Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

β_1 is the change in the mean of Y_i for a 1 unit increase in x_i, β_0 is the mean when $x_i = 0$
$S_{XX} = \sum (x_i - \bar{x})^2$, $S_{YY} = \sum (y_i - \bar{y})^2$,
$S_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y})$

Estimating σ^2

1) Standard deviation of $\hat{\beta}_1$: $\sigma_{\hat{\beta}_1} = \sqrt{var(\hat{\beta}_1)} = \sigma/\sqrt{S_{XX}}$
2) Variance of residuals: $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}$
3) $SSE = S_{YY} - \hat{\beta}_1 S_{XY}$
4) $\hat{\sigma}_{\hat{\beta}_1} = \hat{\sigma}/\sqrt{S_{XX}}$

Inference about β_1

1) When the error terms are normal, $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{XX})$
2) $\mathcal{H}_0 : \beta_1 = 0 \quad vs \quad \mathcal{H}_a : \beta_1 \neq 0$

$$T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{XX}}}, \quad RR = \{T_{obs} > t_{\alpha/2,n-2}\}$$

3) Could get same conclusion from p-value, which illustrates the probability that our results occurred under $\mathcal{H}_0$. That is, the probability that $F > \mathcal{F}$ under $\mathcal{H}_0$.
4) Confidence interval for β_1: $\hat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{XX}}}$.

ANOVA

$SS_{reg} = S_{YY} - SSE = \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2 = \sum (\hat{y}_i - \bar{y})^2$
$T \sim t_v, \quad T^2 \sim \mathcal{F}(1, v)$

Source	df	SS	MS	F	p-value
Model	1	SS_{reg}	$\text{MST} = \text{SS}_{\text{reg}}$	$\frac{\text{MST}}{\text{MSE}}$	$\Pr(F^* > F)$
Error	$n-2$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-2}$		
Total	$n-1$	S_{YY}			

lm summary table

- t-value (slope): $T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$
- F-statistic : T^2_{obs}
- Residual std error: $\hat{\sigma}$

Correlation

1. corr(X,Y) = corr(Y,X)
2. $r = S_{XY}/\sqrt{S_{XX} S_{YY}}$ is an estimator for ρ (the true pop. correlation).
3. $(1-\alpha)100\%$ confidence interval for ρ: transform r to $z = 0.5 \ln(\frac{1+r}{1-r})$. Build an interval:$z \pm \frac{z_{\alpha/2}}{\sqrt{n-3}} = (c_l, c_u)$, where $z_{\alpha/2}$ is from the standard Normal table. Then, the interval is $\left(\frac{e^{2c_L}-1}{e^{2c_L}+1}, \frac{e^{2c_U}-1}{e^{2c_U}+1}\right)$
4. Coefficient of determination: $R^2 = 1 - SSE/S_{YY}$

Estimating response

1. Mean response confidence interval:
   $\hat{y}_0 \pm t_{n-2,\alpha/2} \hat{\sigma} \sqrt{1/n + (x_0 - \bar{x})^2/S_{XX}}$
2. Individual value Y_0 confidence interval:
   $\hat{y}_0 \pm t_{n-2,\alpha/2} \hat{\sigma} \sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{XX}}$

Residual Analysis

1) Assumptions: ϵ_i are independent,
$E(\epsilon_i) = 0,\ var(\epsilon_i) = \sigma^2,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$
2) Check Normality with QQ plot and histogram of the studentized residuals, which have mean 0, all residuals should lie within 3 std deviations.
3) Check $E(\epsilon_i) = 0$ by plotting studentized residuals against fitted values. Points should have equal variance and zero mean, i.e. evenly distributed.

Polynomial Regression

$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p \epsilon_i$, not all intermediate powers need be present.
Higher-order terms are specified using the $I(\cdot)$ function in R.
1. Test that the quadratic term is zero: $H_0 : \beta_2 = 0$.
2. If rejected, use linear and quadratic terms in model.
3. If not rejected, there is no evidence that the quadratic model gives significant improvement over the linear model.

