# Problem 1

calculate the power of a chi-squared

$X_1, \ldots, X_n \sim N(\mu, \sigma)$

$H_0: \sigma \leq 1$

$H_A: \sigma > 1$

$$\frac{1}{n-1}\sum_{i=1}^{2}(X_i - \bar{X})^2 S^2 > \frac{1}{n-1}\chi^2_{\alpha, n-1}$$

$\Rightarrow$ estimate the power of this test if $\sigma = 2$, $n = 10$, $\alpha = .05$

$\Rightarrow$ generate $(m \times n)$ array of $X =$ normal random variable with sd $m_u$, $\sigma = 2$

$\Rightarrow$ calculate critical region

$$C = \frac{1}{n-1} \text{chi2.ppf}(1-a, n-1)$$

⟹ for eg row in X,

calculate $S^2$ np.var(X, axis=1, ddof=1)

S2 =

np.mean(S2 > C)

Problem 3

$H_0: \mu \leq 2.5$

$H_A: \mu > 2.5$

↳ calculating an upper confidence interval and checking if 2.5 ∈ interval

↳ in interval, fail to reject

↳ not in, reject

upper confidence:

(lower bound, ∞)

lower confidence

(-∞, upper bound)

$S^2$ = ccrcy of sample variances

$\left(\frac{1}{n-1}\right)\chi^2_{.05,9}$

$V$

3.0

| | |
|---|---|
| 4.7 | Tre |
| 3.2 | T |
| 2.8 | F |
| 1.7 | F |
| 2.9 | F |
| 3.4. | T |

$S_2 > 3.0$

---

Confidence interval for the expected value

$y|x$

$y_i = \alpha + \beta x_i + \varepsilon_i$

$\hat{\alpha}, \hat{\beta}$ → best guesses

confidence $\hat{\alpha} + \hat{\beta} x_i$

prediction interval accounts for the variation amongst individuals at $x_i$

bootstrap where we resample the pairs and then predict

- $\hat{\alpha} + \hat{\beta} x_i$ → confidence for $\alpha + \beta x_i$

prediction interval given $x_0$

① generate $m$ bootstrap samples $\hat{\alpha}_j, \hat{\beta}_j$ for $j = 1, \dots, m$

② generate a prediction

$$\hat{y}_j = \hat{\alpha}_j + \hat{\beta}_j \cdot x_0 + \hat{\varepsilon}_{0j}$$

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{\alpha}_j - \hat{\beta}_j x_{ij}$$

↳ for each of these $j$ samples, I randomly choose one of these $\hat{\varepsilon}_{ij}$, and I add in

$$\mathrm{Var}(X) = E[(X - \mu_x)^2]$$

$$\mathrm{Cov}(X, Y) = E[(X - \mu_x) \cdot (Y - \mu_y)]$$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{}$$

$SD(X)\,SD(Y)$

$(X_1, \ldots, X_k)$

Covariance matrix $\Sigma$

$k \times k$ matrix

$$\Sigma_{ii} = Var(X_i) = \sigma_i^2$$

$$\Sigma_{ji} = \Sigma_{ij} = Cov(X_i, X_j) = \rho_{ij}\,\sigma_i\,\sigma_j$$

$\rho_{ij} = corr(X_i, X_j)$

$\cdot\, SD(X_i)\cdot SD(X_j)$

---

Multiple regression

$\Longrightarrow$ $y = \alpha + \beta x_i$ $\quad (x_1, y_1), \ldots, (x_n, y_n)$

what if we have more than 1 predictor?

— d predictors

$(x_{11}, x_{12}, \ldots, x_{1d}, y_1)$, $(x_{21}, x_{22}, \ldots, x_{2d}, y_2)$,

