Session 4

Simple Linear Regression (II) & (III):

Inference, Prediction & Assumptions, Diagnostics

Learning Objectives

- What is a simple regression model (SRM) and what are its key assumptions?
- What important diagnostic checks should be run before interpreting regression output?
- How to draw statistical inference about the model parameters?
- How to construct prediction intervals for the response variable?

Simple Linear Regression Model

- Use a linear equation to model the relationship between the variables in the population

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y → Response variable - the variable we are interested in explaining
  - Also referred to as target, dependent or outcome variable
- X → Predictor variable - the variable that is useful in explaining
  - Also referred to as explanatory or independent variable
- β_0 and β_1 → model (population) parameters
- ε → error term (disturbance or noise)

Example: Diamond Prices

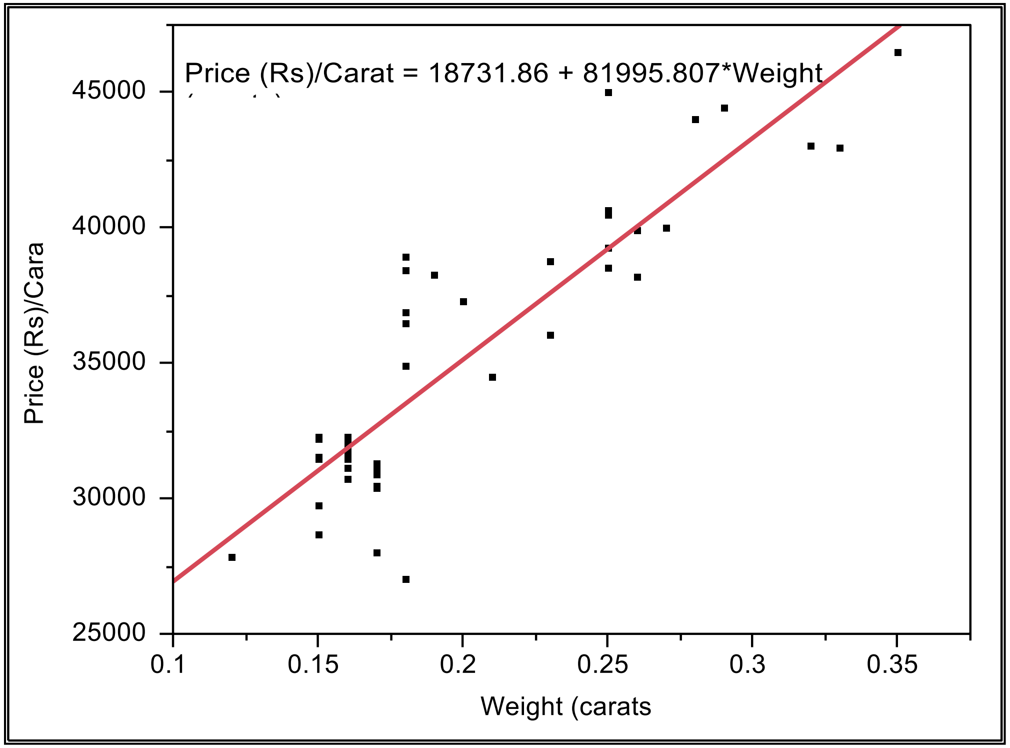
- You are interested in explaining the variation of diamond prices (INR/carat) observed in the marketplace

- After initial discussions and qualitative research, you believe that one of the factors that explains this variation is weight of the diamond (carats)

- Now you would like to establish a relationship between diamond prices and weight of diamonds

- Data on simple random sample of 48 diamonds (Diamonds.xlsx)

Example: Diamond Prices



Summary of Fit

RSquare	0.789389
RSquare Adj	0.784811
Root Mean Square Error	2431.136
Mean of Response	35472.67
Observations (or Sum Wgts	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1019030044	1.019e+9	172.4124
Error	46	271879447	5910422.8	**Prob > F**
C. Total	47	1290909492		<.0001 *

