- Names for $\{f_j(x_i)\}$: (j-th) **feature**, **predictor**, **covariate**, **independent variable**
- Names for $\{y_i\}$: **response**,**response variable**, **dependent variable**, **target**, **label**

- Def. **Observed response variables**: $y_1, y_2, \ldots, y_n$
- Def. **Features**: z_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$
- Def. **Regression coefficients**: $\beta := (\beta_1, \ldots, \beta_p)$
- Def. **Squared error**:
$$S(\beta) := \sum_{i=1}^{n} (y_i - \beta^\top z_i)^2$$
- Def. **Least squares estimate**:
$$\hat{S} := \min_{\beta \in \mathbb{R}^p} S(\beta)$$
- Def. **Least squares regression coefficients**:
$$\hat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p) := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} S(\beta)$$

$$x_i \in \mathbb{R}^d, \qquad y_i \in \mathbb{R},$$
$$y_i = \sum_{j=1}^{p} z_{ij}\beta_j + \epsilon_i,$$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^\top z_i)^2$$

$$\hat{\beta} = (Z^\top Z)^{-1} Z^\top y$$

$$\mathrm{Var}[X] = \mathrm{Var}[\mathbb{E}[X|Y]] + \mathbb{E}[\mathrm{Var}[X|Y]],$$

$$\lim_{n\to\infty} \Pr\left[\sqrt{n}(\bar{X}_n - \mu) \le z\right] = \lim_{n\to\infty} \Pr\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le \frac{z}{\sigma}\right] = \Phi\left(\frac{z}{\sigma}\right)$$

Chi-squared distribution χ^2

- Let $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$. The distribution χ^2_k is defined as
$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$$
- PDF:
$$f(x;\, k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \mathbf{1}_{x \ge 0}$$
- If $X \sim \chi^2_k$, then
$$\mathbb{E}[X] = k, \quad \mathrm{Var}[X] = 2k$$

- For non-random matrices $A \in \mathbb{R}^{*\times p}$ and $B \in \mathbb{R}^{*\times m}$,
$$\mathrm{Cov}(AX, BY) = A\mathrm{Cov}(X, Y)B^\top.$$
- For a constant vector b,
$$\mathrm{Var}[AX + b] = A\mathrm{Var}[X]A^\top$$

- Def. These p equations are known as the **Normal Equation** (bc. normal is a synonym to perpendicular)
- We have
$$(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^\top (z_{1j}, \ldots, z_{n,j}) = 0, \qquad j = 1, \ldots, p$$

where
$$\hat{\epsilon}_i := y_i - \hat{\beta}^\top z_i, \qquad i = 1, \ldots, n$$

are the **residuals**

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

- $\mu(\bar{X}) = \mu$
- $\sigma^2(\bar{X}) = \sigma^2/n$
- $\gamma(\bar{X}) = \gamma/\sqrt{n}$
- $\kappa(\bar{X}) = \kappa/n$

- Taking $h(x) = x^k$ gives the k-th moment of X
- Some special moments of interest have given names:
  - The **mean** $\mu = \mathbb{E}[X]$ corresponds to $h(x) = x$
  - The **variance** of X is $\sigma^2 := \mathbb{E}\left[(X - \mu)^2\right]$
  - The **skewness** of X is $\gamma := \mathbb{E}\left[(X - \mu)^3\right]/\sigma^3$
  - The (excess) **kurtosis** of X is $\kappa := \mathbb{E}\left[(X - \mu)^4\right]/\sigma^4 - 3$
- γ is useful as a measure of symmetry; it is zero for symmetric distributions
- $\kappa = 0$ when $X \sim N(0,1)$. κ is useful in measuring whether the tails of the distribution are heavier ($\kappa > 0$) or lighter ($\kappa < 0$) than the tails of the normal distribution

Suppose $Y \sim \mathcal{N}(\mu, \Sigma)$ and that $\Sigma^{-1} \in \mathbb{R}^{n\times n}$ exists. Then
$$(Y - \mu)^\top \Sigma^{-1} (Y - \mu) \sim \chi^2_n$$

$$\frac{\partial S}{\partial \beta_j} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - \beta^\top z_i)(-z_{ij}) = 0, \qquad j = 1, \ldots, p$$

F-distribution

$$F = \frac{\frac{1}{p-q}\left(SS_{sub} - SS_{full}\right)}{\frac{1}{n-p} SS_{full}} \sim F_{p-q,\, n-p}$$

- The normalized ratio of two Chisquared distribution:
$$F_{d_1,d_2} := \frac{\frac{1}{d_1}\chi^2_{d_1}}{\frac{1}{d_2}\chi^2_{d_2}}$$
- PDF:
$$f(x; d_1, d_2) = \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2 - 1} \left(1 + \frac{d_1}{d_2}x\right)^{-(d_1+d_2)/2}$$
- If $X \sim F_{n_1,n_2}$, then
$$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}, \qquad d_2 > 2$$

