Multiple Regression Analysis (MLR)→ extension of the SLR for investigating how response (y) is affected by several independent variables: $x_1 \ldots x_k$

Model: $y=\beta_0+\beta_1 x_1 \ldots \beta_k x_k+\varepsilon$ where ε = error

The estimated regression equation can be found by minimizing the sum of squared equation residuals <u>least squared method.</u> This equation is: $\hat{y}=b_0+b_1 x_1+b_2 x_2 \ldots b_k x_k$ where $b_1, b_2 \ldots b_k$ are estimates

R-Squared→ describes relative improvement from using prediction equation instead of using sample mean

$SST=\sum(y-ybar)^2$→ sum of squares total, measures variation about the mean in responses

$SSE=\sum(y-\hat{y})^2$→ sum of squares equal, measures spread about regression equation

$SSR=SST-SSE \vee \sum(\hat{y}-ybar)^2$→ measures variation explained by the regression model.

$\mathbf{R^2=\frac{SSR}{SST}}$→ tells us what % of variation in responses is explained by regression model

Properties

1. R^2 is between 0 and 1
2. $R^2=1$ when all residuals are 0
3. $R^2=0$ when each $\hat{y}=\bar{y}$
4. R^2 gets larger or <u>at least</u> stays the same whenever an independent variable is added to the multiple regression model
5. R^2 does not depend on units of measurement

Hypothesis Test

1. $H_0: \beta_i=0 \text{ vs. } H_a: \beta_i \neq 0$
2. **Test Statistics**

$$t=\frac{\beta_i-0}{se(\beta_i)}$$

3. **P-value:** Two-tail probability from t-distribution of values larger than observed t-test stats. The t-distribution has df= n- # of parameters
4. **Conclusion:** compare p-value to significance level if decision needed

Confidence Interval for β_i

Estimate ± Margin of Error

$\beta_i \pm t^{*}_{n-(k+1)} SE(b_i)$

More Test Statistics

$$z=\frac{\bar{p}-p_0}{\sqrt{p_0 ¿ ¿ ¿ ¿}}$$

An estimate for $\sigma \rightarrow se=\sqrt{\frac{\sum \text{ of squared residuals}}{n-(k+1)}}$ or $se=\sqrt{MSE}$

(# of β's)

Hypothesis Test

H_0: $\beta_1=\beta_2=\ldots \beta_k=0$

H_a: At least one $\beta \neq 0$

ANOVA Table

	Degrees of Freedom	Sum of Squares/SSE	Mean Sum of Squares	F-Stat	P-Value
Model (Regression)	**# of β's – 1 or k**	**SSR**	**MSR=**$\frac{SSR}{K}$		
Error (Residuals)	**n - # of β's or n – (k+1)**	**SSE**	**MSE=**$\frac{SSE}{n-(k+1)}$		
Total	**n-1**	**SST**			

F-stat= a test comparing statistical models that have been fitted to a data set

k= number of predictor variables

Test-Statistic

$F=\frac{MSR}{MSE}\ F_{k:n-(k+1)}$

F-Distribution

$F_{a:b}$

P-Value

The area is <u>always</u> going to the right and one-sided

Conclusion

Compare p-value to α

If p-value < α, reject H_0

If p-value > α, fail to reject H_a

R^2 Adjusted: $1-\frac{MSE}{MST}\ where\ MST=\frac{SST}{n-1}$

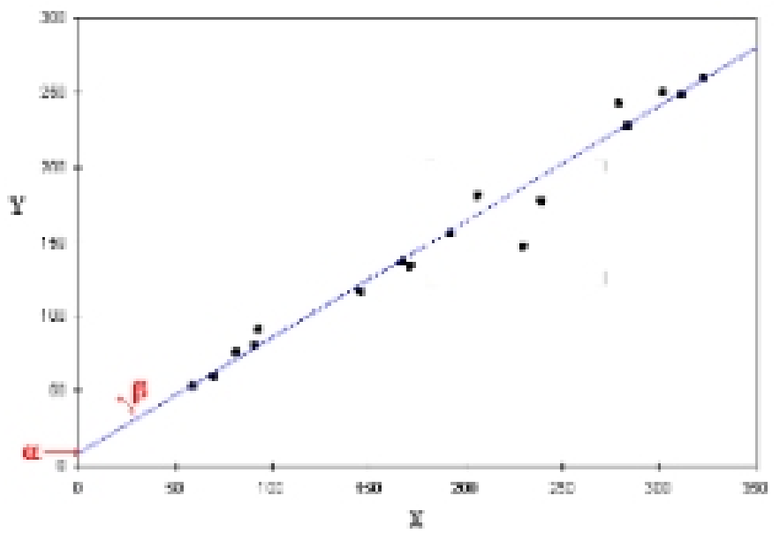
Conditions for Multiple Regression Model