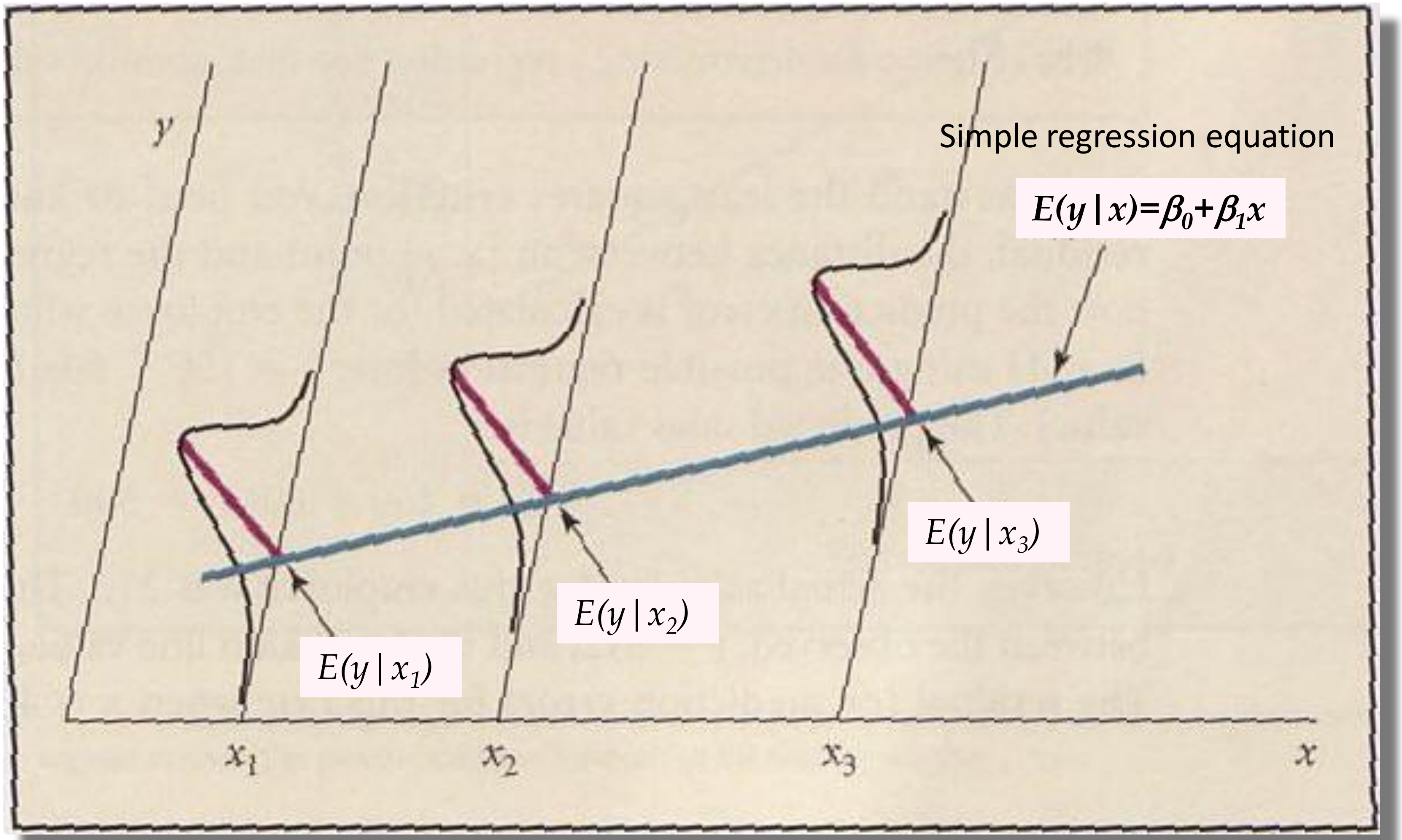
- Claim: In the population, every 1 carat increase in diamond weight is associated with INR 82K increase in price/carat on average
- Note: We have a sample of 48 diamonds and the line might (will) change with another sample

Simple Regression Model: Assumptions

- Assuming that the true relationship between Y and X is indeed given by $Y = \beta_0 + \beta_1 X + \varepsilon$

Assumption	Implication
1. Error term ε is a random variable with an expected value of zero for a given value of X $E[\varepsilon \mid X] = 0$	Since β_0 and β_1 are constants, for a given value of X, the expected value of Y is $E(Y \mid X) = \beta_0 + \beta_1 X$ This also implies that the errors are not correlated (systematically related) to the value of X i.e. $Corr(X, \varepsilon) = 0$
2. Variance of ε is a constant for all values of X $Var[\varepsilon \mid X] = \sigma_\varepsilon^2$	The variance of Y about the regression line is the same for all values of X and equals σ_ε^2 (***Homoskedasticity***)
3. Values of ε_i are independent $Corr[\varepsilon_i, \varepsilon_j] = 0$	The value of Y for a particular value of X is not related to the value of Y for another value of X This condition will generally be satisfied for a SRS
4. Error term is normally distributed $\varepsilon \mid X \sim N(0, \sigma_\varepsilon^2)$	The dependent variable Y is normally distributed for a given value of X, i.e., $y \mid \mid X \sim N(E(\beta_0 + \beta_1 X, \sigma_\varepsilon^2)$

