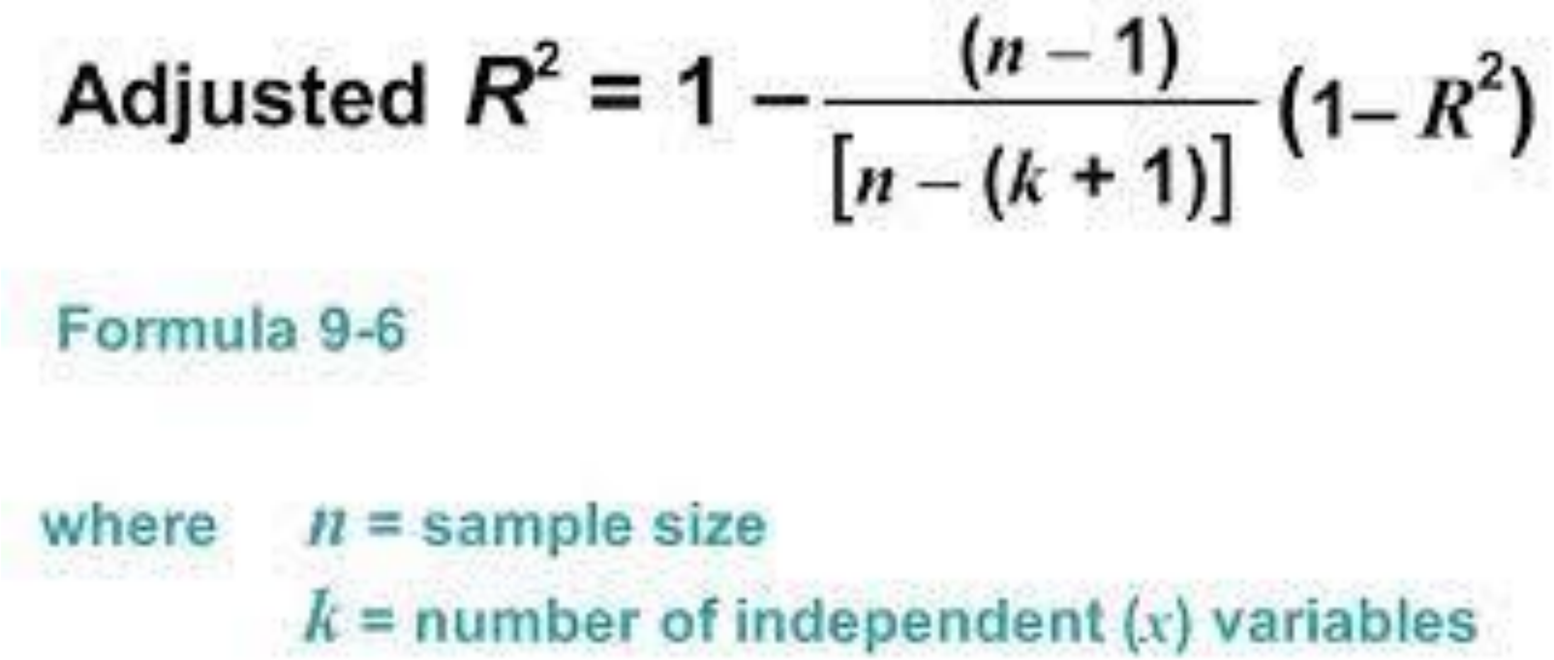
Session 5

Simple Linear Regression (II) & (III):

