Simple Linear Regression. Galaxy Distance and Velocity

Diego Ramos Crespo

March 7, 2025

1 Section I

A 2001 study by Freedman et al. presents data on the relative velocity and distance of 24 galaxies, collected using the Hubble Space Telescope. According to astrophysical theory, this data can be used to estimate the age of the universe.

1.1 Scatter plot

Figure 1 displays the scatter plot of the response variable (Y) 'velocity' versus the predictor variable (X) 'distance'. This plot illustrates the relationship between both variables. The apparent correlation suggests that galaxies farther away tend to move faster, consistent with the expanding universe model. This relationship provides an approximation of the universe's age.

The computed correlation coefficient of **0.8632** indicates a strong positive linear relationship between distance and velocity. This means that as the distance of a galaxy increases, its velocity also tends to increase in a predictable manner.

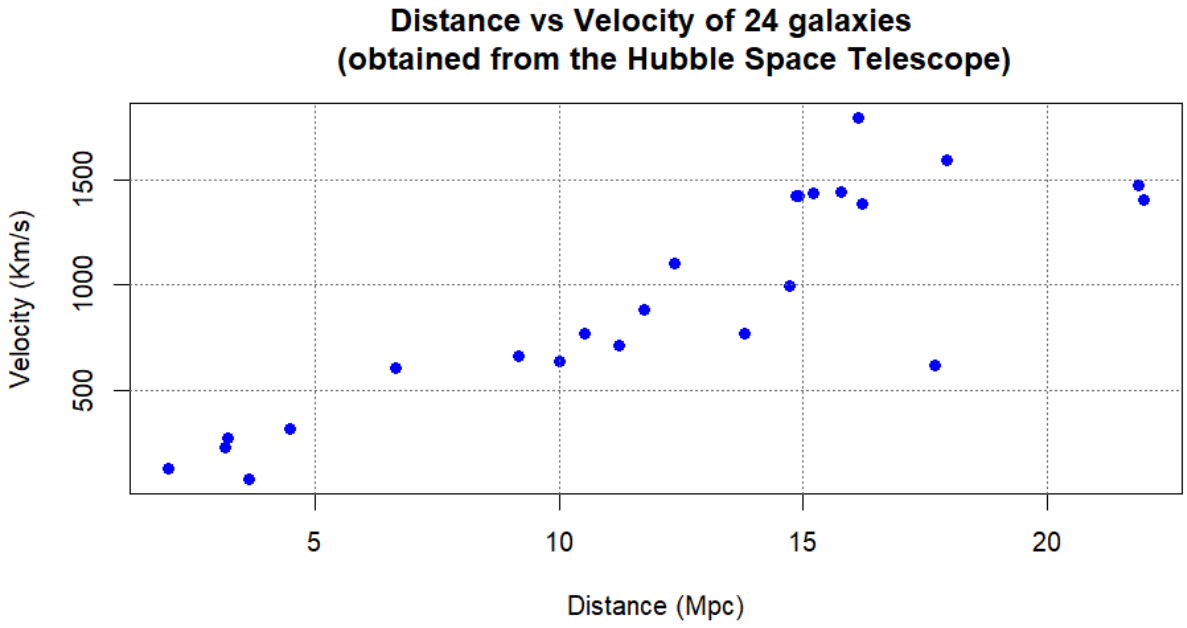

Figure 1: Hubble Space Telescope data: Distance vs. Velocity of 24 Galaxies.

1.2 Regression analysis and summary statistics

To perform the regression analysis, it is necessary to estimate the intercept (β_0) and the slope (β_1). Given that the number of observations is $n = 24$, we first compute the mean of the predictor variable (distance) and the response variable (velocity).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 12.055, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 924.375$$

1.2.1 Estimating the Slope and Intercept

We compute the sum of squares:

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 777.63$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 59198.85$$

From these, the slope ($\widehat{\beta_1}$) and intercept ($\hat{\beta}_0$) are given by:

$$\widehat{\beta_1} = \frac{S_{XY}}{S_{XX}} = 76.12696$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 6.69626$$

Thus, the estimated regression model is:

$$\widehat{y} = 76.12696x + 6.69626 \tag{1.2.1}$$

The formulas above represent the results of the simple linear regression analysis applied to the galaxy distance and velocity data. The estimated slope $\widehat{\beta_1} = 76.12696$ indicates that for each unit increase in distance, the velocity of the galaxies is expected to increase by approximately 76.13 units.

The estimated intercept $\widehat{\beta_0} = 6.69626$, suggests that when the distance is zero, the velocity is approximately6.70.

