DS 853 Case Study 2

Prepared by:

Marcus Nogueira and Chin Ting Wong

Professor:

Leyla Ozsen

Date:

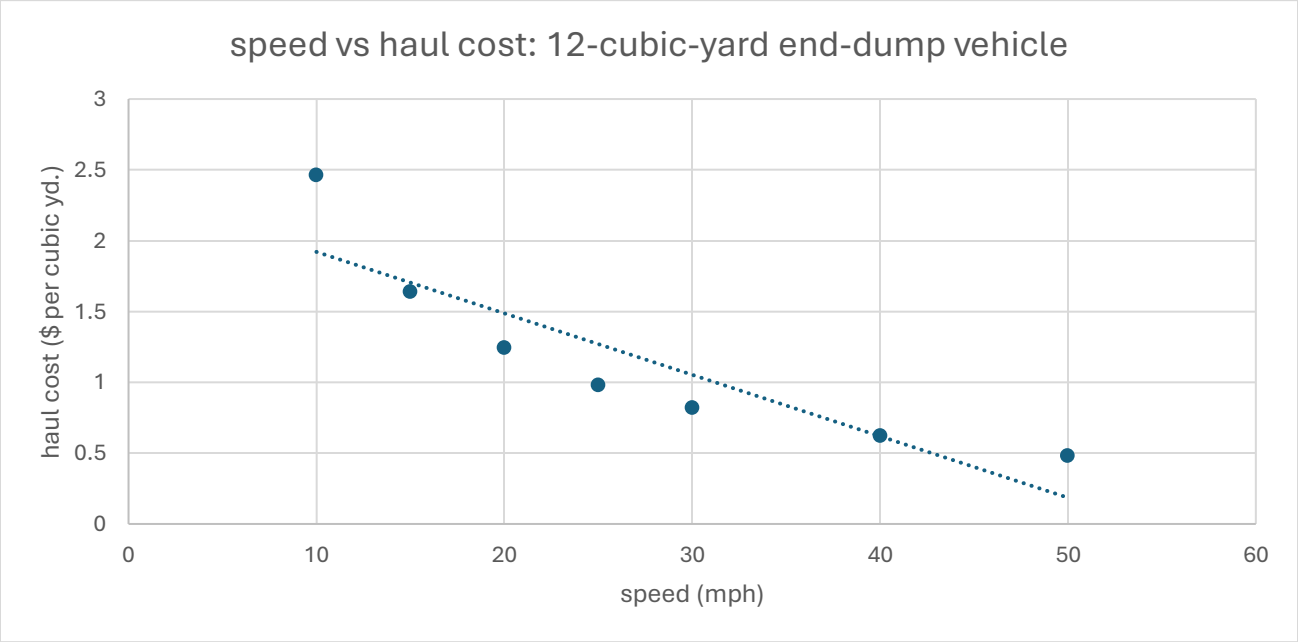
October 2024

Question 1

Part A: 12-cubic-yard end-dump vehicle

1. Create a scatterplot with the independent variable on the x-axis and the dependent variable on the y-axis.

   As we are predicting the haul cost by speed of the vehicle, the independent variable on the x-axis is speed(mph) and the dependent variable on the y-axis is haul cost 12-cubic-yard-end-dump vehicle ($ per cubic yd.).



2. Calculate the correlation using Excel and interpret it.

   Using Excel's correlation function, we found a strong negative correlation of -0.8913 between "12-Cubic Yard" and "Speed." This indicates that as speed increases, the haul cost per cubic yard for the 12-cubic-yard end-dump vehicle decreases.

3. Discuss whether a linear regression analysis seem like a good choice based on your analysis above in step 1 and step 2.

In step 1, the scatterplot reveals that the data points generally cluster around a downward-sloping straight line, suggesting a potential linear relationship between speed and haul cost. However, the point (10, 2.46) may be an outlier, which could influence the regression results. In step 2, the calculated correlation of -0.8913 reinforces the existence of a strong negative relationship between speed and haul cost, indicating that as speed increases, haul costs tend to decrease significantly.

To conclude the analysis regarding the appropriateness of linear regression, it is evident that the strong correlation and the visual representation in the scatterplot suggest a linear relationship between speed and haul cost. Given the generally linear pattern of the data points, applying linear regression could effectively model this relationship. However, it is crucial to consider the outlier at (10, 2.46), as it may skew the results and affect the overall fit of the model.

