Multiple Regression.

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{p-1} X_{i,p-1} + \varepsilon_i$,

$i = 1, 2, \ldots, n$.

ε_i: Independent random error terms that are normally distributed with 0 mean and constant variance σ^2.

- All $(p-1)$ covariates are linearly related to the dependent variable Y.
- p is the number of regression parameters.
- Model parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are estimated by minimizing the least square criterion:

$$f(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1})^2$$

Residuals: $e_i = Y_i - \hat{Y}_i \rightarrow SSE = \sum_{i=1}^{n} e_i^2$. $df = n-p$

Estimate of σ^2: $\hat{\sigma}^2 = MSE = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$ $p = $ # of predictors.

ANOVA table for multiple regression:

Source	SS	DF	MS	F
Regression	SSR	$p-1$	MSR	$F = \frac{MSR}{MSE}$
Error	SSE	$n-p$	MSE	
SSTO	SSTO	$n-1$		

$MSR = \frac{SSR}{p-1}$

$MSE = \frac{SSE}{n-p}$

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

$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = SSTO - SSE$.

Coefficient of determination $R^2 = \frac{SSR}{SSTO}$

F-test. (Global).

Model Usefulness test.

$H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ None of the covariates are linearly related to Y.

H_a: Not all β_i's are equal to 0.

At least some of the covariates are linearly related to Y. $\rightarrow$ Useful predictors of the Y.

The rejection of the H_0 states that the model is statistically sound.

Test statistic: $F = \frac{MSR}{MSE}$ Decision rule: Reject H_0 if $F > F_\alpha$, where α is based on $F_{p-1, n-p}$ distribution.

Conducting a series of t-tests leads to high overall Type I error rate α.

Test for individual regression parameter β_i

$H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$

Test statistic: $t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$

Decision Rule: Reject H_0 if $t > t_{\frac{\alpha}{2}}$ or $t < -t_{\frac{\alpha}{2}}$, where $t_{\frac{\alpha}{2}}$ is based on $n-p$ df.

Correlation plots of the covariates.

Estimation of σ^2: $s^2 = \frac{SSE}{(n - \text{\# of estimated } \beta \text{ parameters})}$

$= \frac{SSE}{n-(k+1)}$ (with k independent variables.)

Multicollinearity.

- When the covariates are highly correlated among themselves, difficulties can occur in carrying out inferences for individual covariates.
- Correlation among covariates makes the interpretations of the coefficients of the covariates questionable.
- The estimated regression coefficients ($\hat{\beta}_i$, $i = 1, 2, \ldots, n$) tend to have large variability.
- This will result: F test shows a significant overall regression relation between covariates, but the t-tests for individual β_i's are not significant.

Model selection Techniques.

1. R^2 Procedure:
   - Carry out regression of the Y on all possible subsets of the covariates, and compute R^2 in each case. (Full factorial).
   - Objective: Find the subset of the covariates such that adding more covariates only leads to a small improvement to R^2.
2. Stepwise Regression.
   - Find the subset of the covariates that best predicts the response variable Y.
   - Advantage: Variable selection is carried out automatically.
   - Involves a series of hypothesis testing procedures that decide whether to include or not to include a variable in the model.

3 variations: 1. Forward selection: Begin with small amount of X, and gradually work up to the largest number.

2. Backward selection: Begin with all X and gradually eliminate until one reaches the point where all remaining X are significantly important.

3. Stepwise/Mixed: Begin in the same way as Forward, but each time a variable is added all variables in the model are examined to see if any should be eliminated at that step. (Computational Demanding).

- If all variables significant (p-values < α), end process - model is fine.
- If F-test significant but some t-tests's p-values high (indicating non-significant covariates), use Forward stepwise regression to refine.

Adjusted R^2:

To account for the additional variables.

- If there are $(p-1)$ covariates in the model, then:

$adj\text{-}R^2 = 1 - (\frac{n-1}{n-p})(1 - R^2)$

It considers both sample size and the # of β parameters.

Akaike Information Criterion.

- Based on Maximum Likelihood and a penalty for total number of parameters in the model.

$AIC = n \log(\frac{SSE}{n}) + 2p + C$

- Choose model with smallest AIC.

BIC: Schwarz's Bayesian Information Criterion.

- Similar to AIC, but assigns a greater penalty for larger models.

$BIC = n \log(\frac{SSE}{n}) + (\log n) p + C$

- Choose model with smallest BIC.

Categorical variable in Regression model.

