## **Men's Fashion Stores**



**Introducing and background:-**

* The aim of the report and analysis is to understand the factors influencing the sales performance of men's fashion stores and to develop a multiple linear regression model to predict the annual sales (tsales) based on various independent variables. By examining the relationships between the dependent variable and the key predictors, the goal is to gain insights into the drivers of sales and identify factors that significantly impact the store's performance.
* Question of Interest:

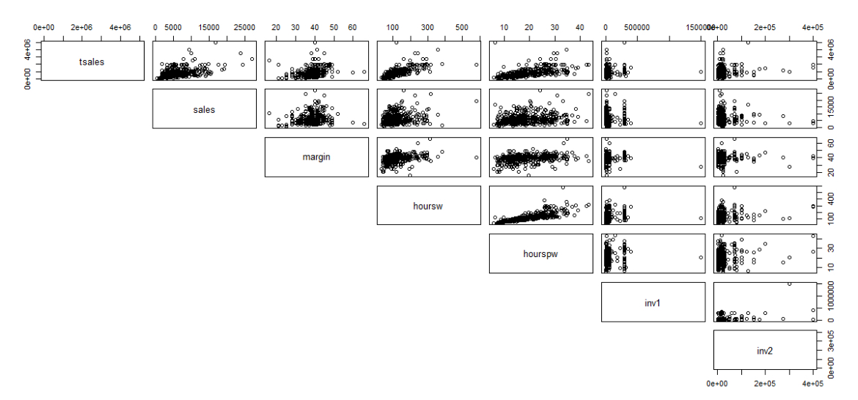
How do the sales per square meter (sales), gross-profit-margin (margin), total number of hours worked (hoursw), number of hours worked per worker (hourspw), investment in shop-premises (inv1), and investment in automation (inv2) relate to the annual sales (tsales) of the men's fashion stores?

**Exploratory analysis:**

* E(Y|x1,x2,x3,x4,x5,x6) =β0+β1x1+β2x2+ β3x3+ β4x4+ β5x5+ β6x6 Where

Y : Annual sales in Dutch guilders/X1: Sales per square meter./X2: Gross-profit-margin/X3: Total number of hours worked/X4: Number of hours worked per worker/X5: Investment in shop-premises./X6: Investment in automation

Β0 : the intercept while β1, β2, β3, β4, β5, β6, are the coefficients associated with each variable.



* First impressions that the linear model couldn’t be appropriate for all the variables like inv1 and inv2
* Also there may be strong linear relation between tsales and ( sales, margin ,hourssw and hourspw) when it may not be strong nor linear with inv1 and inv2

**tsales vs. sales: r=0.47**There is a moderate positive linear relationship between tsales and sales. The variability changes; there appears to be more spread at lower values of sales. The points lie in a fairly tight band indicating that the relationship is quite strong, and we can fit a straight line through the data and use it to predict tsales from sales

**tsales vs. margin: r= 0.24** There is a weak positive linear relationship between tsales and margin. The plot shows a significant amount of variability. The points are not tightly packed together,indicating that the relationship is weak, and we can fit a straight line through the data and use it to predict tsalesfrom margin

**tsales vs. hoursw: r=0.709** There is a strong positive linear relationship between tsales and hoursw. The plot shows a significant amount of variability. The points are not tightly packed together, indicating that the relationship is weak, and we can fit a straight line through the data and use it to predict tsalesfrom hoursw

**tsales vs. hourspw: r=0.552** There is a moderate positive linear relationship between tsales and hourspw. The plot shows a moderate amount of variability. The points are packed together, indicating that the relationship is moderate to strong, and we can fit a straight line through the data and use it to predict tsalesfrom hourspw

**tsales vs. inv1: r=0.19** There is a weak positive linear relationship between tsales and inv1.The plot shows a significant amount of variability. The points are not tightly packed together, indicating that the relationship is weak

