## Appendix A

1. **Derivation of $\hat{\beta}$ for ordinary least squares**

   Let b be an estimator of $\beta$

   $$(y - X_{reg}b)'(y - X_{reg}b) = y'y - 2b'X'_{reg}y + b'X'_{reg}X_{reg}b$$

   Differentiating by a and setting the derivative to 0 to derive the local minimum/maximum,

   $$\frac{d}{db}(y'y - 2b'X'_{reg}y + b'X'_{reg}X_{reg}b) = 0$$
   $$-2X'_{reg}y + 2X'_{reg}X_{reg}b = 0$$
   $$X'_{reg}X_{reg}b = X'_{reg}y$$

   $$\hat{\beta} = b = (X'_{reg}X_{reg})^{-1}X'_{reg}y \text{ (if the inverse of } X'_{reg}X_{reg} \text{ is unique)}$$

2. **Condition for unique $\hat{\beta}$**

   rank(X'X) = column rank of X→if X has full column rank, X'X is full rank and invertible and $\hat{\beta}$ is unique.

3. **Definitions of SST, SSR, SSE and $\hat{\sigma}^2$**

$$\begin{aligned}
\text{Total sum of squares (SST)} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i\bar{y} + \sum_{i=1}^{n}\bar{y}^2 \\
&= \sum_{i=1}^{n} y_i^2 - 2\bar{y}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\bar{y}^2 \\
&= \sum_{i=1}^{n} y_i^2 - 2\bar{y}n\bar{y} + n\bar{y}^2 \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\
&= y'y - n\bar{y}^2
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\epsilon}_i &= j'(y - \hat{y}) \\
&= [1 \quad 0 \quad 0 \quad \ldots \quad 0]\, X'_{reg}(y - X\hat{\beta}) \text{ since the first column of } X_{reg} \text{ is j} \\
&= [1 \quad 0 \quad 0 \quad \ldots \quad 0]\, X'_{reg}(y - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}y) \\
&= [1 \quad 0 \quad 0 \quad \ldots \quad 0]\, X'_{reg}(I - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})y \\
&= [1 \quad 0 \quad 0 \quad \ldots \quad 0]\, (X'_{reg} - X'_{reg}X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})y \\
&= 0
\end{aligned}$$

$$\hat{y} = y - \hat{\epsilon}$$

$$
\begin{aligned}
\bar{\hat{y}} &= \bar{y} - \bar{\hat{\epsilon}} \\
&= \bar{y} - \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i \\
&= \bar{y}
\end{aligned}
$$

$$
\begin{aligned}
\text{Regression sum of squares (SSR)} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}' x_{(i)} - \bar{\hat{y}})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}' x_{(i)})^2 - 2\sum_{i=1}^{n}\hat{\beta}' x_{(i)}\bar{\hat{y}} + \sum_{i=1}^{n}\bar{\hat{y}}^2 \\
&= (X_{reg}\hat{\beta})' X_{reg}\hat{\beta} - 2\bar{\hat{y}}\sum_{i=1}^{n}\hat{\beta}' x_{(i)} + \sum_{i=1}^{n}\bar{\hat{y}}^2 \\
&= \hat{\beta}' X'_{reg} X_{reg}\hat{\beta} - 2\bar{\hat{y}}n\bar{\hat{y}} + n\bar{\hat{y}}^2 \\
&= \hat{\beta}' X'_{reg} X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}y - n\bar{\hat{y}}^2 \\
&= \hat{\beta}' X'_{reg}y - n\bar{y}^2
\end{aligned}
$$

$$
\begin{aligned}
\text{Sum of squared error (SSE)} &= \sum_{i=1}^{n}(\hat{\epsilon}_i - \bar{\hat{\epsilon}})^2 \\
&= \sum_{i=1}^{n}\hat{\epsilon}_i^2 \; (\text{since} \sum_{i=1}^{n}\hat{\epsilon} = 0 \rightarrow \bar{\hat{\epsilon}} = 0) \\
&= (y - X_{reg}\hat{\beta})'(y - X_{reg}\hat{\beta}) \\
&= y'y - 2\hat{\beta}' X'_{reg}y + \hat{\beta}' X'_{reg}X_{reg}\hat{\beta} \\
&= y'y - 2\hat{\beta}' X'_{reg}y + \hat{\beta}' X'_{reg}X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}y \\
&= y'y - \hat{\beta}' X'_{reg}y = y'(I - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})y
\end{aligned}
$$

It can be verified that SST=SSR+SSE.

$$
\hat{\sigma}^2 = \frac{SSE}{n-p-1}
$$

## 4. Derivation of distribution of $\hat{\beta}$

$$
\begin{aligned}
\hat{cov}(\hat{\beta}) &= \hat{cov}((X'_{reg}X_{reg})^{-1}X'_{reg}y) \\
&= (X'_{reg}X_{reg})^{-1}X'_{reg}\hat{var}(y)((X'_{reg}X_{reg})^{-1}X'_{reg})' \\
&= (X'_{reg}X_{reg})^{-1}X'_{reg}\hat{var}(y)X_{reg}(X'_{reg}X_{reg})^{-1} \\
&= (X'_{reg}X_{reg})^{-1}X'_{reg}\hat{\sigma}^2X_{reg}(X'_{reg}X_{reg})^{-1} \\
&= \hat{\sigma}^2(X'_{reg}X_{reg})^{-1}
\end{aligned}
$$

$$\hat{\beta} \sim N(\beta, \sigma^2(X'_{reg}X_{reg})^{-1})$$

5. **Definition and some properties of an idempotent matrix**
   Idempotent
   A square matrix A is idempotent if it satisfies $A^2 = A, A' = A$.
   Properties
   rank(A)=trace(A)
   I-A is idempotent.
   (I-A)A=0 and A(I-A)=0
6. **Derivation of t-statistic**
   Let $P_A = A(A'A)^{-1}A$. $P_A$ is idempotent.

$$
\begin{aligned}
\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2} &= \frac{((I - X_{reg}\hat{\beta})y)'((I - X_{reg}\hat{\beta})y)}{\sigma^2} \\
&= \frac{y'(I - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})'(I - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})y}{\sigma^2} \\
&= \frac{y'(I - P_{X_{reg}})y}{\sigma^2} \text{ since } I - P_{X_{reg}} \text{ is idempotent} \\
&= \frac{y'}{\sigma}(I - P_{X_{reg}})\frac{y}{\sigma}
\end{aligned}
$$

**Theorem 1**

Let $y \sim N_n(\mu, \Sigma)$, let A be a $n \times n$ symmetric matrix. Then $y'Ay \sim \chi^2(rank(A), \frac{1}{2}\mu'A\mu)$ iff $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$ is idempotent.

Since $\frac{y}{\sigma} \sim N_n(\frac{\mu_y}{\sigma}, \frac{1}{\sigma}\sigma^2 I\frac{1}{\sigma} = I)$, $I - P_{X_{reg}}$ is symmetric and $I^{\frac{1}{2}}(I - P_{X_{reg}})I^{\frac{1}{2}}$ is idempotent,

(**Statement 1**)

$$
\begin{aligned}
\frac{y'}{\sigma}(I - P_{X_{reg}})\frac{y}{\sigma} &\sim \chi^2(rank(I - P_{X_{reg}}), \frac{1}{2}\frac{(X_{reg}\beta)'}{\sigma}(I - P_{X_{reg}})\frac{X_{reg}\beta}{\sigma}) \\
&= \chi^2(trace(I - P_{X_{reg}}), \frac{1}{2}\frac{(X_{reg}\beta)'}{\sigma}(X_{reg}\beta - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}X_{reg}\beta)\frac{1}{\sigma}) \text{ since } I - P_{X_{reg}} \text{ is idempotent} \\
&= \chi^2(n - trace(P_{X_{reg}}), 0) \\
&= \chi^2(n - rank(X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})) \text{ since } P_{X_{reg}} \text{ is idempotent} \\
&= \chi^2(n - rank((X'_{reg}X_{reg})^{-1}X'_{reg}X_{reg})) \\
&= \chi^2(n - (p+1)) \\
&= \chi^2(n - p - 1)
\end{aligned}
$$

$$\frac{I - P_{X_{reg}}}{\sigma^2} * ((X'_{reg}X_{reg})^{-1}X'_{reg})' = \frac{X_{reg}(X'_{reg}X_{reg})^{-1} - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}X_{reg}(X'_{reg}X_{reg})^{-1}}{\sigma^2}$$
$$= 0$$

Hence, $\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2}$ and $\hat{\beta}$ are independent.

