# Quantitative Methods

# Simple Linear Regression

## Intro and Exam Focus

- Least squares regression coefficients: intercept and slope calculation
- Assumptions of linear regression: linearity, heteroskedasticity, independence, normality
- Analysis of variance (ANOVA) table
  - *F*-test of regression
  - *T*-test of individual coefficient



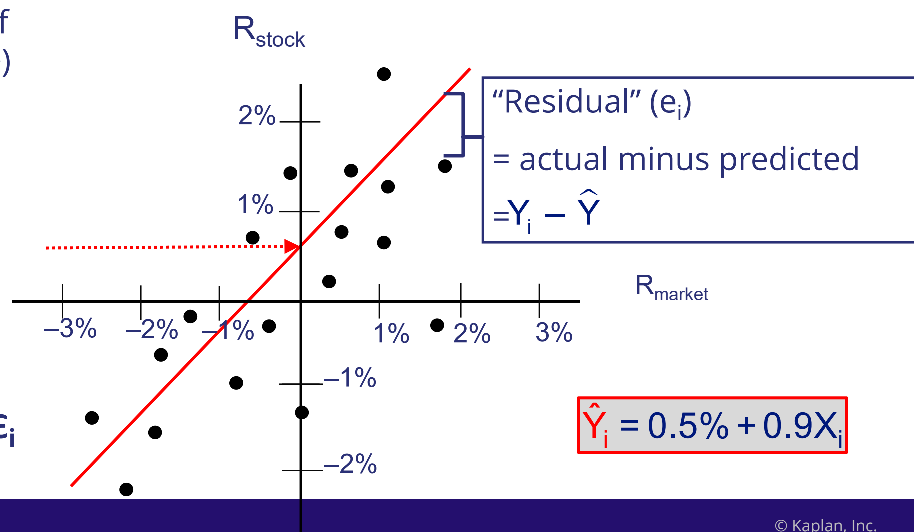
## Least Squares Regression

Minimizes the sum of squared residuals ($e^2$)

Intercept $\hat{b}_0$ = 0.5%

Slope

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$




$$\hat{Y}_i = 0.5\% + 0.9X_i$$

## Regression Coefficients

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}$$

$$\hat{b}_1 = \frac{Cov(X,Y)}{\sigma_x^2}$$

Regression line passes through $(\overline{X}, \overline{Y})$

Variation in X = $\sum_{i=1}^{n} (X_i - \overline{X})^2$

Variation in Y = $\sum_{i=1}^{n} (Y_i - \overline{Y})^2$

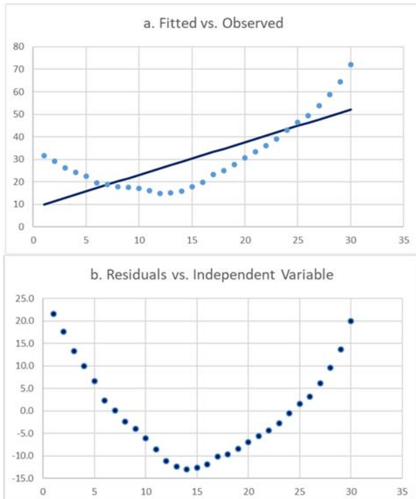


## Assumptions of Linear Regression

1. ***Linear relation*** between dependent and independent variable
2. ***Variance of the residuals*** is constant **(homoskedasticity)**
3. (Y, X) pairs should be ***independent*** to each other (uncorrelated)

   → ***Residuals*** are independently distributed (i.e., uncorrelated) with each other
4. ***Residuals*** are normally distributed