This model implies a positive relationship between distance and velocity, aligning with the theoretical understanding of an expanding universe. The intercept value suggests that at very low distances, the velocity is not significantly different from zero.

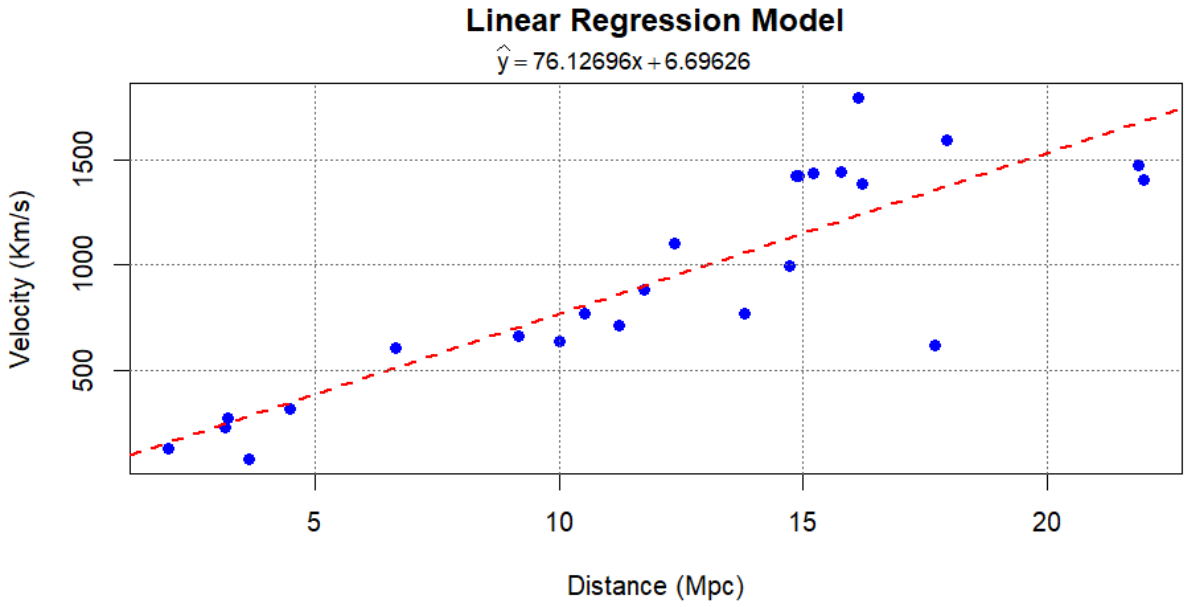

Figure 2: Simple Linear Regression model for predicting velocity of galaxies based on its distance.

1.2.2 Sum of Squared Errors and Mean Squared Error

The predicted values ($\widehat{y_i}$) are:

$$\widehat{y_i} = \widehat{\beta}_1 x_i + \hat{\beta}_0$$

The Sum of Squares due to Error (SSE) or also called Residual Sum of Squares and the Residual Mean Square ($\widehat{\sigma^2}$) are:

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = 1541869$$

$$\widehat{\sigma^2} = \frac{SSE}{n-2} = 70084.97$$

1.2.3 Standard Errors of the estimates

The standard errors of the intercept and slope are:

$$SE_{\widehat{\beta_0}} = \sqrt{\widehat{\sigma^2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)} = 126.557$$

$$SE_{\widehat{\beta_1}} = \sqrt{\frac{\widehat{\sigma^2}}{S_{XX}}} = 9.493$$

1.2.4 Total Sum of Squares and R^2

The total sum of squares (SST) is:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 6048498$$

The coefficient of determination (R^2) is:

$$R^2 = 1 - \frac{SSE}{SST} = 0.7451$$

This means that 74.51% of the variance in the velocity of the galaxy is explained by its distance from Earth, according to the linear regression model. However, there is still a 25.49% of the variance that remains unexplained, this could be due to measurement errors or external factors.

1.2.5 t-Statistics and Hypothesis Testing

For hypothesis tests on the intercept (β_0) and slope (β_1), we compute the t-values:

$$t_{\widehat{\beta_0}} = \frac{\widehat{\beta_0}}{SE_{\widehat{\beta_0}}} = 0.0529$$

$$t_{\widehat{\beta_1}} = \frac{\widehat{\beta_1}}{SE_{\widehat{\beta_1}}} = 8.0189$$

The hypotheses are:

Intercept (β_0)

$H_0 : \beta_0 = 0$

$H_1 : \beta_0 \neq 0$

The p-value is 0.9583, which is very high. Thus, we do not have enough evidence to reject H_0, meaning the model might pass through the origin.