Definition

If $y \sim N(\mu, 1), u \sim \chi^2(p),$ and y and u are independent, $t = \frac{y}{\sqrt{u/p}} \sim t(p, \mu)$

$$t = \frac{\frac{\hat{\beta}_k}{\sigma\sqrt{(X'_{reg}X_{reg})^{-1}_{kk}}}}{\sqrt{\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2}/(n-p-1)}}$$
$$= \frac{\frac{\hat{\beta}_k}{\sigma\sqrt{(X'_{reg}X_{reg})^{-1}_{kk}}}}{\sqrt{\frac{\sum_{i=1}^{n}(\hat{\epsilon}_i - \bar{\hat{\epsilon}})^2}{\sigma^2(n-p-1)}}} \text{ since } \bar{\hat{\epsilon}} = 0$$
$$= \frac{\frac{\hat{\beta}_k}{\sqrt{(X'_{reg}X_{reg})^{-1}_{kk}}}}{\sqrt{\frac{SSE}{n-p-1}}}$$
$$= \frac{\hat{\beta}_k}{\hat{\sigma}\sqrt{(X'_{reg}X_{reg})^{-1}_{kk}}}$$
$$= \frac{\hat{\beta}_k}{\hat{sd}(\hat{\beta}_k)} \sim t_{n-p-1, \frac{\beta_k}{\sigma\sqrt{(X'_{reg}X_{reg})^{-1}_{kk}}}}$$

t-test can be used to determined whether a predictor is significant in the model.

$$H_0 : \beta_k = 0 \text{ then } t = \frac{\hat{\beta}_k}{\hat{sd}(\hat{\beta}_k)} \sim t_{n-p-1}$$

$$H_1 : \beta_k \neq 0 \text{ then } t = \frac{\hat{\beta}_k}{\hat{sd}(\hat{\beta}_k)} \sim t_{n-p-1, \frac{\beta_k}{\sigma\sqrt{(X'_{reg}X_{reg})^{-1}_{kk}}}}$$
$$\text{or } t = \frac{\hat{\beta}_k - \beta_k}{\hat{sd}(\hat{\beta}_k)} \sim t_{n-p-1}$$

7. **Derivation of F-statistic**
   F-test is used to compare 2 models where a model uses a subset of predictors used in the other model
   WLOG, let model B be the model that uses a subset of predictors used in model A. Both models include the intercept term.
   Let $p_A$ be the no. of predictors used in model A (not counting intercept term),
   $p_B$ be the no. of predictors used in model B (not counting intercept term)

$$\begin{aligned} SSR_A - SSR_B &= \hat{\beta}'_A X'_A y - n\bar{y}^2 - (\hat{\beta}'_B X'_B y - n\bar{y}^2) \\ &= \hat{\beta}'_A X'_A y - \hat{\beta}'_B X'_B y \\ &= ((X'_A X_A)^{-1} X'_A y)' X'_A y - ((X'_B X_B)^{-1} X'_B y)' X'_B y \\ &= y'(X_A (X'_A X_A)^{-1} X'_A - X_B (X'_B X_B)^{-1} X'_B) y \\ &= y'(P_{X_A} - P_{X_B}) y \end{aligned}$$

$$(X_A(X'_A X_A)^{-1} X'_A) X_A = X_A$$

$$(X_A(X'_A X_A)^{-1} X'_A) \begin{bmatrix} X_B & X_{A-B} \end{bmatrix} = \begin{bmatrix} X_B & X_{A-B} \end{bmatrix}$$

$$(X_A(X'_A X_A)^{-1} X'_A) X_B = X_B$$

$$P_{X_A} X_B = X_B$$

$$P_{X_A} X_B (X'_B X_B)^{-1} X'_B = X_B (X'_B X_B)^{-1} X'_B$$

$$P_{X_A}P_{X_B} = P_{X_B}$$

$$\begin{aligned}(P_{X_A} - P_{X_B})(P_{X_A} - P_{X_B}) &= P_{X_A}^2 - P_{X_B}P_{X_A} - P_{X_A}P_{X_B} + P_{X_B}^2 \\ &= P_{X_A} - (P'_{X_A}P'_{X_B})' - P_{X_B} + P_{X_B} \text{ since } P_{X_A}, P_{X_B} \text{ are idempotent} \\ &= P_{X_A} - (P_{X_A}P_{X_B})' - P_{X_B} + P_{X_B} \\ &= P_{X_A} - P'_{X_B} - P_{X_B} + P_{X_B} \\ &= P_{X_A} - P_{X_B}\end{aligned}$$

$$\begin{aligned}(P_{X_A} - P_{X_B})' &= P'_{X_A} - P'_{X_B} \\ &= P_{X_A} - P_{X_B}\end{aligned}$$

Hence, $P_{X_A} - P_{X_B}$ is idempotent.

Since rank and trace of an idempotent matrix is equal,

$$\begin{aligned}rank(P_{X_A} - P_{X_B}) &= trace(P_{X_A} - P_{X_B}) \\ &= trace(P_{X_A}) - trace(P_{X_B}) \\ &= trace(X_A(X'_AX_A)^{-1}X'_A) - trace(X_B(X'_BX_B)^{-1}X'_B) \\ &= trace((X'_AX_A)^{-1}X'_AX_A) - trace((X'_BX_B)^{-1}X'_BX_B) \\ &= trace(I_{p_A+1}) - trace(I_{p_B+1}) \\ &= p_A + 1 - (p_B + 1) \\ &= p_A - p_B\end{aligned}$$

Since $\frac{y}{\sigma} \sim N_n(\frac{X_{A-B}\beta_{A-B}}{\sigma}, I)$, $P_{X_A} - P_{X_B}$ is symmetric & $I^{\frac{1}{2}}(P_{X_A} - P_{X_B})I^{\frac{1}{2}}$ is idempotent, according to theorem 1 under appendix A6,

$$\begin{aligned}\frac{y'}{\sigma}(P_{X_A} - P_{X_B})\frac{y}{\sigma} &\sim \chi^2(rank(P_{X_A} - P_{X_B}), \frac{1}{2}\frac{\mu'_y}{\sigma}(P_{X_A} - P_{X_B})\frac{\mu_y}{\sigma}) \\ &= \chi^2(p_A - p_B, \frac{(X_{A-B}\beta_{A-B})'(P_{X_A} - P_{X_B})(X_{A-B}\beta_{A-B})}{2\sigma^2}) \\ &= \chi^2(p_A - p_B, \frac{\beta'_{A-B}X'_{A-B}(X_A(X'_AX_A)^{-1}X'_A - X_B(X'_BX_B)^{-1}X'_B)(X_{A-B}\beta_{A-B})}{2\sigma^2}) \\ &= \chi^2(p_A - p_B, \frac{\beta'_{A-B}(X'_{A-B}X_A(X'_AX_A)^{-1}X'_A - X'_{A-B}X_B(X'_BX_B)^{-1}X'_B)X_{A-B}\beta_{A-B}}{2\sigma^2}) \\ &= \chi^2(p_A - p_B, \frac{\beta'_{A-B}(X'_{A-B} - X'_{A-B}X_B(X'_BX_B)^{-1}X'_B)X_{A-B}\beta_{A-B}}{2\sigma^2}) \\ &= \chi^2(p_A - p_B, \frac{\beta'_{A-B}(X'_{A-B}X_{A-B} - X'_{A-B}X_B(X'_BX_B)^{-1}X'_BX_{A-B})\beta_{A-B}}{2\sigma^2})\end{aligned}$$

$$\frac{y'(I - P_{X_A})y}{\sigma^2} \sim \chi^2(n - p_A - 1)$$

see statement 1 under appendix A6

$$\frac{I - P_{X_A}}{\sigma^2} * \frac{P_{X_A} - P_{X_B}}{\sigma^2} = \frac{0 - (I - P_{X_A})(P_{X_B})}{\sigma^4}$$
$$= 0$$

Hence, $\frac{\hat{\epsilon_A}'\hat{\epsilon_A}}{\sigma^2} = \frac{SSE_A}{\sigma^2}$ and $\frac{y'(P_{X_A} - P_{X_B})y}{\sigma^2}$ are independent.

Definition

If $u \sim \chi^2(p, \lambda), v \sim \chi^2(q)$, with u and v independent, then $F = \frac{u/p}{v/q} \sim F(p, q, \lambda)$

$h = p_A - p_B$,

$\beta_{A-B}$ represents the $\beta$ coefficients of predictors in model A but not in model B.

$X_{A-B}$ be the subset of columns (in the form of a matrix) of $X_A$ excluding the columns in $X_B$

$$F = \frac{\frac{y'(P_{X_A} - P_{X_B})y}{\sigma^2} / (p_A - p_B)}{\frac{y'(I - P_{X_A})y}{\sigma^2} / (n - p_A - 1)}$$
$$= \frac{\frac{SSR_A - SSR_B}{h}}{\frac{SSE_A}{n - p_A - 1}}$$

$$H_0 : \beta_{A-B} = 0 \text{ then } F \sim F_{h, n-p_A-1}$$

$$H_1 : \beta_{A-B} \neq 0 \text{ then } F \sim F_{h, n-p_A-1, \delta}$$

8. **F-test (case for full model vs intercept model)**

$$\begin{aligned}\hat{\epsilon}_{intercept} &= y - X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}y \\ &= y - j(j'j)^{-1}j'y \\ &= y - \frac{1}{n}jj'y \\ &= y - \frac{1}{n}j\sum_{i=1}^{n} y \\ &= y - \bar{y}j\end{aligned}$$

$$\begin{aligned}SSR_{intercept} &= SST_{intercept} - SSE_{intercept} \\ &= \sum_{i=1}^{n}(y - \bar{y})^2 - (y - \bar{y}j)'(y - \bar{y}j) \\ &= 0\end{aligned}$$

$$\begin{aligned}
F &= \frac{\frac{SSR_{full} - SSR_{intercept}}{(p+1)-1}}{\frac{SSE_{full}}{n-p-1}} \\
&= \frac{\frac{SSR_{full}}{p}}{\frac{SSE_{full}}{n-p-1}} \text{ since } SSR_{intercept} = 0 \\
&= \frac{\frac{SSR_{full}}{p}}{\hat{\sigma}^2_{full}}
\end{aligned}$$

9. **Derivation for partial correlation**

$x_j$ and $X_{-j}$ make up all the columns of $X_{reg}$.

$$\begin{aligned}
\text{Partial correlation} &= \hat{\rho}_{x_j y.X_{-j}} \\
&= \frac{\sum_{i=1}^{n}(\hat{\epsilon}_{x_j \text{ on } X_{-j},i} - \bar{\hat{\epsilon}}_{x_j})(\hat{\epsilon}_{y \text{ on } X_{-j},i} - \bar{\hat{\epsilon}}_{y \text{ on } X_{-j}})}{\sqrt{\sum_{i=1}^{n}(\hat{\epsilon}_{x_j \text{ on } X_{-j},i} - \bar{\hat{\epsilon}}_{x_j \text{ on } X_{-j}})^2}\sqrt{\sum_{i=1}^{n}(\hat{\epsilon}_{y \text{ on } X_{-j},i} - \bar{\hat{\epsilon}}_{y \text{ on } X_{-j}})^2}} \\
&= \frac{\sum_{i=1}^{n} \hat{\epsilon}_{x_j \text{ on } X_{-j},i}\hat{\epsilon}_{y \text{ on } X_{-j},i}}{\sqrt{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}\sqrt{\sum_{i=1}^{n} \hat{\epsilon}^2_{y \text{ on } X_{-j},i}}}
\end{aligned}$$

Equivalently (second definition),

$$\hat{\rho}_{x_j y.X_{-j}} = \frac{t_j}{\sqrt{t_j^2 + (n-p-1)}}$$

where $t_j$ is the t-value associated with $x_j$ when regressing y on $X_{reg}$.