Suppose that $c = [0, \ldots, 0, \underset{\uparrow}{1}, 0 \ldots]^T \in \mathbb{R}^p$ 1 in the j-th entry

If we hypothesize that $\beta_j = 0$, we would have $t = \frac{\hat{\beta}_j - 0}{s\sqrt{c^T (Z^T Z)^{-1} c}} \sim t_{n-p}$

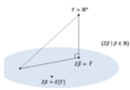
very large or small values of t are evidence against our hypothesis

$$\hat{y} = Hy, \qquad H := Z(Z^\top Z)^{-1} Z^\top$$

(Tukey called H the "hat" matrix)

- Properties of H:
  - Symmetric: $H = H^\top$
  - Idempotent: $H^2 = H$ (a symmetric idempotent matrix such as H is called a perpendicular projection matrix (PPM))
  - The eignevalues of a real PPM are all either 0 or 1
  - If Z is invertible, H has p non-zero eigenvalues
  - $I - H$ is PPM

- Def. **Coefficient of determination**:
$$R^2 := \frac{SS_{Fit}}{SS_{Tot}} = 1 - \frac{SS_{Res}}{SS_{Tot}}$$



  - Proportion of variation accounted for by all variables compared to the sum of squares error under the model $y_i = \beta_0 + \epsilon_i$
  - Measures how well Y is predicted or determined by $Z\hat{\beta}$:
  - $R := \sqrt{R^2}$ is called the **coefficient of multiple correlation** – it measures how well the response y correlates with the p predictors in Z taken collectively
  - When $z_i = (1, x_i) \in \mathbb{R}^2$, R is the Pearson correlation of $\{x_i\}$ and $\{y_i\}$

- In particular, if
$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}\right)$$

then the conditional distribution of Y given X is
$$\mathcal{L}(Y|X) = \mathcal{N}\left(\mu_y + \rho\sigma_y \frac{X - \mu_x}{\sigma_x}, \sigma_y^2(1 - \rho^2)\right)$$

- We can test H_0 using the extra SS principle:
$$F = \frac{\frac{1}{k-1}\left(SS_{tot} - SS_{within}\right)}{\frac{1}{n-k} SS_{within}}$$

reject H_0 at level α if
$$F > F^{1-\alpha}_{k-1,\, n-k}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- $SS_{Tot} := \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the **Total (or centered) sum of square**
- $SS_{Fit} := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ is the **Centered sum of squares of fitted values**
- $SS_{Res} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is the **Residual sum of squares**

$$Y \sim \mathcal{N}(\mu, \Sigma)$$

- If Σ is invertible, then the density of Y is
$$f_Y(y) = \frac{1}{(2\pi)^{m/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu)\right)$$

(here $y = (y_1, \ldots, y_m)$)

$$\mathbb{E}[Y_2|Y_1] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1)$$

variance

$$\mathrm{Var}[Y_2|Y_1] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}$$

$E[Y|X=x] = \beta_0 + \beta_1 x + \beta_2 [x - x_0]_+$

$z_+ := \max\{0, z\} = z \cdot 1_{z>0}$

$E[Y|X=x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1}$

(group 0 has mean β_0, mean of group $j>0$ is $\beta_0 + \beta_j$)

$$SS_{Tot} = SS_{Fit} + SS_{Res}$$

t-distribution

- Suppose that $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_k$, X and Z independent. Then
$$\frac{Z}{\sqrt{\frac{X}{k}}} \sim t_k$$
- PDF:
$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$
- If $Y \sim t_k$, then
$$\mathbb{E}[Y] = 0, \qquad \mathrm{Var}[Y] = \frac{k}{k-2}, \quad k \ge 3$$
- The t-distribution converges to the normal distribution as $k \to \infty$. It has heavier tails than that of the normal distributio

$\alpha/2 \quad 1-\alpha \quad \alpha/2$

$= \Pr\left(-t^{1-\alpha/2}_{n-1} < T < t^{1-\alpha/2}_{n-1}\right)$

$= \Pr\left(-t^{1-\alpha/2}_{n-1} < \frac{\bar{Y} - \mu}{s/\sqrt{n}} < t^{1-\alpha/2}_{n-1}\right)$

$= \Pr\left(\underbrace{\bar{Y} - \frac{s}{\sqrt{n}} t^{1-\alpha/2}_{n-1}}_{L} < \mu < \underbrace{\bar{Y} + \frac{s}{\sqrt{n}} t^{1-\alpha/2}_{n-1}}_{U}\right)$

