SLR MODEL

Population $y = \beta_0 + \beta_1 x + \varepsilon$ or $E(y) = \beta_0 + \beta_1 x$
Estimated model using sample values: $\widehat{E}(y) = \widehat{\beta_0} + \widehat{\beta_1} x$

***y* is the *outcome (response)* variable**
***x* is the *explanatory* variable**

5 Assumptions of the SLR model:

1) **L – *LINEAR* - The relationship between x and y is linear**
   - Validated by checking scatter plot of X vs Y
2) **I – *INDEPENDENCE* - *x*-values are INDEPENDENT of each other so that *y* and ε values are INDEPENDENTLY DISTRIBUTED with means $E(y) = \beta_0 + \beta_1 x$ and $E(\varepsilon) = 0$**

   x-values are OBSERVED WITHOUT ERROR, for both y and ε
   - Plot y vs **residuals (e_i)** to check violations of INDEPENDENCE
3) **N- *NORMAL DISTRIBUTION* - y-values are NORMALLY DISTRIBUTED with N($E(y), \sigma^2$), and for every value of x, ε-values are ≈ N($0, \sigma^2$)**
   - Plot histogram of residual **(e_i)** frequencies to check violations of NORMAL distribution of errors (Skewed indicates problem)
4) **E –*EQUAL VARIANCE* - y-values have CONSTANT VARIANCE, σ^2, for any given value of x; ε values have constant variance, σ^2, for any given value of x.**
   - Plot x vs **residuals (e_i)** to check violations of CONSTANT VARIANCE

z-test – for POPULATION

t-test
H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$
Two-tailed test for α/2
Df = n-2
Reject H_0 if T-test value < - t(n-2;α/2) or T-test value > t(n-2;α/2)
Conclusion: At a 5% level of significance, there is evidence of a linear relationship between x and y.

One-tailed test: **H_0: $E(y) \geq \mu_0$; H_a: $E(y) < \mu_0$; RR:** t < - t(n-2;α/2)
H_0: $E(y) \leq \mu_0$; H_a: $E(y) > \mu_0$; RR: t > - t(n-2;α/2)

CONFIDENCE INTERVALS (β_1)

Make sure that the interval does not BOUND ZERO – otherwise there is a risk the slope could be 0!
If CI does not bound 0 then = evidence of linear relationship

Conclusion: I am 95% confident that in repeated sampling, the true population slope, β_1, would lie in the interval (-a,a)

Notes:

1) Prediction line only valid within the range of X values – do not EXTRAPOLATE beyond the boundaries unless sample size is in the 100's or 1000's
2) Even if we do not reject $\mathbf{H_0}$ (i.e. $\beta 1 = 0$) it does not necessarily mean that x and y are UNRELATED. They might be related in another way (i.e. quadratic or logarithmic). It only means X and y are not LINEARLY related.
3) If we reject $\mathbf{H_0}$, we DO NOT CONCLUDE that the true relationship between X and Y is LINEAR. It only means a linear relationship is reasonable since y may depend on other variables as well
4) If we accept that Ha is true, we do NOT CONCLUDE that here is a CAUSAL relationship between x and y. (there could be a 3rd factor that causes both x and y to increase)

Reject H_0: conclude evidence exists that $\mathbf{H_a}$ could be true
Do NOT Reject H_0: conclude that there is not enough evidence for $\mathbf{H_a}$

ANOVA

Decomposes variation in data into how much comes from ERROR and how much comes from REGRESSION (the model)
PERFECT FIT: SSE=0 & TSS=SSR → there is no error and variation is due entirely to the model
NO Linear Relationship: SSR=0 & TSS=SSE → Variation is due entirely to error

Dividing sum of squares by the degrees of freedom (df) results in MEAN SUM OF SQUARES

$df_{TSS} = n - 1$
df_{SSR} = # of parameters associated with x ∴ 1
df_{SSE} = n – (total # of parameters in the model, including β_0) = n-2

MSE is an unbiased estimator of σ^2

F-TEST STATISTIC

If $\beta_1 = 0$, then F is close to 1.

F-test table lookup:

$F_{\alpha(v1,v2)}$: v1= numerator's df ; v2 = denominators df

	V2	**1**	**2**	**3**
V1				
1				
2				
3				

Reject H_0 for large values of F. if **F-test > $F_{\alpha(v1,v2)}$**
Note that t-test and F-test are EQUIVALENT in that $t^2 = F$, but ONLY IF TESTING 1 PARAMETER, i.e. ($\beta_1=0$ vs $\beta_1 \neq 0$), AND TWO-TAILED TEST. (i.e. not true if doing a one-tailed T-test)

PREDICTION INTERVALS

$$y \in \left(\hat{y} \pm t_{\left(n-2;\frac{\alpha}{2}\right)} * s * \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{s_{xx}}} \right)$$

- **PI should ALWAYS be bigger than CI**
- Minimum value for both should be at $x_p = \bar{x}$
- Plotting endpoints of CI or PI for different values of xp and connecting the points gives a general (1-α)% CONFIDENCE BAND (or PREDICTION BAND) for any value of X in the range of Xp's

CORRELATIONS

-1 <= r <= 1
r : COEFFICIENT OF CORRELATION, or Pearson's Product moment coefficient of correlation – measures the STRENGTH and DIRECTION of the relationship between X and Y for a SAMPLE data set

ρ = (rho) POPULATION CORRELATION COEEFICIENT
H0: ρ = 0 ←→

Conclusion: x and y are (positively/negatively) linearly related with a strength of the relationship at (r)

r^2 : COEFFICIENT OF DETERMINATION -- explains how much of the total variation in the data is explained by the regression and how much is explained by error

- If r^2 = 1 then perfect relationship – 100% of variance is due to model
- You can NEVER TAKE $\sqrt{r^2}$ to obtain r

Conclusion: approximately (100* r^2)% of the total variation in the data is explained by the regression line, which means approximately 1-(100* r^2)% is due to error.