Tedious proof of second definition

Schur complement (definition only)

$$\begin{aligned}
M^{-1} &= \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \\
&= \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \\
&= \begin{bmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{bmatrix} \text{ letting } M/D = A - BD^{-1}C
\end{aligned}$$

$$
\begin{aligned}
\frac{\sum_{i=1}^{n}\hat{\epsilon}_{x_j \text{ on } X_{-j},i}\hat{\epsilon}_{y \text{ on } X_{-j},i}}{\sqrt{\sum_{i=1}^{n}\hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}\sqrt{\sum_{i=1}^{n}\hat{\epsilon}^2_{y \text{ on } X_{-j},i}}} &= \frac{(x_j - X_{-j}\hat{\beta}_{x_j \text{ on } X_{-j}})'(y - X_{-j}\hat{\beta}_{y \text{ on } X_{-j}})}{\sqrt{(x_j'x_j - \hat{\beta}'_{x_j \text{ on } X_{-j}}X'_{-j}x_j)(y'y - \hat{\beta}'_{y \text{ on } X_{-j}}X'_{-j}y)}} \\
&= \frac{(x_j - X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}x_j)'(y - X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}y)}{\sqrt{(x_j'x_j - x_j'X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}x_j)(y'y - y'X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}y)}} \\
&= \frac{x_j'(I - P_{X_{-j}})(I - P_{X_{-j}})y}{\sqrt{(x_j'x_j - x_j'P_{X_{-j}}x_j)(y'y - y'P_{X_{-j}}y)}} \\
&= \frac{x_j'(I - P_{X_{-j}})y}{\sqrt{x_j'x_jy'y - x_j'P_{X_{-j}}x_jy'y - x_j'x_jy'P_{X_{-j}}y - x_j'P_{X_{-j}}x_jy'P_{X_{-j}}y}} \\
&= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{x_j'x_jy'y - x_j'P_{X_{-j}}x_jy'y - x_j'x_jy'P_{X_{-j}}y - x_j'P_{X_{-j}}x_jy'P_{X_{-j}}y}}
\end{aligned}
$$

Let $X_{reg<j>}$ denote the matrix where the first column of $X_{reg}$ is swapped with $x_j$

$$
\begin{aligned}
t_j^2 &= \frac{\hat{\beta}_j^2}{\hat{\sigma}^2(X'_{reg<j>}X_{reg<j>})^{-1}_{jj}} \\
&= \frac{\hat{\beta}_j^2}{\hat{\sigma}^2(\begin{bmatrix} x_j' \\ X'_{-j} \end{bmatrix}[\, x_j \quad X_{-j} \,])^{-1}_{11}} \\
&= \frac{\hat{\beta}_j^2}{\frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{n-p-1}(\begin{bmatrix} x_j'x_j & x_j'X_{-j} \\ X'_{-j}x_j & X'_{-j}X_{-j} \end{bmatrix}^{-1})_{11}} \text{ (Schur complement)} \\
&= \frac{\hat{\beta}_j^2}{\frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{n-p-1}(x_j'x_j - x_j'X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}x_j)^{-1}} \\
&= \frac{\hat{\beta}_j^2}{\frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{n-p-1}\frac{1}{\sum_{i=1}^{n}\hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\beta}_{reg<j>} &= (X'_{reg<j>}X_{reg<j>})^{-1}X'_{reg<j>}y \\
&= \begin{bmatrix} x_j'x_j & x_j'X_{-j} \\ X'_{-j}x_j & X'_{-j}X_{-j} \end{bmatrix}^{-1} \begin{bmatrix} x_j' \\ X'_{-j} \end{bmatrix} y \\
&= \begin{bmatrix} \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} & -\frac{x_j'X_{-j}(X'_{-j}X_{-j})^{-1}}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \\ -\frac{(X'_{-j}X_{-j})^{-1}X'_{-j}x_j}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} & (X'_{-j}X_{-j})^{-1} + \frac{(X'_{-j}X_{-j})^{-1}X'_{-j}x_jx_j'X_{-j}(X'_{-j}X_{-j})^{-1}}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \end{bmatrix} \begin{bmatrix} x_j' \\ X'_{-j} \end{bmatrix} y \\
&= \begin{bmatrix} \frac{x_j'y - x_j'X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}y}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \\ -\frac{(X'_{-j}X_{-j})^{-1}X'_{-j}x_jx_j'y}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} + (X'_{-j}X_{-j})^{-1}X'_{-j}y + \frac{(X'_{-j}X_{-j})^{-1}X'_{-j}x_jx_j'X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}y}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \end{bmatrix} \text{(Schur complement)} \\
&= \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \begin{bmatrix} x_j'(y - X_{-j}\hat{\beta}_{y \text{ on } X_{-j}}) \\ -\hat{\beta}_{x_j \text{ on } X_{-j}}x_j'y + \hat{\beta}_{y \text{ on } X_{-j}} \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i} + \hat{\beta}_{x_j \text{ on } X_{-j}}x_j'X_{-j}\hat{\beta}_{y \text{ on } X_{-j}} \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \begin{bmatrix} x_j'\hat{\epsilon}_{y \text{ on } X_{-j}} \\ -\hat{\beta}_{x_j \text{ on } X_{-j}}x_j'y + \hat{\beta}_{y \text{ on } X_{-j}} \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i} + \hat{\beta}_{x_j \text{ on } X_{-j}}x_j'X_{-j}\hat{\beta}_{y \text{ on } X_{-j}} \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{t_j}{\sqrt{t_j^2 + (n-p-1)}} \\
&= \frac{\frac{\hat{\beta}_j}{\sqrt{\frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{n-p-1} \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}}}{\left( \frac{\hat{\beta}_j^2}{\frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{n-p-1} \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}} + (n-p-1) \right)^{0.5}} \\
&= \frac{\hat{\beta}_j}{\sqrt{\hat{\beta}_j^2 + (n-p-1)\frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{n-p-1} \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}} \\
&= \frac{\hat{\beta}_j}{\sqrt{\hat{\beta}_j^2 + \frac{y'y - \hat{\beta}'X'_{reg<j>}y}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j,i}}}} \\
&= \frac{\frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}{\sqrt{\frac{(x_j'\hat{\epsilon}_{y \text{ on } X_{-j}})^2}{(\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i})^2} + \frac{y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}} \\
&= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{(x_j'\hat{\epsilon}_{y \text{ on } X_{-j}})^2 + (y'y - \hat{\beta}'_{reg<j>}X'_{reg<j>}y) \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}
\end{aligned}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{(x_j'\hat{\epsilon}_{y \text{ on } X_{-j}})^2 + (y'y - \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} \begin{bmatrix} x_j'\hat{\epsilon}_{y \text{ on } X_{-j}} \\ -\hat{\beta}_{x_j \text{ on } X_{-j}} x_j'y + \hat{\beta}_{y \text{ on } X_{-j}} \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i} + \hat{\beta}_{x_j \text{ on } X_{-j}} x_j' X_{-j} \hat{\beta}_{y \text{ on } X_{-j}} \end{bmatrix}' \begin{bmatrix} x_j & X_{-j} \end{bmatrix}' y) \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{(x_j'\hat{\epsilon}_{y \text{ on } X_{-j}})^2 + (y'y - \frac{1}{\sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}} (\hat{\epsilon}'_{y \text{ on } X_{-j}} x_j x_j' y - y' x_j \hat{\beta}'_{x_j \text{ on } X_{-j}} X'_{-j} y + \hat{\beta}'_{y \text{ on } X_{-j}} X'_{-j} y \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i} + \hat{\beta}'_{y \text{ on } X_{-j}} X'_{-j} x_j \hat{\beta}'_{x_j \text{ on } X_{-j}} X'_{-j} y)) \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i}}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{(x_j'(I - P_{X_{-j}})y)^2 + y'y x_j'(I - P_{X_{-j}}) x_j - \hat{\epsilon}'_{y \text{ on } X_{-j}} x_j x_j' y + y' x_j \hat{\beta}'_{x_j \text{ on } X_{-j}} X'_{-j} y - \hat{\beta}'_{y \text{ on } X_{-j}} X'_{-j} y \sum_{i=1}^{n} \hat{\epsilon}^2_{x_j \text{ on } X_{-j},i} - \hat{\beta}'_{y \text{ on } X_{-j}} X'_{-j} x_j \hat{\beta}'_{x_j \text{ on } X_{-j}} X'_{-j} y}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{(x_j'(I - P_{X_{-j}})y)^2 + y'y x_j'(I - P_{X_{-j}}) x_j - y'(I - P_{X_{-j}}) x_j x_j' y + y' x_j x_j' P_{X_{-j}} y - y' P_{X_{-j}} y x_j' (I - P_{X_{-j}}) x_j - y' P_{X_{-j}} x_j x_j' P_{X_{-j}} y}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{x_j' y x_j' y + x_j' P_{X_{-j}} y x_j' P_{X_{-j}} y - x_j' y x_j' P_{X_{-j}} y - x_j' P_{X_{-j}} y x_j' y + y'y x_j' x_j - y'y x_j' P_{X_{-j}} x - y' x_j x_j' y + y' P_{X_{-j}} x_j x_j' y + y' x_j x_j' P_{X_{-j}} y - y' P_{X_{-j}} y x_j' x_j + y' P_{X_{-j}} y x_j' P_{X_{-j}} x_j - y' P_{X_{-j}} x_j x_j' P_{X_{-j}} y}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{A + B - C - D + y'y x_j' x_j - y'y x_j' P_{X_{-j}} x_j - A + C + D - y' P_{X_{-j}} y x_j' x_j + y' P_{X_{-j}} y x_j' P_{X_{-j}} x_j - B}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{y'y x_j' x_j - y'y x_j' P_{X_{-j}} x_j - y' P_{X_{-j}} y x_j' x_j + y' P_{X_{-j}} y x_j' P_{X_{-j}} x_j}}
$$