- $\hat{y}_i = H_i y$ (H_i is the i-th row of H)
- $H_{ij} = z_i^\top (Z^\top Z)^{-1} z_j = H_{ji}$ (the contribution of y_i to $\hat{y}_j$ equals that of y_j to $\hat{y}_i$)
- $H_{ii} = z_i^\top (Z^\top Z)^{-1} z_i \ge 0$ (Exc.)
- H projects vectors onto the **columns** space of Z
$\mathrm{Col}(Z) := \mathcal{M} = \{Z\beta \,|\, \beta \in \mathbb{R}^p\}$
- $I - H$ projects vectors onto the **null** space of Z
$\mathrm{Null}(Z) := \mathcal{M}^\top := \{v \in \mathbb{R}^n, |\, Zv = 0\}$ (the set of vectors orthogonal to vectors in $\mathcal{M}$)

Thm.
$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{\|\hat{\epsilon}\|^2}{\sigma^2} \sim \chi^2_{n-p}$$

One Sample t-test

Data: $y_1, y_2, \dots, y_n$

$$t = \frac{\bar y - \mu}{s/\sqrt{n}} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar y)^2$$

K-Groups (ANOVA):

- Cell mean model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

Effect means model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

- The SS for the cell means:

$$SS_{within} := SS_{res} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar y_{i.})^2$$

- The total SS:

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar y_{..})^2$$

- We can test H_0 using the extra SS principle:

$$F = \frac{\frac{1}{k-1}(SS_{tot} - SS_{within})}{\frac{1}{n-k} SS_{within}}$$

reject H_0 at level α if

$$F > F^{1-\alpha}_{k-1, n-k}$$

- we also define

$$SS_{between} := SS_{fit} = SS_{tot} - SS_{within} = \sum_{i=1}^{k} n_i (\bar y_{i.} - \bar y_{..})^2$$

Source	DoF	SS	MS	F
Groups	$k-1$	$SS_{between}$	$SS_{between}/k-1$	$\frac{MS_{between}}{MS_{within}}$
Error	$n-k$	SS_{within}	$SS_{within}/n-k$	
Total	$n-1$	SS_{tot}		

$$MS_{within} = s^2$$

- In general:

$$\sum_{i=1}^{k} \lambda_i = 0 \qquad \sum_{i=1}^{k} \lambda_i^2 > 0 \qquad t = \frac{\sum_{i=1}^{k} \lambda_i \bar y_{i.}}{s\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2}{n_i}}} \sim t_{n-k}$$

Two-Sample t-test:

$$t = \frac{\bar y_1 - \bar y_0}{s\sqrt{n/n_0 n_1}} = \frac{\bar y_1 - \bar y_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

(the unbiased estimator of σ^2)

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \underbrace{\hat\beta_0 - \hat\beta_1 x_i}_{\hat y_i})^2 = \frac{\sum_{i=1}^{n_0}(y_{0i} - \bar y_0)^2 + \sum_{i=1}^{n_1}(y_{1i} - \bar y_1)^2}{n_0 + n_1 - 2}$$

$$t = \frac{\bar y_A - \frac{2}{3}\bar y_B - \frac{1}{3}\bar y_C}{s\sqrt{\frac{1}{n_A} + \frac{4}{9n_B} + \frac{1}{9n_C}}},$$

$$s^2 = \frac{1}{n-3}\sum_{c\in\{A,B,C\}}\sum_{j=1}^{n_c}(y_{c,j} - \bar y_c)^2$$

Statistical power :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar y)^2$$

$$=: \Pr\left[\frac{(Z + \sqrt{n}(\frac{\mu - \mu_0}{\sigma}))^2}{V/(n-1)} > F^{1-\alpha}_{1,n-1}\right] \ (*)$$

Where $Z \sim N(0,1)$ $\quad Z = \frac{\sqrt{n}(\bar Y - \mu)}{\sigma}$

$V \sim \chi^2_{n-1}$ $\quad V = \frac{s^2}{\sigma^2}(n-1)$

$\lambda = n\left(\frac{\mu - \mu_0}{\sigma}\right)$

$\frac{N^2(\sqrt{\lambda}, 1)}{\frac{1}{k}\chi^2_k}$ is known as the non-central F dist. with 1 DoF over k DoF (denoted $F_{1,k}(\lambda)$)

$$\lambda = \left(\sqrt{n}\left(\frac{\mu - \mu_0}{\sigma}\right)\right)^2$$

Contrasts :

$$t = \frac{\frac{\bar y_{1.} + \bar y_{2.} + \bar y_{3.}}{3} - \frac{\bar y_{4.} + \bar y_{5.} + \bar y_{6.} + \bar y_{7.}}{4}}{s\sqrt{\frac{1}{9}(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3}) + \frac{1}{16}(\frac{1}{n_4} + \dots + \frac{1}{n_7})}}$$