4. Go ahead and do a regression analysis. Include the regression analysis summary output that you gathered using Excel. You can copy/paste it or include a screenshot of the Excel output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.891318508
R Square	0.794448682
Adjusted R Square	0.753338419
Standard Error	0.340893156
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.245702139	2.245702139	19.32482584	0.00704818
Residual	5	0.581040719	0.116208144		
Total	6	2.826742857			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.354850299	0.297277409	7.92139002	0.00051625	1.590674392	3.119026207	1.590674392	3.119026207
Speed	-0.043389222	0.009870157	-4.396001119	0.00704818	-0.068761268	-0.018017175	-0.068761268	-0.018017175

5. What is the R^2 of the model? Interpret it.

The R^2 of the model is 0.7944. This value indicates that about 79.44% of the variability in haul cost can be explained by the speed of the vehicle.

6. What is the p-value associated with the overall model F value and the t test of the slope of the regression model?

The p-value associated with the overall model, and the p-value for the t-test of the slope are both 0.0070.

7. What is the statistical conclusion based on this p-value?

Since the p-value of 0.0070 is less than the common alpha level of 0.05, we would reject the null hypothesis, which in this context typically states that there is no relationship between speed and haul cost. In other words, the p-value indicates that speed is a significant predictor of haul cost.

8. What is the standard error of the regression model?

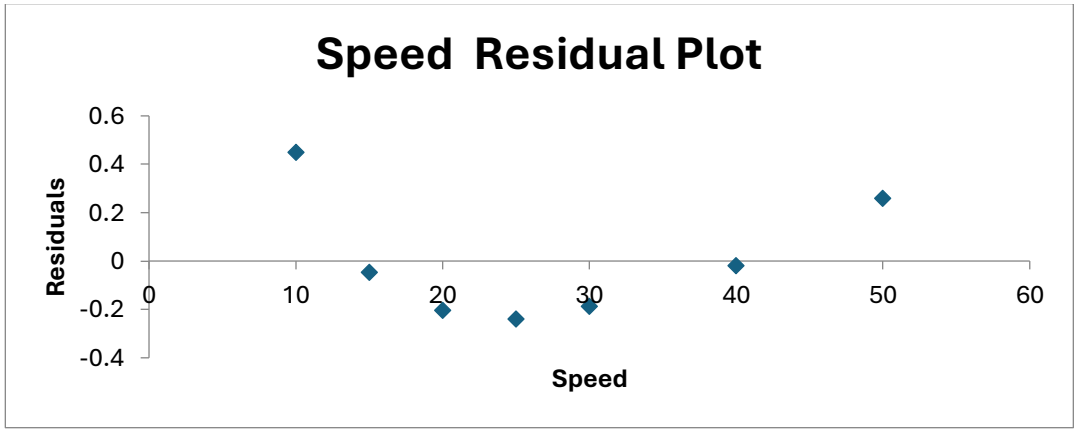
The standard error of the regression model is 0.3409.

9. Examine the residual outputs and calculate what percentage of the residuals are within one standard error.

The residual values range from -0.2901 to 0.5390, with 6 out of 7 residuals, or 85.71%, falling within one standard error, which ranges between -0.3409 and 0.3409.

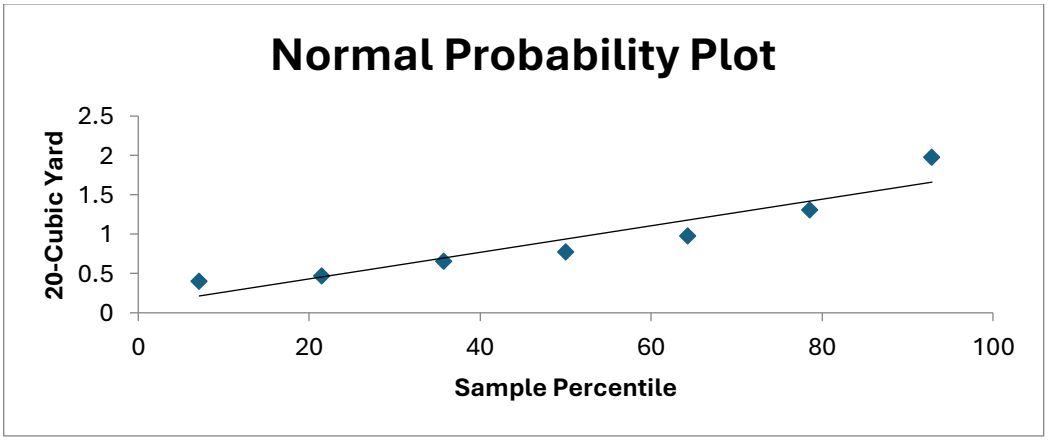
10. Interpret the residual plot.

