STATS 305A Notes

Kenneth Tay

Contents

1 Linear Least Squares: General Case (Chapters 4, 12)

Set-up

The model is $Y = Z\beta + \varepsilon$, with $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$. Here, $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$.

- If Z_i is the **column** vector representing subject i, then we can write $Z^T Z = \sum_{i=1}^{n} Z_i Z_i^T$.

Assumptions

- The data being used in fitting the model is representative of the population.
- The true underlying relationship between Y and Z is linear.
- $\mathbb{E}[\varepsilon_i \mid Z_i] = 0$.
- $\mathbb{E}Z\varepsilon = 0$ (i.e. errors uncorrelated with predictors).
- Errors/residuals are independent of each other.
- The variance of the residuals are constant (homoscedastic).

Fitting the model

- **Normal equation:** $Z^T(Y - Z\hat{\beta}) = 0$.
- (Sec 4.1) **Estimate for coefficients:** $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$. $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (Z^T Z)^{-1}\right)$.
- (Sec 4.2) **Hat matrix** $H = Z(Z^T Z)^{-1} Z^T$. H is symmetric and idempotent, and $\text{tr}(H) = p$. H is a projection onto the column span of Z.
- (Sec 4.3) **Predicted values** $\hat{Y} \sim \mathcal{N}(Z\beta, H\sigma^2)$. Var $\hat{Y}_i = H_{ii}\sigma^2$, $\sum \hat{Y}_i = p\sigma^2$.
- (Sec 4.3) **Residuals** $\hat{\varepsilon} \sim \mathcal{N}(0, (I - H)\sigma^2)$, and $\hat{\varepsilon}$ is independent of $\hat{\beta}$ and $\hat{Y}$.
- (Sec 4.3.1) **Residual sum of squares** $RSS = \|Y - Z\hat{\beta}\|^2 = \|\hat{\varepsilon}\|^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \sim \sigma^2 \chi^2_{n-p}$.
- (Sec 4.4) **Covariance estimate:** Let $s^2 = \frac{1}{n-p}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-p}\|Y - X\hat{\beta}\|^2$. Then $s^2 \sim \frac{\sigma^2 \chi^2_{n-p}}{n-p}$. $\mathbb{E}[s^2] = \sigma^2$.
- (Sec 4.4) For fixed $c \in \mathbb{R}^{1\times p}$ and s^2 as before, $\frac{c\hat{\beta} - c\beta}{s\sqrt{c(Z^T Z)^{-1} c^T}} \sim t_{n-p}$.
- (Sec 4.7) **Gauss-Markov Theorem:** Let $Y \sim (Z\beta, \sigma^2 I)$ (not necessarily normal), and assume $p < n$. If $a \in \mathbb{R}^n$ is such that $\mathbb{E}[a^T Y] = c\beta$, then Var $(a^T Y) \geq V(c\hat{\beta})$.

- **Orthogonal predictors:** Say $Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, with X_1 orthogonal to X_2. Then the regression estimates for the β's in this model would be the same as the estimates obtained when the model only includes one of the regressors.

  More generally, if all the X_i's are orthogonal to each other, then their coefficient estimates don't depend on each other.

- (Sec 4.8) **Computation using SVD:** Computational cost of SVD for an $n \times p$ matrix is $O(\min(n^2 p, np^2))$.

  Do SVD decomposition for Z: $Z_{n\times p} = U_{n\times n}\Sigma_{n\times p}V_{p\times p}^T$, where U and V are orthogonal, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$, where $k = \min(n, p)$ and $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$.

  Note that we can do the skinny version of SVD:

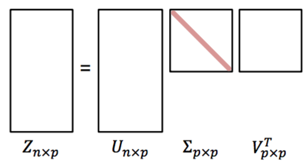

Figure 4.6: The skinny singular value decomposition.

  We can also remove components where $\sigma_i = 0$ in Σ (leaving say, k non-zero σ_i's):

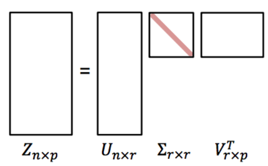

Figure 4.7: An even skinnier singular value decomposition.

  We then have

$$\begin{aligned}
\|Y - Z\beta\|^2 &= \|Y - U\Sigma V^T\beta\|^2 = \|U^T Y - \Sigma V^T \beta\|^2 \\
&= \|Y^* - \Sigma\beta^*\|^2 \qquad\qquad (\text{where } Y^* = U^T Y,\ \beta^* = V^T\beta) \\
&= \sum_{i=1}^{r} (y_i^* - \sigma_i \beta_i^*)^2.
\end{aligned}$$

  We can minimize this easily: $\beta_i^* = y_i^*/\sigma_i$ for $i = 1, \ldots, r$, β_i^* can be anything for $i > r$.

Goodness of fit/Comparing models

- (Sec 4.6) When comparing a full model and a submodel (i.e. a subset of features/columns), we have

$$F = \frac{(RSS_{SUB} - RSS_{FULL})/(p-q)}{RSS_{FULL}/(n-p)} \sim F_{p-q,n-p}.$$

  Null hypothesis H_0: submodel is true. (Intuitively, if difference is big, the submodel does a much worse job of fitting, so we might not trust it.)

- (Sec 4.9) **ANOVA Decomposition:**

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

total sum of squares (SST) = sum of squares of errors (SSE) + sum of "explained" (SSZ)

- (Sec 4.9) $R^2 = \dfrac{SSZ}{SST} = 1 - \dfrac{SSE}{SST}$ is the percentage of variance explained by the model.

  For one predictor, $R^2 = \dfrac{SXY^2}{SXX \cdot SYY} = \hat{\rho}^2$.