$s^2 = MSE_{within}$

$$t \sim t_{n-7}$$

Bonferroni's Union Bound

- we have m tests (e.g. $m = \binom{k}{2}$)
- we conduct each test at level α/m (e.g. for t-tests, we reject based on $t^{1-\frac{\alpha}{2m}}_{n-k}$)

The SVD

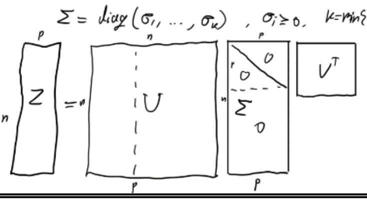
$$Z \in \mathbb{R}^{n\times p}$$

$$Z = U\Sigma V^T$$

where $U \in \mathbb{R}^{n\times n}$ $\quad \Sigma \in \mathbb{R}^{n\times p}$ $\quad V \in \mathbb{R}^{p\times p}$

$U^T U = U U^T = I_n \qquad V^T V = V V^T = I_p$

$\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_k), \quad \sigma_i \geq 0, \quad k = \min\{n, p\}$



overall:

(I) $Z = U\Sigma V^T$

(II) $y^* := U^T y$

(III) $\beta_j^* = \begin{cases} y_j^*/\sigma_j & \sigma_j \neq 0 \\ 0 & \sigma_j = 0 \end{cases}$

(IV) $\hat\beta = V\beta^*$

Computing SVD is done in $O(np^2)$

Theorem (BH '95)

If all p-values are independent, then $FDR \leq \frac{m_0}{m}\alpha \leq \alpha$

FDR controlling using Benjamini & Hochberg (BH)

- Perform each test, sort the p-values from lowest to highest $P_{(1)} \leq P_{(2)} \leq \dots P_{(m)}$
- Define $l_i = \alpha \cdot \frac{i}{m}$ (line with slope $\frac{\alpha}{m}$)
- Define $i^* = \max\{i : P_{(i)} \leq l_i\}$
- Reject all $H_{0,i}$ $\quad P_{(i)} \leq P_{(i^*)}$

If the tests are independent

False-Discovery Rate (FDR)

- Suppose we make m hypotheses tests $\{H_{0,i}\}_{i=1}^m$. Each has either rejected or not.

We summarize the situation in a table:

	# not rej.	# rej.	Total
$H_{0,i}$ true	U	V	m_0
$H_{0,i}$ false	T	S	m_1
Total	$m-R$	R	m

- Def. false discovery proportion is

$$FDP := \begin{cases} V/R & R > 0 \\ 0 & R = 0 \end{cases}$$

Def. false discovery rate is $E[FDP]$

Under general depency structure:

- we should be more conservative
- we use $l_i = \alpha\frac{i}{m}\cdot\frac{1}{C_m}$

$$C_m = \sum_{i=1}^{m}\frac{1}{i} \approx \ln(m) + \gamma - \frac{1}{2m} \approx \ln m$$

γ: Euler constant ≈ 0.57

Suppose that $y \sim N(Z\beta, \sigma^2 I)$ and that $(Z^T Z)^{-1}$ is invertible.

- $\hat\beta \sim N(\beta, \sigma^2(Z^T Z)^{-1})$
- $\hat y = Z\hat\beta \sim N(Z\beta, H\sigma^2)$
- $\hat\varepsilon = y - \hat y \sim N(0, (I - H)\sigma^2)$

- Regression line:

$$\boxed{E[Y \mid X = x]} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

var(g)

The linear Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$Z = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad p = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \hat\beta = (Z^T Z)^{-1} Z^T y$$

We have:

$$\hat\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar x) y_i}{\sum_{i=1}^n (x_i - \bar x)^2} = \frac{\overbrace{\frac{1}{n}\sum_{i=1}^n (x_i - \bar x) y_i}^{S_{xy}}}{\underbrace{\frac{1}{n}\sum_{i=1}^n (x_i - \bar x)^2}_{S_{xx}}}$$

and

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x$$

$$Var(\hat\beta_1) = \frac{\sigma^2}{S_{xx}}$$

$$Var(\hat\beta_0) = \dots = \frac{\sigma^2}{n}\frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{S_{xx}}$$

$$Var(\hat\beta_0 + \hat\beta_1 x) = \dots = \frac{\sigma^2}{n}\left(1 + \left[\sqrt{n}\frac{x - \bar x}{\sqrt{S_{xx}}}\right]^2\right)$$

$$\beta_0 + \beta_1 x \in \left(\hat\beta_0 + \hat\beta_1 x \pm t^{1-\alpha/2}_{n-2}\, s\sqrt{\frac{1}{n} + \frac{(x - \bar x)^2}{S_{xx}}}\right) \text{ w.p. } 1-\alpha$$

$$y_{n+1} \in \left(\hat\beta_0 + \hat\beta_1 x_{n+1} \pm t^{1-\alpha/2}_{n-2}\cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar x)^2}{S_{xx}}}\right) \text{ w.p. } 1-\alpha$$

we get a band if we evaluate this for every $x_{n+1} \in \mathbb{R}$

In general:

$$z_0^T\beta \in \left(z_0^T\hat\beta \pm t^{1-\frac{\alpha}{2}}_{n-p}\cdot s\sqrt{z_0^T(Z^T Z)^{-1} z_0}\right) \text{ w.p. } 1-\alpha \qquad z_0 \in \mathbb{R}^p$$

$$s^2 = \frac{1}{n-p}\sum_{i=1}^n (y_i - \hat y_i)^2 = \frac{\|\varepsilon\|^2}{n-p}$$

