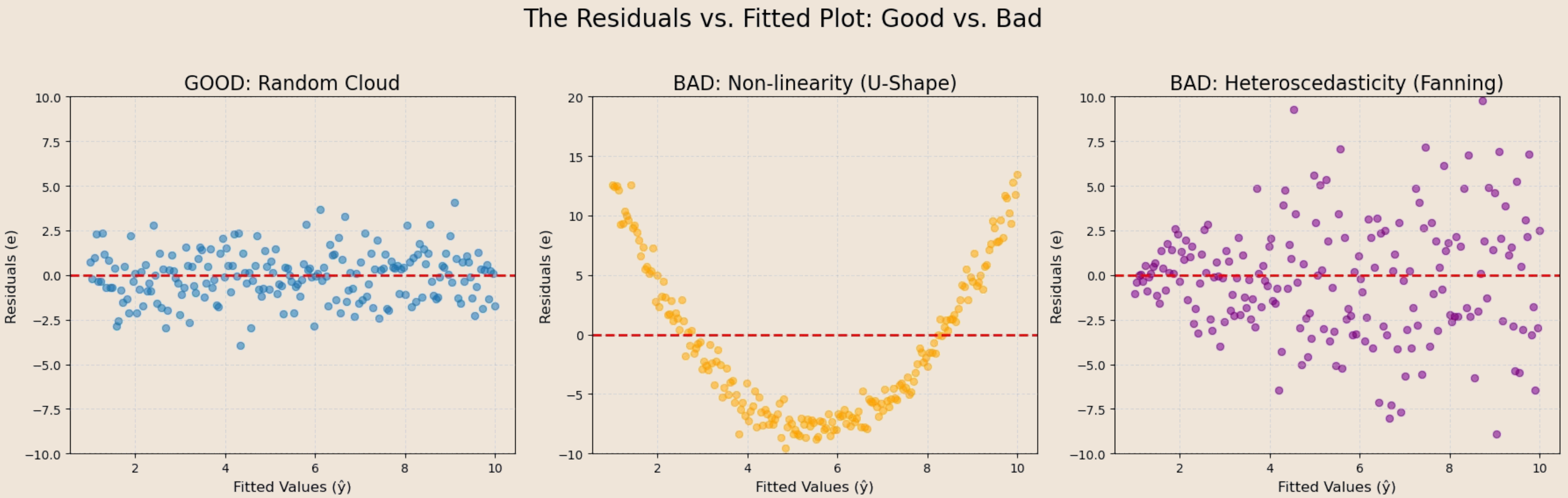


Better fits show smaller residuals (shorter vertical lines), lower RSS and RSE values, and higher R^2 values approaching 1.0.