**tsales vs. inv2 r=0.2** There is a weak positive linear relationship between tsales and inv2. The plot shows a significant amount of variability. The points are not tightly packed together, indicating that the relationship is weak

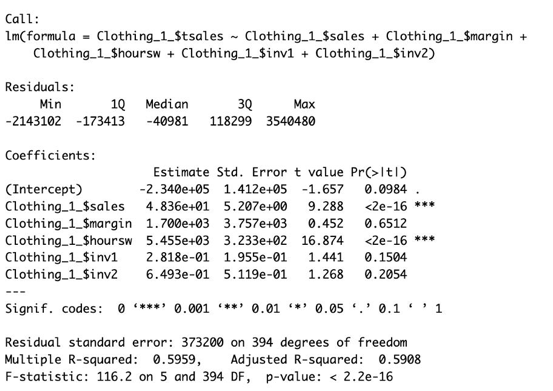
* Analysising these plots it suggests a potential multicollinearity issue between the variables hoursw (total number of hours worked) and hourspw (hours worked per worker). The correlation coefficient between hoursw and hourspw is 0.8077, indicating a strong positive correlation. Additionally, examining their correlations with tsales (annual sales), we find a correlation of 0.709 between tsales and hoursw, and 0.552 between tsales and hourspw.

Given these findings, it is advisable to address multicollinearity concerns by removing one of the correlated variables. In this case, it is recommended to retain hoursw, as it exhibits a higher correlation coefficient with tsales. This decision aims to enhance the model's stability and interpretability.

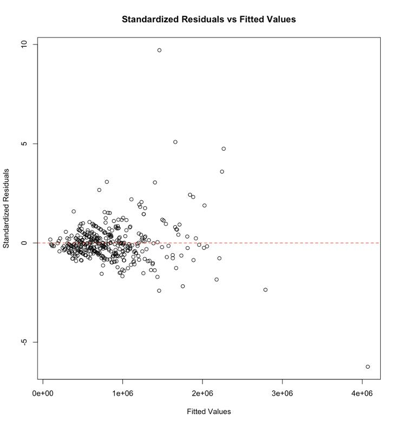
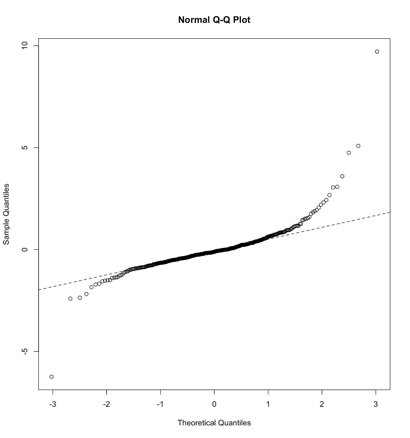
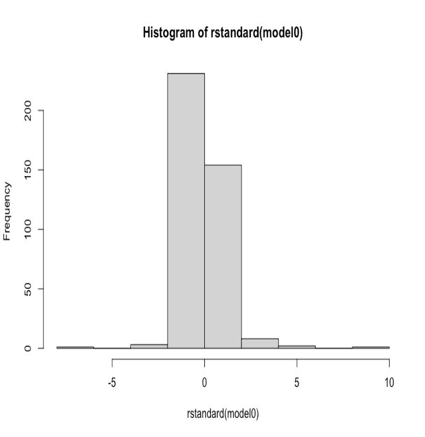
***Model 01*** : “model without hourspw

E(Y|x1,x2,x3 ,x5,x6,x7) =β0+β1x1+β2x2+ β3x3+ β5x5+ β6x6

Tsales =β0+β1(sales)+β2(margin)+ β3(hoursw)+ β5(inv1)+ β6(inv2)

******

Fitting this model we found that the adjusted r^2 ≈0.59 which refers to a moderate relationship



All assumptions regarding the deterministic part of the model are valid

* Check that E(ε*i*) = 0 and V(ε*i*) = σ2 for all *i*= 1,···*n*

It appears that the assumption of E(ε*i*) = 0 ∀*i*= 1, · · · *n*is valid but that of constant variance is not i.e. V(ε*i*) = σ*i*2. i.e. the deterministic part of the model captures the non-random structure in the data but the errors scale of variability is not constant at all values of the covariate.