The residual plot reveals a U-shaped pattern, which suggests that the linear regression model may not adequately capture the relationship between the independent variable (speed) and the dependent variable (haul cost). Ideally, residuals should be randomly scattered around zero, indicating that the model's predictions are unbiased across the range of data. However, the presence of a U-curve, implies that there may be a nonlinear relationship between speed and haul cost that is not being addressed by the linear model.



11. Interpret the normality plot.

The normal probability plot indicates that the residuals from the linear regression model are approximately normally distributed, as evidenced by the points lying close to an upward slopping straight line.



12. What is the linear regression model (the equation)?

The linear regression model (the equation) is y = 2.3549 -0.0434x. In this case, haul cost = 2.3549 – 0.0434 speed

13. Interpret the slope of the model.

The slope of the model is -0.0434. This indicates every 1mph increase in speed would result in $0.04 per cubic yard decrease in haul cost 12-cubic-yard-end-dump vehicle.

14. Construct a 95% confidence interval for the mean haul cost for all vehicles with a speed of 35 mph.

The 95% confidence interval for the mean haul cost for all vehicles with a speed of 35 mph is approximately: Confidence Interval = (0.45, 1.22)

15. Construct a 95% prediction interval for the haul cost for when a single vehicle with a speed of 45 mph.

The 95% prediction interval for the haul cost when a single vehicle with a speed of 45 mph is approximately: Prediction Interval = (-0.64, 1.44)

16. What are your overall conclusions and recommendations as a business consultant?

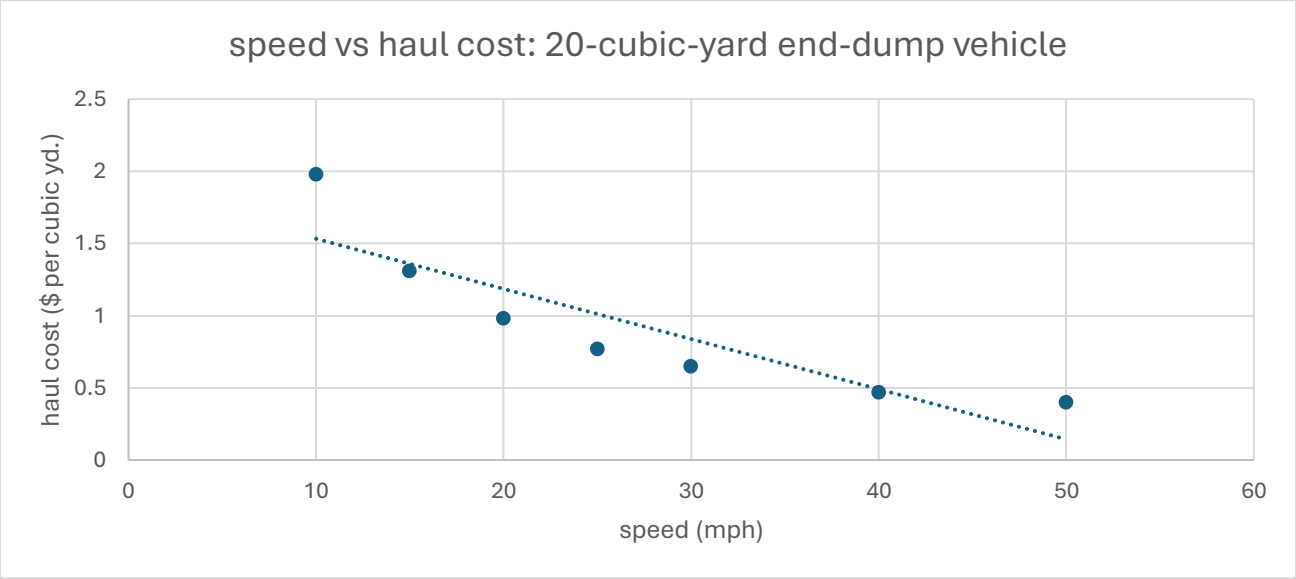
Our analysis reveals a strong negative correlation of -0.8913 between vehicle speed and haul cost, suggesting that as speed increases, haul costs per cubic yard significantly decrease. The model's R^2 of 0.7944 further indicates that approximately 79.44% of the variability in haul cost is explained by speed. Additionally, the p-value of 0.0070 supports the statistical significance of this relationship, allowing us to reject the null hypothesis and confirm that speed is an important predictor of haul cost. The standard error of 0.3409 suggests reasonable predictive accuracy.