Slope (β_1)

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

The p-value is 5.6767×10^{-8}, which is very small. Therefore, we reject H_0 and conclude that the slope is significantly different from zero.

The following Table 1 exhibits a summary of all previous values.

	Estimate	Std. Error	t value	P($<|t|$)
Intercept	6.69626	126.557	0.0529	0.9583
Slope	76.12696	9.493	8.0189	5.6767e-08

Table 1: Regression coefficients, standard errors, t-values, and p-values.

1.2.6 F-Ratio

The F-ratio for the model is:

$$F = \frac{R^2}{(1 - R^2)/(n - 2)} = 0.1329$$

This indicates the proportion of variance explained by the regression model. The F-ratio is low, suggesting that the model does not explain much of the variance in the response variable relative to the residual variance.

1.3 Confidence intervals

Since the estimators provide only an approximation of the true model, we can construct confidence intervals to estimate the range within which the true parameters are likely to be contained in. In this case, we aim to be 95% confident that the interval captures the true values of β_0 and β_1.

We know that the confidence level is $\alpha = 0.95$. Meanwhile, the significance level is $1 - \alpha = 0.05$.

The quantile is calculated using a t distribution; in R, this value can be computed using the $qt()$ function.

```
qt(0.05/2, n-2, lower.tail = FALSE) = 2.0739
```

Thanks to the previous code, we obtain that the critical value is $t_{\alpha/2,n} = 2.0739$ Therefore, the confidence interval for the intercept is computed as follows.

$$\beta_0 \pm (t_{\alpha/2,n} \cdot SE_{\widehat{\beta_0}}) = 6.70 \pm 262.46 = [-255.76, 269.16]$$

After performing the calculations, we observe that the confidence interval (CI) for the intercept includes zero. Consequently, we cannot conclude that the regression line does not pass through the origin. This evidence supports the p-value computed in Section 1.2.5, which did not provide sufficient evidence to reject the null hypothesis.
The confidence interval for the slope is obtained through a similar process.

$$\beta_1 \pm (t_{\alpha/2,n} \cdot SE_{\widehat{\beta_1}}) = 76.13 \pm 19.69 = [56.44, 95.82]$$

Since this interval does not include zero, we can conclude with 95% confidence that the slope is significantly different from zero. This suggests that the predictor variable has a statistically significant relationship with the response variable, reinforcing the validity of the regression model.

1.4 Confidence and Prediction Bands

To estimate the velocity at given distances, we compute both the confidence and prediction intervals at a 90% confidence level.

The critical value from the t-distribution with $n-2$ degrees of freedom is:

$$t_{\alpha/2,n-2} = 1.7531$$

For the given distances $x_1 = 5.5$, $x_2 = 14.0$, and $x_3 = 20.8$. Evaluating these values in the regression model shown in Equation 1.2.1 the values are $y_1 = 425.3945$, $y_2 = 1072.4737$ and $y_3 = 1590.1370$. Furthermore, the standard errors are calculated as follows.

- Confidence band standard error:

$$SE_{\text{mean}} = \sqrt{MSE} \times \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

- Prediction band standard error:

$$SE_{\text{pred}} = \sqrt{MSE} \times \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

The confidence and prediction intervals are given by:

$$\text{Confidence Interval: } \hat{y} \pm t_{\alpha/2,n-2} \cdot SE_{\text{mean}}$$

Prediction Interval: $\hat{y} \pm t_{\alpha/2,n-2} \cdot SE_{\text{pred}}$

- Confidence Interval: The confidence band provides a range in which we expect the *mean* velocity to fall for a given distance with 90% confidence. This interval accounts for the uncertainty in estimating the population mean.
- Prediction Interval: The prediction band provides a range for a *single new observation* at a given distance. Since it includes both: the uncertainty in the mean estimate and the natural variation of individual observations, the prediction interval is always wider than the confidence interval.
- Specific Cases:
  - At $x = 5.5$, the confidence and prediction intervals help determine if the model accurately predicts velocity at short distances.
  - At $x = 14.0$, the intervals show how well the model predicts near the center of the observed data range.
  - At $x = 20.8$, the widening of intervals suggests greater uncertainty in predictions for farther distances.

The following Figure (3) represents the bands. It can be observed that prediction bands for a single observation are wider than confidence bands for the mean. The red dots are plotted according to the model given a value of the predictor variable (distance).