$$
= \frac{x_j'\hat{\epsilon}_{y \text{ on } X_{-j}}}{\sqrt{x_j' x_j y'y - x_j' P_{X_{-j}} x_j y'y - x_j' x_j y' P_{X_{-j}} y - x_j' P_{X_{-j}} x_j y' P_{X_{-j}} y}}
$$

$$
= \hat{\rho}_{x_j y . X_{-j}}
$$

**10. Log-likelihood of residuals (LL)**

According to linear assumption 4, each $e_i$ is normally distributed with $\mu_i = 0$ and $\sigma = \sigma_1 = \sigma_2 = \ldots$ for i=1,2,...,n

$$
P(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\epsilon_i^2}{2\sigma^2}}
$$

$$f_\epsilon(\epsilon_1, \epsilon_2, \ldots, \epsilon_n) = \frac{1}{2\pi^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\epsilon-\mu_\epsilon)'\Sigma^{-1}(\epsilon-\mu_\epsilon)}$$

$$= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^n \sigma_i} e^{-\frac{1}{2}(\epsilon-\mu_\epsilon)' \begin{bmatrix} \sigma_1^2 & 0 & \ldots \\ 0 & \sigma_2^2 & \ldots \\ 0 & \ldots & \sigma_n^2 \end{bmatrix}^{-1} (\epsilon-\mu_\epsilon)}$$

$$= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^n \sigma} e^{-\frac{1}{2}(\epsilon-\mu_\epsilon)' \begin{bmatrix} \sigma^2 & 0 & \ldots \\ 0 & \sigma^2 & \ldots \\ 0 & \ldots & \sigma^2 \end{bmatrix}^{-1} (\epsilon-\mu_\epsilon)}$$

$$= \prod_{i=1}^n \frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{1}{2\sigma^2}(\epsilon_i - 0)^2}$$

$$= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\epsilon_i^2}{2\sigma^2}}$$

$$= \prod_{i=1}^n f(\epsilon_i)$$

This proves independence of error terms, so the log-likelihood can be taken to be,

$$log(\prod_{i=1}^n P(\epsilon_i)) = -nlog(\sigma) - 0.5nlog(2\pi) - \frac{\sum_{i=1}^n \epsilon_i^2}{2\sigma^2}$$

11. **Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Mallows's $C_p$**

Presented below are 3 common sets of definitions for AIC and BIC. $C_p$ is omitted since it is proportional to AIC, and so stepwise selection using $C_p$ or AIC will yield the same model.
Common throughout all definitions is the log(n) term used in the penalty of BIC.
BIC imposes a heavier penalty when 2<log(n) or n>7.

Definition 1
AIC=-2*Log-likelihood of residuals + 2(p+1)
BIC=-2*Log-likelihood of residuals + (p+1)log(n)

Definition 2

$$AIC = \frac{1}{n\hat{\sigma}^2}(SSE + 2p\hat{\sigma}^2)$$

$$BIC = \frac{1}{n\hat{\sigma}^2}(SSE + p\hat{\sigma}^2 log(n))$$

Definition 3

$$AIC = nlog(\frac{SSE}{n}) + 2(p+1)$$

$$BIC = nlog(\frac{SSE}{n}) + (p+1)log(n)$$

**12. Leverage**

$$\begin{aligned} Leverage &= H_{ii} \\ &= h_i \\ &= \gamma_i' X_{reg}(X_{reg}' X_{reg})^{-1} X_{reg}' \gamma_i \\ &= x_{(i)}^{'} (X_{reg}' X_{reg})^{-1} x_{(i)} \end{aligned}$$

where H is the hat matrix= $P_{X_{reg}} = X_{reg}(X_{reg}' X_{reg})^{-1} X_{reg}'$ and $\gamma_i$ is a column vector with 1 on the $i_{th}$ term and 0 otherwise

For only one regressor,

$$\begin{aligned} h_i &= x_{(i)}' (\begin{bmatrix} j' \\ x_1' \end{bmatrix} [j \quad x_1])^{-1} x_{(i)} \\ &= x_{(i)}^{'} \begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix}^{-1} x_{(i)} \\ &= x_{(i)}^{'} \frac{1}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & -\sum_{i=1}^n x_{i1} \\ -\sum_{i=1}^n x_{i1} & n \end{bmatrix} x_{(i)} \\ &= x_{(i)}^{'} \frac{1}{n\sum_{i=1}^n x_{i1}^2 - n^2\bar{x}_{.1}^2} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & -\sum_{i=1}^n x_{i1} \\ -\sum_{i=1}^n x_{i1} & n \end{bmatrix} x_{(i)} \\ &= \frac{1}{n(\sum_{i=1}^n x_{i1}^2 - n\bar{x}_{.1}^2)} [\sum_{i=1}^n x_{i1}^2 - x_{i1}\sum_{i=1}^n x_{i1} \quad \sum_{i=1}^n x_{i1} + x_{i1}n] \, x_{(i)} \\ &= \frac{1}{n(\sum_{i=1}^n x_{i1}^2 - n\bar{x}_{.1}^2)} (\sum_{i=1}^n x_{i1}^2 - 2x_{i1}\sum_{i=1}^n x_{i1} + x_{i1}^2 n) \\ &= \frac{1}{n(\sum_{i=1}^n x_{i1}^2 - n\bar{x}_{.1}^2)} (\sum_{i=1}^n x_{i1}^2 - n\bar{x}_{.1}^2 + n\bar{x}_{.1}^2 - 2x_{i1}n\bar{x}_{.1} + x_{i1}^2 n) \\ &= \frac{1}{n(\sum_{i=1}^n x_{i1}^2 - n\bar{x}_{.1}^2)} (\sum_{i=1}^n x_{i1}^2 - n\bar{x}_{.1}^2 + n(\bar{x}_{.1}^2 - x_{i1})^2) \\ &= \frac{1}{n} + \frac{(x_{i1} - \bar{x}_{.1})^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_{.1})^2} \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{(since there is only 1 regressor)} \end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} h_i &= trace(H) \\ &= trace(X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg}) \\ &= trace((X'_{reg}X_{reg})^{-1}X'_{reg}X_{reg}) \\ &= p+1\end{aligned}$$

$$\bar{h} = \frac{p+1}{n} \approx \frac{p}{n} \text{ when n is large}$$

The practise is if $2p/n < 1$ (additional constraint to ensure n is large enough relative to p), $h_i > \frac{2p}{n}$ is an indication of large leverage.

13. **Internally/Externally studentized residual**

$$\begin{aligned}\hat{\epsilon} &= y - X_{reg}\hat{\beta} \\ &= y - Hy \\ &= (I-H)y\end{aligned}$$

$$\begin{aligned}var(\hat{\epsilon}) &= \sigma^2(I-H)(I-H)' \\ &= \sigma^2(I-H) \text{ since H is idempotent}\end{aligned}$$

$$var(\hat{\epsilon}_i) = \sigma^2(1-h_i)$$

$$i_{th} \text{ internally studentized residuals} = t_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_i)}}$$

The internally studentized residual adjusts residuals for differences in variances, even though variances of true error terms should be equal to each other (according to the 2nd linear assumptions).

$$i_{th} \text{ externally studentized residual} = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2_{(-i)}(1-h_i)}}$$

where $\hat{\sigma}^2_{(-i)}$ is derived from the regression of X against y with observation i removed i.e $\hat{\sigma}^2_{(-i)} = \frac{SSE_{(-i)}}{(n-1)-p-1}$

By excluding the effects of the ith observation on the MSE, it is more evident on whether the ith residual is improbably large for the fitted model.

Equivalently (second definition),

$$i_{th} \text{ externally studentized residual} = t_i\sqrt{\frac{n-p-2}{n-p-1-t_i^2}}$$

where $t_i$ is the ith internally studentized residual (not to be confused with t-statistic here)

Obtaining the externally studentized residuals using the second definition avoids refitting a linear regression for each externally studentized residual.

Tedious proof of second definition
Sherman–Morrison formula (definition only)
Suppose A is invertible and u,v are column vectors.Then if $A + uv'$ is invertible,

$$(A+uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}$$

$$\begin{aligned} X'_{reg}X_{reg} &= \begin{bmatrix} x_{(i)} & X'_{(-i)} \end{bmatrix} \begin{bmatrix} x'_{(i)} \\ X_{(-i)} \end{bmatrix} \\ &= \begin{bmatrix} x_{(i)}x'_{(i)} + X'_{(-i)}X_{(-i)} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} (X'_{(-i)}X_{(-i)})^{-1} &= \begin{bmatrix} X'_{reg}X_{reg} - x_{(i)}x'_{(i)} \end{bmatrix} \\ &= (X'_{reg}X_{reg})^{-1} + \frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1 - x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}} \\ &= (X'_{reg}X_{reg})^{-1} + \frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1-h_i} \end{aligned}$$

$$\begin{aligned} t_i\sqrt{\frac{n-p-2}{n-p-1-t_i^2}} &= \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_i)}}\sqrt{\frac{n-p-2}{n-p-1-\frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2(1-h_i)}}} \\ &= \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_i)}}\sqrt{\frac{n-p-2}{\frac{(n-p-1)\hat{\sigma}^2(1-h_i)-\hat{\epsilon}_i^2}{\hat{\sigma}^2(1-h_i)}}} \\ &= \hat{\epsilon}_i\sqrt{\frac{n-p-2}{SSE(1-h_i)-\hat{\epsilon}_i^2}} \end{aligned}$$

$$
\begin{aligned}
&\frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2_{(-i)}(1-h_i)}} \\
&= \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2_{(-i)}(1-h_i)}} \\
&= \frac{\hat{\epsilon}_i}{\sqrt{\frac{y'_{(-i)}(I-X_{(-i)}(X'_{(-i)}X_{(-i)})^{-1}X'_{(-i)})y_{(-i)}}{(n-1)-p-1}(1-h_i)}} \\
&= \frac{\hat{\epsilon}_i}{\sqrt{\frac{y'_{(-i)}(I-X_{(-i)}((X'_{reg}X_{reg})^{-1}+\frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1-h_i})X'_{(-i)})y_{(-i)}}{n-p-2}(1-h_i)}} \\
&= \frac{\hat{\epsilon}_i}{\sqrt{\frac{y'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}\frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1-h_i}X'_{(-i)}y_{(-i)}}{n-p-2}(1-h_i)}} \\
&= \hat{\epsilon}_i\sqrt{\frac{n-p-2}{(y'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})(1-h_i)-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}}}
\end{aligned}
$$

