Quiz 3 Assignment, SLR Model

Name: Bimarsh Bhusal
Date: March 28, 2025

Introduction

This document contains the solutions for Quiz 3 Assignment on the Simple Linear Regression (SLR) Model.

I have considered *Stories* to be the independent variable and *Height* to be the dependent variable. The number of stories is a design choice that directly determines the height of a building. For example: A building with more stories is constructed to be taller. Height is a physical outcome of stacking a specific number of stories.

In architecture/engineering, the number of stories is decided first (e.g., based on zoning laws or purpose), and the height is calculated from that. Each story adds a roughly fixed height

Data: Height vs. Stories (Row Format)

Below is the data for building height (in feet) and the number of stories, presented in row format.

Stories	54	47	28	38	29	38	80	52	45	40	49	33	50	40	31
Height (ft)	770	677	428	410	371	504	1136	695	551	550	568	504	560	512	448

Stories	40	27	31	35	57	31	26	39	25	23	102	72	57	54	56
Height (ft)	538	410	409	504	777	496	386	530	360	355	1250	802	741	739	650

Stories	45	42	36	30	22	31	52	29	34	20	33	18	23	30	38
Height (ft)	592	577	500	469	320	441	845	435	435	375	364	340	375	450	529

Stories	31	62	48	29	40	30	42	36	33	72	57	34	46	30	21
Height (ft)	412	722	574	498	493	379	579	458	454	952	784	476	453	440	428

Table 1: Data for Building Height (ft) and Number of Stories (Row Format)

SLR Model

The SLR model is given by the formula:

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

where,

- β_0 is the intercept,
- β_1 is the slope,
- ϵ is the error term (assumed to be normally distributed with mean 0).

The estimated regression line is given by the formula:

$$\hat{y} = b_0 + b_1 x$$

where,

- b_0 is the estimate of β_0,
- b_1 is the estimate of β_1,

Given a set of regression data $\{(x_i, y_i), i = 1, 2, \cdots, n\}$ and the fitted model $\hat{y_i} = b_0 + b_1 x_i$, the i^{th} residual e_i is given by:

$$e_i = y_i - \hat{y_i}, \quad i = 1, 2, \cdots, n$$

We find b_0 and b_1 so that sum of the squares of residuals (also called SSE) is minimum.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

To find the minimum, we take partial with respect to b_0 and b_1 and equal to 0. This gives us two equations called normal equations. Equating them we obtain the value of b_0 and b_1.

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}$$

From the given data, we have 60 data points. *i.e.* $n = 60$.

Using R, we get the following values:

- SSE : 197311.01
- b_0 : 90.31
- b_1 : 11.29237

Hence, the estimated regression line is:

$$\hat{y} = 90.31 + 11.29237x$$

Scatterplot and Regression Line

I used the R programming language to create a scatterplot of *Height* vs. *Stories* and overlaid the regression line. The plot is shown below:

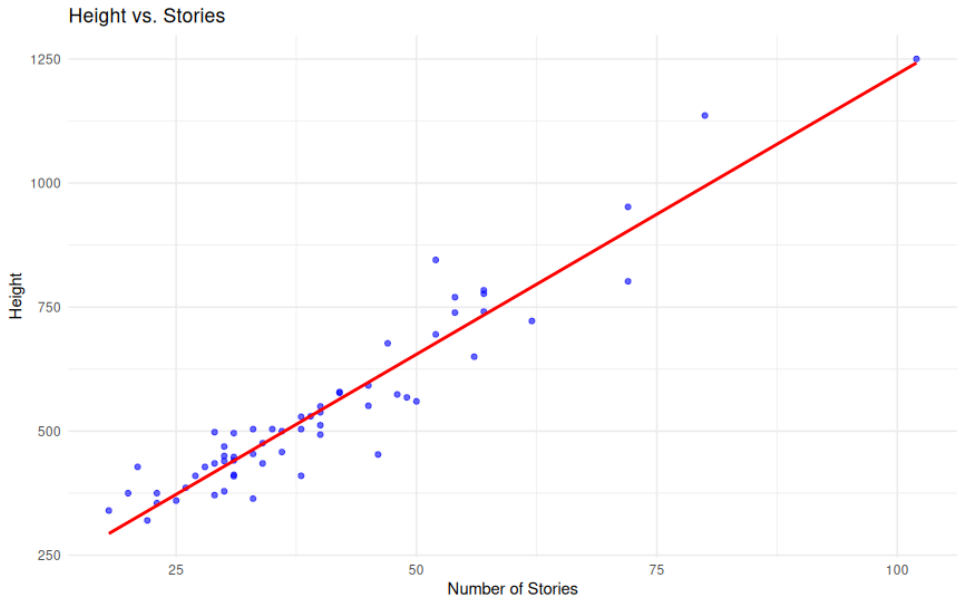

Figure 1: Scatterplot of Height vs. Stories with Regression Line

Observations from Figure 1

- The red regression line shows a linear relationship between Height and Number of Stories.
- The positive slope confirms that height increases as the number of stories increases.
- Each additional story adds a roughly fixed amount of height to the building.
- Data points are close to the regression line, showing a good fit for the linear model.
- Some variability suggests other factors may also influence building height.

Hypothesis Testing on the slope β_1

The null and alternative hypothesis are stated below:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

First, we estimate the model error variance s^2, where $s^2 = \frac{SSE}{n-2}$. s^2 is the unbiased

$\therefore s^2 = 3401.914$

$\therefore s = \sqrt{s^2} = 58.32593$

We use t statistic which is given by:

$$t = \frac{b_1 - \beta_{10}}{\frac{s}{\sqrt{S_{xx}}}}$$

Using R, the t statistic is : $t = 23.310$

The $P-$value is the probability of obtaining a value of the test statistic as extreme or more extreme than the one found when in fact H_0 is true. To test the null hypothesis H_0 that $\beta_1 = \beta_{10}$ against a suitable alternative, use the $t-$ distribution with $n - 2$ degrees of freedom, where n is the number of paired observations.

Using R, the $p-$ value is: 3.835×10^{-31}.

We know that a $p-$value is the lowest level (of significance) at which the observed value of the test statistic is significant.

Using significance level $\alpha = 0.05$.

Since $p-$ value is extremely smaller than the significance level *i.e.* $p - value < 0.05$, we reject the null hypothesis and conclude that there is a significant linear relationship between Stories and Height.

Conclusion:

There is a significant linear relationship between Stories and Height.

Lack-of-fit test

Since we rejected H_0, we want to check if the linear model is adequate. The null and alternative hypothesis for lack-of-fit test are given below:

$$H_0 : \text{The regression is linear in x.}$$
$$H_1 : \text{The regression is not linear in x.}$$

For this test, we use F statistic which is given by the formula:

$$f = \frac{SSE - SSE(pure)}{s^2 \cdot (k-2)}$$

where,

- $s^2 = \frac{SSE(pure)}{n-k}$
- k is the number of distinct x values.

The $f-$ statistic for the test of lack of fit follows an $f-$ distribution with $k-2$ and $n-k$ degrees of freedom.

Using significance level $\alpha = 0.05$.

We use R to calculate f and p value.

- $f-$ statistic: 1.6819
- $p-$ value: 0.089

Since $p-$ value is greater smaller than the significance level *i.e.* $p - value > 0.05$, we fail to reject the null hypothesis and conclude that the linear model is adequate.

Conclusion:

The linear model is adequate.