Visual Interpretation of Assumptions



Diagnostic checks: Using OLS residuals

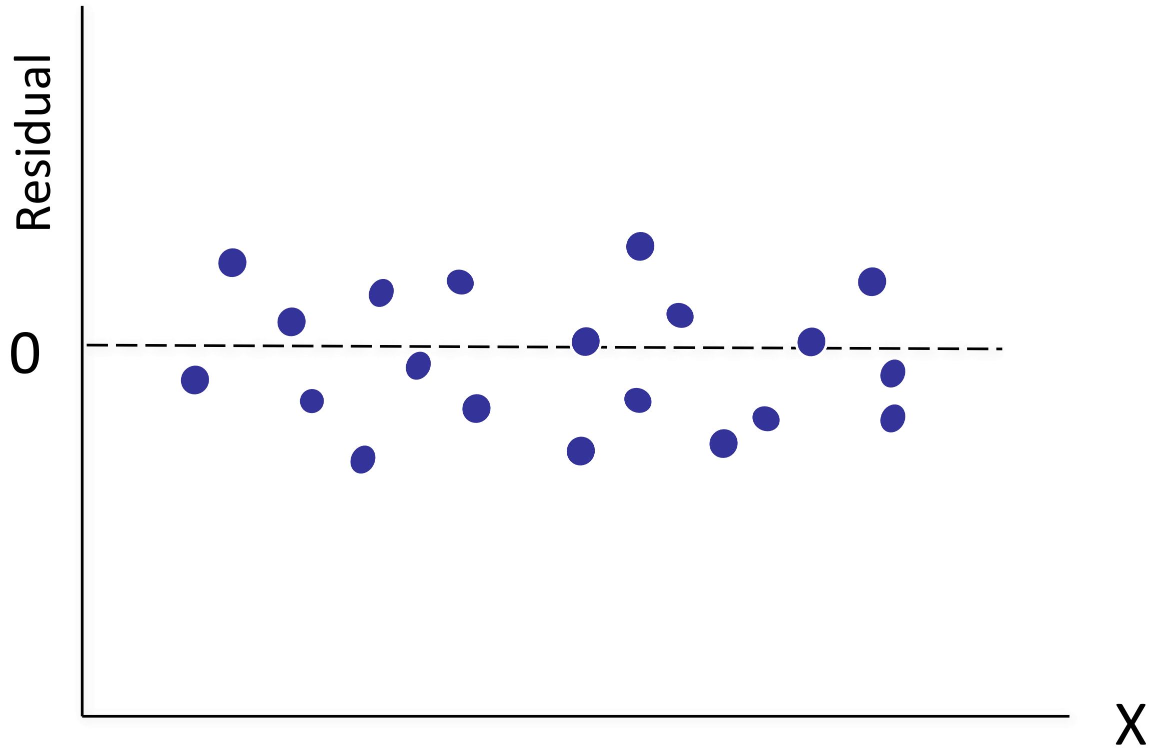
- We need to check the appropriateness of the following assumptions
  1. $E[\varepsilon|X] = 0$
  2. Homoskedasticity: $Var[\varepsilon|X]= \sigma_\varepsilon^2$
  3. $Corr[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$
  4. Normality of errors: $\varepsilon\, |X \sim N(0, \sigma_\varepsilon^2)$