- (Sec 4.9) **Adjusted R^2:**

$$\bar{R}^2 := 1 - \frac{\frac{1}{df}\sum(y_i-\hat{y}_i)^2}{\frac{1}{n-1}\sum(y_i-\bar{y})^2},$$

  where $df = n-p$ if model includes intercept, $df = n-p-1$ if the model doesn't include intercept. (For ordinary R^2, $df = n-1$.) Adjusted R^2 is always smaller than ordinary R^2. It is possible for it to be negative (decent sign of overfitting).

- **Orthogonalization trick:** Say we have the model $Y = X_{-j}\beta_{-j} + X_j\beta_j + \varepsilon$, but we are only interested in the coefficient for X_j. We can orthogonalize: Write $X_j = X_{j,-j} + X_{-j}\gamma$, where $X_{j,-j}$ and X_{-j} are orthogonal (i.e. $X_{j,-j}$ is X_j with all the other columns regressed out).

  We can then rewrite the model as $Y = X_{-j}\tilde{\beta}_{-j} + X_{j,-j}\beta_j + \varepsilon$. Thus, $\hat{\beta}_j = \dfrac{X_{j,-j}^T Y}{\|X_{j,-j}\|^2}$, and $\text{Var}\,\hat{\beta}_j = \dfrac{\sigma^2}{\|X_{j,-j}\|^2}$. Reported standard error would be $\widehat{\text{Var}}\beta_j = \dfrac{\hat{\sigma}^2}{\|X_{j,-j}\|^2}$.

- The t-statistic for β_j is given by $t_j = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X^TX)^{-1}_{jj}}} \sim t_{n-p}$.

- See ESL p52-55 for interpretation of the β_j's.

2 Simple Regression: $Y = \beta_0 + \beta_1 X$ (Chapter 9)

Dealing with the case of 1 predictor variable $X_i \in \mathbb{R}$ and response $Y_i \in \mathbb{R}$. If

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right),$$

then the regression line is

$$Y = \mu_Y + \rho\frac{\sigma_X}{\sigma_Y}(X - \mu_X).$$

For some derivations, see Weisberg Appendix A3 (p293).

- (Sec 9.2) $\hat{\beta}_1 = \dfrac{SXY}{SXX} = \dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \dfrac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$, and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. (For simple regression through the origin, we have $\hat{\beta}_1 = \dfrac{\sum x_i y_i}{\sum x_i^2}$.)

- (Sec 9.3.1) $\mathbb{E}\hat{\beta}_1 = \beta_1$, Var $\hat{\beta}_1 = \dfrac{\sigma^2}{SXX} = \dfrac{\sigma^2}{n\left[\frac{1}{n}\sum(x_i - \bar{x})^2\right]}$.

- (Sec 9.3.2) $\mathbb{E}\hat{\beta}_0 = \beta_0$, Var $\hat{\beta}_0 = \dfrac{\sigma^2 \sum x^2}{n^2 SXX} = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{SXX}\right)$.

- (Weisberg p295) $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \dfrac{\bar{x}}{SXX}$.

- (Sec 9.3.3) Var $(\hat{\beta}_0 + \hat{\beta}_1 x) = \sigma^2\left[\dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{SXX}\right]$.

- (Sec 9.3.3) **Confidence intervals:** We have $\dfrac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 - \beta_1 x)}{s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}} \sim t_{n-2}$, so a confidence interval for $\hat{\beta}_0 + \hat{\beta}_1 x$ is $\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2}^{1-\alpha/2} \cdot s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$. More generally, for $Y = Z\beta + \varepsilon$ where $Z_0 \in \mathbb{R}^{1\times p}$, the confidence band is $Z_0\beta \in Z_0\hat{\beta} \pm t_{n-p}^{1-\alpha/2} \cdot s\sqrt{Z_0(Z^TZ)^{-1}Z_0^T}$.

- (Sec 9.3.3) **Prediction bands:** Prediction bands make a strong assumption of normality. For a single prediction at new x_{n+1}, Var $(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} + \varepsilon_{n+1}) =$ Var $(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) + \sigma^2$.

  If we take the average of m new y's at x_{n+1}, associated t-statistic associated for prediction interval is $\dfrac{\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} - (\beta_0 - \beta_1 x_{n+1}) - \bar{\varepsilon}}{s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{SXX}}} \sim t_{n-2}$. This gives the confidence interval $\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm t_{n-2}^{1-\alpha/2} \cdot s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{SXX}}$.

- (Sec 9.4) **Simultaneous bands:** Contains $\beta_0 + \beta_1 x$ at all $x \in \mathbb{R}$ with probability $1-\alpha$. Basically they are the confidence intervals above, with the $t_{n-2}^{1-\alpha/2}$ term replaced with $\sqrt{2F_{2,n-2}^{1-\alpha}}$. (In p-dimensions, $t_{n-p}^{1-\alpha/2}$ is replaced with $\sqrt{pF_{p,n-p}^{1-\alpha}}$.) (For derivation, see Theory Session 9.)

- (Sec 9.6) For any regression, $R^2 = 1 - \dfrac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$. For the case of simple regression, $R^2 = \dfrac{SXY^2}{SXX \cdot SYY} = \hat{\rho}^2$.

- (HW3) $\hat{\sigma}^2 = \dfrac{1}{n-2}\left(SYY - \dfrac{SXY^2}{SXX}\right)$.

3 Regression through the Origin

Model: $Y_i = \beta X_i + \varepsilon_i$.

- $\hat{\beta} = \displaystyle\sum_{i=1}^{n} \frac{x_i y_i}{x_i^2}$.