Inference & Prediction

Adjusted Squared



$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Formula 9-6

where n = sample size

k = number of independent (x) variables

Diagnostic checks: Using OLS residuals

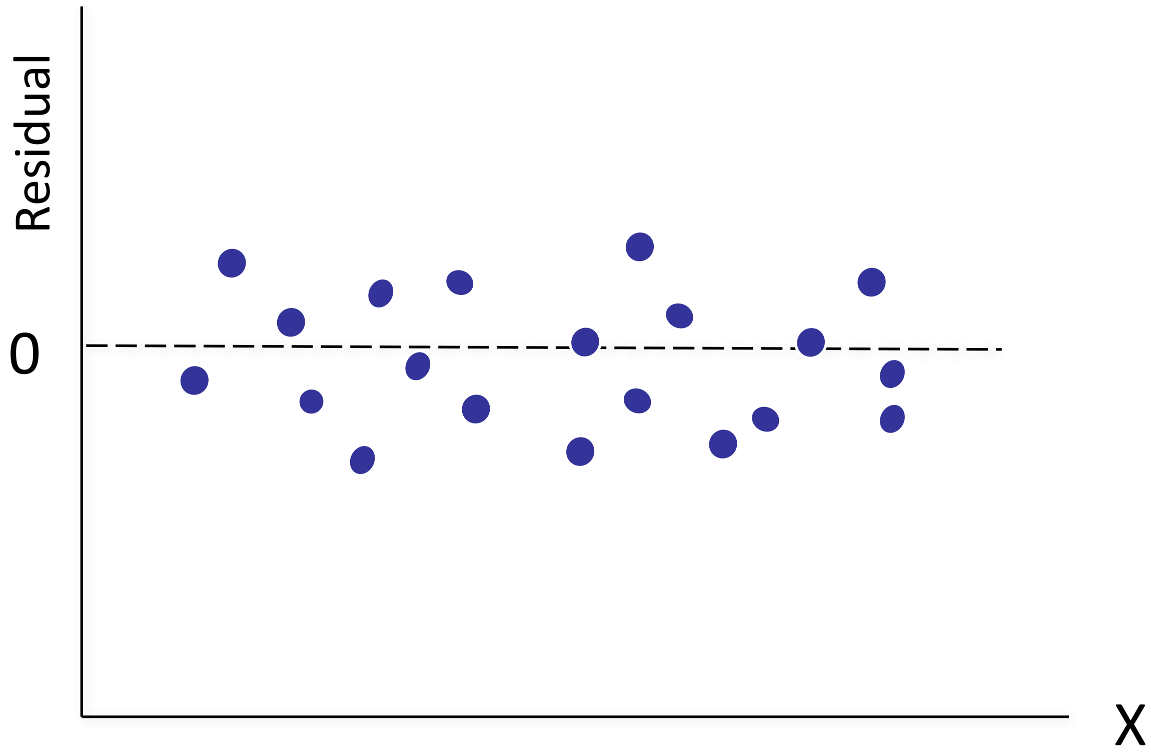
- We need to check the appropriateness of the following assumptions
  1. $E[\varepsilon|X] = 0$
  2. Homoskedasticity: $Var[\varepsilon|X]= \sigma_\varepsilon^2$
  3. $Corr[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$
  4. Normality of errors: $\varepsilon\,|X \sim N(0, \sigma_\varepsilon^2)$

- Other key diagnostic checks include
  - Impact of Outliers
  - Linear relationship between Y and X

- Violations of these assumptions cause problems e.g. bias, inefficiency, incorrect inference

- We can use plots of residuals to get an idea if the assumptions are satisfied

Residuals vs. Predictor Values



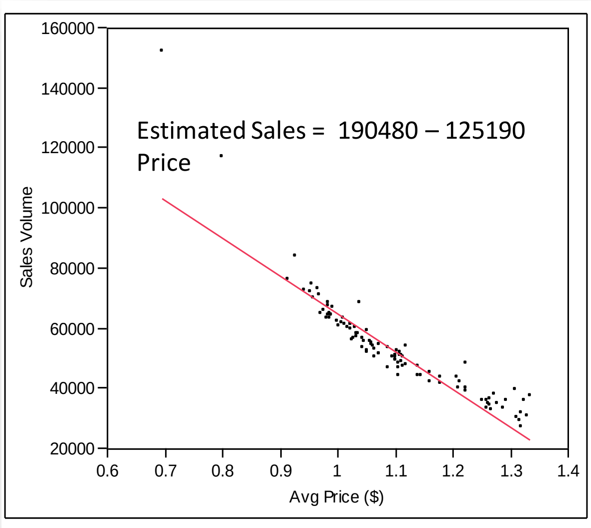
A good pattern of residuals is “no pattern”