By comparison, I will verify that,

$$SSE(1-h_i)-\hat{\epsilon}_i^2 = (y'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})(1-h_i)-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}$$

$$
\begin{aligned}
SSE &= y'(I-P_{X_{reg}})y \\
&= y'(I-X_{reg}(X'_{reg}X_{reg})^{-1}X'_{reg})y \\
&= \begin{bmatrix} y_i & y'_{(-i)} \end{bmatrix}(I-\begin{bmatrix} x'_{(i)} \\ X_{(-i)} \end{bmatrix}(X'_{reg}X_{reg})^{-1}\begin{bmatrix} x_{(i)} & X'_{(-i)} \end{bmatrix})\begin{bmatrix} y_i \\ y_{(-i)} \end{bmatrix} \\
&= y_i^2+y'_{(-i)}y_{(-i)}-(y_ix'_{(i)}+y'_{(-i)}X_{(-i)})(X'_{reg}X_{reg})^{-1}(x_{(i)}y_i+X'_{(-i)}y_{(-i)}) \\
&= y_i^2+y'_{(-i)}y_{(-i)}-y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i-y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&= \left(y'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}\right)+y_i^2-y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&-2y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}
\end{aligned}
$$

$$\text{Let } c = y'_{(-i)}y_{(-i)}-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}$$

$$
\begin{aligned}
&SSE(1-h_i)-\hat{\epsilon}_i^2\\
&=c(1-h_i)+(y_i^2-y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i-2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})(1-h_i)-\hat{\epsilon}_i^2\\
&=c(1-h_i)+(y_i^2-y_i h_i y_i-2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})(1-h_i)-(y_i-x'_{(i)}(X'_{reg}X_{reg})^{-1}\begin{bmatrix}x_{(i)} & X'_{(-i)}\end{bmatrix}\begin{bmatrix}y_i\\ y_{(-i)}\end{bmatrix})^2\\
&=c(1-h_i)+(y_i^2-y_i h_i y_i-2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})(1-h_i)-(y_i-x'_{(i)}(X'_{reg}X_{reg})^{-1}(x_{(i)}y_i+X'_{(-i)}y_{(-i)}))^2\\
&=c(1-h_i)+(y_i^2-y_i h_i y_i-2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})(1-h_i)-(y_i-h_i y_i-x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})^2\\
&=c(1-h_i)+y_i^2-y_i h_i y_i-2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}\\
&-y_i^2 h_i+y_i h_i y_i h_i+2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}h_i\\
&-y_i^2+2y_i h_i y_i+2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}-h_i y_i h_i y_i-2h_i y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}\\
&-x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}\\
&=c(1-h_i)+A-B-2C\\
&-B+D+2E\\
&-A+2B+2C-D-2E\\
&-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}\\
&=c(1-h_i)-y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}
\end{aligned}
$$

Plots

a. Internally/externally studentized residuals against predicted y

Points are expected to be randomly scattered (linear assumption 3) within a band (linear assumption 2) around each level of y (linear assumption 1).

b. Internally/externally studentized residuals against time

Points are expected to be randomly scattered (linear assumption 3) within a band (linear assumption 2) around 0 (linear assumption 1).

c. Normal QQplot of residuals

Check that error terms are normally distributed (assumption 4).

## 14. Cook's distance

Definition

Let $\hat{\beta}_{(-i)}$ be the coefficients obtained from the regression of y on $X_{reg}$ with the ith observation removed,

$$
\begin{aligned}
i_{th}\text{ cook's distance}=D_i&=\frac{(\hat{\beta}-\hat{\beta}_{(-i)})'X'_{reg}X_{reg}(\hat{\beta}-\hat{\beta}_{(-i)})}{(p+1)\sigma^2}\\
&=\frac{(X_{reg}(\hat{\beta}-\hat{\beta}_{(-i)}))'X_{reg}(\hat{\beta}-\hat{\beta}_{(-i)})}{(p+1)\sigma^2}\\
&=\frac{(X_{reg}\hat{\beta}-X_{reg}\hat{\beta}_{(-i)})'(X_{reg}\hat{\beta}-X_{reg}\hat{\beta}_{(-i)})}{(p+1)\sigma^2}\\
&=\frac{\sum_{j=1}^{n}(\hat{y}_j-\hat{y}_{j(-i)})^2}{(p+1)\sigma^2}
\end{aligned}
$$

where $\hat{y}_{j(-i)}$ is the predicted value for $y_j$ with the coefficients derived from the regression of y on $X_{reg}$ with observation i removed.

Equivalently (second definition),

$$D_i = \frac{\hat{\epsilon}_i^2}{(p+1)\hat{\sigma}^2} \frac{h_i}{(1-h_i)^2}$$

Obtaining the cook's distance using the second definition avoids refitting a linear regression for each cook's distance.
If $D_i$ exceeds a certain threshold, the observation is suspected to be influential. Common thresholds used are 1, 4/n and 4/(n-p-1). Alternatively, a percentile of over 50 for the F(p, n-p-1) distribution can be used to indicate a highly influential point.

Proof for second definition

$$\begin{aligned}
\hat{\epsilon}_i^2 &= (y_i - x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{reg}y)^2 \\
&= (y_i - x'_i(X'_{reg}X_{reg})^{-1}\begin{bmatrix} x_{(i)} & X'_{(-i)} \end{bmatrix}\begin{bmatrix} y_i \\ y_{(-i)} \end{bmatrix})^2 \\
&= (y_i - x'_{(i)}(X'_{reg}X_{reg})^{-1}(x_{(i)}y_i + X'_{(-i)}y_{(-i)}))^2 \\
&= (y_i - x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i - x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})^2 \\
&= y_i^2 + (x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i)^2 + (x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})^2 - 2y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&- 2y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} + 2x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&= y_i^2 + (h_iy_i)^2 + (x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})^2 - 2h_iy_i^2 - 2y_ix'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&+ 2h_iy_ix'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}
\end{aligned}$$

$$\begin{aligned}
D_i &= \frac{(\hat{\beta} - \hat{\beta}_{(-i)})'X'_{reg}X_{reg}(\hat{\beta} - \hat{\beta}_{(-i)})}{(p+1)\sigma^2} \\
&= \frac{((X'_{reg}X_{reg})^{-1}X'_{reg}y - (X'_{(-i)}X_{(-i)})^{-1}X'_{(-i)}y_{(-i)})'X'_{reg}X_{reg}((X'_{reg}X_{reg})^{-1}X'_{reg}y - (X'_{(-i)}X_{(-i)})^{-1}X'_{(-i)}y_{(-i)})}{(p+1)\sigma^2} \\
&= \frac{((X'_{reg}X_{reg})^{-1}X'_{reg}y - ((X'_{reg}X_{reg})^{-1} + \frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1-h_i})X'_{(-i)}y_{(-i)})'X'_{reg}X_{reg}}{(p+1)\sigma^2} \\
& \frac{(X'_{reg}X_{reg})^{-1}X'_{reg}y - ((X'_{reg}X_{reg})^{-1} + \frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1-h_i})X'_{(-i)}y_{(-i)}}{1} \\
&= \frac{(y'X_{reg} - y'_{(-i)}X_{(-i)}(I + \frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}}{1-h_i}))((X'_{reg}X_{reg})^{-1}X'_{reg}y - ((X'_{reg}X_{reg})^{-1} + \frac{(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}}{1-h_i})X'_{(-i)}y_{(-i)})}{(p+1)\sigma^2} \\
&= \frac{(1-h_i)y'X_{reg} - y'_{(-i)}X_{(-i)}((1-h_i) + (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)})}{(p+1)\sigma^2(1-h_i)^2} \\
& \frac{(1-h_i)(X'_{reg}X_{reg})^{-1}X'_{reg}y - ((1-h_i)(X'_{reg}X_{reg})^{-1} + (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1})X'_{(-i)}y_{(-i)}}{1}
\end{aligned}$$