- We can take avg. of m measurements at the same x_{n+1}. In this case:

$$\bar y_{n+1} \in \left(\hat\beta_0 + \hat\beta_1 x_{n+1} \pm t^{1-\alpha/2}_{n-2}\cdot s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1} - \bar x)^2}{S_{xx}}}\right) \text{ w.p. } 1-\alpha$$

Simultaneous Bands

- Contain $(\beta_0+\beta_1 x)_{x\in\mathbb{R}}$ with prob. $1-\alpha$
- In p dim, contain $(z_0^T\beta)_{z\in\mathbb{R}^p}$
- From the distribution of $\hat\beta$:

$$P_r\left((\hat\beta-\beta)^T(Z^TZ)^{-1}(\hat\beta-\beta)\le s^2 p\cdot F^{1-\alpha}_{p,n-p}\right)=1-\alpha$$

This define an ellipsoid in $\mathbb{R}^p$

Working Hotelling Bands:

Confidence: $\hat\beta_0+\hat\beta_1 x \pm \sqrt{2F^{1-\alpha}_{2,n-2}}\cdot s\sqrt{\frac{1}{n}+\frac{(x-\bar x)^2}{S_{xx}}}$

Prediction: $\hat\beta_0+\hat\beta_1 x \pm \sqrt{2F^{1-\alpha}_{2,n-2}}\cdot s\sqrt{1+\frac{1}{n}+\frac{(x-\bar x)^2}{S_{xx}}}$

Avg. of m predictions: $\hat\beta_0+\hat\beta_1 x \pm \sqrt{2F^{1-\alpha}_{2,n-2}}\cdot s\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(x-\bar x)^2}{S_{xx}}}$

(these are wider than earlier-mentioned bands)

Competing Variables:

- $\hat\beta_1$ is significant if x_2 is not in the model
- $\hat\beta_2$ is significant if x_1 is not in the model.

Collaborating Variables:

- $\hat\beta_1$ is sig. if x_2 is in the model
- $\hat\beta_2$ is sig if x_1 is in the model

Def. Partial Correlation $\rho_{ij|k}$

Partial correlation of x_i x_j, adjusting for x_k is the correlation of resid. for x_i on x_k and resid. for x_j on x_k

For Gaussian data:

$$\rho_{ij|k}=\frac{\rho_{ij}-\rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}$$

(with or without hats)

$$\hat\rho_{ij}=\frac{x_i^Tx_j}{\|x_i\|\,\|x_j\|}\qquad \rho_{ij}=\frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

- **Forward approach**: we start at ϕ and add the best predictor if the new model is statistically significant; we stop otherwise (using F-test for extra sum of squares)

Backward approach: we start with all predictors, and drop the least significant / the one that leaves you with maximal SS_{res} if the old model is not statistically significantly better than the new one; otherwise, stop

Cross Validation in linear models

- The residuals:

$$y_i-\hat y_i^{(i)}=\frac{y_i-\hat y_i}{1-H_{ii}}=\frac{\hat\varepsilon_i}{1-H_{ii}}$$

$H_{ii}\le 1$

- Overall:

$$CV=\sum_i^n\left(y_i-\hat y_i^{(i)}\right)^2=\sum_{i=1}^n\frac{\hat\varepsilon_i^2}{(1-H_{ii})^2}$$

$$\hat y^{(i)}=\frac{\hat y_i-H_{ii}y_i}{1-H_{ii}}$$

Least trimmed regression (80% smallest residuals example):

$$\hat\beta=\arg\min_\beta\sum_{i=1}^{\lfloor 0.8n\rfloor}|\hat\varepsilon_{(i)}(\beta)|^2$$

The model is robust if less than 20% are outliers

Bias: the prediction error resulting from miss-specifying the model

Variance: the prediction error resulting from variations in the data used for fitting