The assumption of linearity between the predictors and the response variable is somewhat questionable based on this plot. Although there isn't a clear non-linear pattern, the distribution is quite dense near the origin and then fans out as the fitted values increase, which might suggest that the relationship isn't perfectly linear.

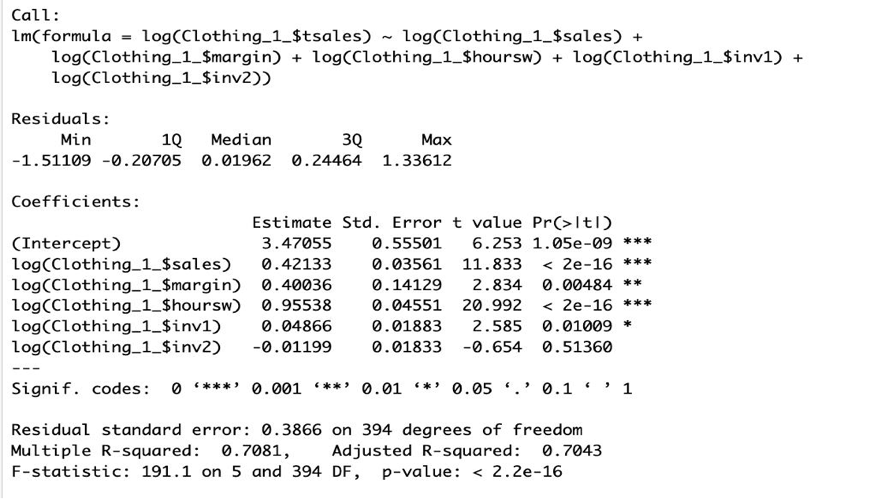
* Check the assumption of independent errors.

there is no intuitive natural order that we know about in the explanatory variable. So, we can assume independence between the errors.

* Check the assumption of normally distributed errors

There are deviations from the equality line and hence normality assumption is dubious

The regression model assumptions are invalid. Hence, the model 01 is not appropriate for the data , for a better model we choose to add log for all variables in the model ( response and explanatories )

**Model 001:” with log**”

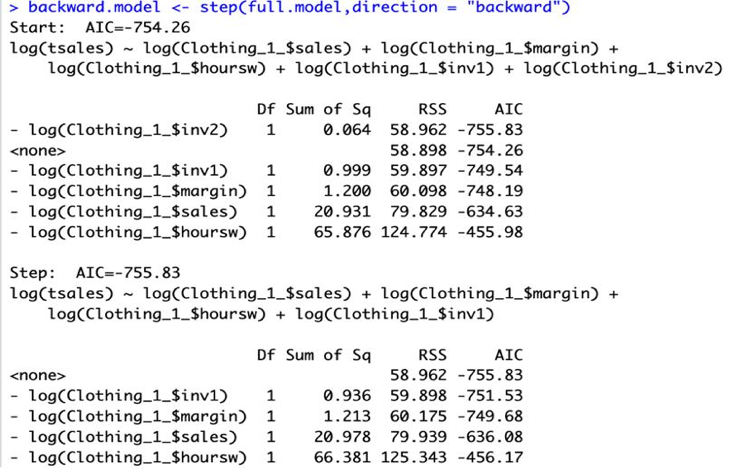
Log(tsales) = 3.47055 + 0.42133 \* Log(sales) + 0.40036 \* Log(margin) + 0.95538 \* Log(hoursw) + 0.04866 \* Log(inv1) -0.0119\*log( inv2)

In this model the adjusted r^2 ≈0,7043 which much better than the adjusted r^2 in model 01