However, the U-shaped pattern observed in the residual plot indicates that the linearity assumption may not hold, suggesting a nonlinear relationship between speed and haul cost. Exploring a polynomial or nonlinear regression could provide a better fit and more accurate predictions, which may further aid in optimizing haul costs.

Part B: 20-cubic-yard end-dump vehicle

1. Create a scatterplot with the independent variable on the x-axis and the dependent variable on the y-axis.

   As we are predicting the haul cost by speed of the vehicle, the independent variable on the x-axis is speed(mph) and the dependent variable on the y-axis is haul cost 20-cubic-yard-end-dump vehicle ($ per cubic yd.).



2. Calculate the correlation using Excel and interpret it.

   Using the correlation function from data analysis in Excel, we obtained a correlation of -0.8836 between the variables "20-Cubic Yard" and "Speed". This indicates a strong negative correlation, and that means that as speed increases, the haul cost per cubic yard for the 20-cubic-yard-end-dump vehicle decreases.

3. Discuss whether a linear regression analysis seem like a good choice based on your analysis above in step 1 and step 2.

In step 1, the scatterplot reveals that the data points generally cluster around a downward-sloping straight line, suggesting a potential linear relationship between speed and haul cost. However, the point (10, 1.98) may be an outlier, which could influence the regression results. In step 2, the calculated correlation of -0.8836 reinforces the existence of a strong negative relationship between speed and haul cost, indicating that as speed increases, haul costs tend to decrease significantly.

To conclude the analysis regarding the appropriateness of linear regression, it is evident that the strong correlation and the visual representation in the scatterplot suggest a linear relationship between speed and haul cost. Given the generally linear pattern of the data points, applying linear regression could effectively model this relationship. However, it is crucial to consider the outlier at (10, 1.98), as it may skew the results and affect the overall fit of the model.

4. Go ahead and do a regression analysis. Include the regression analysis summary output that you gathered using Excel. You can copy/paste it or include a screenshot of the Excel output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.883574316
R Square	0.780703571
Adjusted R Square	0.736844285
Standard Error	0.284506927
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.440821899	1.440821899	17.80018887	0.008335592
Residual	5	0.404720958	0.080944192		
Total	6	1.845542857			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.880479042	0.248105545	7.579351137	0.000634415	1.242703436	2.518254648	1.242703436	2.518254648
Speed	-0.034754491	0.008237561	-4.219027006	0.008335592	-0.055929815	-0.013579167	-0.055929815	-0.013579167

5. What is the R^2 of the model? Interpret it.

   The R^2 of the model is 0.7807. This value indicates that about 78.07% of the variability in haul cost can be explained by the speed of the vehicle.

6. What is the p-value associated with the overall model F value and the t test of the slope of the regression model?

   The p-value associated with the overall model, and the p-value for the t-test of the slope are both 0.0083.

7. What is the statistical conclusion based on this p-value?

   Since the p-value of 0.0083 is less than the common alpha level of 0.05, we would reject the null hypothesis, which in this context typically states that there is no relationship between speed and haul cost. In other words, the p-value indicates that speed is a significant predictor of haul cost.