By comparison, I will verify that

$$((1-h_i)y'X_{reg} - y'_{(-i)}X_{(-i)}((1-h_i) + (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}))((1-h_i)(X'_{reg}X_{reg})^{-1}X'_{reg}y- ((1-h_i)(X'_{reg}X_{reg})^{-1} + (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1})X'_{(-i)}y_{(-i)}) = \hat{\epsilon}_i^2 h_i$$

$$\begin{aligned}
&((1-h_i)y'X_{reg} - y'_{(-i)}X_{(-i)}((1-h_i) + (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}))((1-h_i)(X'_{reg}X_{reg})^{-1}X'_{reg}y \\
&- ((1-h_i)(X'_{reg}X_{reg})^{-1} + (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1})X'_{(-i)}y_{(-i)}) \\
&= ((1-h_i)\begin{bmatrix} y_i & y'_{(-i)} \end{bmatrix}\begin{bmatrix} x'_{(i)} \\ X_{(-i)} \end{bmatrix} - (1-h_i)y'_{(-i)}X_{(-i)} - y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}) \\
&((1-h_i)(X'_{reg}X_{reg})^{-1}\begin{bmatrix} x_{(i)} & X'_{(-i)} \end{bmatrix}\begin{bmatrix} y_i \\ y_{(-i)} \end{bmatrix} - (1-h_i)(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&- (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}) \\
&= ((1-h_i)y_i x'_{(i)} - y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)})((1-h_i)(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&- (X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}) \\
&= (1-h_i)^2 y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i - (1-h_i)y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&- (1-h_i)y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&+ y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&= (1-h_i)^2 y_i h_i y_i - (1-h_i)y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}h_i y_i \\
&- (1-h_i)y_i h_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&+ y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}h_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&= y_i h_i y_i - 2h_i y_i h_i y_i + h_i^2 y_i h_i y_i - y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}h_i y_i + h_i y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}h_i y_i \\
&- y_i h_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} + h_i y_i h_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&+ y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}h_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&= h_i(y_i^2 - 2h_i y_i^2 + h_i^2 y_i^2 - y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i + h_i y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&- y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} + h_i y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&+ y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)}) \\
&= h_i(y_i^2 - 2h_i y_i^2 + h_i^2 y_i^2 - 2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} + 2h_i y'_{(-i)}X_{(-i)}(X'_{reg}X_{reg})^{-1}x_{(i)}y_i \\
&+ (x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})^2) \\
&= h_i(y_i^2 - 2h_i y_i^2 + h_i^2 y_i^2 - 2y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} + 2h_i y_i x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)} \\
&+ (x'_{(i)}(X'_{reg}X_{reg})^{-1}X'_{(-i)}y_{(-i)})^2) \\
&= \hat{\epsilon}_i^2 h_i
\end{aligned}$$

**15. Condition number**

$$\text{Condition number} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

where $\lambda_{max}$ is the maximum eigenvalue of X'X
and $\lambda_{min}$ is the minimum eigenvalue of X'X
Common interpretation
$\text{condition number} < 100$ implies no serious multicollinearity is present
$100 \leq \text{condition number} < 1000$ implies moderate to strong multicollinearity is present
$\text{condition number} \geq 1000$ implies strong multicollinearity is present

16. **Condition index**

$$i_{th} \text{ condition index} = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

Estimated number of near-linear dependencies = No. of condition indices more than 1000

17. **Variance inflation factor (VIF)**

Intuitively, multicollinearity leads to larger $var(\hat{\beta})$ since a feature is closely substitutable by other features, leading to $\hat{\beta}$.

WLOG, let $X_{reg<j>}$ denote the matrix where the first column of $X_{reg}$ is swapped with $x_j$

$$\begin{aligned}
v\hat{a}r(\hat{\beta}_j) &= \sigma^2((X'_{reg<j>}X_{reg<j>})^{-1})_{11} \\
&= \sigma^2((\begin{bmatrix} x'_j \\ X'_{-j} \end{bmatrix}[\, x_j \quad X_{-j}\,])^{-1})_{11} \\
&= \sigma^2(\begin{bmatrix} x'_j x_j & x'_j X_{-j} \\ X'_{-j}x_j & X'_{-j}X_{-j} \end{bmatrix}^{-1})_{11} \\
&= \sigma^2(x'_j x_j - x'_j X_{-j}(X'_{-j}X_{-j})^{-1}X'_{-j}x_j)^{-1} (\text{Schur complement}) \\
&= \sigma^2(x'_j x_j - \hat{\beta}'_{j \text{ on } -j}X'_{-j}x_j)^{-1} \text{ where } \hat{\beta}'_{j \text{ on } -j} \text{ is the coefficients of regressing } x_j \text{ on } X_{-j} \\
&= \frac{\sigma^2}{SSE_{j \text{ on } -j}} \\
&= \frac{\sigma^2}{SST_{j \text{ on } -j} - SSR_{j \text{ on } -j}} \\
&= \frac{\sigma^2}{SST_{j \text{ on } -j}(1 - \frac{SSR_{j \text{ on } -j}}{SST_{j \text{ on } -j}})} \\
&= \frac{\sigma^2}{SST_{j \text{ on } -j}} * \frac{1}{(1 - R^2_{j \text{ on } -j})}
\end{aligned}$$

where $R^2_{j \text{ on } -j}$ is obtained from the regression of $x_j$ on the remaining p-1 predictors and intercept

$$VIF_j = C_{jj}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = \frac{1}{1 - R^2_{j \text{ on } -j}}, j = 1, 2, \ldots, k$$

where $C = (X'_{reg}X_{reg})^{-1} \propto c\hat{o}v(\hat{\beta})$

Common interpretation
$VIF < 1$ implies no serious multicollinearity is present
$1 \leq VIF < 5$ implies moderate to strong multicollinearity is present
$VIF \geq 5$ implies strong multicollinearity is present

**18. Derivation of $\hat{\beta}$ for ridge regression**

$X_{reg}$ is assumed to be standardised.
Let b be an estimator of $\beta$

$$(y - X_{reg}b)'(y - X_{reg}b) + \lambda b'b = y'y - 2b'X'_{reg}y + b'X'_{reg}X_{reg}b + \lambda b'b$$

Differentiating by b and setting the derivative to 0 to derive the local minimum/maximum,

$$\frac{d}{db}(y'y - 2b'X'_{reg}y + b'X'_{reg}X_{reg}b + \lambda b'b) = 0$$
$$-2X'_{reg}y + 2X'_{reg}X_{reg}b + 2\lambda b = 0$$
$$X'_{reg}X_{reg}b + \lambda b = X'_{reg}y$$

$$\hat{\beta} = b = (X'_{reg}X_{reg} + \lambda I)^{-1}X'_{reg}y \text{ (if the inverse of } X'_{reg}X_{reg} + \lambda I \text{ is unique)}$$

**19. Finding $\hat{\beta}$ for lasso regression**

$X_{reg}$ is assumed to be standardised.
Proximal gradient descent
Let $r(\beta)$ denote a regularization function separable in $\beta$, b be an estimator of $\beta$.

$$
\begin{aligned}
\text{Loss function } f(b) &= \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda r(b) \\
&= \frac{1}{2n}||y - X_{reg}b||_2^2 + \lambda r(b) \\
&= \frac{1}{2n}||(y - X_{reg}\beta^{(k)}) + (X_{reg}\beta^{(k)} - X_{reg}b)||_2^2 + \lambda r(b) \\
&= \frac{1}{2n}\left[||y - X_{reg}\beta^{(k)}||_2^2 + ||X_{reg}\beta^{(k)} - X_{reg}b||_2^2 + 2(y - X_{reg}\beta^{(k)})'(X_{reg}\beta^{(k)} - X_{reg}b)\right] + \lambda r(b) \\
&= \frac{1}{2n}\left[||y - X_{reg}\beta^{(k)}||_2^2 + ||X_{reg}(\beta^{(k)} - b)||_2^2 + 2(y - X_{reg}\beta^{(k)})'(X_{reg}\beta^{(k)} - X_{reg}b)\right] + \lambda r(b) \\
&\leq \frac{1}{2n}\left[||y - X_{reg}\beta^{(k)}||_2^2 + ||X_{reg}||_{op}^2||\beta^{(k)} - b||_2^2 + 2(y - X_{reg}\beta^{(k)})'X_{reg}(\beta^{(k)} - b)\right] + \lambda r(b) \\
&\text{where } ||X_{reg}||_{op} \text{ is the operator norm of } X_{reg} \text{ i.e. square root of the largest eigenvalue of } X_{reg}'X_{reg} \\
&< \frac{1}{2n}\left[||y - X_{reg}\beta^{(k)}||_2^2 + \frac{1}{\tau}||\beta^{(k)} - b||_2^2 + 2(y - X_{reg}\beta^{(k)})'X_{reg}(\beta^{(k)} - b)\right] + \lambda r(b) \\
&\text{where } 0 < \tau < \frac{1}{||X_{reg}||_{op}^2}, \frac{1}{\tau} > ||X_{reg}||_{op}^2 \\
&= \frac{1}{2n\tau}(\tau||y - X_{reg}\beta^{(k)}||_2^2 + ||\beta^{(k)} - b||_2^2 + 2\tau(y - X_{reg}\beta^{(k)})'X_{reg}(\beta^{(k)} - b) + 2n\tau\lambda r(b)) \\
&= \frac{1}{2n\tau}(\tau||y - X_{reg}\beta^{(k)}||_2^2 - \tau^2||X_{reg}'(y - X_{reg}\beta^{(k)})||_2^2 + ||\tau X_{reg}'(y - X_{reg}\beta^{(k)}) + (\beta^{(k)} - b)||_2^2 + 2n\tau\lambda r(b)) \\
&= g(b)
\end{aligned}
$$

$$
\begin{aligned}
argmin_b(\text{Loss function } g(b)) &= \beta^{(k+1)} \\
&= argmin_\beta(||\tau X_{reg}'(y - X_{reg}\beta^{(k)}) + (\beta^{(k)} - b)||_2^2 + 2n\tau\lambda r(b))
\end{aligned}
$$

$$
\text{Let z} = \tau X_{reg}'(y - X_{reg}\beta^{(k)}) + \beta^{(k)} = \beta^{(k)} - \tau X_{reg}'(X_{reg}\beta^{(k)} - y)
$$

Since $||z - b||_2^2$ and $r(b)$ are separable in b,

**(Statement 2)**

Differentiating $||z - b||_2^2 + \tau\lambda r(b))$ with respect to $b_j$ for j=1,2,...p and setting derivative to 0,

$$
-2(z_j - b_j) + 2n\tau\lambda\frac{dr(b)}{db_j} = 0
$$

In the case of lasso,

$$
0 = \begin{cases} -2(z_j - b_j) + 2n\tau\lambda sign(b_j) \text{ if } b_j \neq 0 \\ [-2z_j - 2n\tau\lambda, -2z_j + 2n\tau\lambda] \text{ if } b_j = 0 \end{cases}
$$