- Var $\hat{\beta} = \dfrac{\sigma^2}{\sum x_i^2}$ if σ^2 known. If σ^2 unknown, use the plug-in estimator $\hat{\sigma}^2 = \dfrac{\sum \hat{u}_i^2}{n-1}$.

4 One-Group Model: $Y_i = \mu + \varepsilon_i$ (Chapters 5-6)

- (Sec 5.1) **Unbiased estimate for variance** σ^2 of the Y_i's: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. When $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, then $s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$.

  In general, if kurtosis exists, then Var $s^2 = \sigma^4 \left(\frac{2}{n-1} + \frac{\kappa}{n}\right)$.

- (Sec 5.3.1) **1-Sample t-test:** Assume $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$. Test $\mu = \mu_0$ using t-statistic $t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$. Under the null distribution, $t \sim t_{n-1}$.

- (Sec 5.4) p-value $p \doteq e^{-k \times n}$, where n is sample size and k depends on μ, μ_0 and σ.

- **Bootstrap t confidence intervals:** Draw $Y_1^{*b}, \ldots Y_n^{*b} \overset{iid}{\sim} \hat{F}$, i.e. the empirical CDF. Compute $t^{*b} = \frac{(\bar{Y}^{*b} - \bar{Y})}{s^{*b}/\sqrt{n}}$. Do this B times.

  Then $P(L \le t \le U) \approx P(L \le t^* \le U) \approx \frac{1}{B}\#\{t^{*b} : L \le^{*b} \le U\}$.

- (Sec 6.2) **Power for the standard t-test** is related to the non-central F-distribution:

$$\text{Power} = 1 - \beta = P\left(F'_{1,n-1}\left(n\left(\frac{\mu - \mu_0}{\sigma}\right)^2\right) > F^{1-\alpha}_{1,n-1}\right).$$

- (Sec 6.2) $\Delta = \frac{\mu - \mu_0}{\sigma}$ is called the **effect size**. α increasing, n increasing or Δ increasing leads to power increasing.

- (Sec 6.3) If $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ and $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, then $\mathbb{E}[s^2] = \sigma^2$ and Var $s^2 = \frac{2\sigma^4}{n-1}$. From this, can get confidence interval for σ^2:

$$P\left(\frac{s^2}{\left(\chi^2_{n-1}\right)_{1-\frac{\alpha}{2}}/(n-1)} \le \sigma^2 \le \frac{s^2}{\left(\chi^2_{n-1}\right)_{\frac{\alpha}{2}}/(n-1)}\right) = \alpha.$$

5 Two-Sample Tests (Chapter 7)

Setup: n_0 observations from Group 0, n_1 observations from Group 1. Let X be the group an observation is from. Model is $\mathbb{E}[Y \mid X = 0] = \beta_0$, $\mathbb{E}[Y \mid X = 1] = \beta_0 + \beta_1$ or β_1 (different parametrizations). In the rest of this section, we assume $\mathbb{E}[Y \mid X = 1] = \beta_0 + \beta_1$.

- (Sec 7.1) **t-statistic** for $\hat{\beta}_1$ is $t = \frac{\hat{\beta}_1 - 0}{s\sqrt{(Z^T Z)^{-1}_{22}}} = \frac{\bar{Y}_1 - \bar{Y}_0}{s\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$, which has t_{n-2} distribution if $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$.

- (Sec 7.1) Estimate for s^2: $s^2 = \dfrac{1}{n-2}\left[\sum_{i:X_i=0}(Y_i-\bar{Y}_0)^2 + \sum_{i:X_i=1}(Y_i-\bar{Y}_1)^2\right]$. This gives $\mathbb{E}[s^2] = \dfrac{1}{n-2}[(n_0-1)\sigma_0^2 + (n_1-1)\sigma_1^2]$.

- (Sec 7.2) **Welch's t:** statistic $t' = \dfrac{\bar{Y}_1 - \bar{Y}_0 - \Delta}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}}$, where $s_j^2 = \dfrac{1}{n_j - 1}\sum_{j=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$.

  $t' \to \mathcal{N}(0,1)$ if $\mu_1 - \mu_0 = \Delta$ and $\min n_j \to \infty$. For small samples, appropriate degrees of freedom lies between $\min n_j - 1$ and $n_0 + n_1 - 2$.

- (Sec 7.3) **Permutation Test:** Say $Y_i \overset{ind}{\sim} F_j$ for $j = 0, 1$ and $i = 1, \ldots, n_j$. Null hypothesis is $H_0 : F_0 = F_1$.

  Pool the n observations together, randomly pick n_0 and assign to group 0, rest assign to group 1. Test statistic
  $$p = \frac{\text{no. of permutations with } (\bar{Y}_1 - \bar{Y}_0) \geq |\text{observed} \bar{Y}_1 - \bar{Y}_0|}{\binom{n_0+n_1}{n_0}}.$$

  Asymptotically, permutation test approaches the two-sample t-test.

- **Two-sample bootstrap:** Let the data be $(0, Y_{01}), \ldots (0, Y_{0n_0}), (1, Y_{11}), \ldots (1, Y_{1n_1})$. There are at least 3 methods to do the bootstrap:

  - Independently sample n_0 and n_1 observations from empirical CDFs $\hat{F}_0$ and $\hat{F}_1$ respectively. Compute $T(\hat{F}_1^*) - T(\hat{F}_0^*)$. Do this B times, draw histogram.
  - Compute $t'^{*b} = \dfrac{\bar{Y}_1^{*b} - \bar{Y}_0^{*b}}{\sqrt{\frac{s_1^{*b2}}{n_1} + \frac{s_0^{*b2}}{n_0}}}$ B times. Draw histogram.
  - Put all the data together. Resample $n_0 + n_1$ rows from the data. Compute $T(\hat{F}_1^*) - T(\hat{F}_0^*)$. Do this B times, draw histogram.