We also checked the assumptions that were all valid “ we didn't add graphs for abbreviation " as they are somehow similar to the graphs for the final model

We have here 6 explanatory variables, this makes a very large number of models to be fitted if we use the all-subset regression procedure (64 models)

So for a better model we used the backward method as the follows

It seems that we will remove inv2 as a result of backward method for a better model

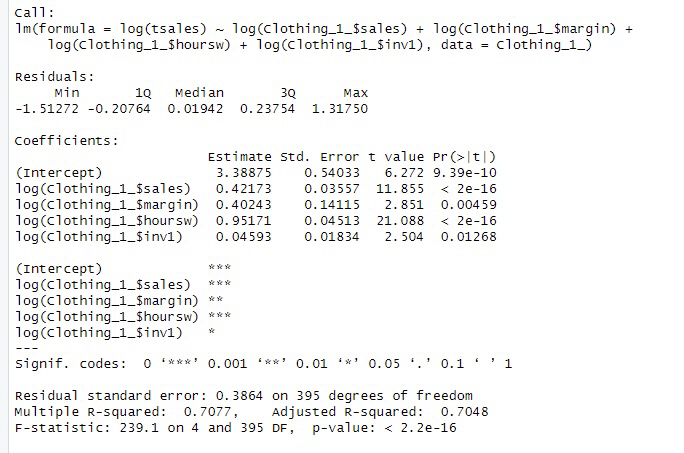
“Less explanatory variables is better to make the study of the model easier “

Final model:

Log(tsales) = β0 + β1 \* Log(sales) + β2 \* Log(margin) + β3 \* Log(hoursw) + β5 \* Log(inv1)

It is shown that to study the relation between annual sales we will study the relation with sales ,margin, hoursw and inv1

Answer: yes there is a relation between them and each explanatory variable affects the annual sale which we will explain later in the statistical analysis

**Statistical analysis**

Log(tsales) = 3.38875 + 0.42173 \* Log(sales) + 0.4243 \* Log(margin) + 0.95171 \* Log(hoursw) + 0.04593 \* Log(inv1)

* The regression analysis indicates that the natural logarithm of tsales is predicted by a linear combination of the natural logarithms of sales, margin, hoursw, and inv1. The coefficients associated with each predictor reveal the estimated impact of a one-unit change in the logarithm of each independent variable on the logarithm of tsales.
* We can interpret the relationship between the independent variables and the dependent variable as follows:

1. Sales per square meter (sales):

- A one-unit increase in the logarithm of sales is associated with an estimated increase of 0.42173 Dutch guilders in the logarithm of tsales (annual sales), holding other variables constant.

- This suggests that an increase in sales per square meter is positively associated with an increase in annual sales.

2. Gross-profit-margin (margin):

- A one-unit increase in the logarithm of margin is associated with an estimated increase of 0.4243 Dutch guilders in the logarithm of tsales, holding other variables constant. - This indicates that a higher gross-profit-margin is positively associated with higher annual sales.

3. Total number of hours worked (hoursw): - A one-unit increase in the logarithm of hoursw is associated with an estimated increase of 0.95171 Dutch guilders in the logarithm of tsales, holding other variables constant. - This suggests that an increase in the total number of hours worked is positively associated with higher annual sales.which is the most affective variable.

4. Investment in shop-premises (inv1):

- A one-unit increase in the logarithm of inv1 is associated with an estimated increase of 0.04593 Dutch guilders in the logarithm of tsales, holding other variables constant.

- This implies that higher investment in shop-premises is positively associated with increased annual sales but with a weak effect

5- the average annual sales is~3.39 for Dutch guilderszero sales,margin,hoursw and inv1

In summary, according to the regression equation, all the independent variables (sales, margin, hoursw, and inv1) have positive coefficients, indicating a positive relationship with annual sales. This suggests that higher sales per square meter, gross-profit-margin, total hours worked, and investment in shop-premises are associated with increased annual sales in men's fashion stores.