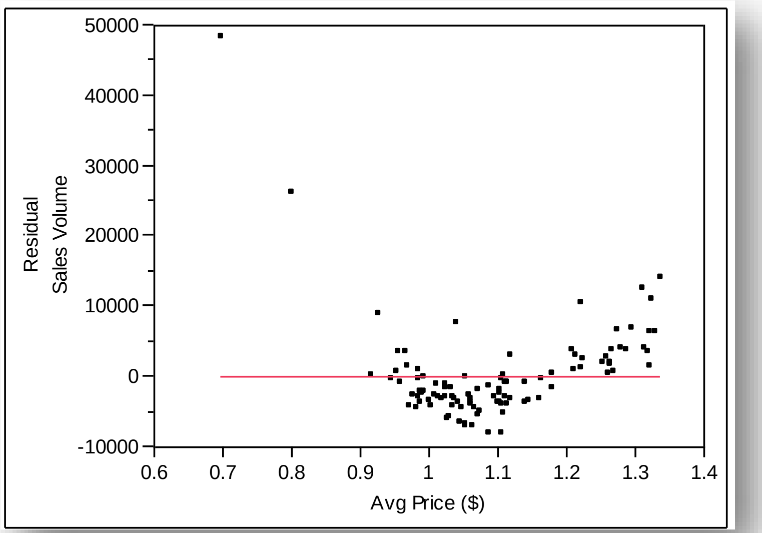
Problem 5: Linearity Assumption

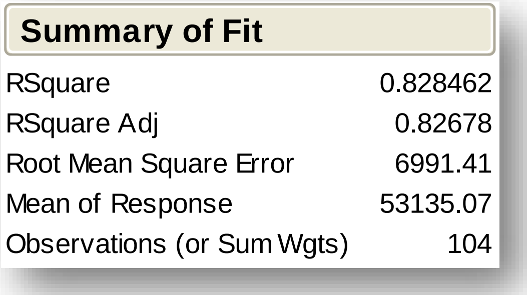
- In many applications of interest the relationship between the dependent and the independent variable might be nonlinear
- Look for transformation of the data (either X or Y or both variables) such that the relationship between transformed variables is approximately linear
- The transformed data should satisfy the OLS assumptions

Example: Pet Food Demand Curve Estimation

- The manager would like to estimate a demand curve, i.e., relationship between demand and price using data on the prices and weekly sales over a period of two years (Petfood.xlsx)

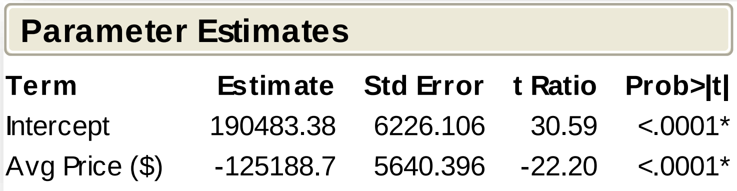

Estimated Sales = 190480 – 125190 Price





Summary of Fit

RSquare	0.828462
RSquare Adj	0.82678
Root Mean Square Error	6991.41
Mean of Response	53135.07
Observations (or Sum Wgts)	104