Penalty on number of variables:

$$AIC:=n\cdot\log\left(\frac{SS_{res}}{n}\right)+2p$$

$$BIC=n\log\frac{SS_{res}}{n}+p\log n$$

Regularization : Ridge regression (L2) was created originally to handle the case when Z.TZ is (nearly)

$$\tilde\beta_\lambda=(Z^TZ+\lambda I)^{-1}Z^Ty\qquad \lambda>0$$

Variation : if we don't want to shrink the intercept:

$$\tilde\beta_\lambda=\left(Z^TZ+\lambda\begin{pmatrix}0&\\&I_{p-1}\end{pmatrix}\right)^{-1}Z^Ty\ ,\quad\lambda>0$$

$$\ell(\beta;y,Z,\lambda)=\|y-Z\beta\|^2+\lambda\|\beta\|^2$$

Advantages :

- More accurate with predictions
- More stable when predictors are correlated

Disadvantages: gives non zero values to all predictors

L1 Lasso :

$$\ell(\beta;y,Z,\lambda)=\|y-Z\beta\|^2+\lambda\|\beta\|_1$$

$$\|\beta\|_1=\sum_{j=1}^p|\beta_j|$$

- If we regress y on Z

$$\hat\beta_{OLS}=(Z^TZ)^{-1}Z^T(Z\beta+\tilde z\cdot\tilde\beta)$$
$$=\hat\beta+(Z^TZ)^{-1}Z^T\tilde z\cdot\tilde\beta$$

- Suppose $E[Y]=Z\beta+\tilde z\cdot\tilde\beta$, $\tilde z\in\mathbb{R}^n$, $\tilde\beta\in\mathbb{R}$

- Conditions for CLT:
  1) $\lambda_{min}(Z^TZ)\xrightarrow{p}\infty$, Z^TZ is a matrix version of n
  2) No z_{ij} is "too large"
  3) ε_i is not heavy tailed

Detection QQ plot :

1) Sort $\hat\varepsilon_i$ so that $\hat\varepsilon_{(1)}\le\hat\varepsilon_{(2)}\dots\le\hat\varepsilon_{(n)}$
2) Plot $\hat\varepsilon_{(i)}$ vs. $\Phi\left(\frac{i-\frac{1}{2}}{n}\right)$

Detecting Heteroscadasticity:

- If we know V, we can use generalized LS:

$$V=P^T\Lambda P\qquad P^TP=PP^T=I$$
$$\Lambda=diag(\lambda_1,\dots,\lambda_n)$$

let: $D=\Lambda^{-\frac{1}{2}}P$

- we have:

$$\underbrace{Dy}_{\tilde y}=D(Z\beta+\varepsilon)=\underbrace{DZ}_{\tilde Z}\beta+\underbrace{D\varepsilon}_{\tilde\varepsilon}$$

$$\boxed{\tilde y=\tilde Z\beta+\tilde\varepsilon}$$

Where: $Var(\tilde\varepsilon)=D\underbrace{Var(\varepsilon)}_{V}D^T$
$=DP^T\Lambda PD$
$=\Lambda^{-\frac{1}{2}}\underbrace{PP^T}_{I}\Lambda\underbrace{PP^T}_{I}\Lambda^{-\frac{1}{2}}=I$

- we take:

$$\hat\beta_{GLS}=(\tilde Z^T\tilde Z)^{-1}\tilde Z^T\tilde y$$

so that $\hat\beta_{GLS}$ minimizes $\|Dy-DZ\beta\|^2$

- If $V=diag(\sigma_1^2,\dots,\sigma_n^2)$ then we can use

$$\hat\sigma_i^2=\hat u_i^2\ ,\quad \hat u_i=y_i-z_i^T\hat\beta_{OLS}$$

We can also compute the correlation between the residuals:

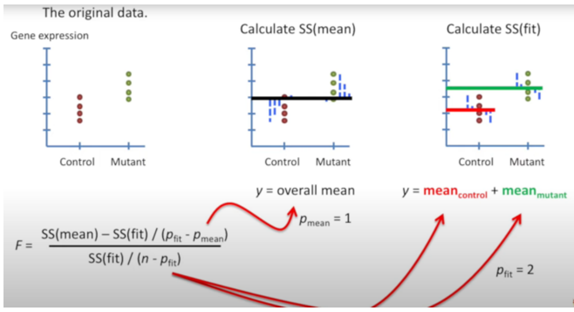
$$\hat\rho_k=\frac{\frac{1}{n}\sum_{i=k+1}^n\hat\varepsilon_i\hat\varepsilon_{i-k}}{\frac{1}{n}\sum_{i=1}^n\hat\varepsilon_i^2}$$

- The p-value:

$$p=\frac{\#\text{ of permutations with }\bar Y_1-\bar Y_0\ge|\bar Y_1-\bar Y_0|_{\text{observed}}}{\binom{n_0+n_1}{n_0}}$$

- if the number of permutations is too large, consider a random sample of N permutations.

$$p=\frac{1+(\#\text{ of sampled perm. with }\bar Y_1-\bar Y_0\ge\text{observed}|\bar Y_1-\bar Y_0|)}{N+1}$$