Multiple Regression (2+ covariates)

$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i$
The model is linear in the parameters (β_i), not necessarily in the covariates (x_i). Same assumptions are made about the residuals.
β_j is the change in the mean of Y_i for a 1 unit increase of x_{ij} when holding all other variables constant.
1) $\hat{\sigma}^2 = (n - (K+1))^{-1} \sum (y_i - \hat{y}_i)^2 = SSE/(n - (K+1))$ where (K+1) is the number of coefficients β_i in the model.
2) Can test each coefficient individually with same hypothesis as in simple regression. In which case, we test for e.g. β_j after adjusting for all other variables.
3) Confidence interval for β_j : $\hat{\beta}_j \pm t_{n-(K+1),\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_j}$
4) Global Fit:
$R_a^2 = 1 - \frac{n-1}{n-(K+1)} \left(\frac{SSE}{S_{YY}}\right) = 1 - \frac{n-1}{n-K-1}(1 - R^2)$
$\mathcal{H}_0 : \beta_1 = \beta_2 = ... = 0 \quad \mathcal{H}_a :$ at least one $\beta_j \neq 0$
$F = \frac{(S_{YY} - SSE)/K}{SSE/(n-(K+1))} = \frac{R^2/K}{(1-R^2)/(n-(K+1))}$
$\mathcal{H}_0$ is rejected for $F > \mathcal{F}_{\alpha,K,n-(K+1)}$.

Interaction

if an interaction is suspected between X_1 and X_2, we incorporate the interaction by setting
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i =$
$\beta_0 + \beta_1 x_{i1} + (\beta_2 + \beta_3 x_{i1}) x_{i2} + \epsilon_i = \beta_0 + (\beta_1 + \beta_3 x_{i2}) x_{i1} + \beta_2 x_{i2} + \epsilon_i$
In the above model, a 1-unit increase in x_2 for a fixed x_1 corresponds to an estimated $\hat{\beta}_2 + \hat{\beta}_3 x_1$ increase in Y_i.
1) Fit the model including the covariates and interaction.
2) Conduct a global F-test with $\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$
3) If rejected, test for an interaction by using a Student t-test to test $\mathcal{H}_0 : \beta_3 = 0$. If rejected, stop. Otherwise, re-fit the model without the interaction.

Qualitative

Set $Z_i = 0\ \forall i$ for reference group and $Z_i = \begin{cases} 1 & \text{if condition i} \\ 0 & \text{otherwise} \end{cases}$

$$Y_i = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \epsilon_i$$

Where $\hat{\beta}_0 = \mu_0$, $\hat{\beta}_1 = \mu_1 - \mu_0$, and $\hat{\beta}_2 = \mu_2 - \mu_0$, and
$\mathcal{H}_0 : \beta_1 = \beta_2 = 0 \iff \mathcal{H}_0 : \mu_2 = \mu_1 = \mu_0$, $\mathcal{H}_a : \geq 1\ \beta_i \neq 0$

Qualitative and quantitative:
The model is $Y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i$.
1) $z_i = 0$: $Y_i = \beta_0 + \beta_2 x_i$
2) $z_i = 1$: $Y_i = \beta_0 + \beta_1 + \beta_2 x_i$
So the slope is the same, only y-intercept changes.
Interaction: $Y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \beta_3 z_1 x_i$. Then, slopes vary:
1) $z_i = 0$: $Y_i = \beta_0 + \beta_2 x_i$
2) $z_i = 1$: $Y_i = \beta_0 + \beta_1 + (\beta_2 + \beta_3) x_i$
Where $\beta_1 + \beta_3 x_i$ is the difference in Y between $z_i = 1$ and $z_i = 0$. Should always test for existence of an interaction. If no evidence to reject $\mathcal{H}_0$ of no interaction, must re-fit model without interaction. If evidence of interaction, slopes are different and interpret the results accordingly.