Diagnostic Plots: Residual vs Fitted Plot

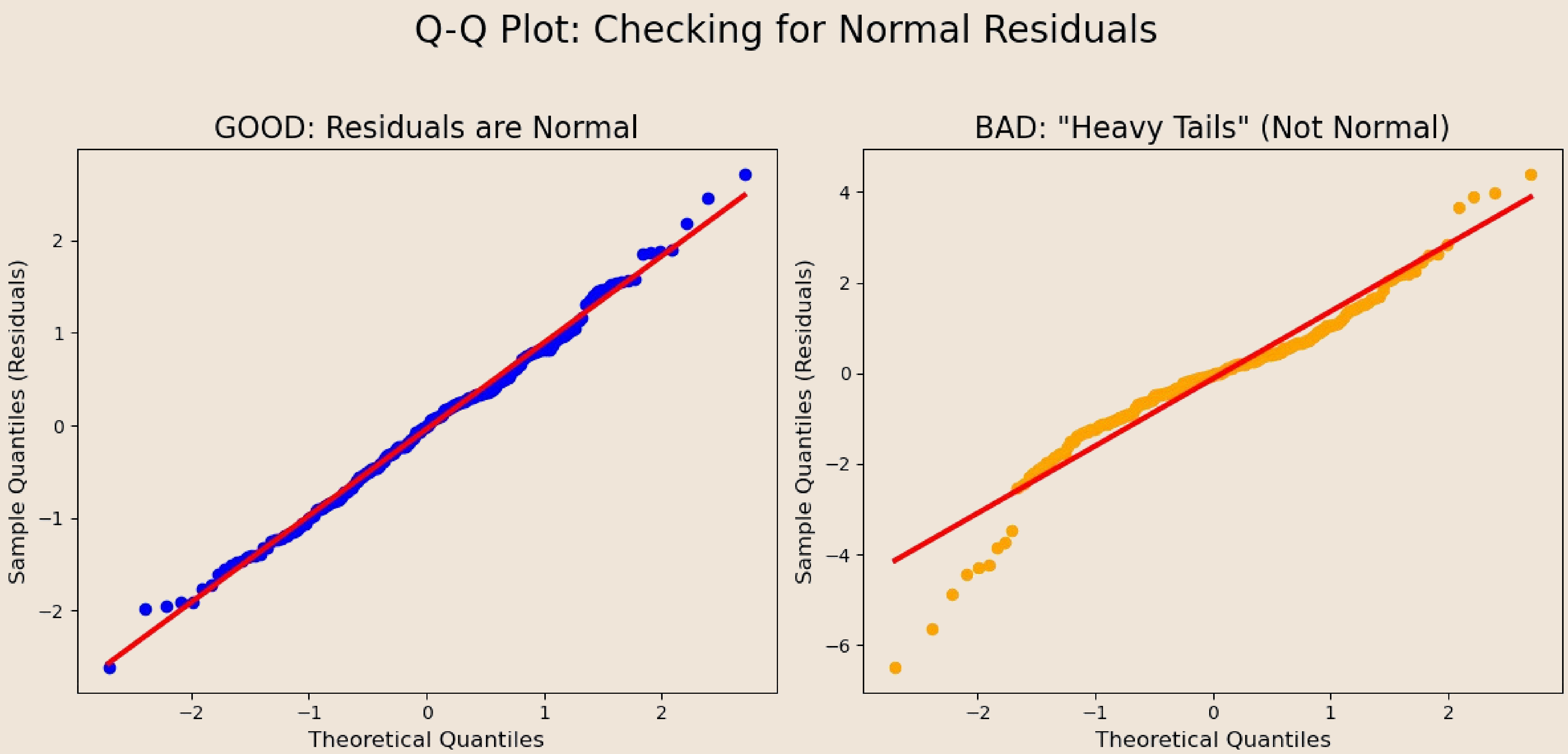
Scoring metrics like R^2 don't reveal *why* errors occur or if the model is fundamentally flawed. We use residual plots as diagnostic tools to validate model assumptions.

> We plot the **Residuals (e_i)** on the y-axis against the **Fitted (Predicted) Values ($\hat{y}_i$)** on the x-axis. Any deviation from a random cloud indicates a deviation from a linear model.