8. What is the standard error of the regression model?

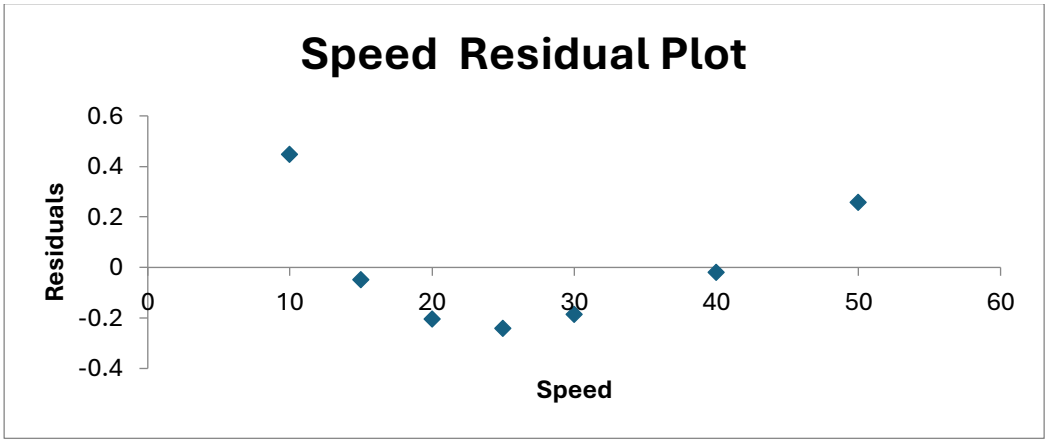
   The standard error of the regression model is 0.2845.

9. Examine the residual outputs and calculate what percentage of the residuals are within one standard error.

   The residual outputs range between -0.2416 and 0.4471, with 6 out of 7 residuals, or 85.71%, falling within one standard error, which ranges between -0.2845 and 0.2845.

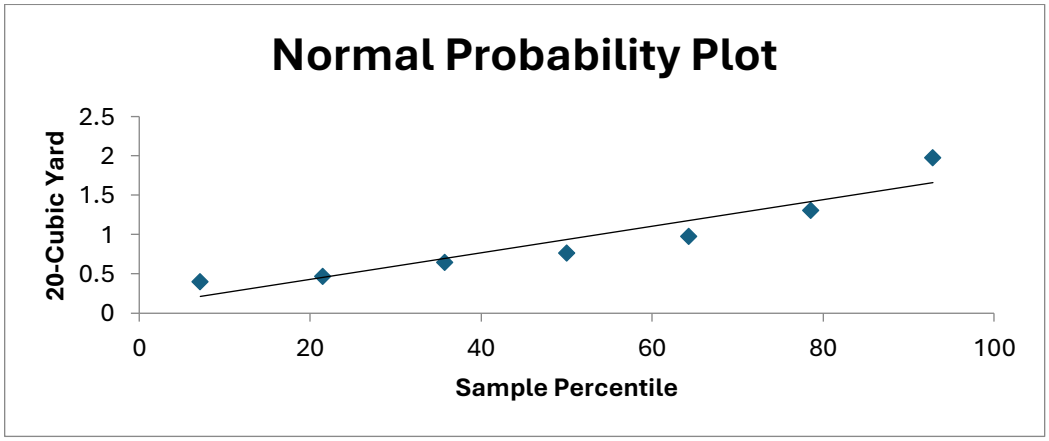
10. Interpret the residual plot.

The residual plot reveals a U-shaped pattern, which suggests that the linear regression model may not adequately capture the relationship between the independent variable (speed) and the dependent variable (haul cost). Ideally, residuals should be randomly scattered around zero, indicating that the model's predictions are unbiased across the range of data. However, the presence of a U-curve, implies that there may be a nonlinear relationship between speed and haul cost that is not being addressed by the linear model.



11. Interpret the normality plot.

The normal probability plot indicates that the residuals from the linear regression model are approximately normally distributed, as evidenced by the points lying close to an upward slopping straight line.



12. What is the linear regression model (the equation)?

The linear regression model (the equation) is y = 1.8805 – 0.0348x. In this case, haul cost = 1.8805 – 0.0348 speed

13. Interpret the slope of the model.

The slope of the model is -0.0348. This indicates that every 1mph increase in speed would result in $0.03 per cubic yard decrease in haul cost of 20-cubic-yard-end-dump vehicle.