6 k Groups (Chapter 8)

Let there be n_j observations in group j, $N = n_1 + \cdots + n_k$.

Cell means model: $\beta = \begin{bmatrix}\mu_1 & \ldots & \mu_k\end{bmatrix}^T$.

- (Sec 8.2) For testing $C\beta = 0$, we have $\dfrac{\frac{1}{r}(\hat{\beta} - \beta)^T C^T [C(Z^T Z)C^T]^{-1} C(\hat{\beta} - \beta)}{s^2} \sim F_{r,N-k}$, where $r = \text{rank}(C)$.

- (Sec 8.2) Alternative to the above: H_0: group means all equal. Let SS_{SUB} be the sum of squared errors under the common mean model, SS_{FULL} be the sum of squared errors under the individual group means model. Then
  $$\frac{\frac{1}{k-1}(SS_{SUB} - SS_{FULL})}{\frac{1}{n-k}SS_{FULL}} \sim F_{k-1,N-k}.$$

- (Sec 8.2) **ANOVA identity:**
  $$\sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2.$$

- (Sec 8.4) **General contrast:**

$$\frac{\sum_{i=1}^k \lambda_i \bar{Y}_{i\cdot}}{s\sqrt{\sum_{i=1}^k \frac{\lambda_i^2}{n_i}}} \sim t_{N-k},$$

  where $\sum \lambda_i = 0$, $\sum \lambda_i^2 \neq 0$, an s^2 is the pooled variance estimate, i.e. $\dfrac{SSE}{N-k}$.

- (Sec 8.4.2) We can compute (with some algebra)

$$t^2 = \frac{\left(\sum_{i=1}^k \lambda_i \bar{Y}_{i\cdot}\right)^2}{s^2 \sum_{i=1}^k \frac{\lambda_i^2}{n_i}} \sim F'_{1,N-k}\left(\frac{\left(\sum_i \lambda_i \alpha_i\right)^2}{\sigma^2 \sum_i \frac{\lambda_i^2}{n_i}}\right).$$

  The larger the non-centrality parameter, the more power we have. Thus, the most powerful contrast is the one s.t. $\lambda \propto \alpha$.

- (Sec 8.5) 2 contrasts $C_1 = \sum \lambda_i \bar{Y}_i$ and $C_2 = \sum \eta_i \bar{Y}_i$ are orthogonal if $\text{Cov}(C_1, C_2) = \sigma \sum \dfrac{\lambda_i \eta_i}{n_i} = 0$.

- **Effects model:** $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, with the constraint that $\sum \alpha_i = 0$. Assume that $i = 1, 2, \ldots, I$, and that there are n_i observations from group i. Then we have the estimates

$$\hat{\mu} = \frac{1}{I}\sum_{i=1}^I \bar{Y}_{i\cdot}, \qquad \hat{\alpha}_i = \bar{Y}_{i\cdot} - \frac{1}{I}\sum_{i=1}^I \bar{Y}_{i\cdot}.$$

  If $n_1 = \cdots = n_I = n$, then the above simplify to $\hat{\mu} = \bar{Y}_{\cdot\cdot}$, $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$.

- In the effects model, the variance estimate for group i is the usual $s_i^2 = \dfrac{1}{n_i - 1}\displaystyle\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$. The overall (pooled) variance estimate for within groups is $\hat{\sigma}^2 = \displaystyle\sum_{i=1}^I \frac{(n_i - 1)s_i^2}{n_1 + \cdots + n_I - I}$.

7 Random Effects (Chapter 11)

11.1-11.2: Single random effects model

- **Model:** $Y_{ij} = \mu + a_i + \varepsilon_{ij}$, with $a_i \sim \mathcal{N}(0, \sigma_A^2)$, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_E^2)$.

- $\text{Corr}(Y_{ij}, Y_{i'j'}) = \dfrac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} 1\{i = i'\}$.

- **ANOVA identity:**

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2}_{SSE/SS_{within}} + \underbrace{\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}_{SSA/SS_{between}}.$$

- In this case, $SSE \sim \sigma_E^2 \chi_{N-k}^2$. If $n = n_i$ (all groups same size), then $\dfrac{1}{n}SSA \sim \left(\sigma_A^2 + \dfrac{\sigma_E^2}{n}\right)\chi_{k-1}^2$, and

$$F = \frac{\frac{1}{k-1}SSA}{\frac{1}{N-k}SSE} \sim \left(1 + \frac{n\sigma_A^2}{\sigma_E^2}\right)F_{k-1,N-k}.$$

- To estimate effect of group a_i, we have

$$\tilde{a_i} := \mathbb{E}[a_i \mid Y_{ij}, i=1,\dots,k, j=1,\dots,n_i] = \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}(\bar{Y}_{i\cdot} - \mu),$$

$$\mu + \tilde{a_i} = \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}\bar{Y}_{i\cdot} + \left(1 - \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}\right)\mu.$$

11.3.1: Fixed $\times$ fixed model

- **Model:** $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, where $k = 1,\dots,n_{ij}$, $i = 1,\dots,I$, $j = 1,\dots,J$. Assume $n_{ij} = n$ for all i, j.
- **ANOVA decomposition:**

$$\sum_{i,j,k}(Y_{ijk} - \bar{Y}_{\cdots})^2 = \underbrace{\sum_{i,j,k}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2}_{SSA} + \underbrace{\sum_{i,j,k}(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots})^2}_{SSB} + \underbrace{\sum_{i,j,k}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2}_{SSAB} + \underbrace{\sum_{i,j,k}(Y_{ijk} - \bar{Y}_{ij\cdot})^2}_{SSE}.$$

The 4 terms are also called sum of row variance, column variance, $SS_{between}$ and SS_{within}.