1.5 ANOVA Test

To analyze the variance in velocity explained by the regression model, we construct the ANOVA table. The sum of squares are defined as follows.

- **Total Sum of Squares (SST)**: Measures the total variation in the response variable (velocity).
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- **Sum of Squares for Regression (SSR)**: Measures the variation explained by the regression model.
$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

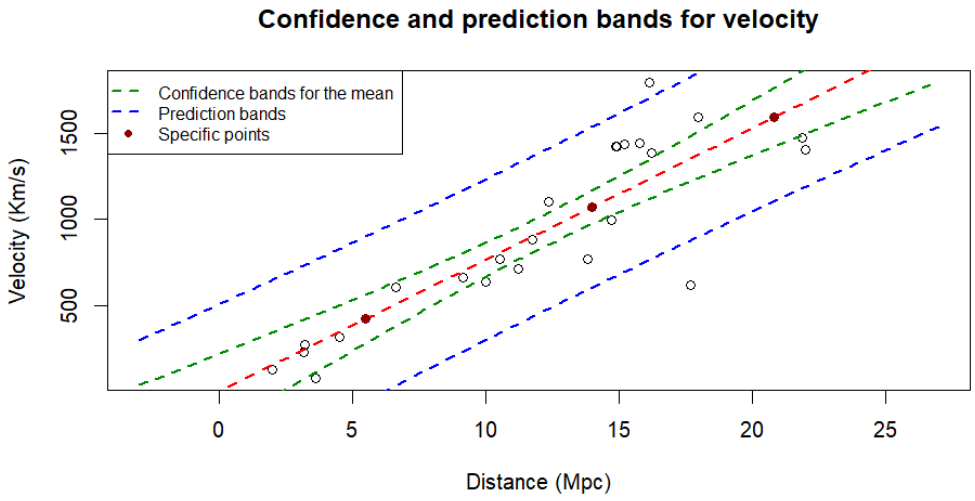

Figure 3: Confidence bands for the mean and prediction bands for specific values in the simple linear regression model.

- **Sum of Squares for Error (SSE)**: Measures the unexplained variation (residuals).

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The relationship between these quantities is given by:

$$SST = SSR + SSE$$

The degrees of freedom (DoF) associated with each source of variation are:

- **Regression DoF**: $df_{\text{regression}} = 1$ (since there is one predictor variable).
- **Error DoF**: $df_{\text{error}} = n - 2$.
- **Total DoF**: $df_{\text{total}} = n - 1$.

The **Mean Squares** are computed as:

$$MSR = \frac{SSR}{df_{\text{regression}}}$$

$$MSE = \frac{SSE}{df_{\text{error}}}$$

The **F-statistic** is given by.

$$F = \frac{MSR}{MSE}$$

and its corresponding p-value (*p-value*) is used to test the null hypothesis that the slope of the regression model is zero.

The resulting ANOVA table is shown below.

Source	**DoF**	**Sum of Squares**	**Mean Square**	**F-ratio**	**p-value**
Regression	1	4,506,628	4,506,628.20	64.30235	5.68×10^{-8}
Residuals	22	1,541,869	70,084.97	-	-
Total	23	6,048,498	-	-	-

Table 2: ANOVA table for the regression model

The F-ratio evaluates whether the regression model explains a significant portion of the variability in velocity. We can state two hypotheses:

- **Null hypothesis** (H_1): The predictor variable (distance) does not significantly explain the variability in the response variable (velocity). In other words, the regression model with distance as a predictor is no better than the model without it. It can be thought that distance does not significantly affect velocity.

- **Alternative hypothesis** (H_1): The predictor variable (distance) significantly explains the variability in the response variable (velocity). The regression model with distance as a predictor explains a significant portion of the variability in velocity, in other words, distance significantly affects velocity.

A large **F-value** and a small **p-value** indicate that the predictor variable (distance) significantly influences the response variable (velocity). In this case, there is tremendous evidence that the predictor contributes to explaining the response.

1.6 Diagnostic plots

The evaluation of the model is reach by using residuals and diagnostic plots as shown in Figure 4. The first plot, Scale-Location, presents the square root of standardized residuals against the fitted values. In this plot, the residuals are not entirely dispersed; instead, they tend to accumulate around the value of 1300, which suggests the presence of heavy tails in the distribution of residuals.

Despite this, the Q-Q plot shows that the residuals closely follow a straight line, indicating that the distribution of the errors is approximately normal. This is a crucial assumption for linear regression, and the Q-Q plot provides reassurance that the residuals align with normality, which strengthens the model's validity.