Comparing Nested Models

M_0 and M_1 are nested models if one contains a subset of the other. $M_0 = \beta_0 + ... + \beta_g x_g$, $M_1 = M_0 + \beta_{g+1} x_{g+1} + ... + \beta_k x_k$

$$\mathcal{H}_0 : \beta_{g+1} x_{g+1} = ... = \beta_k x_k = 0 \quad \mathcal{H}_a : \text{ at least 1 } \beta_i \neq 0$$

Note: we always have $SSE_{M_0} \geq SSE_{M_1}$.
1) $SSE_{M_0} - SSE_{M_1}$ large $\implies$ M_1 explains more variance than just using M_0.
2)$SSE_{M_0} - SSE_{M_1}$ small $\implies$ additional terms in M_1 don't contribute to model fit.
To determine how "large" the difference is:

$$F = \frac{(SSE_{M_0} - SSE_{M_1})/(k-g)}{SSE_{M_1}/(n-(k+1))}, \quad RR = \{F > \mathcal{F}(\alpha, k-g, n-(k+1))\}$$

Multicollinearity

When two covariates in a regression analysis are highly correlated with each other (their coefficient of correlation is high), the analysis is said to be subject to multicollinearity. The covariates "compete" for the explanatory power in the association with the response.

Multinomial Distribution

One qualitative variable C can take k possible values $\{c_1, ..., c_k\}$. Let X_i the # of times c_i occurs. The set of X_i has a multinomial distribution.

$$P(X_1 = n_1, ..., X_k = n_k) = \frac{n!}{n_1! ... n_k!} p_1^{n_1} ... p_k^{n_k}$$

where $n_1 + ... + n_k = n$. $E(X_i) = np_i$.

Chi-square

$\mathcal{H}_0 : p_1 = p_1^*, ..., p_k = p_k^*$ $\quad \mathcal{H}_a : p_i \neq p_i^*$ for at least one i
Given by Pearson's chi-square statistic :

$$X_{obs}^2 = \sum_{i=1}^{k} \frac{(n_i - np_i^*)^2}{np_i^*} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Where O_i :=observed and E_i :=expected. Distribution of χ^2 under $\mathcal{H}_0$ is $\chi^2_{(k-1)}$, $(k-1)$:=(deg. of freedom). Given α, $RR = \{X_{obs}^2 > \chi^2_{\alpha,(k-1)}\}$ and $p = Pr\{\chi^2_{(k-1)} > X^2_{obs}\}$. Every expected count must be ≥ 5 for this test.

Contingency tables

$n_{j\bullet} = n_{j1} + ... + n_{jc}$, $n_{\bullet k} = n_{1k} + ... + n_{rk}$.
$n_{1\bullet} + ... + n_{r\bullet} = n_{\bullet 1} + ... + n_{\bullet c} = n$ (sum of all entries).
We want to test :
$\mathcal{H}_0$: X, Y independent *vs* $\mathcal{H}_a$: X, Y not independent. The expected counts are given by $\hat{E}_{jk} = n\hat{p}_{j\bullet}\hat{p}_{\bullet k} = n_{j\bullet}n_{\bullet k}/n$, where $\hat{p}_{j\bullet} = n_{j\bullet}/n$, $\hat{p}_{\bullet k} = n_{\bullet k}/n$

$$X^2 = \sum_{j=1}^{r}\sum_{k=1}^{c} \frac{(n_{jk} - (n_{j\bullet}n_{\bullet k}/n))^2}{n_{j\bullet}n_{\bullet k}/n} = \sum_{j=1}^{r}\sum_{k=1}^{c} \frac{(n_{jk} - \hat{E}_{jk})^2}{\hat{E}_{jk}}$$

Under $\mathcal{H}_0$: $X^2 \sim \chi^2_{(r-1)(c-1)}$, $RR = \{X^2 > \chi^2_{\alpha,(r-1)(c-1)}\}$
Caveats: must have expected cell count ≥ 5 for all cells, observations must be mutually independent and identically distributed.