- **Distributions:**

$$\begin{aligned} SSA &\sim \sigma^2\chi'^2_{I-1}\left(\frac{nJ\sum_i \alpha_i^2}{\sigma^2}\right), \\ SSB &\sim \sigma^2\chi'^2_{J-1}\left(\frac{nI\sum_j \beta_j^2}{\sigma^2}\right), \\ SSAB &\sim \sigma^2\chi'^2_{(I-1)(J-1)}\left(\frac{n\sum_{i,j}(\alpha\beta)_{ij}^2}{\sigma^2}\right), \\ SSE &\sim \sigma^2\chi^2_{IJ(n-1)}. \end{aligned}$$

- **Testing:**

$$\begin{aligned} &H_0: \alpha_i = 0, && F_A = \frac{MSA}{MSE} \sim F_{(I-1),IJ(n-1)}, \\ &H_0: \beta_j = 0, && F_B = \frac{MSB}{MSE} \sim F_{(J-1),IJ(n-1)}, \\ &H_0: (\alpha\beta)_{ij} = 0, && F_{AB} = \frac{MSAB}{MSE} \sim F_{(I-1)(J-1),IJ(n-1)}. \end{aligned}$$

11.3.2: Random $\times$ random model

- **Model:** $Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$, where $k = 1,\dots,n$, $i = 1,\dots,I$, $j = 1,\dots,J$. Distributional assumptions: $a_i \sim \mathcal{N}(0,\sigma_A^2)$, $b_j \sim \mathcal{N}(0,\sigma_B^2)$, $(ab)_{ij} \sim \mathcal{N}(0,\sigma_{AB}^2)$, $\varepsilon_{ijk} \sim \mathcal{N}(0,\sigma_E^2)$, all distributions independent.
- **Distributions:**

$$\begin{aligned} SSA &\sim (\sigma_E^2 + n\sigma_{AB}^2 + nJ\sigma_A^2)\chi^2_{I-1}, \\ SSB &\sim (\sigma_E^2 + n\sigma_{AB}^2 + nI\sigma_B^2)\chi^2_{J-1}, \\ SSAB &\sim (\sigma_E^2 + n\sigma_{AB}^2)\chi^2_{(I-1)(J-1)}, \\ SSE &\sim \sigma_E^2\chi^2_{IJ(n-1)}. \end{aligned}$$

- Proper test for $H_0 : \sigma_A = 0$: $F_A = \dfrac{MSA}{MSAB} \sim \left(1 + \dfrac{nJ\sigma_A^2}{\sigma_E^2 + n\sigma_{AB}^2}\right) F_{(I-1),(I-1)(J-1)}$.

11.3.3: Random × fixed model (mixed effects)

- **Model:** $Y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk}$, where $k = 1, \ldots, n$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Assumptions: $\sum \alpha_i = 0$, $b_j \sim \mathcal{N}(0, \sigma_B^2)$, $(\alpha b)_{ij} \sim \mathcal{N}(0, \sigma_{AB}^2)$, $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_E^2)$, all distributions independent. Additional constraint: $\displaystyle\sum_{i=1}^{I} (\alpha b)_{ij} = 0$ for all j with probability 1.
- **Distributions:**

$$\begin{aligned}
SSA &\sim (\sigma_E^2 + n\sigma_{AB}^2)\chi_{I-1}'^2 \left(\frac{nJ\sum_i \alpha_i^2}{\sigma_E^2 + n\sigma_{AB}^2}\right), \\
SSB &\sim (\sigma_E^2 + nI\sigma_B^2)\chi_{J-1}^2, \\
SSAB &\sim (\sigma_E^2 + n\sigma_{AB}^2)\chi_{(I-1)(J-1)}^2, \\
SSE &\sim \sigma_E^2 \chi_{IJ(n-1)}^2.
\end{aligned}$$

- **Testing:** To test A, use $\dfrac{MSA}{MSAB}$. To test B, use $\dfrac{MSB}{MSE}$. To test AB, use $\dfrac{MSAB}{MSE}$. All have (potentially non-central) F distributions.

8 Interplay between Variables (Chapter 13)

- **Simpson's paradox:** See diagram in notes.
- **Competition:** $\hat{\beta}_2$ is significant if X_1 not in model, and vice versa. Occurs when X_1 and X_2 have high correlation.
- **Collaboration:** $\hat{\beta}_2$ is significant if X_2 is in the model, and vice versa.
- **Partial correlation** of X_i, X_j adjusting for X_k is Corr(residual for X_i vs. X_k, residual for X_j vs. X_k). Sometimes written as $\rho_{ij|k}$.
- For Gaussian population, we have $\rho_{ij|k} = \dfrac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}$.

9 Automatic Variable Selection (Chapter 14)

- (Sec 14.1) **Forward stepwise:** Start with $\emptyset$. Add the best predictor if it is statistically significant, otherwise stop.
- (Sec 14.1) **Backward stepwise:** Start with all predictors. Drop the least significant predictor if it is not statistically significant, otherwise stop.
- (Sec 14.2) **Mallow's C_p:** Say we have q regressors. An unbiased estimate for expected squared error $ESE = \mathbb{E}\left[\displaystyle\sum_{i=1}^{n} (\hat{Y}_i - \mathbb{E}[Y_i])^2\right]$ is $RSS - (n-2p)\hat{\sigma}^2$, where $\hat{\sigma}^2 = \dfrac{1}{n-q}\displaystyle\sum_i (Y_i - \hat{Y}_i)^2$.