Indicator/Dummy variable:

$X_2 = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female.} \end{cases}$

Thus the fitted model $(\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2)$:

$\hat{Y} = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1 & \text{if Male} \\ \beta_0 + \beta_1 X_1 & \text{if Female.} \end{cases}$

Indicators with more than two values

$X_2 = \begin{cases} 1 & \text{if SM} \\ 0 & \text{Otherwise} \end{cases}$

$X_3 = \begin{cases} 1 & \text{if SS} \\ 0 & \text{Otherwise.} \end{cases}$

$\hat{Y} = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1 & \text{if SM} \\ (\beta_0 + \beta_3) + \beta_1 X_1 & \text{if SS} \\ \beta_0 + \beta_1 X_1 & \text{if HM} \end{cases}$

When we have a categorical variable with k levels, we need to create (k-1) indicator variables.

- Idea: Based on a reference category, where one of the categories serves as a reference and the effects of other categories are measured in relation to it.

Interaction Term.

- The effect of one covariate on the response variable may depend on the value of another.

$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$.

- If X_2 is an indicator variable (say, $=1$ if Male and 0 if Female):

$\hat{Y} = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 & \text{if Male} \\ \beta_0 + \beta_1 X_1 & \text{if Female.} \end{cases}$

Nonlinear regression.

Polynomial regression of order 2 (second order model with one covariate).

$Y_i = \beta_0 + \beta_1 (X_i - \bar{X}) + \beta_2 (X_i - \bar{X})^2 + \varepsilon_i$, $i = 1, 2, \ldots, n$.

- Viewed as a multiple regression model with two variables: $X_1 = (X - \bar{X})$ and $X_2 = (X - \bar{X})^2$

ANOVA Part I

: Covariates are Nominal. Dependent Y: Continuous.

Factors: Covariates included in the experiment.

The treatment method is the factor.

Factor Level: Specific forms of factor used in the experiment.

3 forms of treatment method: Medication, Exercise, Diet.

Treatment: Combinations of factor levels.

- In one factor ANOVA, the factor levels are the treatments.
- In multifactor ANOVA, specific combinations of factor levels from various factors are the treatments.

Total number of treatments: Multiply the number of levels for each factor together.

Factor A has 3 levels and Factor B has 2 levels $\rightarrow 3 \times 2 = 6$.

Replications: Number of observations within each treatment.

Balanced: If the number of replications in all treatments are equal. (Equal number of observations).

Side-by-side boxplot: Get an understanding of the differences between the observations in various factor levels.

Implications of the Model: The observations in the ith factor level describe a random sample of size n_i from a normal population with mean μ_i and variance σ^2, for $i = 1, 2, \ldots, k$. $\rightarrow$ k independent samples from k normal populations with mean μ_i and a common variance σ^2. (Normal distribution with different means but constant variance).

Test all three populations are identical: $H_0: \mu_1 = \mu_2 = \mu_3$ vs. H_a: Not all 3 μ_i's are equal.

- One factor ANOVA is a generalization of two-sample inference. (k-sample inference).

Objective of one-factor ANOVA: Verify whether factor levels (treatments) have any effect on the response variable.

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ vs. H_a: Not all μ_i's are equal.

If we reject H_0, then we need to find the "best" treatment, or the best group of treatments. $\rightarrow$ Pairwise Comparison Techniques.

Useful notations: Y_{ij}: jth observation in ith factor level

$n = \sum_{i=1}^{k} n_i$: Total sample size.

$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$: Average of all observations in ith factor level.

$\bar{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$ $\quad S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$: Sample variance of all observations in ith factor level.

$SSTO = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$

$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$

$SSTR = SSTO - SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2$

Computational formula: $SSTO = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{(\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij})^2}{n}$

$SSE = \sum_{i=1}^{k} (n_i - 1) S_i^2$ $\quad SSR = SSTO - SSE$

ANOVA Table for one factor Model

Source	SS	DF	MS	F
Treatment	SSTR	$k-1$	MSTR	$F = \frac{MSTR}{MSE}$
Error	SSE	$n-k$	MSE	
Total	SSTO	$n-1$		

$R^2 = \frac{SSTR}{SSTO}$ F-test: $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

H_a: Not all μ_i's are equal. Reject H_0 if $F > F_\alpha$ (F_α is based on $(k-1, n-k)$ df.)

Example: We will need "Average" and "Variance" for each treatment. $\sum_i^3 \sum_{j=1}^5 y_{ij}^2 = 1082$

$\sum_i^3 \sum_{j=1}^5 y_{ij} = 112$ Also: $\bar{Y} = \frac{112}{15} = 7.47$

$SSTO = 1082 - \frac{112^2}{15} = 245.73$ $SSE = 4 \times 5.7 + 4 \times 10.2 + 4 \times 4.7 = 82.4$. $SSTR = SSTO - SSE$

Tukey multiple Comparison Procedure.