* Adjusted R-squared:

The adjusted R-squared value is an indicator of how well the model fits the data, taking into account the number of predictors and sample size. In this case, the adjusted R-squared is 0.7048.

An adjusted R-squared of 0.7048 suggests that approximately 70.48% of the variation in the logarithm of tsales can be explained by the logarithm of the explanatory variables included in the model (sales, margin, hoursw, and inv1). This indicates a reasonably good fit, as the adjusted R-squared value is relatively high

* The point and interval prediction:

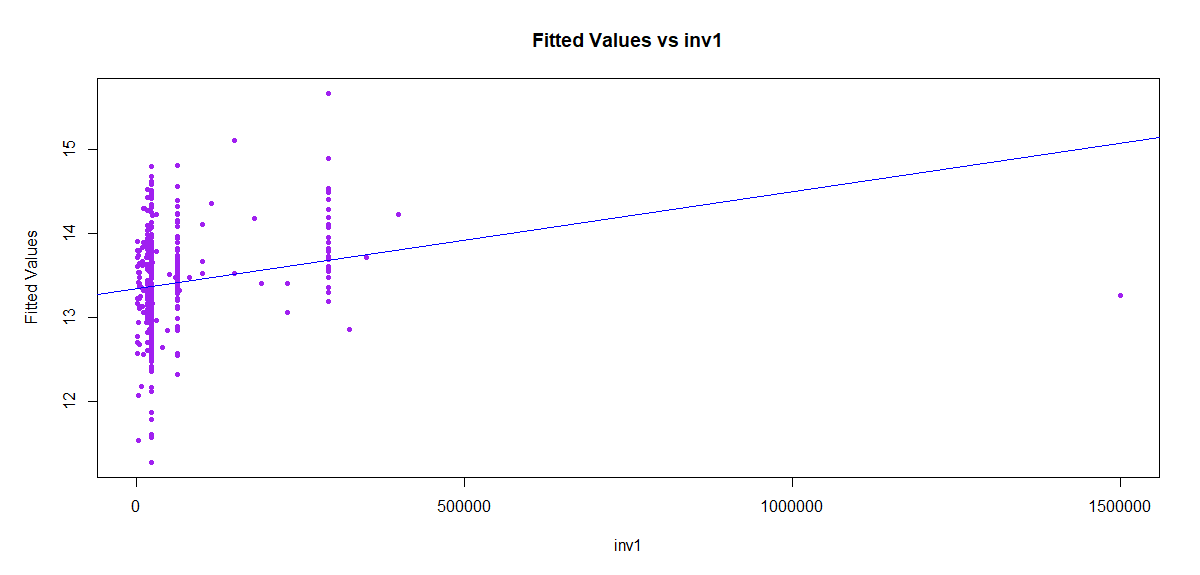
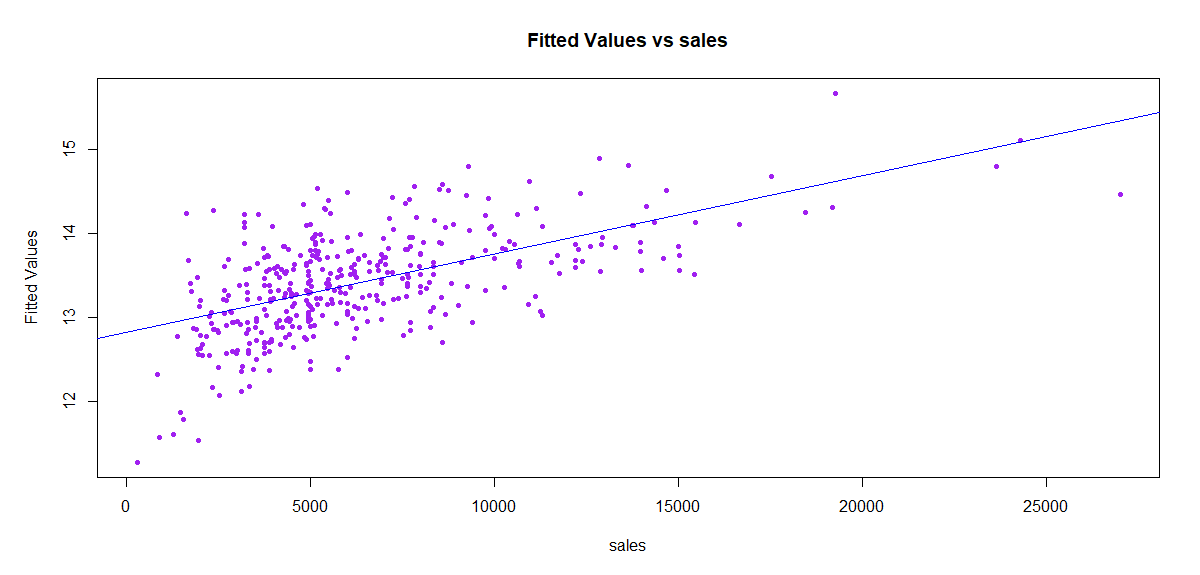
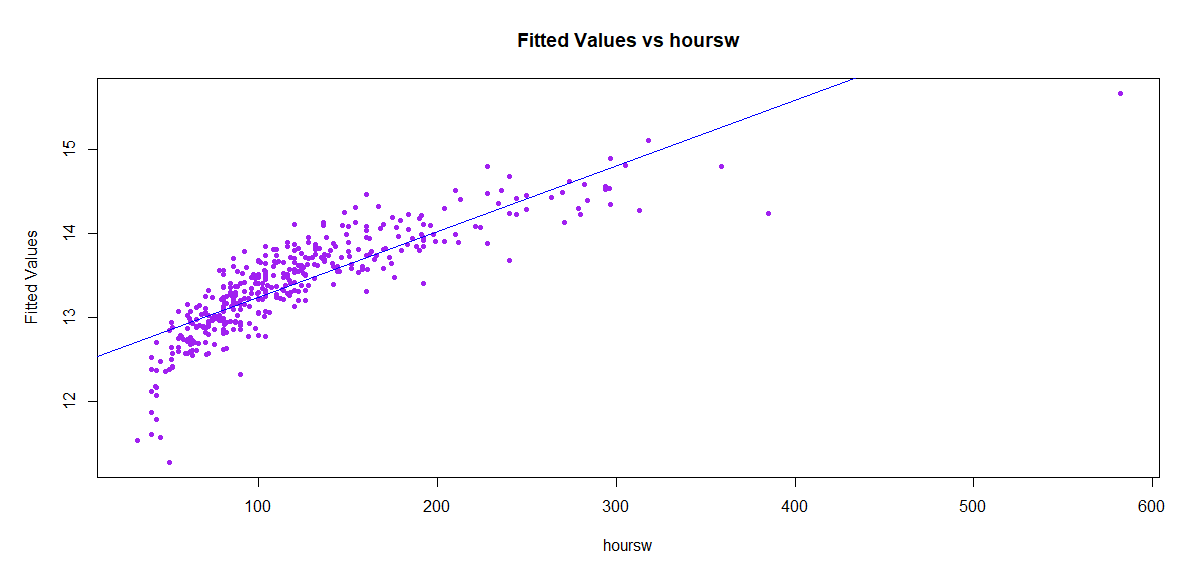
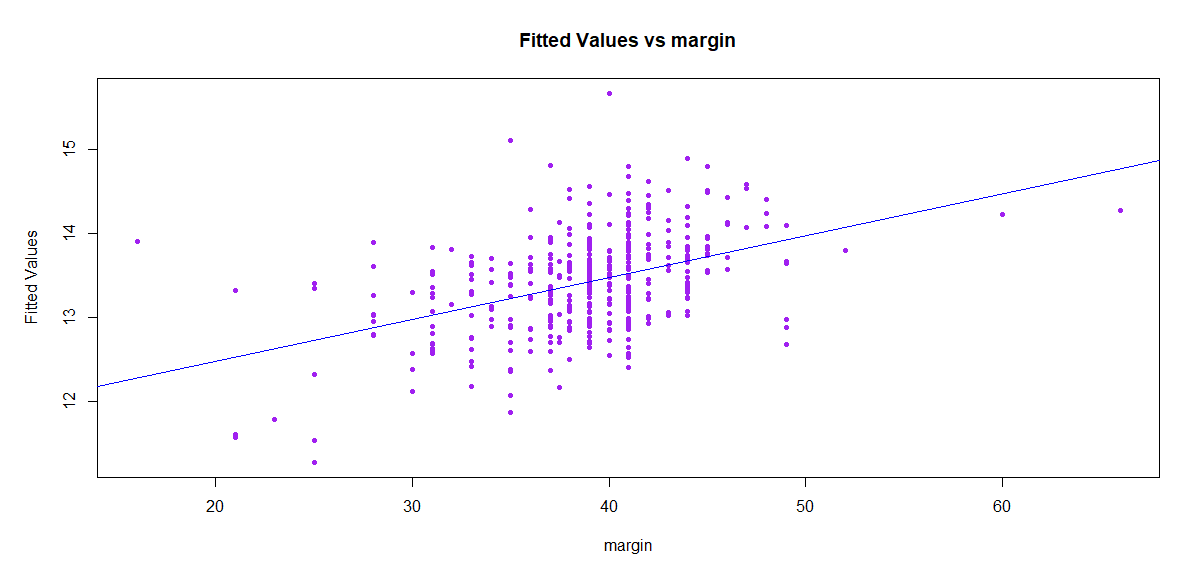
Sales=700 margin=50 hoursw=400 inv1=10000