Diagnostic Plots: Q-Q Plot

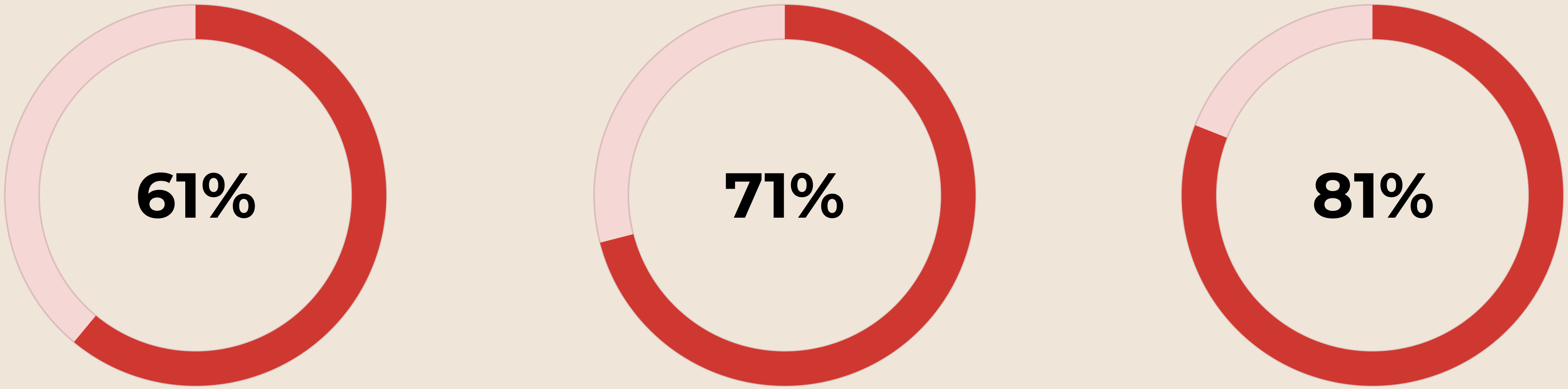
If the relationship is linear, the errors are distributed according to a Gaussian. We can check this with a Q-Q plot comparing sample quantiles to theoretical ones as shown below:



Multiple Linear Regression

Our simple model explains only 61% of MPG variance. The remaining 39% may be due to stochasticity, non-linearity, or missing variables. We can improve predictions by including additional predictors:

$$mpg = \beta_0 + \beta_1 \cdot horsepower + \beta_2 \cdot weight$$



Simple Model R^2

Horsepower alone

Two-Variable R^2

Adding weight

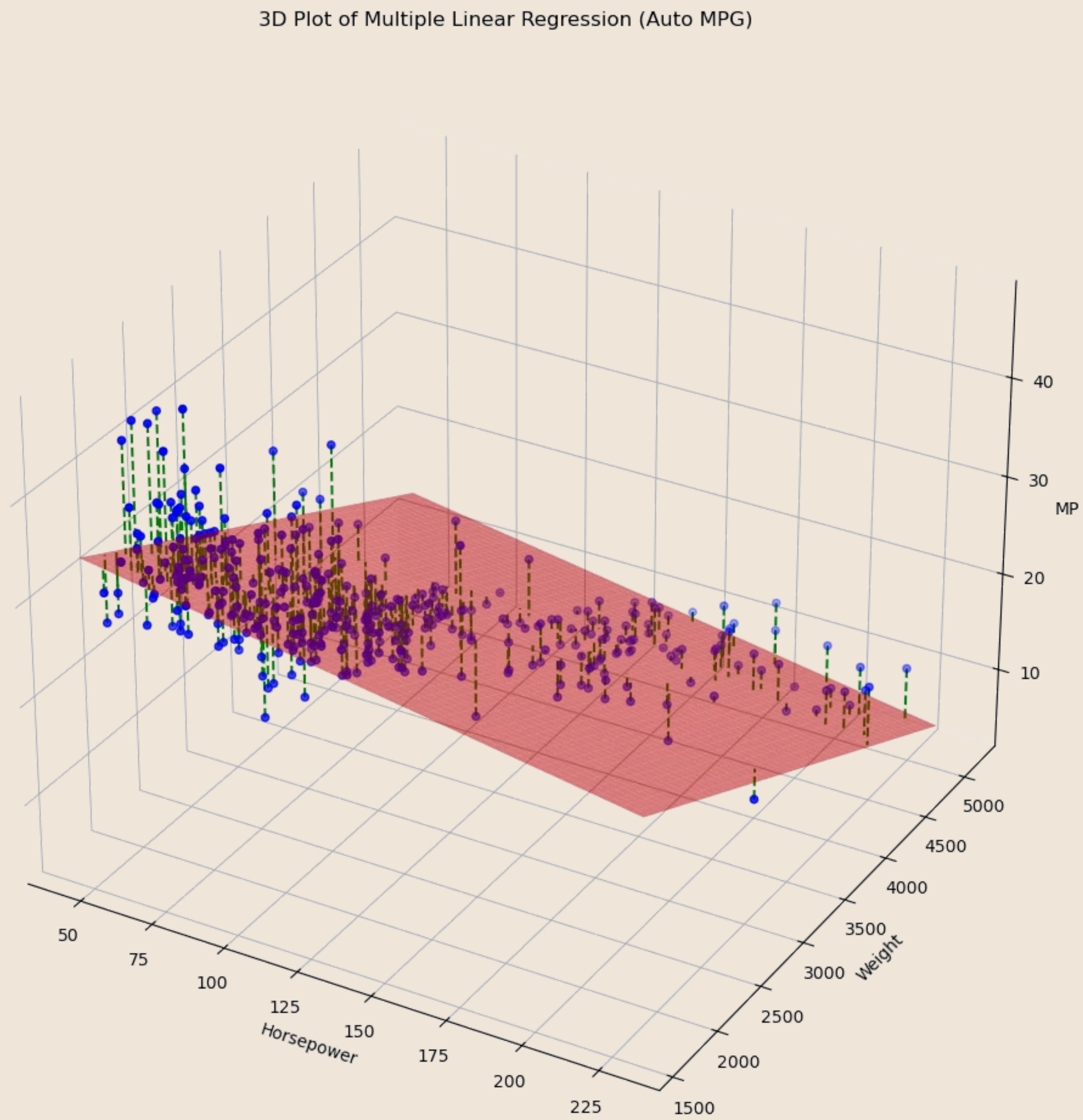
Three-Variable R^2

Adding model year

Each additional relevant variable improves our model's explanatory power. The general form accommodates any number of predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Geometric Interpretation in 3D



Dashed lines show residuals—vertical distances from observations to the plane. OLS finds the plane minimizing the sum of squared residuals. With n variables, the model becomes an (n-1) dimensional hyperplane.

With two predictors, our regression model defines a plane in 3D space. The equation $mpg = \beta_0 + \beta_1 \cdot horsepower + \beta_2 \cdot weight$ creates a surface where predictions lie.

Interpreting Multiple Regression Coefficients

The statistical interpretation of a multiple linear regression model is very similar to the interpretation of a simple linear regression model. Given the general model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \ldots + \beta_n x_n$$

we can interpret the coefficients as follows:

- The value of β_0 indicates the value of y when all independent variables are set to zero;
- *The value of β_i indicates the increment of y that we expect to see when x_i increments by one unit,* **provided that all other values $x_j | j \neq i$ are constant**.

For the model $mpg = \beta_0 + \beta_1 horsepower + \beta_2 weight$, we obtain:

β_0	β_1	β_2
45.64	-0.05	-0.01

1

Intercept ($\beta_0 = 45.64$)

Expected MPG when both horsepower and weight equal zero

2

Horsepower Effect ($\beta_1 = -0.05$)

Each additional horsepower unit decreases MPG by 0.05, **holding weight constant**

3

Weight Effect ($\beta_2 = -0.01$)

Each additional pound decreases MPG by 0.01, **holding horsepower constant**

Note how coefficients differ from the simple model (β_1 was -0.16). Adding variables changes how the model explains variance.

Estimating Multiple Regression Coefficients

Building on the concept of Ordinary Least Squares, we extend the method to estimate coefficients for models with multiple predictors. The objective remains to find the set of coefficients that minimizes the sum of squared residuals.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \ldots + \beta_n x_n$$

Our cost function, the Residual Sum of Squares (RSS), now includes all `n` predictors:

$$RSS(\beta_0, \ldots, \beta_n) = \sum_{i=1}^{m} (y_i - (\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_n x_n^{(i)}))^2$$

The values $\hat{\beta}_0, \ldots, \hat{\beta}_n$ that minimize this loss function are the **multiple least square coefficient estimates**. To find these optimal values, it is convenient to express the system in matrix notation.