LINE

1. Linearity→ look at all scatterplots of y vs. x_i to see if they are linear
2. Independence→ assume observations come from an SRS
3. Normality→ make histogram of residuals; make a ggplot of residuals
4. Equal Spread→ residual plots are equally far away from zero

Simple Linear Regression (SLR)

- Inference for SLR



The data in a scatterplot are a random sample from a population that may exhibit a linear relationship between x and y.
DIFFERENT SAMPLE= DIFFERENT PLOT

$y=\beta+p_1 x+\varepsilon\ where\ \varepsilon\ \ (0,\sigma)\vee y\ \ \mu(\beta_0+\beta_1 x,\sigma)$

In the regression $\mu_y=\beta_0+\beta_1 x$ Population Model Above **population, linear equation is**

<u>Sample Data</u> then fits the model

Date= fit + residual

$y_i=(\beta_0+\beta_1 x_i)+\varepsilon_i$ *where* ε_i *is independent* $\wedge$ *normally distributed* $N(0,\sigma)$

Linear Regression assumes equal variance of y (σ is same for all values of x)

Conditions for SLR

1. Linearity
2. Independence
3. Normality
4. Equal Spread

Estimating the Parameters

$\mu_y=\beta_0+\beta_1 x$ *where* $\beta_0=$ *intercept* $\beta_1=$ *slope*

Least squares regression line ($\hat{y}=b_0+b_1 x$¿ is the best estimate of the true population regression line. The population standard deviation, **σ**, for **y** at any given value of **x** represents the spread of the normal distribution of the ε_i around mean, μ_y.

$$se=\sqrt{\frac{(\sum \text{ of residuals})^2}{n-2}}$$

Standard error about the regression line

Confidence Interval for Regression Parameters

Estimating regression parameters $\beta_0\beta_1$ is a case of one-sample inference with unknown population variance.

Rely on the T-distribution with n-2 degrees of freedom

$$SE(b_1)=\frac{s_e}{\sqrt{\sum ¿¿¿}} \qquad S_x=\sqrt{\sum ¿¿¿}$$

therefore, $SE(b_1)=\frac{s_e}{s_x\sqrt{n-1}}$

Hypothesis Test:

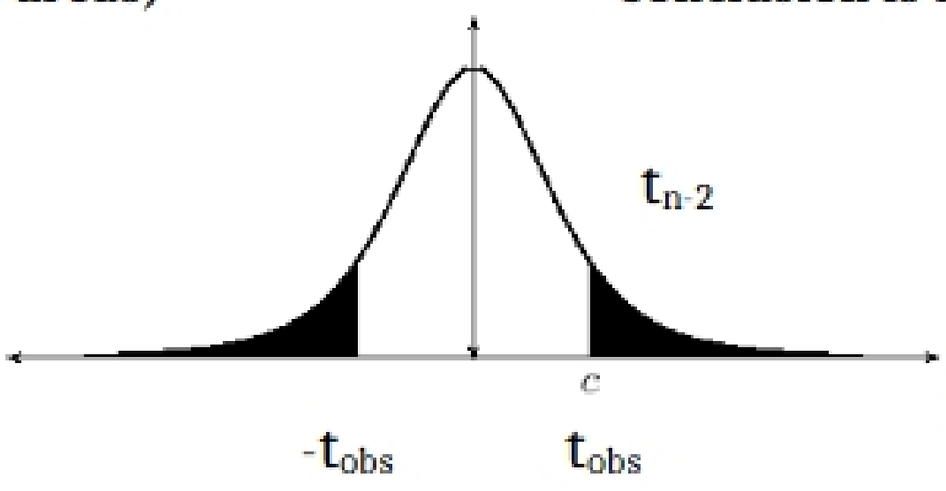
H_0: $\beta_1=0$ → (x and y are **not** linearly related)

H_a: $\beta_1\neq 0$ → (x and y are linearly related)

T-Test: $t=\frac{b_1-0}{se¿¿}$

P-Value (sum of the areas)

Conclusion is the same as usual



Confidence Interval for μ_y (mean response)

$CI=\hat{y}\pm t^{¿}_{n-2}SE(\hat{\mu}y)$

$SE(\hat{\mu}y)=se\sqrt{\frac{1}{n}+¿¿¿¿}$

Predicting Individual Response