RESIDUAL ANALYSIS

TESTS FOR LACK OF FIT of data to MODEL

2 ways to do residual analysis:

1) Graphical
2) Numerical

GRAPHICAL RESIDUAL ANALYSIS

a) **Scatter Plot of X vs Y** – shows OUTLIERS, validates **ASSUMPTION OF LINEARITY**
b) Plot **PREDICTED VALUES vs RESIDUALS** ($\hat{y}_i \; vs \; e_i$)
   - No pattern (points randomly scattered above and below 0) = NO PROBLEM
   - Tests **ASSUMPTION OF INDEPENDENCE** and **ASSUMPTION OF LINEARITY**
   - If curved you may need to add a quadratic to the model
c) Plot **X vs RESIDUALS** ($x_i \; vs \; e_i$)
   - No pattern = NO VIOLATION
   - Checks for VIOLATIONS of the **ASSUMPTION OF CONSTANT VARIANCE**
d) Plot **HISTOGRAM of frequency of errors**
   - Should be BELL-SHAPED (normal curve)
   - Checks for VIOLATIONS of the **ASSUMPTION OF NORMALCY OF ERRORS**
   - Does not have to be PERFECTLY symmetric
   - If SKEWED, check to see if normal curve can be obtained by a) reducing bin width, and/or REMOVING OUTLIERS

If we have NON-LINEARITY AND NOT INDEPENDENT, then square all x values and retry the model
If we have NON-NORMALCY AND NON-CONSTANT VARIANCE then use transformation of RESPONSE VARIABLE I.E. $\sqrt{y} \;\; \ln(y) \;\; or \frac{1}{y}$

In SAS:
plot R.*P.;
plot R.*;

NUMERICAL RESIDUAL ANALYSIS

Can be done if we have any data set where there is MORE THAN 1 observation (y-value) for a given X value

Divides the error into two parts:

1) Pure experimental error
2) Error due to LACK OF FIT

i.e. SSE = SSPE + SSLF ; Where SSPE = Pure Experimental error, and SSLF = Sum of Squares for Lack of Fit

Steps:

1) Group Y Data by X values
2) K = number of "groups" of x-values (or number of UNIQUE VALUES OF X)
   a. i.e. in a data set where x=(20, 40, 40,20,10) and y=(86,78,84,33,64), there are 3 groups of unique X values: (10, 20, 40) so K=3
3) Calculate $\bar{y}$ for every value of k:

   $x_1 = 10,\ y_{11}=64 \quad ; \overline{y_1} = 64$

   $x_2 = 20,\ y_{21}=86,\ y_{22}=33 \quad ; \overline{y_2} = (86+33)/2$

   $x_3 = 40,\ y_{31}=78,\ y_{32}=84 \quad ; \overline{y_3} = (78+84)/2$

4) Determine df for SSE, SSPE, and SSLF:
   $df_{SSE} = (n-2)$; $df_{SSPE} = \sum_{i=1}^{K}(n_i - 1)$; $df_{SSLF} = n - 2 - \sum_{i=1}^{K}(n_i - 1)$
   Where n = total number of samples
   And K = number of unique X groups

$$SSPE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y_i}\right)^2$$

SSLF = SSE-SSPE

$$MSPE = \frac{SSPE}{\sum_{i=1}^{K}(n_i - 1)} \rightarrow SSPE/df_{SSPE}$$

$$MSLF = \frac{SSLF}{n-2-\sum_{i=1}^{K}(n_i - 1)} \rightarrow SSLF/df_{SSLF}$$

H_0: Linear model **IS APPROPRIATE**
H_a: Linear model is **NOT APPROPRIATE**

F-test: $F = \frac{MSLF}{MSPE}$

Reject **H_0 if $F > F_{\alpha(df_{SSLF}, df_{SSPE})}$**

Conclusion: If F-test value is > F_α, we DO NOT REJECT **H_0** and conclude that at a 5% level of significance, there is NOT ENOUGH EVIDENCE to say that our model is not appropriate. (i.e. the model is fine)

Steps In SLR analysis

1) Use a SCATTER PLOT as a preliminary check. If ther are no obvious non-linearities, go on to step 2
2) State a SIMPLE LINEAR REGRESSION MODEL for the POPULATION: $\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$
   Along with its ASSUMPTIONS
3) Use the sample data to find the LEAST SQUARES FITTED LINE $\boldsymbol{\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x}$
4) Perform RESIDUAL ANALYISIS to check for any violations of the assumptions of the model. If satisfied, go to step 5, if not try remedies and go to step 2 and start over
5) Test whether a Linear Relationship between X and Y exists : $\mathbf{H_0: \beta_1 = 0}$; $\mathbf{H_a: \beta_1 \neq 0}$
   And also calculate r^2 → explains how much variation is due to model and how much is due to error
6) If $\mathbf{H_0}$ is rejected (i.e. linear relationship exists), use $\hat{y}$ to estimate $\boldsymbol{E(y)}$, or prediction of future value.