If $b_j = 0, 0 \in [-2z_j - 2n\tau\lambda, -2z_j + 2n\tau\lambda]$

$0 > -2z_j - 2n\tau\lambda, 0 < -2z_j + 2n\tau\lambda$

$-n\tau\lambda < z_j < n\tau\lambda$

$$
\beta_j = \begin{cases} \frac{2z_j + 2n\tau\lambda}{2} = z_j + n\tau\lambda \text{ if } b_j < 0 \text{ or } z_j < -n\tau\lambda \\ 0 \text{ if } -n\tau\lambda < z_j < n\tau\lambda \\ \frac{2z_j - 2n\tau\lambda}{2} = z_j - n\tau\lambda \text{ if } b_j > 0 \text{ or } z_j > n\tau\lambda \end{cases}
$$

More compactly,

$$
b_j = ST(z^{(k)}, n\tau\lambda) = max(|z^{(k)}| - n\tau\lambda, 0) * sign(z^{(k)})
$$

Procedure

Step 1: $\beta^{(0)} = 0, 0 < \tau < \frac{1}{||X_{reg}||^2_{op}}$

Step 2: Repeat until $||\beta^{(k+1)} - \beta^{(k)}||_2$ is small or for a fixed number of iterations. For each iteration k,

$$(a)\ z^{(k)} = \beta^{(k)} - \tau X'_{reg}(X_{reg}\beta^{(k)} - y)$$
$$(b)\ \beta^{(k+1)} = max(|z^{(k)}| - n\tau\lambda, 0) * sign(z^{(k)})$$

Addition of Nesterov's accelerated gradient (NAG) step increases the convergence rate of the algorithm from $O(\frac{1}{k})$ to $O(\frac{1}{k^2})$.

In particular the loops in step 2 becomes,

$$(a)\ v = \beta^{(k-1)} + \frac{j-2}{j+1}(\beta^{(k-1)} - \beta^{(k-2)})$$
$$(b)\ z^{(k)} = v - \tau X'_{reg}(X_{reg}v - y)$$
$$(c)\ \beta^{(k)} = max(|z^{(k)}| - n\tau\lambda, 0) * sign(z^{(k)})$$

**20. Finding $\hat{\beta}$ for elastic net regression**

$X_{reg}$ is assumed to be standardised.

$$\text{Loss function} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\rho\sum_{j=1}^{p}|\beta_j| + \frac{\lambda(1-\rho)}{2}\sum_{j=1}^{p}\beta_j^2$$

b be an estimator of $\beta$

Following statement 2 (derivative) under appendix A19, it can be shown that

$$-2(z_j - b_j) + 2n\tau\lambda\rho\frac{d\sum_{k=1}^{p}|b_k|}{db_j} + 2n\tau\lambda\frac{(1-\rho)}{2}\frac{d\sum_{k=1}^{p}b_k^2}{db_j} = 0$$
$$-2(z_j - b_j) + 2n\tau\lambda\rho\frac{d\sum_{k=1}^{p}|b_k|}{db_j} + 2n\tau\lambda(1-\rho)b_j = 0$$
$$0 = \begin{cases} -2(z_j - b_j) + 2n\tau\lambda\rho sign(b_j) + 2n\tau\lambda(1-\rho)b_j \text{ if } b_j \neq 0 \\ [-2z_j - 2n\tau\lambda\rho, -2z_j + 2n\tau\lambda\rho] \text{ if } b_j = 0 \end{cases}$$
$$\text{If } b_j = 0, 0 \in [-2z_j - 2n\tau\lambda\rho, -2z_j + 2n\tau\lambda\rho]$$
$$0 > -2z_j - 2n\tau\lambda\rho, 0 < -2z_j + 2n\tau\lambda\rho$$
$$-n\tau\lambda\rho < z_j < n\tau\lambda\rho$$
$$b_j = \begin{cases} \frac{2z_j + 2n\tau\lambda\rho}{2 + 2n\tau\lambda(1-\rho)} = \frac{z_j + n\tau\lambda\rho}{1 + n\tau\lambda(1-\rho)} \text{ if } b_j < 0 \text{ or } z_j < -n\tau\lambda\rho \\ 0 \text{ if } -n\tau\lambda\rho < z_j < n\tau\lambda\rho \\ \frac{2z_j - 2n\tau\lambda\rho}{2 + 2n\tau\lambda(1-\rho)} = \frac{z_j - n\tau\lambda\rho}{1 + n\tau\lambda(1-\rho)} \text{ if } b_j > 0 \text{ or } z_j > n\tau\lambda\rho \end{cases}$$

More compactly,

$$b_j = \frac{ST(z^{(k)}, n\tau\lambda\rho)}{1+n\tau\lambda(1-\rho)} * sign(z^{(k)}) = \frac{max(|z^{(k)}| - n\tau\lambda\rho, 0)}{1+n\tau\lambda(1-\rho)} * sign(z^{(k)})$$

Procedure (with Nesterov's accelerated gradient step)

Step 1: $\beta^{(0)} = 0, 0 < \tau < \frac{1}{||X_{reg}||_{op}^2}$

Step 2: Repeat until $||\beta^{(k+1)} - \beta^{(k)}||_2$ is small or for a fixed number of iterations. For each iteration k,

$$(a)\ v = \beta^{(k-1)} + \frac{j-2}{j+1}(\beta^{(k-1)} - \beta^{(k-2)})$$

$$(b)\ z^{(k)} = v - \tau X'_{reg}(X_{reg}v - y)$$

$$(c)\ \beta^{(k)} = \frac{max(|z^{(k)}| - n\tau\lambda\rho, 0)}{1+n\tau\lambda(1-\rho)} * sign(z^{(k)})$$

**21. Finding $\hat{\beta}$ for SCAD regression**

$$J_{\lambda,a}(\beta_j) = \begin{cases} \lambda|\beta_j| \text{ if } |\beta_j| \le \lambda \\ -\frac{\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)} \text{ if } \lambda < |\beta_j| \le a\lambda \\ \frac{(a+1)\lambda^2}{2} \text{ if } |\beta_j| > a\lambda \end{cases}$$

where $\lambda > 0$, $a \ge 1$

$$\frac{dJ_{\lambda,a}(\beta_j)}{d\beta_j} = \begin{cases} \lambda sign(\beta_j) \text{ if } |\beta_j| \le \lambda, \beta_j \ne 0 \\ [-\lambda, \lambda] \text{ if } \beta_j = 0 \\ -\frac{2\beta_j - 2a\lambda sign(\beta_j)}{2(a-1)} = \frac{a\lambda sign(\beta_j) - \beta_j}{a-1} \text{ if } \lambda < |\beta_j| \le a\lambda \\ 0 \text{ if } |\beta_j| > a\lambda \end{cases}$$

$$\text{Loss function} = \frac{1}{2(n-1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{j=1}^{p} J_{\lambda,a}(\beta_j)$$

Differentiating $\frac{1}{2(n-1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{j=1}^{p} J_{\lambda,a}(b_j)$ with respect to $b_j$ for j=1,2,...p and setting derivative to 0,

$$-\frac{1}{n-1}(y - b_0 - \sum_{k=1}^{p} b_k x_k)' x_j + \frac{dJ_{\lambda,a}(b_j)}{db_j} = 0$$

$$-\frac{1}{n-1}(y - b_0 - \sum_{k=1, k \ne j}^{p} b_k x_k)' x_j + \frac{1}{n-1} b_j x_j' x_j + \frac{dJ_{\lambda,a}(b_j)}{db_j} = 0$$

Let $z_j$ denote $\frac{1}{n-1}(y - b_0 - \sum_{k=1, k\ne j}^{p} b_k x_k)' x_j$

$$-z_j + \frac{1}{n-1} b_j x_j' x_j + \frac{dJ_{\lambda,a}(b_j)}{db_j} = 0$$

$$b_j = \begin{cases} \frac{z_j - \lambda sign(b_j)}{\frac{1}{n-1}x_j'x_j} \text{ if } |b_j| \leq \lambda \\ 0 \text{ if } z_j \in [-\lambda, \lambda] \\ \frac{\frac{1}{n-1}(a-1)r'x_j - a\lambda sign(b_j)}{\frac{1}{n-1}(a-1)x_j'x_j - 1} = \frac{z_j - \frac{a}{a-1}\lambda sign(b_j)}{\frac{1}{n-1}x_j'x_j - \frac{1}{a-1}} \text{ if } \lambda < |b_j| \leq a\lambda \\ \frac{z}{\frac{1}{n-1}x_j'x_j} \text{ if } |b_j| > a\lambda \end{cases}$$

Since $x_j$ is standardised,

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)^2 = \frac{\langle x_j, x_j \rangle}{n-1} = 1$$

$$\langle x_j, x_j \rangle = x_j'x_j = n-1$$

$$b_j = \begin{cases} z_j - \lambda sign(b_j) \text{ if } |b_j| \leq \lambda \\ \frac{z_j - \frac{a}{a-1}\lambda sign(b_j)}{1 - \frac{1}{a-1}} \text{ if } \lambda < |b_j| \leq a\lambda \\ z_j \text{ if } |b_j| > a\lambda \end{cases}$$