$\ldots (x_{n1}, x_{n2}, \ldots, x_{nd}, y_n)$

$$Y_i = \beta_1 X_{i1} + \dots + \beta_d X_{id} + \varepsilon_i$$

$X_{i1} = 1$ for every $i = 1, \dots, N$ $\quad \varepsilon_i \sim N(0, \sigma)$

$$= \beta_1 X_{i1} + \dots + \beta_d X_{id} + \varepsilon_i \quad (1)$$

$X$ is $n \times d$ matrix

$Y$ is a $n \times 1$ column vector

$\beta$ is a $d \times 1$ column vector

$\varepsilon$ is a $n \times 1$ colm vector

$$Y = \underline{\underline{X}}\beta + \varepsilon$$ ← design matrix

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ X_{21} & & & \\ \vdots & & & \\ X_{n1} & & \cdots & X_{nd} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

$$+ \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n}$$

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left( Y_i - \beta_1 X_{i1} - \ldots - \beta_d X_{id} \right)^2$$

$$\underset{\beta}{\text{minimize}} \left( Y - X\beta \right)^T \left( Y - X\beta \right)$$

$$\frac{d}{d\beta} \left( Y - X\beta \right)^T \left( Y - X\beta \right)$$

$$2 X^T (Y - X\beta) = 0$$

$$X^T Y = X^T X \beta$$

$$\hat{\beta} = (X^T X)^{-1} X^T \beta$$

$X^T X \Rightarrow$ has to be invertible

column of $X$ need to be linearly independent

$\hat{\beta}$ is a random variable

d-dimensional normal random variable

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

↳ this can be problematic if there are columns $X_i$ and $X_j$ that are highly correlated

columns of $X$ have standardized so

$$X^T X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\hat{\beta} \sim N(\beta, \sigma^2 I)$$

$$X^T X = \begin{pmatrix} 1 & .95 \\ .95 & 1 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{1 - .95^2} \begin{pmatrix} 1 & -.95 \\ -.95 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -.95 \\ \end{pmatrix}$$

$\approx 11$ $\begin{pmatrix} & = \\ -.95 & 1 \end{pmatrix}$

$X^T X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\quad y_i = 0.5\, x_{i1} + 0.5\, x_{i2} + \varepsilon_i$

$\hat{\beta}_1 = 0.55 \qquad \hat{\beta}_2 = 0.49$

$X^T X = \begin{pmatrix} 1 & .95 \\ .95 & 1 \end{pmatrix}$ $\quad \hat{\beta}_1 = 3.1 \qquad \hat{\beta}_2 = -1.8$

↳ tends to really create issues with out-of-sample predictive

d is relatively low → spurious correlations are going to arise between the variables

idea: penalize the magnitude of the regression coefficients

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left( y_i - \beta_1 x_{i1} - \dots - \beta_d x_{id} \right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2$$

$\lambda$ is what we call a "hyperparameter"

↳ not a parameter that determines a distin

$$\min_{\beta} \; (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

↳ for this to make sense, we need to standardize the cols of $X \Rightarrow$ all have unit $\sum_{i=1}^{n} x_{ij}^2 = 1$

$$\hat{\beta}_\lambda = (X^T X + \lambda \underline{\underline{I}}_p)^{-1} X^T Y$$

$$X^T X = \begin{pmatrix} 1 & .95 \\ .95 & 1 \end{pmatrix}$$

$$(X^T X)^{-1} \approx \frac{1}{=} \begin{pmatrix} 1 & -.95 \\ -.95 & 1 \end{pmatrix}$$

$\lambda = 1$

$$(X^T X + I) = \begin{pmatrix} 2 & .95 \\ .95 & 2 \end{pmatrix}$$

$$(X^T X + I)^{-1} = \frac{1}{3.1} \begin{pmatrix} 2 & -.95 \\ -.95 & 2 \end{pmatrix}$$

reduced version of $\beta_i$ by about a factor of 15

What is the downside of this?

$$\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T Y$$

$$(X^T X + \lambda I)\hat{\beta}_\lambda = X^T Y$$

$$(X^T X)^{-1}(X^T X + \lambda I)\hat{\beta}_\lambda = (X^T X)^{-1} X^T Y = \hat{\beta}_0$$

$$(I + (X^T X)^{-1})\hat{\beta}_\lambda = \hat{\beta}_0$$

$$\hat{\beta}_\lambda = (I + (X^T X)^{-1})^{-1}\hat{\beta}_0 .$$

↳ this is a biased estimator of $\beta$

How do we pick $\lambda$?

↳ trying to evaluate which $\lambda$ will generalize best out-of-sample