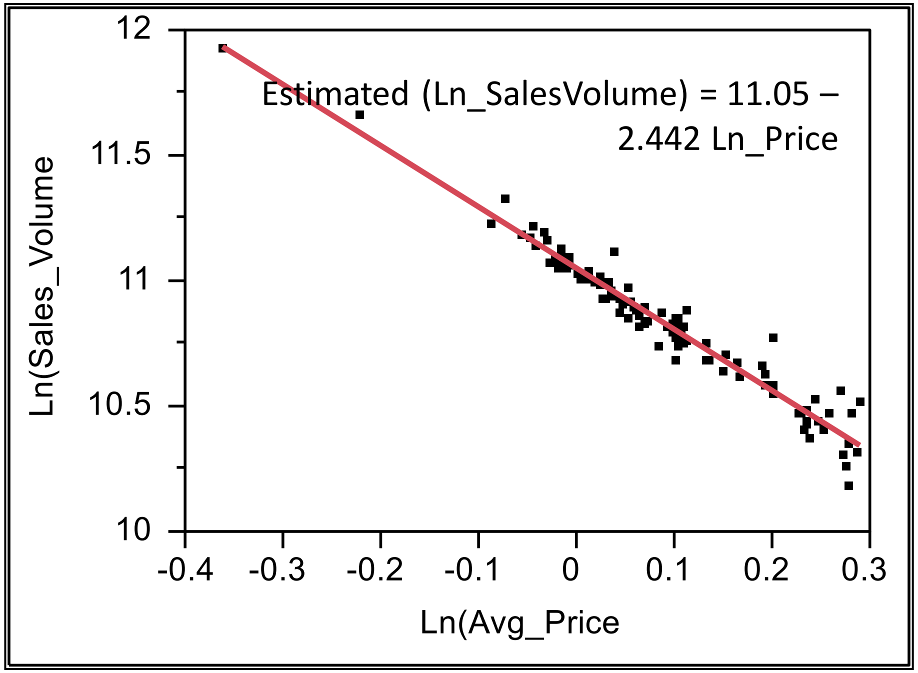


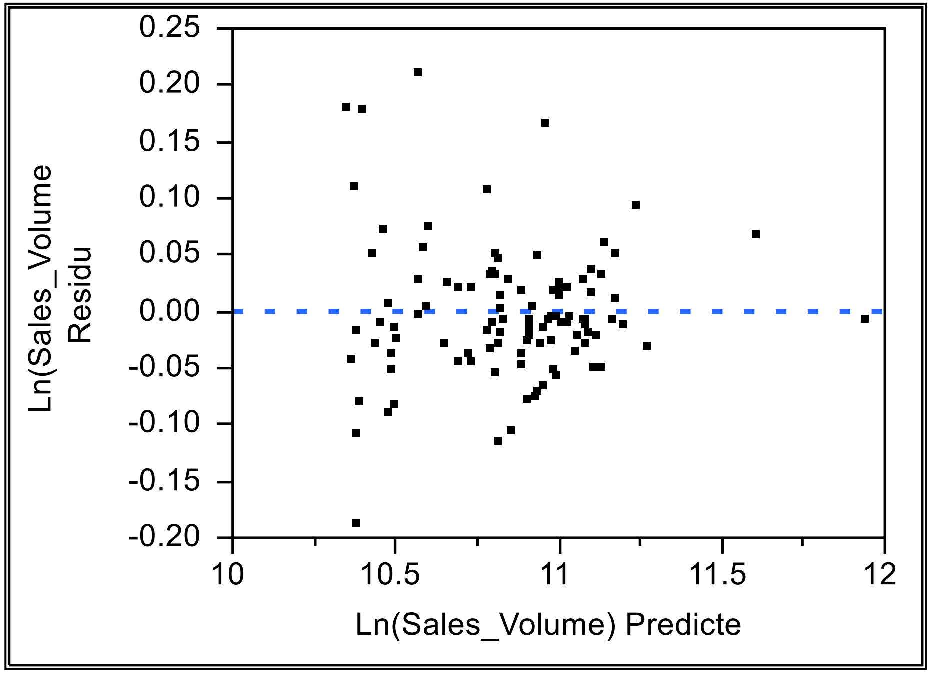
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	190483.38	6226.106	30.59	<.0001*
Avg Price ($)	-125188.7	5640.396	-22.20	<.0001*

Log-Log Transformation: Is there a pattern now?

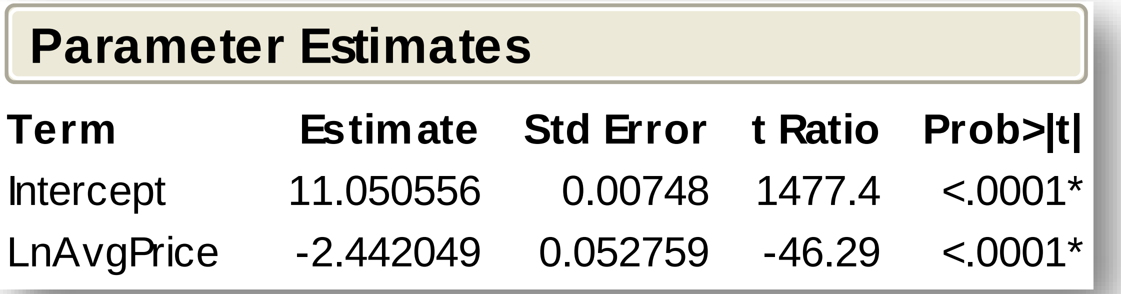
- Obtain the logarithm transform for both sales and average price
- Calculate OLS estimates for the transformed data
- Examine the residual plot for any nonlinear patterns