In the Residuals vs Leverage plot, we can observe that no points exceed the dashed lines, which suggests that there are no influential points or outliers that might disproportionately affect the model's estimates. The absence of extreme values in this plot indicates that the model is not unduly influenced by a small number of points, which is a positive sign for the model's stability and robustness. However, despite the overall good fit, the diagnostic plots indicate potential issues such as the slight accumulation of residuals in the Scale-Location plot and the presence of heavy tails. These features suggest that certain transformations could improve the model further. For instance, we could apply a log transformation to the dependent or independent variables to address any skewness in the residuals or a square root transformation to help stabilize the variance and reduce the impact of extreme residuals. By applying such transformations, we could potentially achieve more homoscedasticity and improve the overall accuracy of the regression model.

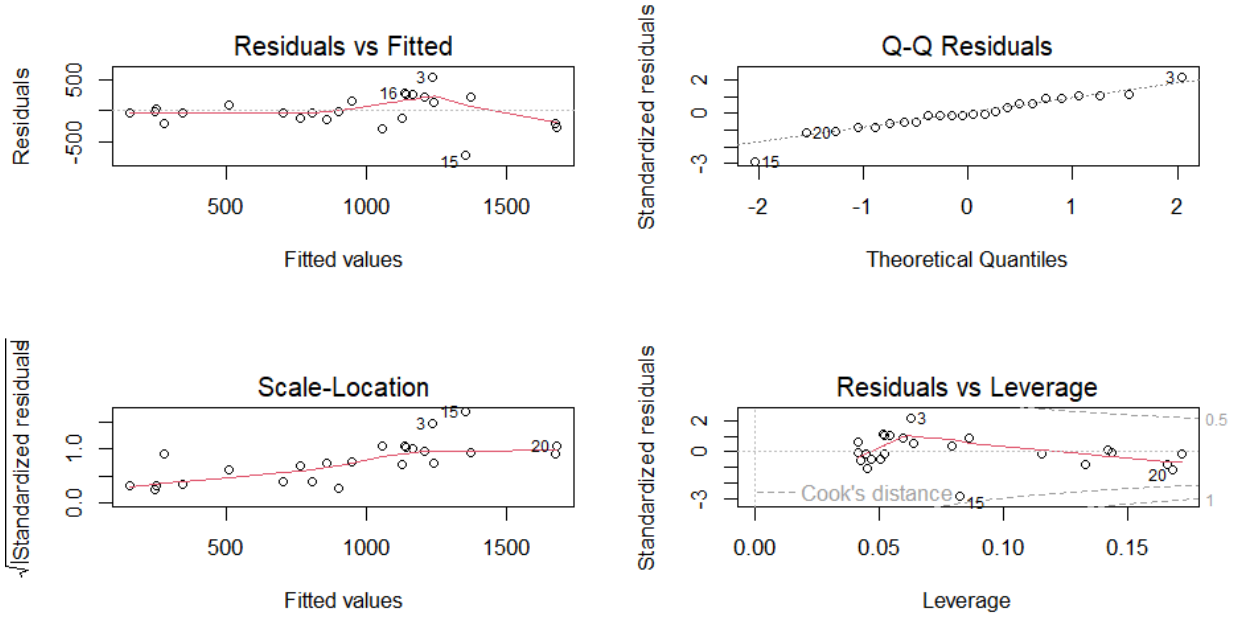

Figure 4: Residuals and diagnostic plots for model validation.

2 Section II

Sometimes, it is necessary to fit a linear regression model, where it is known that the intercept is zero. This model is expressed as follows.

$$y_i = \beta_1 x_i + e_i, \quad \text{for } i = 1, 2, \ldots, n. \tag{2.0.1}$$

2.1 Show that in this context, the least squares estimator of β_1 is $\widehat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$

We want to derive an estimate for $\widehat{\beta}_1$. The goal is to get the estimator as close as possible to the true relationship. The following equation is the predicted model.

$$\widehat{y}_i = \widehat{\beta}_1 x_i \tag{2.1.1}$$

The main idea is to minimize the difference between the true model and the predicted model.

$$y_i - \widehat{y}_i \tag{2.1.2}$$

We square the difference to prevent the values from canceling out.

$$(y_i - \widehat{y}_i)^2 \tag{2.1.3}$$

Substitute $\widehat{y}_i$ from equation 2.1.1.

$$(y_i - \widehat{\beta}_1 x_i)^2 \tag{2.1.4}$$

All values are added together over the n observations.

$$\sum_{i=1}^{n} (y_i - \widehat{\beta}_1 x_i)^2 \tag{2.1.5}$$

Once we have equation 2.1.5, we proceed to minimize it. To find the value of $\widehat{\beta}_1$ that minimizes the sum of squared errors, we take the derivative of the expression with respect to $\widehat{\beta}_1$. This allows us to find the critical points that will give the minimum value of the error.