- 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

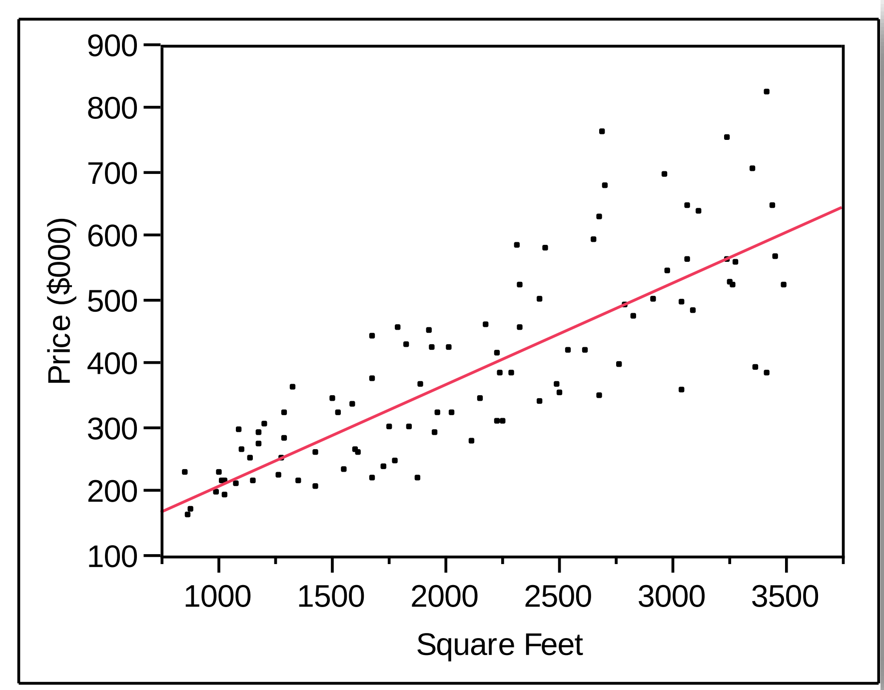
Residuals vs. Predictor Values

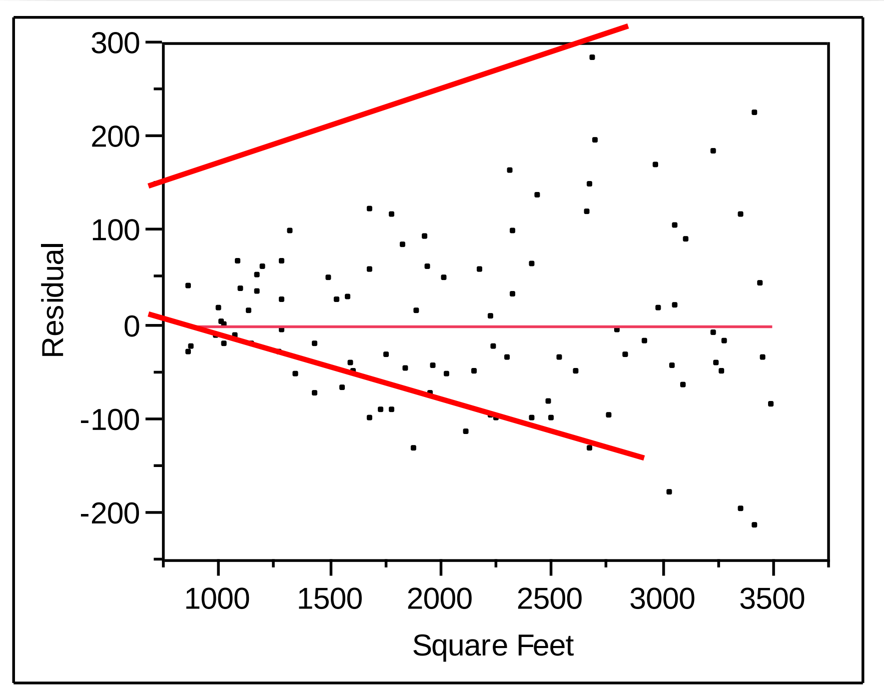


A good pattern of residuals is “no pattern”

Problem 1: Heteroskedasticity

- Variance of the residuals increases/decreases with the value of the predictor variable



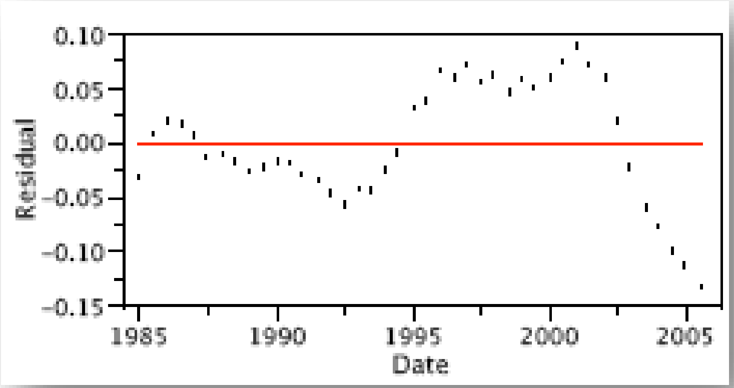


- OLS estimates are unbiased but inefficient
- Actual standard errors will be higher than the reported ones