2. ANOVA table for $k=2$:

Source	DoF	SS	MS	F
Groups	$k-1=1$	$SS_{between}=n_0(\bar y_{0\cdot}-\bar y_{\cdot\cdot})^2+n_1(\bar y_{1\cdot}-\bar y_{\cdot\cdot})^2$	$MS_{between}=\frac{SS_{between}}{k-1}$	$\frac{MS_{between}}{MS_{within}}$
Error	$n-k=n_0+n_1-2$	$SS_{within}=\sum_{j=1}^{n_0}(y_{0j}-\bar y_{0\cdot})^2+\sum_{j=1}^{n_1}(y_{1j}-\bar y_{1\cdot})^2$	$MS_{within}=\frac{SS_{within}}{n-k}$	
Total	$n-1=n_0+n_1-1$	$SS_{tot}=\sum_{j=1}^{n_0}(y_{0j}-\bar y_{\cdot\cdot})^2+\sum_{j=1}^{n_1}(y_{1j}-\bar y_{\cdot\cdot})^2$		

Increasing the number of measurements improves the accuracy in estimating the LS coefficients β_j using $\hat{\beta}_j$ in the sense that the confidence interval for β_j (centered at $\hat{\beta}_j$) gets smaller with n (this is the formal way of saying that the accuracy is increased). Connecting this back to test of significance against $\beta_j = 0$: if 0 is not in the $1 - \alpha$ confidence interval of β_j , then the t-test P-value must be smaller than α.
The ability to handle singular design matrices is one of the motivations of using ridge regression.
We assume that the variations of all samples around their group mean are normally distributed, independent, and of equal variance. Under the null distribution H0 we have $t \sim t_{n-3}$. We reject H0 at significance level $\alpha = 0.05$ if |t| exceeds $t^{0.975}_{n-3}$.

$n = n_0 + n_1$

$$SS\ between = \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = n_0 (\bar{y}_{0\cdot} - \bar{y}_{\cdot\cdot})^2 + n_1 (\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$SSwithin = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{j=1}^{n_0} (y_{0j} - \bar{y}_{0\cdot})^2 + \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{1\cdot})^2$$

$$SStot = \sum_{i=1}^{k}\sum_{j=1}^{n_j} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{j=1}^{n_0} (y_{0j} - \bar{y}_{\cdot\cdot})^2 + \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{\cdot\cdot})^2$$

Answer: (i): We set H_0 as the hypothesis that $y_{new} \sim \mathcal{N}(\mu, \sigma^2)$. Under H_0,

$$y_{new} - \bar{y} \sim \mathcal{N}(0, \sigma^2(1 + 1/n)),$$

so

$$t = \frac{y_{new} - \bar{y}}{s\sqrt{1 + 1/n}} \sim t_{n-1}, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Our test rejects H_0 at significance level $\alpha = 0.05$ if $|t| > t_{n-1}^{0.975}$

(another acceptable answer is to incorporate y_{new} into s^2. However, this may inflate s^2 under the alternative hence result in a test of lesser power)

(ii) $y_{new} \in \bar{y} \pm s\sqrt{1 + 1/n} \cdot t_{n-1}^{.975}$

(iii) The best option is to use the two-sample t-test with $n_1 = n$ and $n_2 = 2$. In this case,

$$t = \frac{\bar{y}_{new} - \bar{y}}{s\sqrt{\frac{1}{2} + \frac{1}{n}}}, \qquad s^2 = \frac{1}{n+2-2}\left(\sum_{t=1}^{n}(y_i - \bar{y})^2 + (y_{new1} - \bar{y}_{new})^2 + (y_{new2} - \bar{y}_{new})^2\right)$$

We reject the null if $|t|$ exceeds $t_n^{0.975}$.

Another option is to use $\bar{y}_{new} = (y_{new1} + y_{new2})/2$ in a procedure similar to (i). With this option, under the null we have

$$\bar{y}_{new} - \bar{y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{2} + \frac{1}{n}\right)\right),$$

so

$$t = \frac{\bar{y}_{new} - \bar{y}}{s\sqrt{\frac{1}{2} + \frac{1}{n}}}, \qquad s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_i - \bar{y})^2.$$

We reject the null if $|t|$ exceeds $t_{n-1}^{0.975}$.