The derivative of the sum of squared errors with respect to $\widehat{\beta}_1$ is:

$$\frac{d}{d\widehat{\beta}_1} \sum_{i=1}^{n} \left(y_i^2 - 2 y_i \widehat{\beta}_1 x_i + \widehat{\beta_1^2} x_i^2 \right) \tag{2.1.6}$$

Taking the derivative term by term:

- The derivative of y_i^2 with respect to $\widehat{\beta}_1$ is zero because it does not depend on $\widehat{\beta}_1$.
- The derivative of $-2y_i\widehat{\beta}_1 x_i$ with respect to $\widehat{\beta}_1$ is $-2y_i x_i$.
- The derivative of $\widehat{\beta_1^2} x_i^2$ with respect to $\widehat{\beta}_1$ is $2\widehat{\beta}_1 x_i^2$.

Thus, the derivative becomes:

$$\sum_{i=1}^{n} \left(-2y_i x_i + 2\widehat{\beta}_1 x_i^2\right) \tag{2.1.7}$$

Now, we set this derivative equal to zero in order to minimize the expression. This yields the normal equation for the least squares estimator:

$$\sum_{i=1}^{n} \left(-2y_i x_i + 2\widehat{\beta}_1 x_i^2\right) = 0 \tag{2.1.8}$$

Solving this equation for $\widehat{\beta}_1$, we have

$$-2\sum_{i=1}^{n} x_i y_i + 2\widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{2.1.9}$$

This gives us the least squares estimator for the slope of the regression line.

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{2.1.10}$$

2.2 Show that $\widehat{\beta}_1$ is an unbiased estimator for β_1 and that $Var\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_i x_i^2}$

Recall an unbiased estimator is $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.1}$$

Substituting y_i from equation 2.0.1.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i(\beta_1 x_i + e_i)}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.2}$$

Developing the numerator.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\frac{\beta_1 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.3}$$

Simplifying within the expected value.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\frac{\beta_1 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.4}$$

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\beta_1 + \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.5}$$

Applying expected value property: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, where X and Y are random variables.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}[\beta_1] + \mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.6}$$

We know that $\mathbb{E}[c] = 0$, where c is a constant.
Furthermore, since x_i are given data points, they are treated as constants. The error terms e_i are the random variables in the model. Therefore, we apply the expectation operator.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \frac{\sum_{i=1}^{n} x_i \mathbb{E}[e_i]}{\sum_{i=1}^{n} x_i^2} \tag{2.2.7}$$

The mean of the errors is assumed to be zero: $\mathbb{E}[e_i] = 0$.

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 \tag{2.2.8}$$

Equation 2.2.8 shows that $\widehat{\beta_1}$ is an unbiased estimator.
Now, the variance of the estimator is computed $\left(Var\left[\hat{\beta}_1\right]\right)$
Applying variance to the estimator β_1

$$Var\left[\hat{\beta}_1\right] = Var\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.9}$$

Substituting y_i from equation 2.0.1.

$$Var\left[\hat{\beta}_1\right] = Var\left[\frac{\sum_{i=1}^{n} x_i(\beta_1 x_i + e_i)}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.10}$$

Developing the numerator.

$$Var\left[\hat{\beta}_1\right] = Var\left[\frac{\beta_1 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.11}$$

Simplifying within the variance.

$$Var\left[\hat{\beta}_1\right] = Var\left[\frac{\beta_1 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.12}$$

$$Var\left[\hat{\beta}_1\right] = Var\left[\beta_1 + \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.13}$$

Recalling that $Var[c] = 0$, when c is a constant we have the following expression.

$$Var\left[\hat{\beta}_1\right] = Var\left[\frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}\right] \tag{2.2.14}$$

Furthermore, x_i are given data points, so they are treated as constants; besides, the error terms e_i are the random variables in the model that have constant variance ($Var[e_i] = \sigma^2$).
We can also simplify the expression by using the following property of variance: $Var[aX] = a^2 Var[X]$, where a is constant and X is a random variable.
Bearing in mind the previous facts, the expression can be developed as follows.

$$\text{Var}\left[\hat{\beta}_1\right] = \frac{\sum_{i=1}^{n} x_i^2 \cdot \text{Var}\left[e_i\right]}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \tag{2.2.15}$$

$$\text{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \tag{2.2.16}$$

Equation 2.2.16 shows the variance of the estimator $\widehat{\beta_1}\beta_1$.