14. Construct a 95% confidence interval for the mean haul cost for all vehicles with a speed of 35 mph.

The 95% confidence interval for the mean haul cost for all vehicles with a speed of 35 mph is approximately: Confidence Interval = (0.34, 0.99)

15. Construct a 95% prediction interval for the haul cost for when a single vehicle with a speed of 45 mph.

The 95% prediction interval for the haul cost when a single vehicle with a speed of 45 mph is approximately: Prediction Interval = (-0.55, 1.19)

16. What are your overall conclusions and recommendations as a business consultant?

Our analysis reveals a strong negative correlation of -0.8836 between vehicle speed and haul cost, indicating that as speed increases, the haul cost per cubic yard for the 20-cubic-yard-end-dump vehicle decreases significantly. The model's R^2 value of 0.7807 further indicates that approximately 78.07% of the variability in haul cost is explained by speed. Additionally, the p-value of 0.0083 supports the statistical significance of this relationship, allowing us to reject the null hypothesis and confirm that

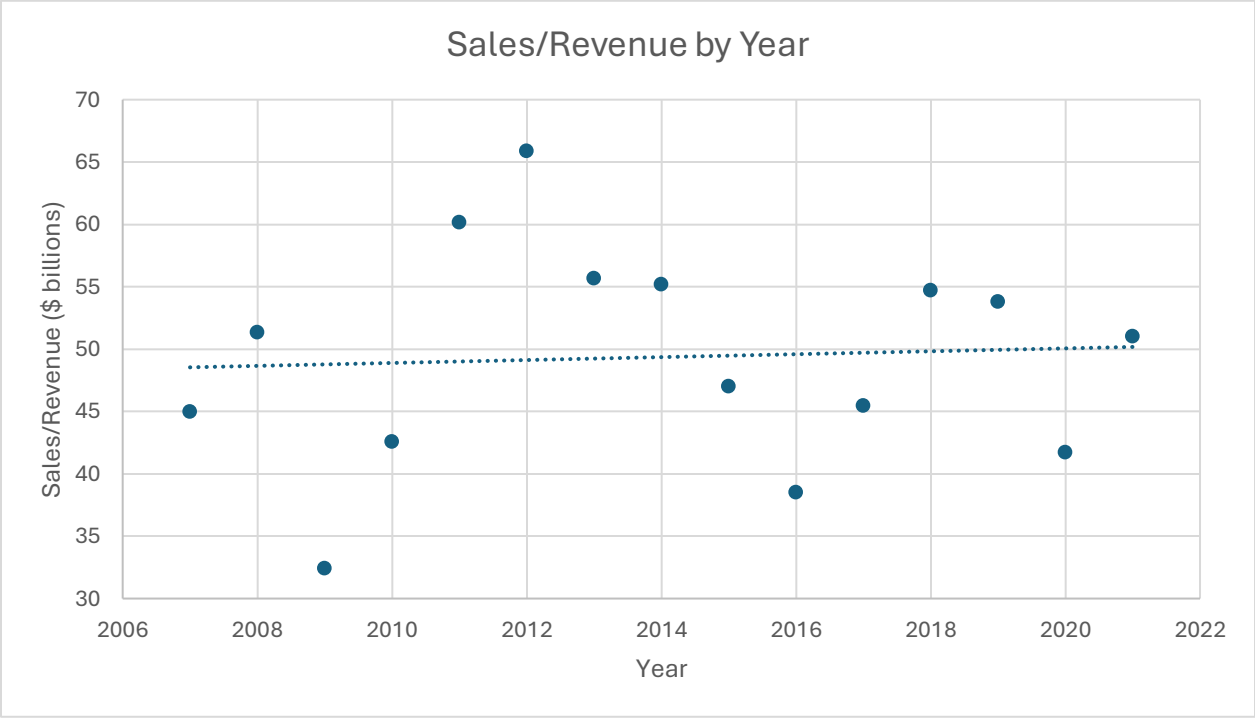
speed is a meaningful predictor of haul cost. The standard error of 0.2845 suggests a reasonable degree of predictive accuracy.

However, the U-shaped pattern observed in the residual plot suggests that the linearity assumption may not hold, indicating a possible nonlinear relationship between speed and haul cost. Exploring polynomial or other nonlinear regression models could provide a better fit and more accurate predictions, potentially enhancing strategies for optimizing haul costs.