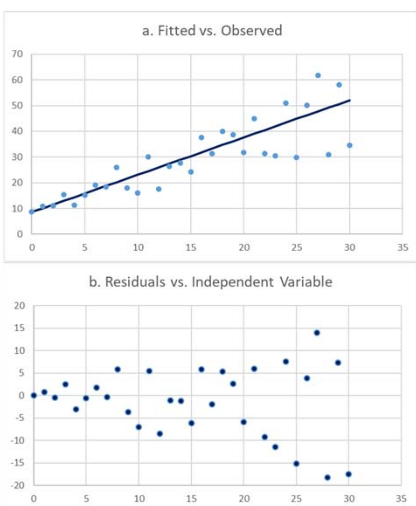
## Linearity




Residuals not independent to independent variable X

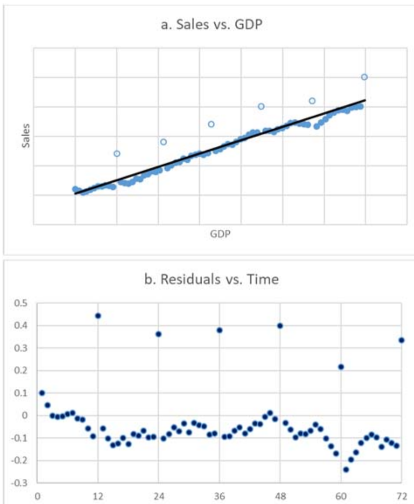


## Heteroskedasticity




Variance of residuals increasing with X → heteroskedasticity

# Independence and Normality

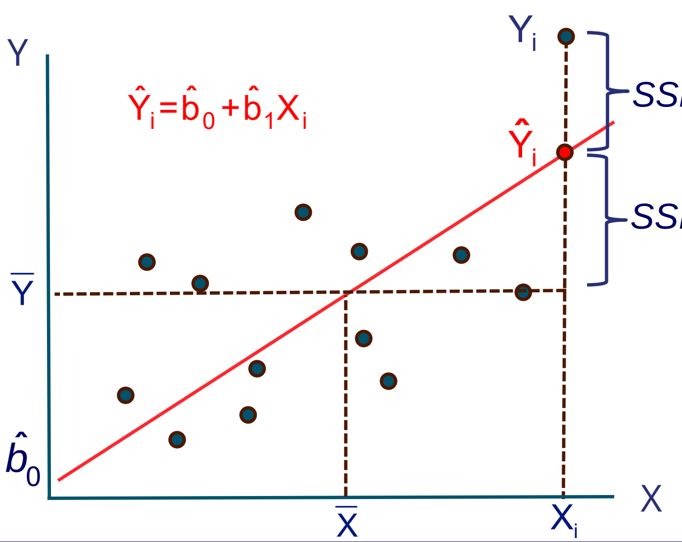



- Hollow dots are December sales
- Large residuals for December
- Suggests seasonality (**not independent, or autocorrelation**)

Residuals can be tested for **normality**. With large sample sizes, the normality assumption can be relaxed.



# Analysis of Variance (ANOVA)




$$\left.\begin{aligned} SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ SSR &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \end{aligned}\right\} SST$$

SSE + SSR = SST

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

-6

## ANOVA Table for Regression With $n = 5$

| Source | df | SumSquares | Mean Square | *F*-stat | $R^2$ |
|---|---|---|---|---|---|
| Regression | | 88.0 | | | |
| Residual | | 7.2 | | | |
| Total | | | | | |

Degrees of freedom (df)? Regression = Residual = Total =

Total sum of squares? =

Mean square regression (MSR)? = SSR / df = =

Mean square error (MSE)? = SSE / df = =

-4



## ANOVA Table for Regression With $n = 5$

| Source | df | SumSquares | Mean Square | *F*-stat | $R^2$ |
|---|---|---|---|---|---|
| Regressio n | 1 | 88.0 | 88.0 | | |
| Residual | 3 | 7.2 | 2.4 | | |
| Total | 4 | 95.2 | | | |

$F\text{-stat?} = \frac{MSR}{MSE} = \quad =$

Critical *F*-value from statistical tables (95th percentile, df numerator = 1, df denominator = 3):

→ null hypothesis that all coefficients are zero

-3

## ANOVA Table for Regression With $n = 5$

| Source | df | SumSquares | Mean Square | F-stat | $R^2$ |
|---|---|---|---|---|---|
| Regression | 1 | 88.0 | 88.0 | 36.67 | |
| Residual | 3 | 7.2 | 2.4 | | |
| Total | 4 | 95.2 | | | |

Coefficient of determination ($R^2$)? $\frac{SSR}{SST} = \quad =$

Standard error of estimate, SEE ($s_e$)? $\sqrt{MSE} = \quad =$

-2



## Regression Coefficient t-Test: Example

Estimated slope ($b_1$) = 1.50 with $s_{\hat{b}_1} = 0.46$. Assuming n = 32, determine if the slope is equal to zero at the 5% significance level. Test the hypothesis that $b_1 = 0$.

$$t_{b_1} = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \quad =$$

Critical two-tailed *t*-stat (30 df and 5% significance) is 2.042

: $b_1 = 0$.

-3

## Predicted Values for "Y" Variable: Example

Given $\hat{b}_0$ = 1.8%, $\hat{b}_1$ = 0.76, and a value of X of 8%, what is the predicted value of Y according to this regression model?

$$\hat{Y} = \qquad\qquad\qquad =$$

-1



## Prediction Intervals

Given a predicted value of Y of 7.88% and standard error of the forecast ($s_f$) of 1.22% for a regression with 27 observations, construct a 95% prediction interval (critical value with 25 df is 2.06).