2.3 Find an expression for $\hat{\sigma}^2$

Recall that $\hat{\sigma}^2$ is an estimator for the population variance σ^2. The general expression for $\hat{\sigma}^2$ is derived from the sum of squared residuals. In this case, we assume the intercept is zero, so the model is.

$$y_i = \beta_1 x_i + e_i$$

where y_i is the observed value, β_1 is the model parameter, and e_i are the random errors with mean 0 and variance σ^2.

We define the residuals e_i as the difference between the observed and predicted values:

$$e_i = y_i - \widehat{y}_i \tag{2.3.1}$$

where $\hat{y}_i = \widehat{\beta}_1 x_i$ is the predicted value.

The sum of squared residuals is:

$$SSE = \sum_{i=1}^{n} e_i^2 \tag{2.3.2}$$

To estimate σ^2, we divide the sum of squared residuals by the degrees of freedom, which is $n-1$ for this model (since we are estimating 1 parameter $[\beta_1]$).

$$\hat{\sigma}^2 = \frac{SSE}{n-1} \tag{2.3.3}$$

Using the residual sum of squares SSE.

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Substitute the expression for $\hat{y}_i$:

$$SSE = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_1 x_i \right)^2 \tag{2.3.4}$$

Thus, the estimator for $\widehat{\sigma}^2$ becomes:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2}{n-1} \tag{2.3.5}$$

The degrees of freedom associated with $\hat{\sigma}^2$ are the number of independent pieces of information used to estimate the variance. In this model, we only one parameter (β_1). Therefore, the degrees of freedom for the residuals are:

$$\text{Degrees of freedom} = n - 1 \tag{2.3.6}$$

Thus, the estimator $\hat{\sigma}^2$ has $n-1$ degrees of freedom.

The expression for $\hat{\sigma}^2$ is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2}{n-1} \tag{2.3.7}$$

The degrees of freedom associated with $\hat{\sigma}^2$ are $n-1$.

2.4 SLR through the origin

We fit a regression model to the data using the formula:

$$y_i = \beta_1 x_i$$

This model assumes that there is no intercept. We obtain the estimated slope, denoted as $\hat{\beta}_1$, and the standard error for the slope.

$$\hat{\beta}_1 = 76.12696, \quad \text{Standard Error for } \hat{\beta}_1 = 9.28$$

Next, we calculate the residual sum of squares (SSE) and the mean square error (MSE). The formula for SSE is:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $y_i = \beta_1 x_i$. The MSE is given by:

$$\text{MSE} = \frac{\text{SSE}}{n-1}$$

We calculate:

$$\text{SSE} = 67084.59, \quad \text{MSE} = 134.11$$

2.4.1 Hypothesis Testing for β_1

To test whether the estimated slope is significantly different from zero, we perform a hypothesis test for $\beta_1 \neq 0$. The test statistic is:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = 8.198575$$

The corresponding p-value is:

$$p = 2 \times \text{pt}(|t|, df = n-1) = 2.809103 \times 10^{-8}$$

Given that the p-value is very small, we reject the null hypothesis and conclude that the slope is significantly different from zero.

2.4.2 Confidence Interval for β_1

We compute the 95% confidence interval for β_1. The critical value for a 95% confidence interval, using $n-2$ degrees of freedom, is:

$$t_{0.025,n-2} = 2.0687$$

Thus, the 95% confidence interval for β_1 is:

$$[56.91867, 95.33525]$$

This interval suggests that, with 95% confidence, the true value of the slope parameter β_1 lies within this range. Since the interval does not contain zero, we can conclude that there is strong evidence to suggest that the slope is significantly different from zero. implying a meaningful relationship between distance and velocity.

2.4.3 Regression model and plot

The fitted regression model is:

$$\hat{y} = 76.12696 \times x$$

The regression line is plotted along with the scatter plot of the data points in the figure below.