Confidence Interval for β_1

A $100(1-\alpha)\%$ confidence interval for the parameter β_1 in the regression line for $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Using R, the computed confidence interval is:

$$10.323 \leq \beta_1 \leq 12.262$$

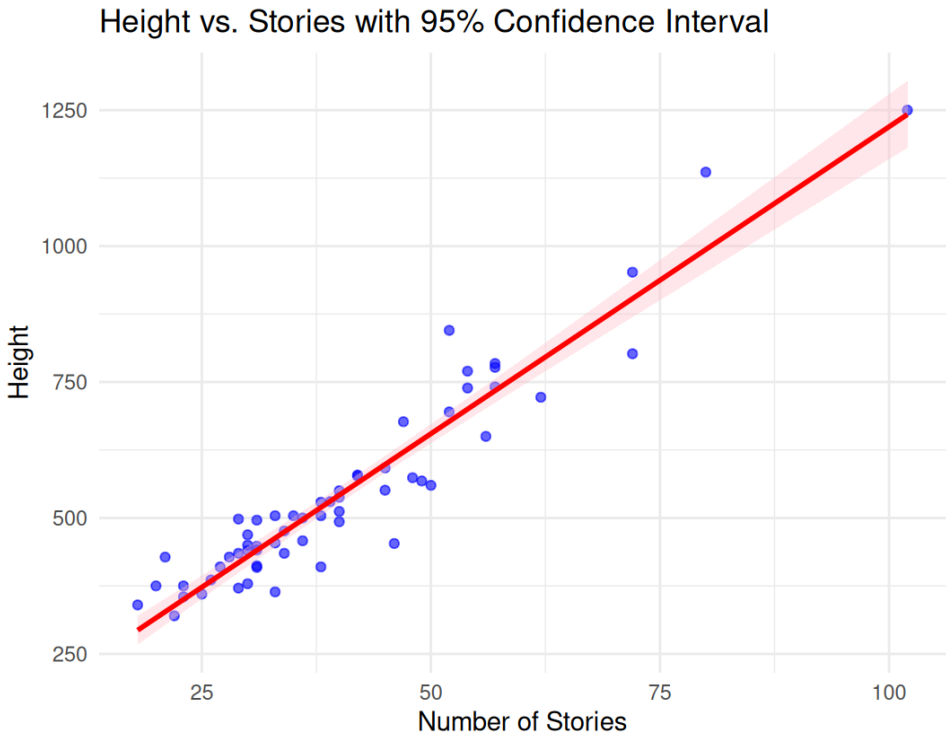

Figure 2: Height vs. Stories with 95% confidence interval

Hypothesis Testing on the slope β_0

The null and alternative hypothesis are stated below:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

We use the t-statistic which is given by:

$$t = \frac{b_0 - \beta_{00}}{s\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}.$$

Using R, we get $t = 4.308$

Using R, the $p-$ value is: 6.44×10^{-5}.

We know that a $p-$value is the lowest level (of significance) at which the observed value of the test statistic is significant.

Using significance level $\alpha = 0.05$.

Since $p-$ value is extremely smaller than the significance level *i.e.* $p - value < 0.05$, we reject the null hypothesis and conclude that the intercept is significantly different from zero.

This means that even when the number of stories is zero, the model predicts a non-zero height.

Conclusion:

The intercept is significantly different from zero.

Confidence Interval for β_0

A $100(1-\alpha)\%$ confidence interval for the parameter β_0 in the regression line for $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_0 - t_{\alpha/2}\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} < \beta_0 < b_0 + t_{\alpha/2}\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Using R, the computed confidence interval is:

$$48.349 \leq \beta_1 \leq 132.27$$

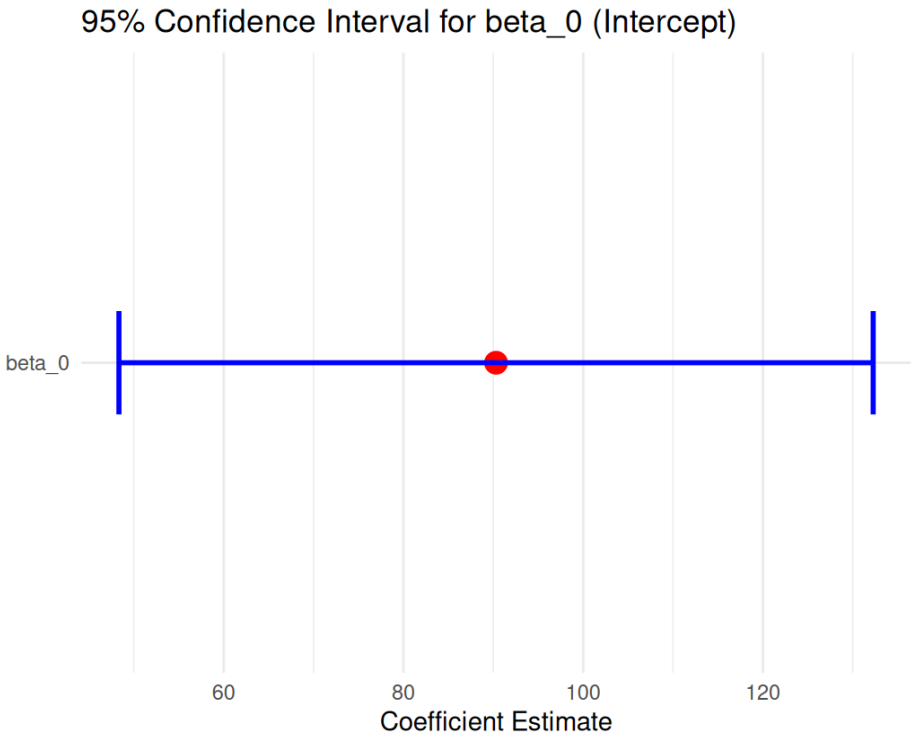

Figure 3: 95% confidence interval for β_0

Measure of Quality of Fit

R^2, is called the coefficient of determination. This quantity is a measure of the proportion of variability explained by the fitted model.

We know R^2 is given by:

$$R^2 = \frac{SSR}{SST}$$

Using R, we get the following values:

- SSR : 1848520.326
- SST : 2045831.333
- R^2 : 0.904

From the above R^2 value, we understand that 90.4% of the variability in building height is explained by the number of stories. This is a very strong fit.

Since we failed to reject the null hypothesis during lack-of-fit test, we can confirm that there is no non-linear relationship.