- Identify the best treatment by comparing all possible pairs of treatment means

$H_0: \mu_i = \mu_j$ vs. $H_a: \mu_i \neq \mu_j$

For k means involves $kC_2 = k(k-1)/2$ tests.

Based on 'ind' hypothesis testing each with α.

P(at least one of the inferences is incorrect)

$= 1 - P(\text{None of the inferences is incorrect})$

$= 1 - (1-\alpha)^m$

Constructing Simultaneous CI for the difference $(\mu_i - \mu_j)$ with Family confidence coefficient $(1-\alpha)$

- The proportion of correct families of estimates when repeated samples are selected and specified confidence intervals for the entire family are calculated for each sample.

TCI for all distinct pairwise differences $(\mu_i - \mu_j)$ with family CI $(1-\alpha)$: $(\bar{y}_i - \bar{y}_j) \pm \frac{q_\alpha}{\sqrt{2}} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

$\hat{\sigma} = \sqrt{MSE}$ and q_α is the $(1-\alpha)$ percentile of the studentized range distribution with df $(k, n-k)$.

- HSD helps control the Type I error rate when conducting multiple comparisons.

$C(n,k) = n!/[k!(n-k)!]$

- Accept H_0 if zero is included in the interval, and conclude that the difference between the means is not statistically significant. (And vice versa).

Presentation of the results: Arrange the treatments in the increasing order of their means, and identify pairs of treatments that are not significantly different by joining them with lines below the treatments.

3.8	6.8	11.8	Tukey group.
Exercise	Diet	Medication	

{Exercise, Diet} and {Medication}

- Tukey's HSD test controls the probability of the type I error well. While SNK, REGWQ can be more efficient than Tukey when the sample sizes in each factor level are equal.

ANOVA Part II: Two-way ANOVA.

Factorial Experiment and Interaction: Includes all possible Combinations of factor levels of all factors. If A and B are two factors with factor levels a and b respectively, then the experiment is called an a×b factorial experiment.

Quadratic Model: $y = a + bx + cx^2$

If a factorial experiment, if the effect one factor has on the response variable depends on the factor level of the other factors, we say that the factors A and B interact.

Determine the interaction from Response plot.

- No interact: The lines in the response plot move in the same direction.
- Do interact: The direction of the lines are substantially different.

ANOVA Table for two-factor model.

- Assume that there are a levels of Factor A, b levels of factor B, and r replications within each treatment. → Total sample size: $n = a \cdot b \cdot r$.

Source	SS	DF	MS	F
Treatment	SSTR	$ab-1$	MSTR	$F_{TR} = \frac{MSTR}{MSE}$
Error	SSE	$ab(r-1)$	MSE	
Total	SSTO	$abr-1$		

Breakdown of SSTR

Source	SS	DF	MS	F
A	SSA	$a-1$	MSA	$F_A = \frac{MSA}{MSE}$
B	SSB	$b-1$	MSB	$F_B = \frac{MSB}{MSE}$
AB	SSAB	$(a-1)(b-1)$	MSAB	$F_{AB} = \frac{MSAB}{MSE}$
Total	SSTR	$ab-1$		

Interpretation of the ANOVA.

Step 1: H_0: There is no treatment effect.

H_a: There is a significant treatment effect.

Decision Rule: Reject H_0 if $F_{TR} > F_\alpha$; $df = \{ab-1, ab(r-1)\}$

- Acceptance of H_0 implies that the factors are not effective in explaining the variability of the response variable.
- So, we will stop and conclude that the model is not significantly significant. Otherwise, we proceed to next step.

Step 2: Test for interaction: H_0: There is no possible interaction between factors A and B. H_a: There is a significant interaction between factors A and B.

Decision Rule: Reject H_0 if $F_{AB} > F_\alpha$; $df = \{(a-1)(b-1), ab(r-1)\}$

- Rejection of H_0 implies strong interaction between factors A and B. So, individual influence of factor A and B are not meaningful. Do not carry out the F-test for main effects. Instead, we can carry a pairwise comparison procedure on each treatments to identify the best treatments. | Otherwise, if the interaction is not significant, we proceed to the next step.

Step 3: Test for main effect A.

H_0: There is no significant effect of effect A.

H_a: There is a significant effect of factor A.

Decision Rule: Reject H_0 if $F_A > F_\alpha$; $df = \{(a-1), ab(r-1)\}$

- If we reject H_0, it implies that the factor levels of factor A are significant, and we can use multiple comparison procedures to identify the equivalent groups among the factor levels of factor A.

Step 4: Test for main effect B

H_0: There is no significant effect of effect B.

H_a: There is a significant effect of factor B.