Log(tsales) = 3.38875 + 0.42173 \* Log(700) + 0.4243 \* Log(50) + 0.95171 \* Log(400) + 0.04593 \* Log(10000)

Log(tsales)=1159.798

Sales=3000 margin=40 hoursw=50 inv1=400000

Log(tsales)=19705.13



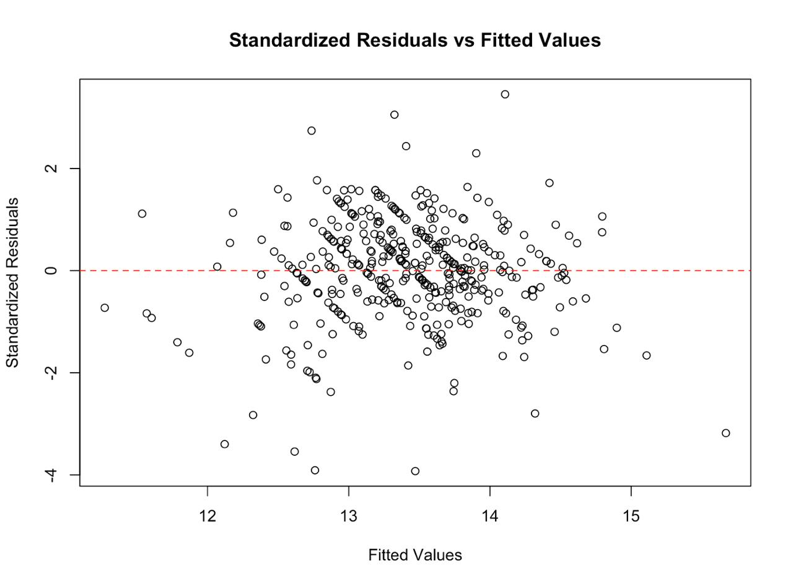
"Figure 9 a moderate linear relationship is apparent, although the data points are not well-scattered.