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}))^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_{i\cdot})^2 + 2(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2)$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 + \underbrace{2\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}_{*} + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$* = 2 \cdot \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})\right)$$
$$= 2 \cdot \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\left(\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \bar{y}_{i\cdot}\right)$$
$$= 2 \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\left(\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\right)$$
$$= 2 \sum_{i=1}^{k} (y_{i\cdot} - \bar{y}_{\cdot\cdot})\left(\sum_{j=1}^{n_i} y_{ij} - \frac{n_i}{n_i}\sum_{j=1}^{n_i} y_{ij}\right)$$
$$= 2 \sum_{i=1}^{k} (y_{i\cdot} - \bar{y}_{\cdot\cdot})(0)$$

$$Pr(\min p_i \leq \alpha/n \mid H_0 = true)$$
$$= 1 - Pr(\min p_i > \tfrac{\alpha}{n} \mid H_0 \text{ is true})$$
$$= 1 - Pr(p_1 > \tfrac{\alpha}{n}, p_2 > \tfrac{\alpha}{n} \ldots p_n > \tfrac{\alpha}{n} \mid H_0 = true)$$
$$= 1 - \prod_{i=1}^{n} Pr(p_i > \tfrac{\alpha}{n} \mid H_0 \text{ is true})$$
$$= 1 - (1 - \tfrac{\alpha}{n})^n .$$

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population. Although type I and type II errors can never be avoided entirely, the investigator can reduce their likelihood by increasing the sample size

$$1 - \alpha = Pr\left(-t_{n-2}^{1-\alpha/2} \leq \frac{y_{n+1} - (\hat{\beta}_0 + \hat{\beta}_1 x_{n+1})}{s\sqrt{\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}} \leq t_{n-2}^{1-\alpha/2}\right)$$

If R2 is closed to 1, the predictor explains the response well. Increasing the number of predictors never decreases R2.

```
n = len(y_t)
p = 3
q = 1
SSful = np.sum(np.square(y_t - np.dot(beta_hat1, Z1.T)))
SSsub = np.sum(np.square(y_t - np.dot(beta_hat2, Z2.T)))
F = (1/(p-q) * (SSsub - SSful))/(1/(n-p) * SSful)

pvalue = stats.f.sf(F, 2, n-p)
print(pvalue)
```

9.631240818725029e-11

Since the pvalue is really small, we can say that we reject the null hypothesis that says that the two models are similar. Therefore, we conclude that the fitted model significantly improves the trivial model.

The two sample t test quantifies the difference between the arithmetic means of the two months over the years. The p-value quantifies the probability of observing as or more extreme values assuming that the rainfall doesn't change over the years.The paired t test measures whether the average score differs significantly across samples (years). If we observe a large p-value then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, then we reject the null hypothesis of equal averages. Therefore, from previous questions, the two sample ttest would be better since we don't want to generalize that if we reject the null hypothesis for one year, then the rainfall do change over the years.

```
for i in wineries1:
    mean1.append(np.mean(wine_df[wine_df.winery == i]["points"]))
    n1s.append(len(wine_df[wine_df.winery == i]["points"]))

for i in wineries2:
    mean2.append(np.mean(wine_df[wine_df.winery == i]["points"]))
    n2s.append(len(wine_df[wine_df.winery == i]["points"]))

n1s = 1/np.array(n1s)
n2s = 1/np.array(n2s)
```

```
wineries_sets = ["Bazelet HaGolan", "Gamla", "Golan Heights Winery", "Katlav", "P
```

```
def ssquares(x):
    return np.sum((x - np.mean(x)) ** 2)

ss_wit = wine_df.groupby('winery')[variable].agg(ssquares).sum()

n = len(wine_df)
k = len(wineries_sets)

MSE_wit = ss_wit / (n-k)
s = np.sqrt(MSE_wit)
```

```
up = (np.sum(mean1)/len(mean1)) - (np.sum(mean2)/len(mean2))
low = np.sqrt((1/len(mean1)**2) * np.sum(n1s) + (1/len(mean2)**2) * np.sum(n2s))
t = up/(s * low)
print(t)
```

In this example we want to ask whether the number of fireplaces affects positively on the price of a house, so that we know to build some in order to increase the value of ours. However, what if the number of fireplaces is merely a function of the number of rooms which responsible to the increase. In this case, adding additional fireplaces would not affect the price (becasue we did not changed the number of rooms). To account for the effect of fireplaces, we can adjust for the number of rooms.

```
import statsmodels.api as sm

varX = 'Fireplaces'
varY = 'SalePrice'
varZ = 'TotRmsAbvGrd' # total rooms above ground level

model_LotFrontageYearBuilt = smf.ols(formula= f"{varX} ~ {varZ}", data=data2).fit()
model_SalePriceYearBuilt = smf.ols(formula= f"{varY} ~ {varZ}", data=data2).fit()

X = sm.add_constant(model_LotFrontageYearBuilt.resid)
y = model_SalePriceYearBuilt.resid
model_res = sm.OLS(y, X).fit()
plt.scatter(model_LotFrontageYearBuilt.resid, model_SalePriceYearBuilt.resid)
plt.title(fr"Regressing residuls. Partial Correlation ({varX},{varY}), adjusting for {varZ} is $R^2 = {m
```