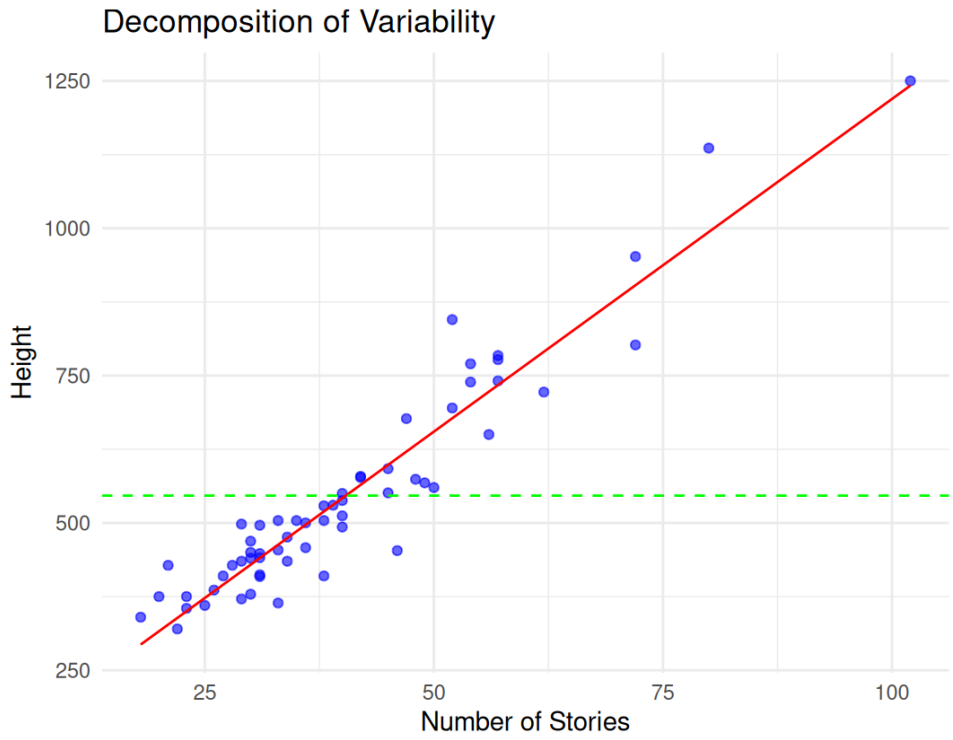

Figure 4: Mean of $Y(\overline{y})$ and Predicted Values $\hat{y}$

Residuals Plot

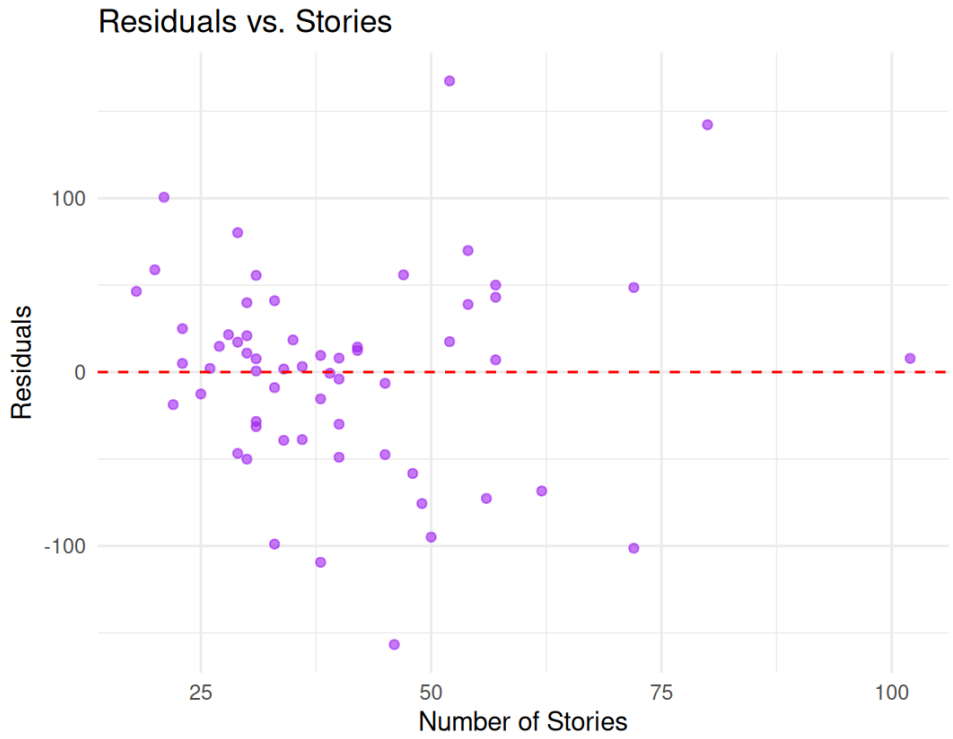

Figure 5: Residuals vs Number of Stories

The residuals appear to fluctuate randomly around the horizontal red line (which represents zero). This suggests that the residuals do not exhibit any systematic pattern, which is consistent with the assumption that the errors are random.

The spread of the residuals appears relatively constant across different values of Stories. There is no noticeable increase or decrease in the variability of residuals as the number of stories increases.

There are a few points with large residuals (both positive and negative), particularly at higher values of Stories (e.g., near 100 stories). These could be potential outliers.

There are no clear patterns such as curvature or trends in the residuals.

Conclusion:

The absence of patterns (e.g., curves) suggests that the relationship between Stories and Height is adequately captured by a linear model.