Interpreting the Estimates in Log Models

Model	Specification	Interpretation of β_1
Log-Log	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + \varepsilon$	Elasticity: A 1% change in X is associated with a β_1% change in Y
Log-Linear	$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon$	A one unit change in X is associated with a 100 β_1% increase in Y
Linear-Log	$Y = \beta_0 + \beta_1 \ln(X) + \varepsilon$	A 1% change in X is associated with a 0.01β_1 change in Y



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.050556	0.00748	1477.4	<.0001*
LnAvgPrice	-2.442049	0.052759	-46.29	<.0001*

Statistical Inference

- If certain assumptions hold, we can find the distributions for B_0 and B_1

$$\frac{B_0 - \beta_0}{\mathrm{SE}(B_0)} \sim T_{n-2} \qquad \frac{B_1 - \beta_1}{\mathrm{SE}(B_1)} \sim T_{n-2}$$

- We can use these distributions for making inferences about the relationship of X and Y in the population
  - Confidence intervals
  - Hypothesis tests

- If assumptions don't hold, a different model or a different method may be preferable
  - Assess the appropriateness of the assumptions after estimating the model
  - Understand the impact of “typical” violations on the estimates

Inference (I): Hypothesis Tests

- We can use the point estimate b_1 and $se(b_1)$ to test hypothesis about specific value of slope parameter

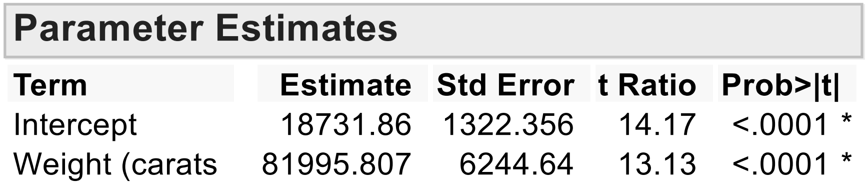
- For the slope parameter, we can test the hypotheses

$$H_0: \beta_1 = \theta \qquad H_a: \beta_1 \neq \theta$$

- This hypothesis can be tested using a t-statistic with df = n-2:

$$t_{1-\alpha,n-2} = \frac{b_1 - \theta}{SE(B_1)}$$

- By default, most software test for $\theta = 0$



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18731.86	1322.356	14.17	<.0001 *
Weight (carats	81995.807	6244.64	13.13	<.0001 *

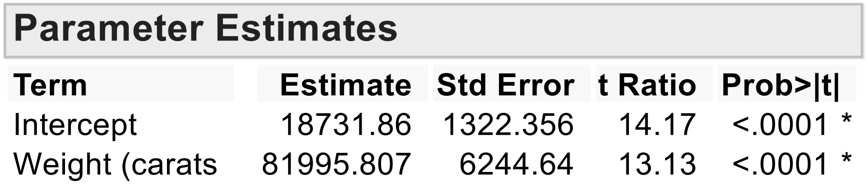
- Is there strong enough evidence to conclude that weight is associated with price per carat?

Inference (II): Confidence Intervals

- We can use the point estimate b_1 and $se(b_1)$ to construct confidence interval for the slope parameter β_1

- Given a confidence level $(1-\alpha)$, the corresponding interval is given by

$$b_1 \pm t_{1-\alpha,n-2}\, SE(B_1)$$



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18731.86	1322.356	14.17	<.0001 *
Weight (carats	81995.807	6244.64	13.13	<.0001 *

- What is the 95% confidence interval for the slope parameter?

Prediction Using the OLS Equation

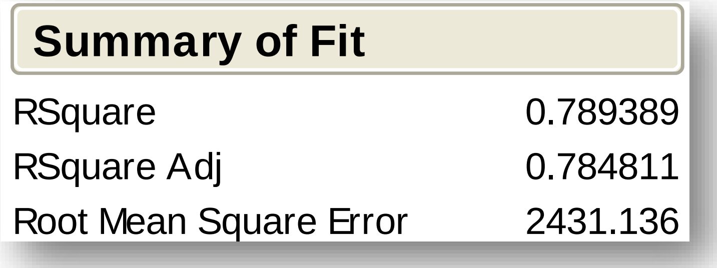
Interval Estimate:	Point Estimate	±	Margin of Error
Confidence interval: What is our estimate for the expected value of y, given x? e.g. Average price of 2.4 carat diamonds	$\hat{y}$ from the regression line		Uncertainty in the estimate of the regression line
Prediction interval: What is our estimate for an individual value of y, given x? e.g. Price of a specific 2.4 carat diamond	$\hat{y}$ from the regression line		Additional uncertainty from the idiosyncratic errors