Problem 2: Dependence and Autocorrelation

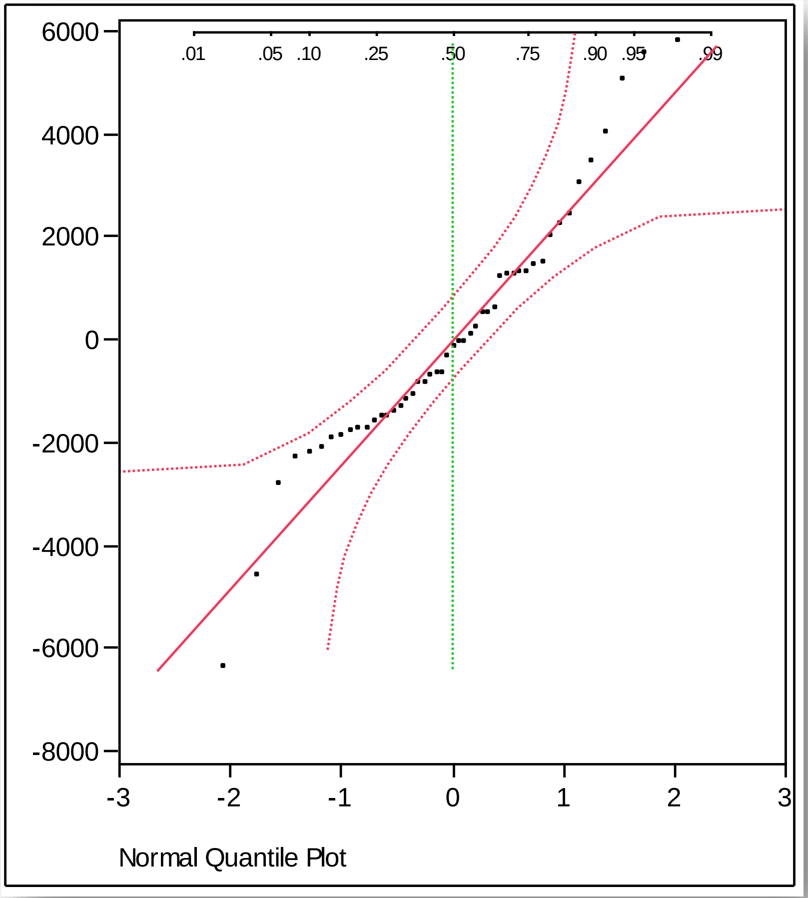
- The errors may be correlated to each other if data were collected over time
  - e.g. return of stock over time

- Often shows up as a pattern in the residuals, if plotted in chronological order



- Errors are can also correlated when the data structure is hierarchical or nested
  - e.g. Salary of MBA students across different b-schools and GMAT scores

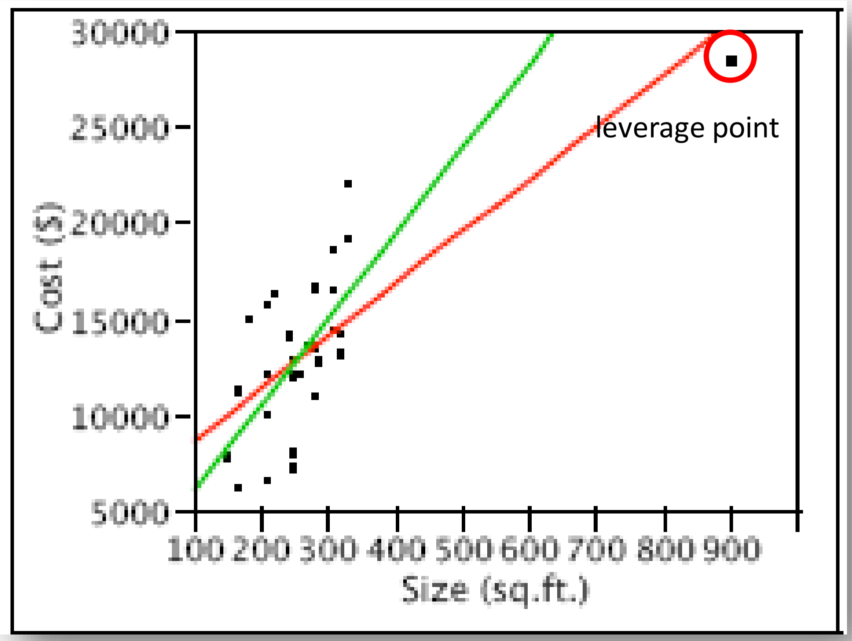
Problem 3: Departures from Normality



- Construct a quantile plot of the residuals instead of original variables
- Inferences (hypothesis tests and confidence intervals) work pretty well even when residuals are not strictly normal

Problem 4: Outliers, leverage points, influential observations

- In the case of regression, outliers (unusual observations) can occur in the y or x variables
- Unusual observations in the x variable are called leverage points
- Typically leverage points are suspect for being influential observations as OLS penalizes large errors more (due to squaring)



Green line - Best fit without the leverage point
Red line – Best fit with the leverage point