We can represent the `m` observations and their corresponding equations in a compact matrix form:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

Where the components are defined as:

The **response vector $\mathbf{y}$** (`m`x`1`):

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

The **coefficient vector β** (`n+1`x`1`):

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

The **design matrix $\mathbf{X}$** (`m`x`n+1`) includes a column of ones for the intercept and the **error vector $\mathbf{e}$** (`m`x`1`):

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \ldots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \ldots & x_n^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \ldots & x_n^{(m)} \end{bmatrix}$$

$$\mathbf{e} = \begin{bmatrix} e^{(1)} \\ e^{(2)} \\ \vdots \\ e^{(m)} \end{bmatrix}$$

In the notation above, we want to minimize:

$$RSS(\beta) = \sum_{i=1}^{m} (e^{(i)})^2 = \mathbf{e}^T \mathbf{e}$$

This can be done using least squares:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

The F-Test

When fitting a multiple linear regressor, we use an F-test to determine if at least one regression coefficient is significantly different from zero. This statistical test helps validate the overall significance of the model.

We define the **null and alternative hypotheses** as follows:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$$

$$H_a : \exists j \ s.t. \ \beta_j \neq 0$$

The null hypothesis states that all predictor coefficients (excluding the intercept, β_0) are zero, implying the model has no explanatory power. The alternative hypothesis suggests that at least one predictor is significant.

An **F-statistic** is computed, which follows an **F-distribution**. The resulting **p-value** indicates the probability of observing such an F-statistic if the null hypothesis were true.

For our `mpg` regression example (using `horsepower` and `weight`), the results are:

R^2	F-statistic	Prob(F-statistic)
0.706	467.9	3.06e-104

With a p-value (`Prob(F-statistic)`) of `3.06e-104`, which is extremely small (far below the common significance level of $\alpha = 0.05$), we can confidently reject the null hypothesis. This indicates that our multiple regression model is statistically significant and useful for prediction.

Variable Selection: Backward Elimination

When we include all variables, not all may be statistically significant. Backward elimination systematically removes irrelevant predictors:

01

Fit Full Model

Include all potential predictors

02

Identify Highest p-value

Find predictor with $p > 0.05$

03

Remove Variable

Drop the least significant predictor

04

Refit and Repeat

Continue until all $p < 0.05$

Starting with all seven predictors (displacement, cylinders, horsepower, weight, acceleration, model_year, origin), we iteratively removed acceleration ($p=0.415$) and cylinders ($p=0.128$). The final model maintains $R^2 = 0.82$ with only significant predictors.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-17.2184	4.644	-3.707	0.000	-26.350	-8.087
horsepower	-0.0170	0.014	-1.230	0.220	-0.044	0.010
weight	-0.0065	0.001	-9.929	0.000	-0.008	-0.005
displacement	0.0199	0.008	2.647	0.008	0.005	0.035
cylinders	-0.4934	0.323	-1.526	0.128	-1.129	0.142
acceleration	0.0806	0.099	0.815	0.415	-0.114	0.275
model_year	0.7508	0.051	14.729	0.000	0.651	0.851
origin	1.4261	0.278	5.127	0.000	0.879	1.973

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-15.5635	4.175	-3.728	0.000	-23.773	-7.354
horsepower	-0.0239	0.011	-2.205	0.028	-0.045	-0.003
weight	-0.0062	0.001	-10.883	0.000	-0.007	-0.005
displacement	0.0193	0.007	2.579	0.010	0.005	0.034
cylinders	-0.5067	0.323	-1.570	0.117	-1.141	0.128
model_year	0.7475	0.051	14.717	0.000	0.648	0.847
origin	1.4282	0.278	5.138	0.000	0.882	1.975

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-16.6939	4.120	-4.051	0.000	-24.795	-8.592
horsepower	-0.0219	0.011	-2.033	0.043	-0.043	-0.001
weight	-0.0063	0.001	-11.124	0.000	-0.007	-0.005
displacement	0.0114	0.006	2.054	0.041	0.000	0.022
model_year	0.7484	0.051	14.707	0.000	0.648	0.848
origin	1.3853	0.277	4.998	0.000	0.840	1.930

Understanding Multicollinearity

Why did horsepower become non-significant in the full model despite being significant alone? This is **multicollinearity**—high correlation between predictors.