- Confidence intervals (mean values) are narrower near the mean of the predictor and when the sample size is very large

- However, no matter how large the sample size is, there is always an error in prediction intervals (individual values)

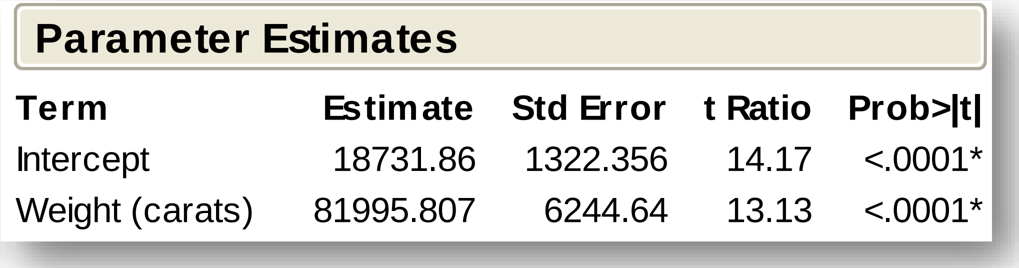
“Approximate” Prediction Interval (Individual Values)

- What is the 95% prediction interval for the price of a diamond weighing 0.2 carats?



Summary of Fit

RSquare	0.789389
RSquare Adj	0.784811
Root Mean Square Error	2431.136



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18731.86	1322.356	14.17	<.0001*
Weight (carats)	81995.807	6244.64	13.13	<.0001*

- The point prediction for price of a 0.2 carat diamond is Rs. 35131

- RMSE, best point estimate of σ_ε, is Rs. 2431

- Since 0.2 is fairly close to the mean of X, we can approximate the prediction interval by
  - $\hat{y} \pm 2(RMSE)$
  - 35131 ± 2 (2431) = [30268, 39993]

Summary

- What is a simple regression model (SRM) and what are its key assumptions?
  - A Simple Regression Model is a claim about the relationship between the dependent and the independent variable comprising a systematic linear part and an idiosyncratic error part
- How to draw statistical inference about the model parameters?
  - A number of key assumptions regarding the error term in the linear regression model are required to ensure that the OLS estimates are unbiased, efficient and reliable
- How to construct prediction intervals for the response variable?
  - We can conduct hypothesis tests and construct confidence intervals for the coefficients in the simple regression model
- What important diagnostic checks should be run before interpreting regression output?
  - Prediction intervals and confidence intervals can be constructed for the average and individual values of the response variable
  - Key diagnostic checks include inspection of the residual plots to ensure that there is no discernible pattern in the plots, which would indicate heteroskedasticity, autocorrelation, non-normality, nonlinearity
  - Transformations of the data (e.g. logarithm) enable the modeling of non-linear relationships using linear regression framework

Exercise

Cool Caffeine Deals (CCD) is a chain of budget coffee shops that is targeted at smaller towns in India. The low-price format was an instant hit with college going teenagers with tight pocket money budgets. Consequently, the company went on an expansion spree and added 53 stores within a span of a couple of years. However, management feels that the sales performance of the stores is below par.