(This is different from that in 300C. There, Mallow's C_p is an unbiased estimate for prediction risk, which is $ESE + n\sigma^2$.)

- (Sec 14.3) Fit the model without observation i to obtain regression coefficients $\hat{\beta}_{-i}$. Let $\hat{Y}_{-i}$ be the predictor of Y_i when we fit the model without observation i, i.e. $\hat{Y}_{-i} = Z_i^T \hat{\beta}_{-i}$. **Cross validation of a model** is defined as $\text{CV(model)} = \sum_{i=1}^n (Y_i - \hat{Y}_{-i})^2$.

- (Sec 14.3) **For linear models**, $\hat{Y}_i = H_{ii} Y_i + (1 - H_{ii}) \hat{Y}_{-i}$, where $H = Z(Z^T Z)^{-1} Z^T$ is the hat matrix. (Proof in Sec 14.3.2.) Hence,

$$\hat{Y}_{-i} = \frac{\hat{Y}_i - H_{ii} Y_i}{1 - H_{ii}}, \quad \text{so residual} \quad Y_i - \hat{Y}_{-i} = \frac{Y_i - \hat{Y}_i}{1 - H_{ii}} \text{ and } CV = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}} \right)^2 = \sum_{i=1}^n \frac{\hat{\varepsilon_i}^2}{(1 - H_{ii})^2}.$$

- (Sec 14.4) **Generalized CV:** Actual a special case of CV with $H_{ii} = \frac{p}{n}$ for all i.

- (Sec 14.5) **Akaike's Information Criterion** For a model with p regressors, $AIC = 2p - 2 \log \hat{L}$, where $\hat{L}$ is the log-likelihood of the data under the best model with p regressors.

- (Sec 14.5) **Bayes Information Criterion:** For a model with p regressors and sample size n, $BIC = p \log n - 2 \log \hat{L}$.

- (Sec 14.5) AIC is better for prediction, BIC is better at getting the "right" model. Asymptotically, $AIC \approx CV$.

- **Elastic net:** Minimize $\sum_{i=1}^n (Y_i - Z_i^T \beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$.

- (Sec 14.7) **Principal Components Regression:** Suppose $Z_i \in \mathbb{R}^d$. PC regression attempts to choose dimensions with the highest variance for Z, i.e. maximize Var $(Z^T U)$ subject to $U^T U = 1$.
  - This is the same thing as $\text{argmax}_{\|u\|=1} (\|Zu\|^2) = \text{argmax}_{\|u\|=1} (u^T Z^T Z u)$.
  - To find the k^{th} principal component, we can first subtract the first $k-1$ principal components from Z: $\hat{Z}_k = Z - \sum_{s=1}^{k-1} Z w_{(s)} w_{(s)}^T$, then do the same as for the first component, i.e. $\text{argmax}_{\|u\|=1} (\|\hat{Z}_k u\|^2) = \text{argmax}_{\|u\|=1} (u^T \hat{Z}_k^T \hat{Z}_k u)$.
  - Alternatively, we can do SVD: $Z = U \Sigma W^T$. The columns of W are the principal components. (This only works if X is centered!)

Ridge Regression

- (Sec 14.6) **Ridge regression:** Good for nearly singular $Z^T Z$ matrices. Estimate for β is $\tilde{\beta} = (Z^T Z + \lambda I)^{-1} Z^T Y$, $\lambda > 0$.

- (Sec 14.6) Ridge regression is the solution to the minimization problem $\text{argmin}_\beta \sum_{i=1}^n (Y_i - Z_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2$.

- (Sec 14.6) Bayesian connection: If $Y \sim \mathcal{N}(Z\beta, \sigma^2 I_n)$, prior $\beta \sim \mathcal{N}(0, \tau^2 I_p)$, then maximizing the posterior density of β is the same as solving the ridge regression problem with $\lambda = \dfrac{\sigma^2}{\tau^2}$.

- (Sec 14.6) If we don't want to put a penalty on the intercept, we can center the data (Z and Y, then do a regression through the origin to get $\tilde{\beta}$ and use $\bar{Y}$ as the intercept.

- (Sec 14.6) **Calculation for ridge regression:** Append data: $Y^* \in \mathbb{R}^{n+p}$ such that Y^* is Y with p zeros appended below. New design matrix $Z^* = \begin{bmatrix} Z \\ \sqrt{\lambda} I_p \end{bmatrix}$. Then solve $\min \|Y^* - Z^*\beta\|^2$ by SVD.

- If we have SVD decomposition $Z = UDV^T$, $D = \text{diag}(d_1, \ldots, d_p)$, then we can rewrite

$$\begin{aligned} \tilde{\beta}_\lambda &= V \text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) U^T Y, \\ \tilde{Y}_\lambda &= U \text{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right) U^T Y = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda} U_j U_j^T Y. \end{aligned}$$

(For OLS, we have $\hat{Y} = \sum_{j=1}^p U_j U_j^T$ instead.) Hence, **effective degrees of freedom** for ridge regression is $\text{tr}[X(X^TX + \lambda I)^{-1} X^T] = \sum_{j=1}^p \dfrac{d_j^2}{d_j^2 + \lambda}$.

- With the formulas above, we can see that SVD only has to be done once, even when fitting multiple values of λ.

- Sacrifices unbiasedness for reduced variance.

- Implicit assumption of ridge regression: The response will tend to vary most in the directions of high variance of the inputs.

LASSO

- Find β which minimizes $\sum_{i=1}^n (Y_i - Z_i^T\beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$.

- LASSO can get coefficient estimates exactly equal to 0 (while ridge regression usually won't).

