OLS Model Diagnostics

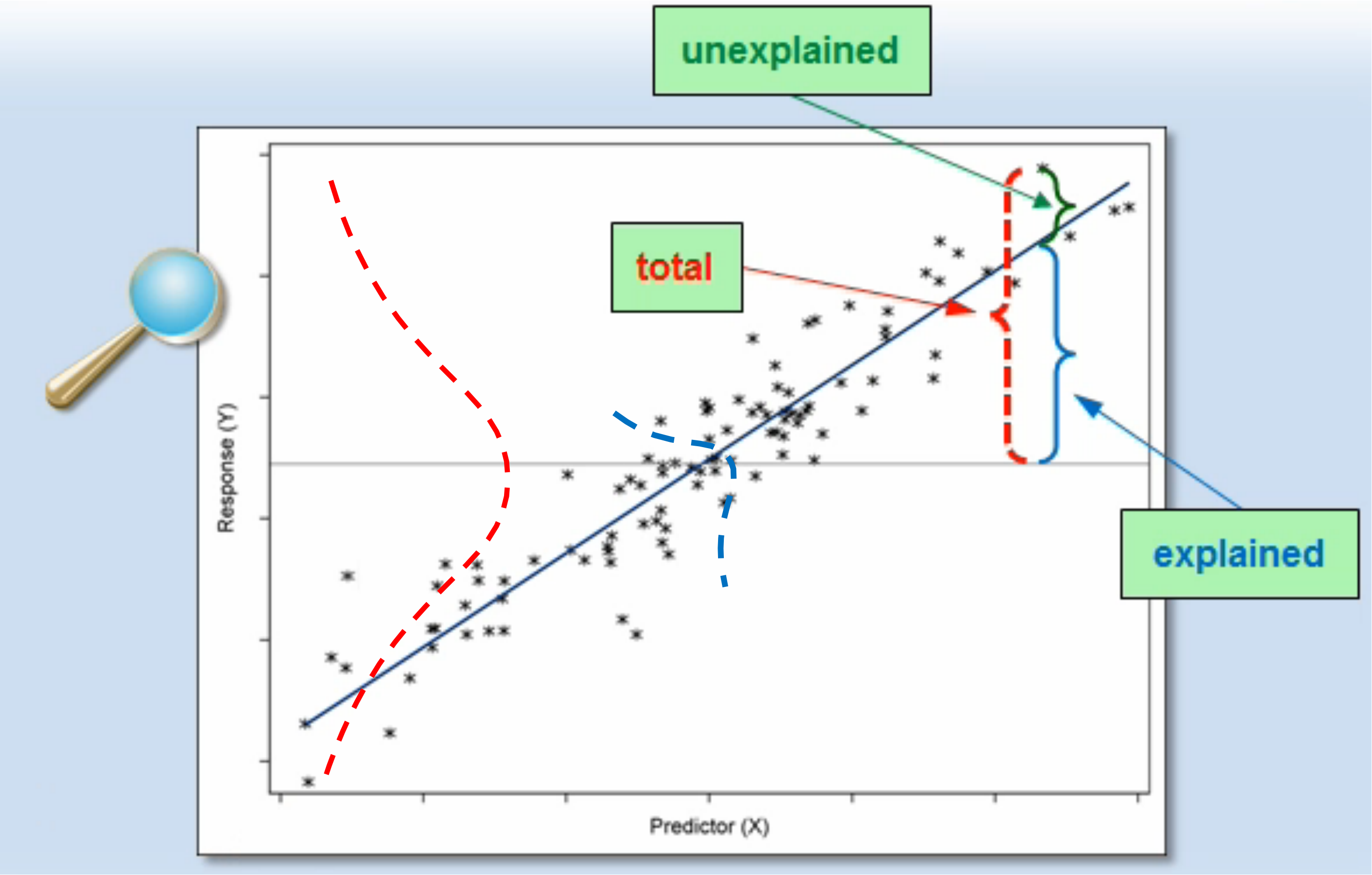
OLS: Fit Statistics

- Total sum of squares (*SST*) – i.e., relative to $\bar{y}$
- Regression sum of squares (*SSR*) – i.e., $\hat{y}$ *relative to* $\bar{y}$
- Sum of squared errors (*SSE*) – i.e., deviations from the regression line or the part of SST unexplained by $SSE = SST - SSR$
- Degrees of freedom (*df*): the number of observations in a sample minus the number of parameters that restrict the data
- n = number of observations → $df_{Total} = n - 1$
- p = number of predictors in the model → $df_{Regression} = p$
- $df_{Residual} = n - p - 1$
- Mean Squares: $MST = \frac{SST}{n-1}$ $MSR = \frac{SSR}{p}$ $\boxed{MSE = \frac{SSE}{n-p-1}}$
- Root Mean Square: $RMS = \sqrt{MSE}$

OLS Predictive Power and Fit

- **R^2** (coefficient of determination) = proportion of **variance** of Y (the predicted variable) **explained** by the regression line:
  $R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - (\frac{SSE}{SST})$
  → Closer to **1** → More **variance explained** by the regression model
- Think of it as the **squared correlation statistic** for the whole model – in fact, for simple regression models **R^2** is equal to **$\rho(y,x)^2$**
- **ANOVA** in regression compares the variance explained by the regression line to the variance of the errors and yields an **F statistic**, which translates into a **p-value** for the regression as a whole, indicating whether the full regression model has significant **predictive power** (i.e., **$p<0.05$** or smaller)
- Every time we add **one more variable** to a model it explains more variance, so the **R^2 goes up**. Thus, comparing R^2's between two models can be misleading. **Adjusted R^2** corrects for this.
- $Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$ → **R^2** is penalized by **p**

Understanding R^2 Further



$$R^2 = 1 - \frac{Unexplained\ Variability}{Total\ Variability}$$

Kogod School of Business