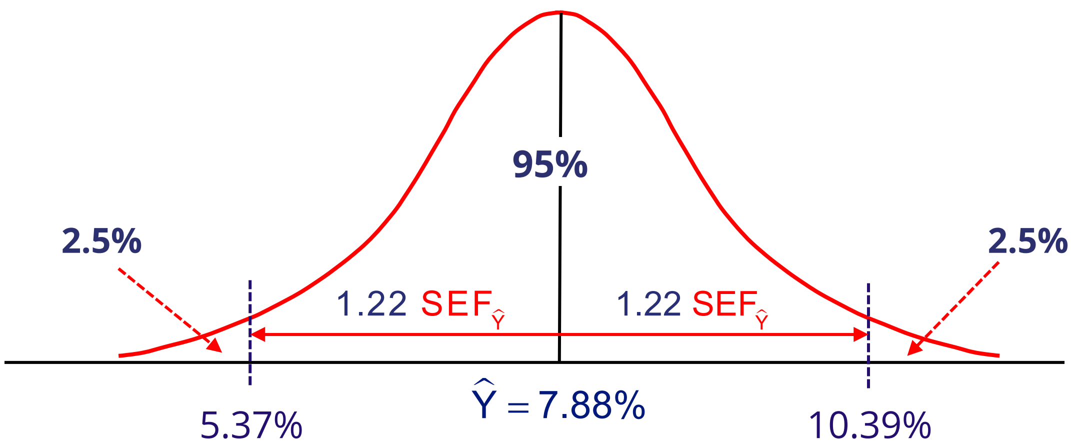
$$7.88 \pm (\qquad\qquad) =$$

$$s_f^2 = SEE^2\left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2}\right]$$

$s_f$ can be approximated with SEE for large samples.

-1

# Prediction Intervals






# Functional Forms

When the relationship between X and Y is not linear, fitting a linear model would result in biased predictions.

| Model | Dependent Variable | Independent Variable | Slope Interpretation |
|---|---|---|---|
| Log-lin | Ln (Y) | X | Relative change in Y, absolute change in X<br>Forecast Y is $e^{\ln Y}$ |
| Lin-log | Y | Ln(X) | Absolute change in Y, relative change X |
| Log-log | Ln(Y) | Ln(X) | Relative change in Y, relative change X<br>Forecast Y is $e^{\ln Y}$ |

Check for increased $R^2$ and *F*-stat, and lower SEE, with transformed variable(s).

## Solutions

## ANOVA Table for Regression With $n = 5$

| Source | df | SumSquares | Mean Square | $F$-stat | $R^2$ |
|---|---|---|---|---|---|
| Regression | 1 | 88.0 | 88 | | |
| Residual | 3 | 7.2 | 2.4 | | |
| Total | 4 | 95.2 | | | |

Degrees of freedom (df)? Regression = 1, Residual = n – 2, Total = n – 1

Total sum of squares? = 88.0 + 7.2 = 95.2

Mean square regression (MSR)? = SSR / df = 88.0 / 1 = 88.0

Mean square error (MSE)? = SSE / df = 7.2 / 3 = 2.4

-4

## ANOVA Table for Regression With $n = 5$

| Source | df | SumSquares | Mean Square | *F*-stat | $R^2$ |
|---|---|---|---|---|---|
| Regression | 1 | 88.0 | 88.0 | 36.67 | |
| Residual | 3 | 7.2 | 2.4 | | |
| Total | 4 | 95.2 | | | |

F-stat? $= \frac{\text{MSR}}{\text{MSE}} = \frac{88.0}{2.4} = 36.67$

Critical *F*-value from statistical tables (95th percentile, df numerator = 1, df denominator = 3): 10.1

→ **Reject** null hypothesis that all coefficients are zero

-3



## ANOVA Table for Regression With $n = 5$

| Source | df | SumSquares | Mean Square | *F*-stat | $R^2$ |
|---|---|---|---|---|---|
| Regression | 1 | 88.0 | 88.0 | 36.67 | 0.924 |
| Residual | 3 | 7.2 | 2.4 | | |
| Total | 4 | 95.2 | | | |

Coefficient of determination ($R^2$)? $\frac{\text{SSR}}{\text{SST}} = \frac{88.0}{95.2} = 0.924$

Standard error of estimate, SEE ($s_e$)? $\sqrt{\text{MSE}} = \sqrt{2.4} = 1.549$

-2

## Regression Coefficient *t*-Test: Example

Estimated slope ($b_1$) = 1.50 with $s_{\hat{b}} = 0.46$. Assuming n = 32, determine if the slope is equal to zero at the 5% significance level. Test the hypothesis that $b_1$ = 0.

$$t_{b_1} = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \frac{1.50 - 0}{0.46} = 3.26$$

Critical two-tailed *t*-stat (30 df and 5% significance) is 2.042

Reject $H_0$: $b_1$ = 0.

-3



## Predicted Values for "Y" Variable: Example

Given $\hat{b}_0$ = 1.8%, $\hat{b}_1$ = 0.76, and a value of X of 8%, what is the predicted value of Y according to this regression model?

$$\hat{Y} = 1.8\% + 0.76(8\%) = 7.88\%$$

-1

## Prediction Intervals

Given a predicted value of Y of 7.88% and standard error of the forecast (**$s_f$**) of 1.22% for a regression with 27 observations, construct a 95% prediction interval (critical value with 25 df is 2.06).

$$7.88 \pm (2.06 \times 1.22) = 5.37\% - 10.39\%$$

$$s_f^2 = SEE^2\left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2}\right]$$

**$s_f$** can be approximated with SEE for large samples.

-1