- Sacrifices unbiasedness for reduced variance.

10 Violations of Assumptions (Chapter 16)

Bias/Lack of fit

- **Detection:**
  - Plots: Plot against other variables that you have but were not in the model, etc.

- Say there are n_i observations with $X = X_i$. We can test our model $Y_{ij} = Z_i^T\beta + \varepsilon_{ij}$ against $Y_{ij} = \mu(X_i) + \varepsilon_{ij}$, where $\mu(X_i)$ is the mean of the responses at $X = X_i$.

$$\text{pure error sum of squares } = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2,$$

$$\text{lack of fit sum of squares } = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - Z_i^T\hat{\beta})^2,$$

$$\text{test statistc } F = \frac{\frac{1}{N-p}(\text{lack of fit sum of squares})}{\frac{1}{\sum(n_i-1)}(\text{pure error sum of squares})}.$$

- **Transformations:**
  - **Cobb-Douglas model:** $Y = \beta_0 X_1^{\beta_1}\dots X_p^{\beta_p}(1+\varepsilon)$. Then take logs and fit.
  - **Box-Cox transformations:** Let $\tilde{Y}(\lambda) = \dfrac{Y^\lambda - 1}{\lambda}$ for $\lambda \neq 0$, $= \log Y$ when $\lambda = 0$. Fit $\tilde{Y}(\lambda) = Z\beta + \varepsilon$.

Heteroskedasticity

Suppose $Y \sim \mathcal{N}(Z\beta, \sigma^2 V)$, where V is full rank and not necessarily I.

- **Detection:**
  - Plot $\hat{\varepsilon}_i$ vs. X_i. When there are many X's, we can plot $\hat{\varepsilon}_i$ vs. $\hat{Y}_i$.
  - Plot $\hat{\varepsilon}_i$ vs. $\hat{\varepsilon}_{i-1}$ to see if there is dependence in errors.
  - Compute autocorrelations at lag k: $\hat{\rho}_k = \dfrac{\frac{1}{n}\sum_{i=k+1}^n \hat{\varepsilon}_i\hat{\varepsilon}_{i-k}}{\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2}$.
- **Correction:**
  - In this case, $\hat{\beta}$ is still an unbiased estimate for β, but Var $\hat{\beta}$ will not be correct, hence giving wrong CIs and p-values.
  - Can deal with this with **generalized least squares**. Basically, whiten the noise:
    * Let $V = P^T\Lambda P$, where P orthogonal and Λ diagonal. Let $D = \Lambda^{-1/2}P$, so that $D^TD = V^{-1}$.
    * Multiply our model on both sides: $DY = DZ\beta + D\varepsilon$, and rewrite $\tilde{Y} = DY$, $\tilde{Z} = DZ$ and $\tilde{\varepsilon} = D\varepsilon$. Then we are back in the usual OLS case ($\tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$).

Non-normality

- **Detection:** Use QQ-plots of residuals vs. normal quantiles. Instead of using residuals $\hat{\varepsilon}_i$, we can also use $\dfrac{\hat{\varepsilon}_i}{s\sqrt{1-H_{ii}}}$ or $\dfrac{\hat{\varepsilon}_i}{s_{-i}}$.
- **Correction:** Usually this is not an issue as the CLT will correct it for us. 3 conditions for the CLT to take effect:
  1. Need the eigenvalues of Z^TZ to go to ∞.
  2. No Z_{ij} can be too large.
  3. ε_i cannot be too heavy-tailed.

Outliers

- **Detection:**
  - Large $|\hat{\varepsilon}_i|$ is a sign.
  - Can also look at the **leave-one-out residual** $\frac{|\hat{\varepsilon}_{-i}|}{s_{-i}}$.
- **Masking:** 2 outliers make each other look like non-outliers.
- **Swamping:** Outliers make good data points look bad.
- **Correction:**
  - One possibility is to remove it from the data (but we should have a good reason to do so!)
  - Use robust regression methods (e.g. minimize absolute sums of errors, least trimmed means, least median of squares).

11 Bootstrapped Regressions (Chapter 17)

Bootstrapped pairs

- Resample pairs of data, drawing (X^{*b}, Y^{*b}), and estimate $\hat{\beta}^{*b} = (Z^{*bT}Z^{*b})^{-1}Z^{*bT}Y^{*b}$ repeatedly. We can then estimate $\text{Cov}(\hat{\beta})$ by using the sample covariance of $\hat{\beta}^*$.
- This method is especially good for t-statistics.
- This method corrects for unequal variance across observations.
- This method can break if $Z^{*T}Z^*$ is singular, but as long as this is an uncommon occurence, this method is fine.

Bootstrapped residuals

- Fit the data, obtain residuals $\hat{\varepsilon}_i$. Resample ε_i^{*b}'s from the $\hat{\varepsilon}_i$'s, then take $Y_i^{*b} = Z_i^T\beta + \varepsilon_i^{*b}$ and $\hat{\beta}^{*b} = (Z^TZ)^{-1}Z^TY^{*b}$.
- This method always uses Z^TZ, so we don't have to worry about singular Z^TZ.
- This method is good because the X_i's are fixed.
- This method wires in the assumption that the ε_i are i.i.d., and especially that they have equal variance. Hence, it does not correct for unequal variance.