Confidence Interval for Mean Response

A $100(1-\alpha)\%$ confidence interval for the mean response $\mu_{Y|x_0}$ is:

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Using R, we graph the 95% confidence interval for mean response.

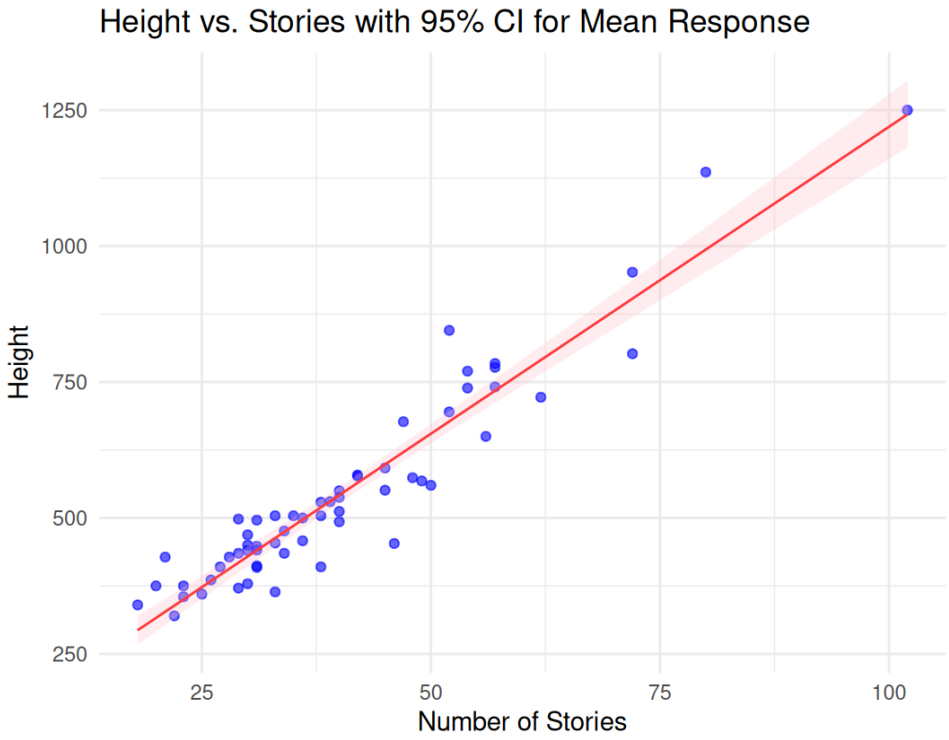

Figure 6: Height vs Number of Stories with 95% confidence interval

The narrow confidence band near the center confirms high precision in estimating the mean height for typical values of Stories.

The widening at the extremes highlights the increased uncertainty when predicting heights for buildings with very few or many stories (e.g., < 20 or > 80 stories).

Final Comments

After performing all the statistical analysis on the data, I would like to provide some of my final comments on the developed SLR model.

- High $R^2 = 0.904$ suggests 90.4% of the variability in building height is explained by the number of stories, indicating a very strong linear relationship.
- Pearson correlation $r = \sqrt{R^2} = 0.95$ confirms a near-perfect positive association between stories and height.
- Both the slope (β_1) and intercept (β_0) were statistically significant ($p \ll 0.05$), confirming the model's relevance.
- The linear model was deemed adequate by lack-of-fit test.
- The residuals showed no systematic patterns, supporting the assumption of linearity.
- A few high-story buildings had residuals outside the 95% confidence band.
- The confidence band for the mean response was narrowest around the average number of stories (40–60), reflecting higher precision.
- Predictions for very short or tall buildings (> 80 stories) have wider intervals, signaling increased uncertainty.

Final Recommendation

The SLR model can still be improved. We can incorporate domain knowledge. We can also test for non-linear relationships in the data.

Code

The entire analysis was performed using R studio. The code is available in my github repo: `https://github.com/beemarsh/slr_story_heights`