The Mathematical Problem

OLS estimates coefficients using:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

When predictors are highly correlated, the $X^T X$ matrix becomes nearly singular (determinant ≈ 0). Computing its inverse involves "dividing by almost zero," causing coefficient standard errors to explode and p-values to become unreliable.

Practical Impact

- Coefficient estimates become unstable
- Standard errors inflate dramatically
- Previously significant variables appear non-significant
- Model interpretation becomes unreliable

Adjusted R-squared

Standard R-squared always increases with more predictors, even insignificant ones. This can lead to complex models that may overfit the training data.

The Adjusted R-squared addresses the **bias-variance tradeoff** by penalizing the inclusion of unnecessary variables.

It re-balances model fit with complexity and is calculated as:

$$\overline{R}^2 = 1 - \frac{m-1}{m-n-1}(1 - R^2)$$

Where m is the number of data points and n is the number of independent variables.

For our last model, $R^2 = 0.820$ and $\overline{R}^2 = 0.818$.

When performing variable selection, always prioritize the adjusted R-squared over the standard R-squared to ensure a more parsimonious and robust model.

Qualitative Predictors

Many real-world features are categorical, not numerical. These **qualitative predictors**, like car origin (e.g., Europe, Asia, North America), need special handling to be included in a regression model.

Dummy Variables

We convert each category into a binary "dummy" variable (0 or 1). For example, a car from Europe would have `is_europe = 1` and other origin dummies as `0`.

Reference Category

To avoid perfect multicollinearity (the dummy variable trap), one category is chosen as the "reference" and its dummy variable is omitted from the model.

Interpreting Coefficients

The coefficient for a dummy variable represents the average difference in the response variable compared to the reference category, assuming other predictors are constant.

`income`	`income[T.low]`	`income[T.medium]`
`low`	1	0
`medium`	0	1
`high`	0	0

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	38.8638	1.234	31.504	0.000	36.431	41.297
fuelsystem[T.2bbl]	-1.6374	1.127	-1.453	0.148	-3.860	0.585
fuelsystem[T.4bbl]	-12.0875	2.263	-5.341	0.000	-16.551	-7.624
fuelsystem[T.idi]	-0.3894	1.300	-0.299	0.765	-2.954	2.175
fuelsystem[T.mfi]	-5.8285	3.661	-1.592	0.113	-13.049	1.392
fuelsystem[T.mpfi]	-5.4942	1.202	-4.570	0.000	-7.865	-3.123
fuelsystem[T.spdi]	-5.2446	1.612	-3.254	0.001	-8.423	-2.066
fuelsystem[T.spfi]	-6.1522	3.615	-1.702	0.090	-13.282	0.978
horsepower	-0.0968	0.009	-11.248	0.000	-0.114	-0.080

Linear Regression in Python

	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	18.0	8	307.0	130.0	3504	12.0	70	usa
1	15.0	8	350.0	165.0	3693	11.5	70	usa
2	18.0	8	318.0	150.0	3436	11.0	70	usa
3	16.0	8	304.0	150.0	3433	12.0	70	usa
4	17.0	8	302.0	140.0	3449	10.5	70	usa

Conclusions and Next Steps

We Have Explored:

- Simple Linear Regression
- Statistical Interpretation of coefficients
- Geometric Interpretation
- Multiple Linear Regression
- Backward Elimination
- Dummy Variables

In the next lectures, we will look at the extensions to the linear regression

References

- Chapter 3 of [1]
- Parts of chapter 11 of [2]

[1] Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.

[2] James, Gareth Gareth Michael. An introduction to statistical learning: with applications in Python, 2023.https://www.statlearning.com