Wild bootstrap

- Model $Y_i^{*b} = Z_i^T\beta + \varepsilon_i^{*b}$, with ε_i^{*b}'s independent, $\mathbb{E}[\varepsilon_i^{*b}] = 0$ and Var $\varepsilon_i^{*b} = \hat{\varepsilon}_i^2$.
- Variation: Model as above, but with ε_i^{*b}'s independent, $\varepsilon_i^{*b} = a_i$ w.p. p_i, $= b_i$ w.p. $1 - p_i$ such that $\mathbb{E}[\varepsilon_i^{*b}] = 0$, Var $\varepsilon_i^{*b} = \hat{\varepsilon}_i^2$ and $\mathbb{E}[\varepsilon_i^{*b3}] = \hat{\varepsilon}_i^3$.
- These models have fixed Z_i's and allow for unequal variances. However, this method is not good at dealing with lack of it.

Weighted likelihood bootstrap

- The typical MLE $\hat{\beta}$ puts equal weights $\frac{1}{n}$ on each of the n observations. We could put random multinomial weights on the observations.
- We could also reweight with exponentially distributed random variables: If $N_i^* \sim \text{Exp}(1)$, weights $W_i^* = \dfrac{N_i^*}{\sum_{k=1}^n N_k^*}$, then

$$\hat{\beta}^* = \left(\sum_{i=1}^n W_i^* Z_i Z_i^T\right)^{-1} \left(\sum_{i=1}^n W_i^* Z_i Y_i\right).$$

12 Instrumental Variables

Basic model: $Y = Z\beta + \varepsilon$, ε i.i.d., $\varepsilon_i \sim (0, \sigma^2)$. What if ε is correlated with Z?

This can happen if there are other variables affecting Y which are not included in Z. Consider the simple cases where we only have one regressor S. Assume all of the other variables affecting Y can be wrapped up in variable A.

- Model 1: $Y = \beta_0 + \beta_1 S + \varepsilon^{(1)}$.
- Model 2: $Y = \beta_0 + \beta_1 S + \beta_2 A + \varepsilon^{(2)}$.

We want the β_1 from model 2, not model 1. **Idea:** Suppose we have a variable W s.t. $\text{corr}(W, S) \neq 0$, but $\text{corr}(W, \varepsilon^{(1)}) = 0$ (or in other words, $\text{corr}(W, A) = 0$). Then we can perform the following:

1. Regress Y on W to get $\hat{\beta}_{Y \sim W}$.
2. Regress S on W to get $\hat{\beta}_{S \sim W}$.
3. Compute $\hat{\beta}_{IV} = \dfrac{\hat{\beta}_{Y \sim W}}{\hat{\beta}_{S \sim W}}$.

13 Extra from Weisberg

- **Marginal plot:** Plot Y against just one regressor X_i.
- **Added-variable plot:** Say we have Y against X_1 and we are thinking of adding another regressor X_2. The added-variable plot is the plot of the residuals from Y against X_1 against the residuals from X_2 against X_1. It shows the relationship of Y and X_2 adjusting for X_1.

14 Other Models

- (Sec 6.4.1) **Autoregressive AR(1) model:** $X_t = \delta + \phi_1 X_{t-1} + \varepsilon_t$, where $\varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $|\phi_1| < 1$.

  In this model, $\mathbb{E}X_t = \dfrac{\delta}{1 - \phi_1}$, $\text{Var } X_t = \dfrac{\sigma^2}{1 - \phi_1^2}$, and $\text{Corr}(X_i, X_j) := \rho_{ij} = \rho^{|i-j|}$

- **AR(p) model:** $X_t = \delta + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$, where $\varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and ϕ_i's are such that the roots of $z^p - \sum_{i=1}^p \phi_i z^{p-i}$ all have norm less than 1.

- (Sec 6.4.2) **Moving average MA(1) model** (Owen's version): $\rho_{ij} = 1$ if $i = j$, ρ if $|i - j| = 1$, 0 otherwise. $Y_i = U_i + \gamma U_{i-1}$.

  In the moving average model, $\mathbb{E}[\bar{Y}] = \mu$, Var $[\bar{Y}] = \frac{\sigma^2}{n}\left[1 + 2\rho\frac{n-1}{n}\right]$, $\mathbb{E}s^2 = \sigma^2\left(1 - \frac{2\rho}{n}\right)$, and $t = \sqrt{n}\frac{\bar{Y} - \mu}{s} \to \mathcal{N}(0, 1 + 2\rho)$.

- **MA(1) model:** $X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}$, where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

  In this model, $\mathbb{E}X_t = \mu$, Var $X_t = \sigma^2(1 + \theta_1)^2$, autocorrelation function is $\rho_1 = \frac{\theta_1}{1 + \theta_1^2}$, $\rho_h = 0$ for $h \geq 2$.

- **MA(q) model:** $X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$, where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

- **Autoregressive Moving Average ARMA(p, q) model:** A model with p autoregressive terms and q moving-average terms: $X_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$.

  This is a parsimonious description of a weakly stationary stochastic process in terms of two polynomials, one for the autoregression and one for the moving average.

  ARMA is appropriate when a system is a function of a series of unobserved shocks as well as its own behavior.

- **Autoregressive Integrated Moving Average (ARIMA) model:** A generalization of the ARMA model. ARIMA(p, d, q) means that the d^{th} order difference follows an ARMA(p, q) model.

- **Autoregressive Conditional Heteroskedasticity (ARCH) Model:** To model time-varying volatility. X_t is an ARCH(q) process if it is stationary and if $X_t = \sigma_t Z_t$, where $Z_t \sim \mathcal{N}(0, 1)$ and $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2$ with $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i \geq 1$.

  (X_t is usually the error term in a time series regression model.)

- **Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Model:** X_t is a GARCH(p, q) process if it is stationary and if $X_t = \sigma_t Z_t$, where $Z_t \sim \mathcal{N}(0, 1)$ and $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-h}^2$ with $\alpha_0 > 0$, $\alpha_i, \beta_j \geq 0$ for $i, j \geq 1$.