Question 2

1. Create a scatterplot with the independent variable on the x-axis and the dependent variable on the y-axis.

As we are predicting the sales and revenue streams ($ billions) by year, the independent variable on the x-axis is year and the dependent variable on the y-axis is sales and revenue streams ($ billions).



2. Calculate the correlation using Excel and interpret it.

Using the correlation function from data analysis in Excel, we obtained a correlation of 0.0598 between the variable "Year" and "Sales/Revenue Streams ($ billions)". This indicates a very weak positive correlation, suggesting that there is almost no relationship between the year and sales or revenue streams. In practical terms, this means that changes in the year have little to no impact on sales or revenue streams levels, implying that other factors may play a more significant role in influencing sales over time.

3. Discuss whether a linear regression analysis seem like a good choice based on your analysis above in step 1 and step 2.

The scatterplot reveals a fluctuating trend, with data points showing an alternating pattern of increases and decreases around the trendline. This suggests that while there may be some underlying relationship between "Year" and "Sales/Revenue Streams", and it is not consistently linear. The correlation of 0.0598 further reinforces this observation, indicating a very weak positive relationship.

Given these characteristics, a linear regression analysis may not be the most appropriate choice for modeling the relationship between these variables. The non-linear pattern of the data suggests that factors other than the year could be influencing sales or revenue levels more significantly.

4. Go ahead and do a regression analysis. Include the regression analysis summary output that you gathered using Excel. You can copy/paste it or include a screenshot of the Excel output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.059844649
R Square	0.003581382
Adjusted R Square	-0.073066204
Standard Error	9.043268914
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.82122893	3.82122893	0.04672531	0.83221763
Residual	13	1063.14926440	81.78071265		
Total	14	1066.97049333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-185.92102	1088.44602	-0.17081	0.86700	-2537.36569	2165.52364	-2537.36569	2165.52364
Year	0.11682	0.54044	0.21616	0.83222	-1.05073	1.28437	-1.05073	1.28437

5. What is the R^2 of the model? Interpret it.

The R^2 of the model is 0.0036. This value indicates that about 0.36% of the variability in sales/revenue streams can be explained by the year variable. This extremely low R^2 value suggests that there is virtually no explanatory power in the model regarding the relationship between year and sales/revenue streams.

6. What is the p-value associated with the overall model F value and the t test of the slope of the regression model?

The p-value associated with the overall model, and the p-value for the t-test of the slope are both 0.8322.

7. What is the statistical conclusion based on this p-value?

Since the p-value of 0.8322 is greater than the common alpha level of 0.05, we fail to reject the null hypothesis, which in this context states that there is no significant relationship between the year and sales/revenue. This means that the year variable is not a significant predictor of sales or revenue. In other words, the p-value indicates that changes in the year do not have a meaningful impact on sales/revenue streams levels, suggesting that other factors are likely more influential in driving sales performance.

8. What is the standard error of the regression model?

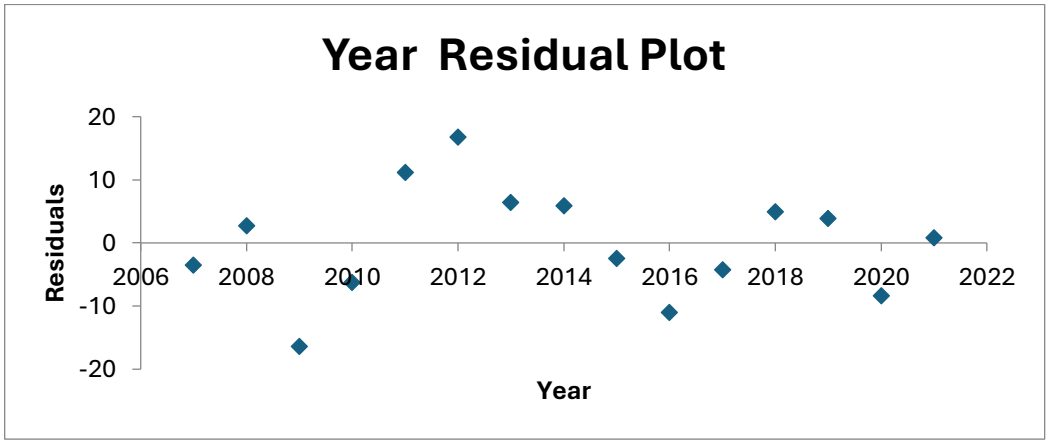
The standard error of the regression model is 9.0433.