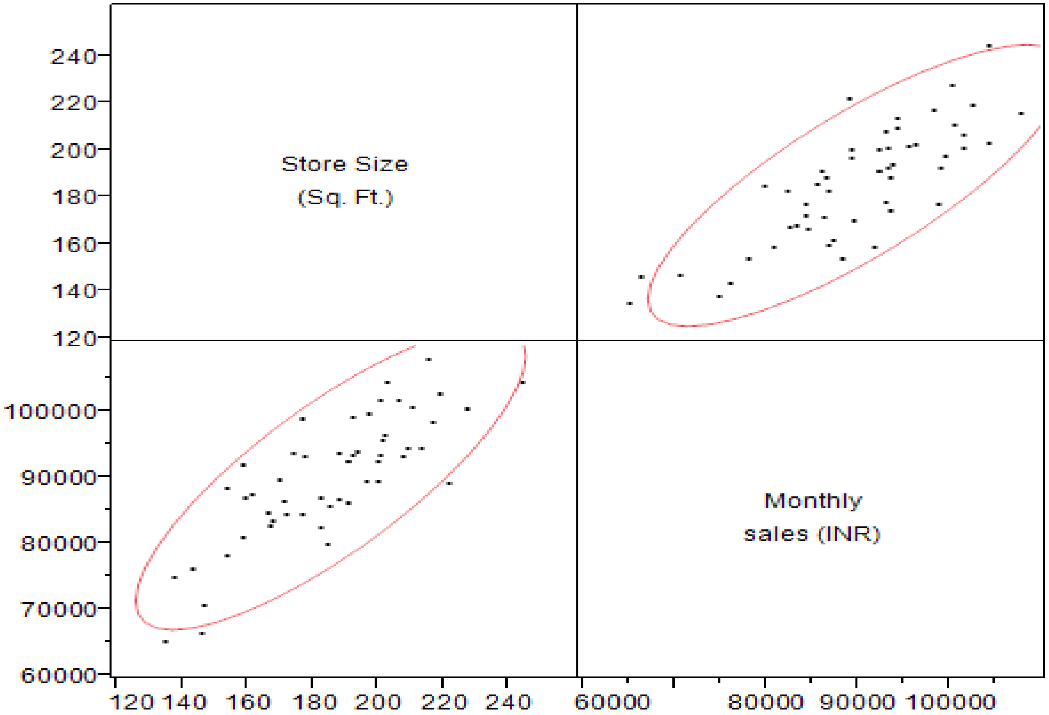
One of the potential explanations put forward by store managers is that their stores have smaller space (measured in sq. ft.) and hence can serve fewer customers, which explains the lower monthly sales (measured in Rs). They also cite the retail industry wisdom that every square foot of store generates Rs. 500 per month in sales.

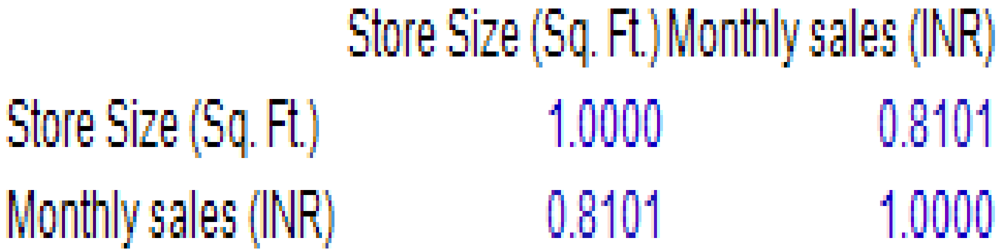
The COO of the company is interested in investigating these claims before making further changes to the network, either in terms of expansion or rationalization. She has asked you, GM Store Operations, to study the performance of stores for a month (summarized in dataset CCD.xls) and submit a detailed report based on it. Specifically, you have been asked to address the following questions. (Use α=0.05)

Questions

1. How would you characterize the relationship between the area of the store and the monthly sales in terms of (i) form, (ii) direction, (iii) strength?
2. Estimate a linear relationship between the area of the store and monthly sales and provide a managerial interpretation.
3. Based on the evidence presented in the dataset, what is your assessment of the claim that store sales are associated with store size?
4. Is the retail industry wisdom of Rs. 500 per month per sq. ft. applicable to CCD?
5. You would like to use the analysis as a forecasting tool for the sales performance of new stores. What is the 95% prediction interval of sales performance for a store with an area of 200 sq. ft.? Similarly, what is this interval for a store of 500 sq. ft.?