In Figure 10, illustrates a strong positive linear relationship, with some values notably deviating from the fitted line."

Figure 11 displays a positive linear relationship, potentially of moderate strength, as the data deviates from the line.

Finally, Figure 12 exhibits a weak, somewhat linear relationship with values concentrated towards the beginning of the line.

**Checking assumptions** A graph of a normal plot

Description automatically generated with medium confidenceFor the final model : Log(tsales) = 3.38875 + 0.42173 \* Log(sales) + 0.4243 \* Log(margin) + 0.95171 \* Log(hoursw) + 0.04593 \* Log(inv1)

* All the assumptions concerning the deterministic partare valid , nothing is mentioned about possibility of violation and the model is linear in the parameters and ( n=400) > ( p=5)
* It appears that the assumptions of The expected value of the random error is zero and The variability in the random error is constant (Homoscedasticity)are valid. i.e the deterministic part of the model captures all the non-random structure in the data and the scale of variability of the errors is constant at all values of the explanatory variables. The plot shows a fairly random scatter of residuals across the range of fitted values, without any obvious increasing or decreasing pattern in the spread of residuals. This suggests that the variance of the residuals is relatively constant for all levels of the fitted values.

It also appears that the residuals are randomly dispersed around the horizontal axis, suggesting that they are independent, indicating that the assumption of independence is not violated.

Therefore, it does not seem that there is a violation of the key assumptions of regression analysis

* The data points are closely aligned with the dashed line, indicating that the residuals are approximately normally distributed. This alignment suggests that the residuals of this regression model are approximately normally distributed
* However, there are a few points that deviate from the line, especially in the tails this could suggest the presence of outliers or heavy-tailed distributions, which might violate the assumption of normally distributed errors. But for a 400 observations it can be accepted compared with the histogram so Normality assumption is not violated
* A graph of a bar graph

  Description automatically generatedit appears that the residuals are approximately normally distributed, as the shape of the histogram resembles the bell shape of a normal distribution. However, there are some deviations, particularly in the tails, which could suggest slight skewness or kurtosis. This shows that Normality assumption is valid

A graph with a red line

Description automatically generated

* Most of the data points are clustered around the center, indicating a good fit for most observations. However, there are a few outliers, which could potentially influence the model’s performance.

The red line indicating Cook’s distance shows an increase as leverage increases. Cook’s distance measures the effect of deleting a given observation. Points with a large Cook’s distance are considered to merit closer examination in the context of the analysis.

* We think keeping all data points is more useful for the analysis.

**Conclusion:**

Our analysis commenced with an assessment of multicollinearity among explanatory variables, revealing a substantial correlation between the variables hoursp and hoursw. Consequently, a decision was made to retain only the variable hoursw.Initially, a linear model was fitted to the data, and the ensuing diagnostic checks revealed violations of model assumptions. To address this, various transformations, such as logarithmic and square root transformations, were applied. Employing a backward elimination approach, variables were systematically removed to enhance model fit. Subsequent checks on the assumptions were performed, and a forward regression was also executed. Remarkably, both approaches converged on a common model.

The resulting model demonstrated adherence to key assumptions, thereby achieving a more robust representation of the relationship between explanatory variables and annual sales (tsales). Notably, both the formal analysis and initial impressions concurred on the presence of a strong positive linear relationship between tsales and hoursw, a moderate relationship between tsales and sales, and a weak relationship between tsales and inv1.

Despite an initial perception of a weak association between tsales and margin, the formal analysis revealed a transformation of this relationship from weak to moderate. Additionally, through the process of backward elimination, the variable inv2 was removed from the model, aligning with the goal of refining the model and addressing potential multicollinearity.

This systematic and iterative approach, involving both variable transformations and model refinement, facilitated the creation of a model that not only aligns with initial impressions but also upholds key regression assumptions. The analytical journey, marked by thoughtful consideration of transformations and systematic variable selection, contributes to the robustness and reliability of the final model.