9. Examine the residual outputs and calculate what percentage of them are within one standard error.

The residual outputs range between -16.3732 and 16.7563, with 11 out of 15 residuals, or 73.33%, falling within one standard error, which ranges between -9.0433 and 9.0433.

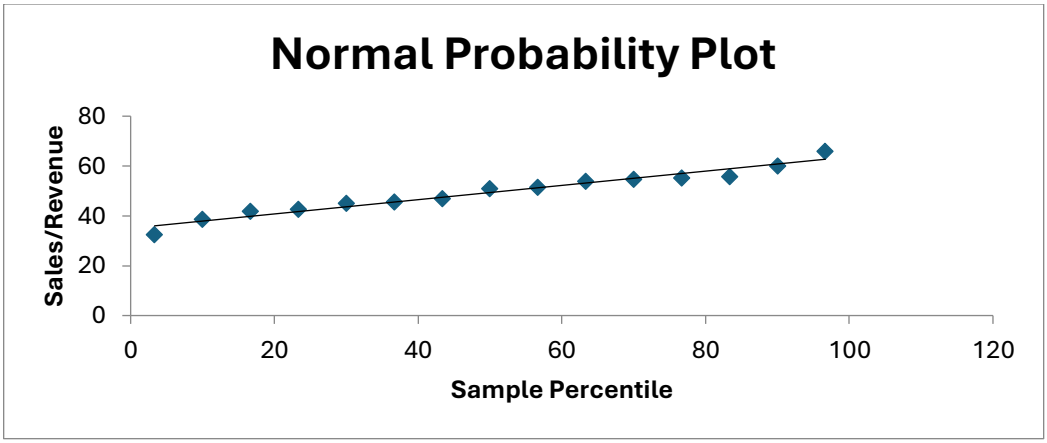
10. Interpret the residual plot.

The residual plot reveals a fluctuating pattern, with points moving up and down rather than remaining close to zero. This suggests that the linear regression model may not adequately capture the relationship between the independent variable (year) and the dependent variable (sales/revenue streams). Ideally, residuals should be randomly scattered around zero, indicating that the model's predictions are unbiased across the range of data. However, the observed pattern implies that there may be underlying nonlinear relationships or other influential factors affecting sales/revenue that are not accounted for in the model.



11. Interpret the normality plot.

The normal probability plot indicates that the residuals from the linear regression model are approximately normally distributed, as evidenced by the points lying close to a straight line.



12. What is the linear regression model (the equation)?

The linear regression model (the equation) is y = -185.92 + 0.1168 x. In this case, revenue/sales streams = -185.92 + 0.1168 year.

13. Interpret the slope of the model.

The slope of the model is 0.1168. This indicates that every unit increase in year would result in $0.1168 billion increase in sales/revenue streams.

14. Predict the Sales and Revenue Streams for year 2022.

The predicted Sales and Revenue Streams for year 2022 is $50.29 billion.

15. What are your overall conclusions and recommendations as a business consultant?

Our analysis reveals a very weak positive correlation (0.0598) between year and sales/revenue streams, with an R^2 of 0.0036, indicating that only 0.36% of the variability in sales can be attributed to changes in the year. Additionally, the high p-value of 0.8322 suggests that year is not a statistically

significant predictor of sales/revenue, underscoring that other factors likely play a more substantial role.

The residual plot displays a fluctuating pattern, indicating that the linear regression model may not adequately capture the relationship between year and sales/revenue. Ideally, residuals should be randomly scattered around zero, suggesting that the model's predictions are unbiased. The observed pattern implies potential nonlinear relationships or other influential factors affecting sales/revenue that are not accounted for in the model.

Given these insights, the business should focus on investigating additional variables—such as market trends, consumer preferences, or economic conditions—that may better explain sales performance. Exploring nonlinear relationships may also provide a more accurate fit for the data. This expanded approach could yield actionable insights, aiding strategic decision-making and optimizing revenue growth.