Fisher's exact test

If expected cell count is not ≥ 5 for all cells.

McNemar's test

Matched pairs experiments, e.g.

Response 1/2	Yes	No	Total
Yes	n_{11}	n_{12}	$n_{1\bullet}$
No	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Want to test whether the proportions are the same before and after, i.e. $\mathcal{H}_0 : p_1 = p_2$, $\mathcal{H}_a : p_1 \neq p_2$.

$$Q_M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}, \qquad RR = \{Q_M > \chi^2_{\alpha,1}\}$$

Non-Parametric statistics

Wilcoxon test

To test the hypothesis that the probability distributions of both pop. are equivalent ($D_0 = D_1$).
Conditions: independent samples, cts distributions.
1) order together the observations from both samples
2) assign a rank to each, if equality, take average of ranks.
3) take sum of ranks of each group, let T the sum of sample with smaller size.
$\mathcal{H}_0 : D_0 = D_1$, $\mathcal{H}_a$: (T_U, T_L are table values),
1) D_1 left of D_2: $RR = \{T \leq T_L\}$ if $T = T_1$, $\{T \geq T_U\}$ o.w.
2) D_1 right of D_2: $RR = \{T \geq T_U\}$ if $T = T_1$, $\{T \leq T_L\}$ o.w.
3) one or two.

Normal approx. of Wilcoxon

If $n_1, n_2 \geq 10$. $Z = \frac{T_1 - (n_1(n_1+n_2+1)/2)}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}$, where T_1 sum of ranks corresponding to D_1, n_i sample size. Then, $Z \sim \mathcal{N}(0,1)$ and, letting z_α the value from normal table:
(1) $RR = \{Z < -z_\alpha\}; p-value = Pr(Z < Z_{obs})$
(2) $RR = \{Z > z_\alpha\}; p-value = Pr(Z > Z_{obs})$
(3) $RR = \{|Z| > z_{\alpha/2}\}; p-value = 2 \times Pr(Z > |Z_{obs}|)$

Wilcoxon's signed rank test (for paired data)

Let $X_1, ..., X_n$, $Y_1, ..., Y_n$ random samples of paired observations. $Diff_1 = X_1 - Y_1, ..., Diff_n = X_n - Y_n$
1) order absolute values of differences, take out the zeros
2) rank the differences, ties handled as usual
3) let T_+, T_- sum of ranks of positive and negative differences
$\mathcal{H}_0 : D_1 = D_2$, $\mathcal{H}_a$: same as before, with T_0 table value:
1) $RR = \{T_+ \leq T_0\}$ 2) $RR = \{T_- \leq T_0\}$
3) $RR = \{T \leq T_0\}$ where T = the smallest of T_-, T_+

Large sample Wilcoxon's signed rank test

If the number of pairs $n \geq 25$ (after excluding zeros), then $Z = \frac{T_+ - (n(n+1)/4)}{\sqrt{n(n+1)(2n+1)/24}}$, $Z \sim \mathcal{N}(0,1)$, RR same as 2 sections above.

Kruskal-Wallis test (assume CRD)

Def. (CRD): treatments are assigned randomly so that each experimental unit gets the same chance of receiving any one treatment. Rank-based non-parametric test to test difference in distribution among ≥ 2 groups. $\mathcal{H}_0$: the k distributions are identical, $\mathcal{H}_a$: at least one differs.
1) Take ranks, as with *Wilcoxon*.
2) Let $\overline{R}_j$ be the rank average of group j, $\overline{R}$ overall avg.
Under $\mathcal{H}_0$, we expect $\overline{R}_1 \approx ... \approx \overline{R}_k$.

$$KW = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j(\overline{R}_j - \overline{R})^2 \qquad RR = \{KW > \chi^2_{\alpha,k-1}\}$$

Friedman test (assume RBD)