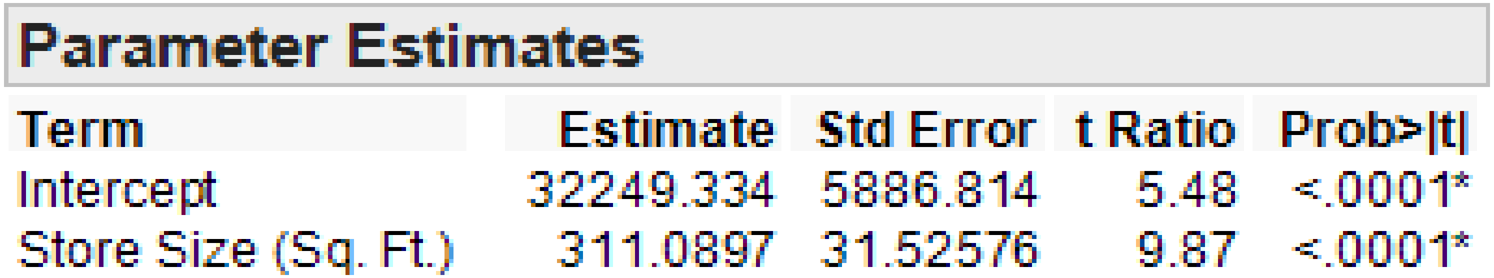
Q1





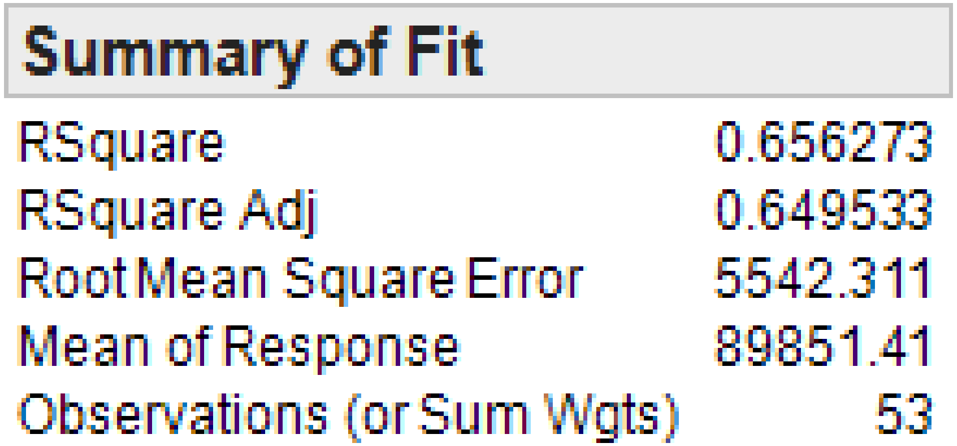
	Store Size (Sq. Ft.)	Monthly sales (INR)
Store Size (Sq. Ft.)	1.0000	0.8101
Monthly sales (INR)	0.8101	1.0000

Q2



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	32249.334	5886.814	5.48	<.0001*
Store Size (Sq. Ft.)	311.0897	31.52576	9.87	<.0001*



Summary of Fit

RSquare	0.656273
RSquare Adj	0.649533
Root Mean Square Error	5542.311
Mean of Response	89851.41
Observations (or Sum Wgts)	53

Q3

The associated p-value is less than 0.001. Because this is less than α=0.05, we can reject the null hypothesis that this coefficient is zero. Thus, we can conclude that store sales are associated with store size

Q4

We need to test the null hypothesis that the coefficient of store size is INR500/sq. ft. The corresponding t-statistic is (500-311.1)/31.52 = 5.99 and the p-value (corresponding to a two-tailed test) is less than 0.001. Thus, we can reject the nul hypothesis that the slope of the regression line is INR500/sq. ft. In other words, at the significance level of more than 99.999% we can claim that the retail industry thumb rule is not applicable to CCD.

Q5

The predicted value of average monthly sales for a store of 200 sq ft. is given by 32249.1 + 311.1*200 = 94469.1 Since the RMSE = 5542.311, the 95% prediction interval is approximately [94469.1 – 2*5542.31, 94469.1 + 2*5542.31], which is equal to [83384.48, 10553.72].

Summary

- Two Sample Comparison
  - Paired t-test
  - Independent sample t-test

- More than 2 Sample Comparison
  - ANOVA – One categorical variable and the other continuous variable
  - Chi-Square – Two categorical variables

- Simple Linear Regression (one variable case; both X and Y continuous)
  - Assumptions
  - Diagnostic Checks
  - Fitting the best line
  - Inference
  - Prediction