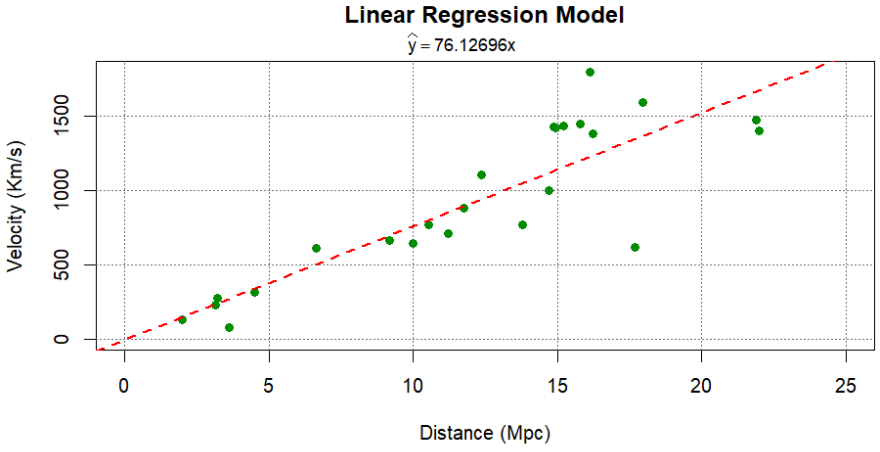

Figure 5: Linear regression model with the slope passing through the origin.

We evaluate the model fit using residuals and diagnostic plots, as shown in Figure 6. The first plot (top-right) shows no distinct pattern, with the residuals randomly scattered around zero, suggesting that the variance of the errors is constant (homoscedasticity). The Scale-Location plot below it displays the square root of standardized residuals against the fitted values. The points appear randomly scattered without any clear trend, which further reinforces the conclusion from the previous plot.

The Q-Q plot closely follows a straight line, indicating that the residuals are normally distributed, which is a key assumption in linear regression. Finally, the Residuals vs Leverage plot helps identify outliers and influential points that could disproportionately affect the model's results. In this case, no extreme points are observed, except for one at position 15, which exceeds the dashed line. This point should be reviewed to determine whether it should be removed or retained. Nevertheless, the dataset appears to be good and meets the assumptions of simple linear regression.

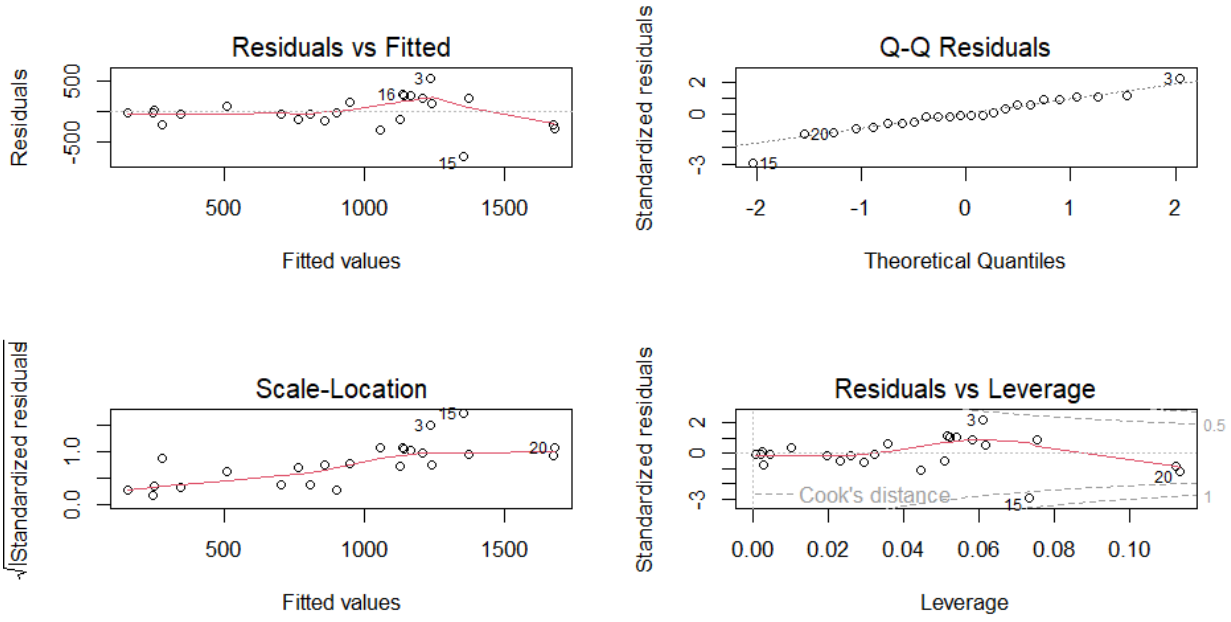

Figure 6: Residuals and diagnostic plots for model validation.