A matched set of B blocks are formed with each consisting of K experimental units. One experimental unit from each block is randomly assigned to each treatment. B or $K \geq 5$
1) Rank the observations within blocks.
2) $\overline{R}_j$ average of ranks within treatment j.
Total average of ranks is $K(K+1)/2$, under $\mathcal{H}_0$, we expect $\overline{R}_1 \approx ... \approx \overline{R}_K$.

$$F_r = \frac{12B}{K(K+1)} \sum_{j=1}^{K} (\overline{R}_j - \overline{R})^2 \qquad RR = \{F_r > \chi^2_{\alpha,K-1}\}$$

Spearman's rank correlation coefficient

Consider n mutually indep. and indentically distributed pairs of variables $(X_1, Y_1), ..., (X_n, Y_n)$.
1) Rank the observations within $X : u_i$ and $Y : v_i$
2) if no ties in both, $r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (v_i - u_i)^2$
3) else, $r_s = \frac{SS_{uv}}{\sqrt{SS_{uu}SS_{vv}}}$, where $SS_{uv} = \sum (u_i - \bar{u})(v_i - \bar{v})$
A confidence interval for ρ is built the same way as earlier.

ANOVA

To test the equality of means in K populations.
$\mathcal{H}_0 : \mu_1 = ... = \mu_K = \mu$, $\mathcal{H}_a$: at least one differs. Assume homoscedasticity ($\sigma_1^2 = ... = \sigma_k^2$), normally distributed response. Let $\overline{Y}_k = 1/n_k(Y_{k1} + ... + Y_{kn_k}) = \hat{\mu}_k$, $\hat{\mu} = \overline{Y}$.
$SSE = \sum_{k=1}^{K}\sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_k)^2 = \sum_{k=1}^{K} (n_k - 1)S_k^2$, where S_k^2 the sample variance in group k, $SST = \sum_{k=1}^{K} n_k(\hat{\mu}_k - \hat{\mu})^2$.

$$F = \frac{MST}{MSE} = \frac{SST/(K-1)}{SSE/(n-K)}, \quad RR = \{F > \mathcal{F}_{\alpha,K-1,n-K}\}$$

Source	df	SS	MS	F	p-value
Treat	$K-1$	SST	MST	$\frac{\text{MST}}{\text{MSE}}$	$P\left(\mathcal{F}_{K-1,n-K} > F_{obs}\right)$
Error	$n-K$	SSE	MSE		
Total	$n-1$	SS_{tot}	$= SSE$	$+SST$	

Comparing means

Want to compare differences between groups two at a time to see how different they are. Assume equal var, normality. Conf. int for diff. in means:
$(\bar{Y}_i - \bar{Y}_j) \pm t_{(\alpha/2,n-K)}\sqrt{MSE/n_i + MSE/n_j}$
Bonferroni: experiment-wise error $\alpha_E \implies$ comparison-wise error α_E/C where C is the # of comparisons.
Turkey's HSD: all groups of equal size, only pair-wise comp.
Scheffe's: for linear comb. of means : $\mu_1 + \mu_2 - \mu_3$

ANOVA for RBD

Source	df	SS	MS	F
Treat	$K-1$	SST	MST $= \frac{\text{SST}}{K-1}$	$\frac{\text{MST}}{\text{MSE}}$
Blocks	$B-1$	SSB	MSB $= \frac{\text{SSB}}{B-1}$	$\frac{\text{MSB}}{\text{MSE}}$
Error	$n-K-B+1$	SSE	MSE $= \frac{\text{SSE}}{n-K-B+1}$	
Total	$n-1$	SS_{Tot}		

Anova F-test (treat) : $\mathcal{H}_0 : \mu_1 = ... = \mu_k$, $\mathcal{H}_a : \geq 1$ differs.

$$F = \frac{SST/(K-1)}{SSE/(n-K-B+1)} = \frac{MST}{MSE}, RR = \{F > \mathcal{F}_{\alpha,K-1,n-K-B+1}\}$$

Anova F-test (block) : same as above, but with the means within blocks. Replace SST with SSB, $(K-1)$ with $(B-1)$. SSB is the same as SST, but within blocks.