Decision Rule: Reject H_0 if $F_B > F_\alpha$; $df = \{(b-1), ab(r-1)\}$

- If we reject H_0, it implies that the factor levels of factor B are significant, and we can use multiple comparison procedures to identify the equivalent groups among the factor levels of factor B.

- In a two-factor ANOVA, the hypothesis to test first is: Overall treatment effect is significant.
- If the overall treatment effect is significant, the hypothesis to test next is the interaction is significant.
- If the interaction is statistically significant, the next step is do not test the significance of the main effects, carry out Pairwise Comparisons between all possible treatment means.
- If the interaction is not significant, the next step is test for significance of main effects and carry out Pairwise Comparisons of factor level means only for the significant main effects.

Multifactor ANOVA.

- For the three factor experiment, the interactions are AB, AC, BC, and ABC.
- Typically assume that higher order interactions are negligible and not include those terms in the models.
- In three factor model, we may want to include terms up to second order interaction.

Formula (Combinations): $C(n,r) = \frac{n!}{[(n-r)!\,r!]}$

A $100(1-\alpha)\%$ CI for a β parameter

$\hat{\beta}_i \pm (t_{\frac{\alpha}{2}}) s_{\hat{\beta}_i}$. t is based on $n-(k+1)$ df, n = # of observations, k+1 = # of β parameters in the model.

A Quadratic (Second-order) Model in a Single Quantitative Independent Variable.

$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$

β_0 = y-Intercept of the Curve. β_1 is a shift parameter. β_2 is the rate of curvature.

Extreme Multicollinearity: 1. Extreme Multi.: $|r| \geq 0.8$

Moderate Multi.: $0.2 \leq |r| < 0.8$ Low Multi: $|r| < 0.2$

- The best one-variable predictor of y is typically the variable that has the highest absolute value of the coefficient $(\hat{\beta}_i)$ relative to its standard error $(S_{\hat{\beta}_i})$. (Stepwise Regression).
- Include the highest absolute t-statistic at the start, then continuing the addition of the remaining variables one at a time.

Experimental unit: The smallest entity to which a treatment (or control) is applied in a study or experiment.

Multiple pairwise comparisons: Formula: $k \cdot (k-1) \div 2$, k: groups.

Choosing 2 groups (r) from a set of 4 groups (n) for comparison.

$4C_2 = 6$

Categorical Data Analysis. (Goodness of Fit Test)

- Frequency table. (CDA is appropriate when both X and Y are categorical.)
- Observed Frequency: f_i, Expected Frequency: F_i
- H_0: The theoretical model fits the data. H_a: The theoretical model doesn't fit the data.
- The "closeness" between the observed and expected frequency is measured by computing the Pearson's Chi-square statistic:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - F_i)^2}{F_i} = \sum_{i=1}^k \frac{(\text{Obs. freq.} - \text{Exp. freq.})^2}{\text{Exp. freq.}}$$

Decision Rule: Reject H_0 if $\chi^2 > \chi^2_\alpha$ ($df = k-1$)

Category	f_i	F_i	$f_i - F_i$
Round Y	315	312.75	2.25
Wrinkled Y	101	104.25	-3.25
Round G	108	104.25	3.75
Wrinkled G	32	34.75	-2.75.
Total	556	556	

$$\chi^2 = \frac{(2.25)^2}{312.75} + \frac{(-3.25)^2}{104.25} + \frac{(3.75)^2}{104.25} + \frac{(-2.75)^2}{34.75} = 0.47$$

With 3 df (= 7.81) ∴ Do not reject H_0.

Two-way Classification: The population is divided into categories according to the factor levels of two different factors. The main objective is to explore the relationship between the factors.

Contingency Table: Shows the frequencies of occurrence of events classified by the combination of factor levels of categorical variables. r rows and c columns → r×c contingency table.

H_0: Two factors are independent; H_a: Two factors are not independent.

Test statistic: $\chi^2 = \sum_{\text{all the cells}} \frac{(\text{Obs. freq.} - \text{Exp. freq.})^2}{\text{Exp. freq.}}$

Reject H_0 if $\chi^2 > \chi^2_\alpha$ $(df\ (r-1)(c-1))$

Two events are independent: $P(A \text{ and } B) = P(A) \cdot P(B)$.

Expected Frequency = $\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$

Two-way Classification

p_1 = P(rate of cancer|smoker) and p_2 = (rate of cancer|nonsmoker).

$H_0: p_1 = p_2$ and $H_a: p_1 > p_2$ | $\hat{p}_1 = \frac{155}{650}$, $\hat{p}_2 = \frac{55}{350}$, $\hat{p} = \frac{155+55}{650+350}$

$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$, $Z = 1.92$, p-value $= P(Z > 1.92) = 0.0274$. Reject H_0 at $\alpha = 0.05$.