If a>2,

$$\lambda < \frac{a-1}{a-2}(z_j - \frac{a}{a-1}\lambda) \leq a\lambda$$

$$\frac{a-2}{a-1}\lambda < z_j - \frac{a}{a-1}\lambda \leq \frac{a-2}{a-1}a\lambda$$

$$\frac{a-2}{a-1}\lambda + \frac{a}{a-1}\lambda < z_j \leq \frac{a-2}{a-1}a\lambda + \frac{a}{a-1}\lambda$$

$$\frac{2a-2}{a-1}\lambda < z_j \leq \frac{a^2-a}{a-1}\lambda$$

$$2\lambda < z_j \leq a\lambda$$

Similarly,

$$-a\lambda \leq \frac{a-1}{a-2}(z_j + \frac{a}{a-1}\lambda) < -\lambda$$

$$-\frac{a-2}{a-1}a\lambda \leq z_j + \frac{a}{a-1}\lambda < -\frac{a-2}{a-1}\lambda$$

$$-\frac{a-2}{a-1}a\lambda - \frac{a}{a-1}\lambda \leq z_j < -\frac{a-2}{a-1}\lambda - \frac{a}{a-1}\lambda$$

$$\frac{a-a^2}{a-1}\lambda \leq z_j < \frac{2-2a}{a-1}\lambda$$

$$-a\lambda \leq z_j < -2a\lambda$$

$$b_j = \begin{cases} z_j + \lambda \text{ if } -\lambda \le b_j \le 0 \text{ or } -2\lambda \le z_j < \lambda \\ 0 \text{ if } z_j \in [-\lambda, \lambda] \\ z_j - \lambda \text{ if } 0 < b_j \le \lambda \text{ or } \lambda < z_j \le 2\lambda \\ \frac{a-1}{a-2}(z_j - \frac{a}{a-1}\lambda) \text{ if } \lambda < b_j \le a\lambda \text{ or } 2\lambda < z_j \le a\lambda \\ \frac{a-1}{a-2}(z_j + \frac{a}{a-1}\lambda) \text{ if } -a\lambda \le b_j < -\lambda \text{ or } -a\lambda \le z_j < -2\lambda \\ z_j \text{ if } |b_j| > a\lambda \text{ or } |z_j| > a\lambda \end{cases}$$

where a is additionally constrained on a>2

More concisely,

$$b_j = \begin{cases} ST(z_j, \lambda) \text{ if } |z_j| \le 2\lambda \\ \frac{ST(z_j, \frac{a}{a-1}\lambda)}{1 - \frac{1}{a-1}} \text{ if } 2\lambda < |z_j| \le a\lambda \\ z_j \text{ if } |z_j| > a\lambda \end{cases}$$

Let $r = y - b_0 - \sum_{k=1}^{p} b_k x_k$

$$z_j = \frac{1}{n-1}(r + b_j x_j)' x_j = \frac{1}{n-1} r' x_j + \frac{1}{n-1} b_j x_j' x_j = \frac{1}{n-1} r' x_j + b_j$$

Procedure

Step 1: $\beta^{(0)} = 0$,

Step 2: $\beta_0^{(0)} = \bar{y}, r = y - \bar{y}$

Step 3: Repeat until $||\beta^{(k+1)} - \beta^{(k)}||_2$ is small or for a fixed number of iterations. For each iteration k,

(a) For j=1,...,p,

$$\text{(i) } z_j = \frac{1}{n-1} r' x_j + \beta_j^{(k-1)}$$

$$\text{(ii) } \beta_j^{(k)} = \begin{cases} ST(z_j, \lambda) \text{ if } |z_j| \le 2\lambda \\ \frac{ST(z_j, \frac{a}{a-1}\lambda)}{1 - \frac{1}{a-1}} \text{ if } 2\lambda < |z_j| \le a\lambda \\ z_j \text{ if } |z_j| > a\lambda \end{cases}$$

$$\text{(iii) Set } r \leftarrow r - (\beta_j^{(k)} - \beta_j^{(k-1)}) x_j$$

## Appendix B

### 1. Residual mean deviance of decision trees

$$\text{Residual mean deviance} = \frac{RSS}{n - |T|}$$

where |T| is the number of terminal nodes

2. **Correlation between principal component and original feature**
   X is assumed to be standardised.

$$
\begin{aligned}
\rho_{z_j,x_k} &= \frac{cov(\sum_{j=1}^{p} e_{ij}x_j, x_k)}{\sqrt{\lambda_j}\sqrt{\sigma_{kk}}} \\
&= \frac{\sum_{j=1}^{p} e_{ij}cov(x_j, x_k)}{\sqrt{\lambda_j}\sqrt{\sigma_{kk}}} \\
&= \frac{\text{kth row of } \Sigma e_j}{\sqrt{\lambda_j}\sqrt{\sigma_{kk}}} \\
&= \frac{\text{kth row of } \lambda e_j}{\sqrt{\lambda_j}\sqrt{\sigma_{kk}}} \\
&= \frac{e_{jk}\lambda_j}{\sqrt{\lambda_j}\sqrt{\sigma_{kk}}} \\
&= \frac{e_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}
\end{aligned}
$$

3. **Explanation and proof for PLS**
   Restating the procedure of PLS,
   Procedure
   Step 1. Standardize/demean X
   Step 2. Set $x_j^{(0)} = x_j, j = 1, ..., p$
   Step 3. For m=1,...,M

$$\text{(a) Compute } \phi_{mj} = \langle x_j^{(m-1)}, y \rangle \text{ for each j}$$

$$\text{(b) Construct } z_m = \sum_{j=1}^{p} \phi_{mj} x_j^{(m-1)}$$

$$\text{(c) } x_j^{(m)} = x_j^{(m-1)} - \frac{\langle z_m, x_j^{(m-1)} \rangle}{\langle z_m, z_m \rangle} z_m$$

$$\text{Step 4. } \hat{y} = \bar{y}1 + \sum_{m=1}^{M} \frac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle} z_m$$

For each standardized/demeaned predictor,

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_j^{(i)} - \bar{x}_j)^2 = \frac{\langle x_j, x_j \rangle}{n-1}$$

Regressing y on $x_j$,

$$\begin{aligned}\text{regression coefficient of } x_j &= (x_j^{'} x_j)^{-1} x_j^{'} y \\ &= \frac{\langle x_j, y\rangle}{\langle x_j, x_j\rangle} \\ &= \frac{\langle x_j, y\rangle}{n-1}\end{aligned}$$

The weights of the first PLS component are equally proportional to the dot product of the corresponding demeaned predictor and the response variable. Exact coefficients are not required since the scaling of PLS components in the final regression model does not affect prediction outcome.

For your information only,

$$\begin{aligned} corr(x_j, y) &= \frac{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j) y_i - \sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)\bar{y}}{\sqrt{(n-1)\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\langle x_j, y\rangle}{\sqrt{(n-1)\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &\propto \text{regression coefficient of } x_j \text{ when y is regressed on } x_j \end{aligned}$$

At the end of the first iteration,

$$\begin{aligned} residuals_{x_j^{(0)}} &= x_j^{(0)} - \frac{\langle z_1, x_j^{(0)}\rangle}{\langle z_1, z_1\rangle} z_1 \\ &= x_j^{(1)} \end{aligned}$$

Assuming $\bar{x}_j^{(k)} = 0$,

$$\begin{aligned} \bar{x}_j^{(k+1)} &= \bar{x}_j^{(k)} - \frac{\langle z_{k+1}, x_j^{(k)}\rangle}{\langle z_{k+1}, z_{k+1}\rangle} \bar{z}_{k+1} \\ &= 0 - \frac{\langle z_{k+1}, x_j^{(k)}\rangle}{\langle z_{k+1}, z_{k+1}\rangle} \sum_{j=1}^{p} \phi_{(k+1),j} \bar{x}_j^{(k)} \\ &= 0 \end{aligned}$$

Iteratively, this shows that $\bar{x}_j^{(k)} = 0$ for k=0,1,2,...,M by mathematical induction (base step is direct from demeaning of predictors). This simplifies the construction of weights of all PLS components (not just applicable to the first PLS component) to the dot product between y and $x_j^{(k)}$ at the kth iteration.

For each $t \in Z_{\geq 1}$, let P(t) be the proposition that

$$\langle x_j^{(t)}, z_i \rangle = 0 \text{ for i=1,2,...,t and j=1,2,...,p}$$

(Base step)
P(1) is true because

$$\begin{aligned}
\langle x_j^{(1)}, z_1 \rangle &= \langle x_j^{(0)} - \frac{\langle z_1, x_j^{(0)} \rangle}{\langle z_1, z_1 \rangle} z_1, z_1 \rangle \\
&= \langle x_j^{(0)}, z_1 \rangle - \frac{\langle z_1, x_j^{(0)} \rangle}{\langle z_1, z_1 \rangle} \langle z_1, z_1 \rangle \\
&= 0
\end{aligned}$$

(Induction step)
Let $k \in \mathbb{Z}_{\geq 1}$ such that P(k) is true i.e.

$$\langle x_j^{(k)}, z_i \rangle = 0 \text{ for i=1,2,...,k, j=1,2,...,p}$$

$$\begin{aligned}
\langle x_j^{(k+1)}, z_i \rangle &= \langle x_j^{(k)} - \frac{\langle z_{k+1}, x_j^{(k)} \rangle}{\langle z_{k+1}, z_{k+1} \rangle} z_{k+1}, z_i \rangle \\
&= \langle x_j^{(k)}, z_i \rangle - \frac{\langle z_{k+1}, x_j^{(k)} \rangle}{\langle z_{k+1}, z_{k+1} \rangle} \langle z_{k+1}, z_i \rangle \\
&= -\frac{\langle z_{k+1}, x_j^{(k)} \rangle}{\langle z_{k+1}, z_{k+1} \rangle} \langle \sum_{j=1}^{p} \phi_{(k+1),j} x_j^{(k)}, z_i \rangle \\
&= 0
\end{aligned}$$

When i=k+1,

$$\begin{aligned}
\langle x_j^{(k+1)}, z_{k+1} \rangle &= \langle x_j^{(k)} - \frac{\langle z_{k+1}, x_j^{(k)} \rangle}{\langle z_{k+1}, z_{k+1} \rangle} z_{k+1}, z_{k+1} \rangle \\
&= \langle x_j^{(k)}, z_{k+1} \rangle - \frac{\langle z_{k+1}, x_j^{(k)} \rangle}{\langle z_{k+1}, z_{k+1} \rangle} \langle z_{k+1}, z_{k+1} \rangle \\
&= 0
\end{aligned}$$

Hence, $\forall n \in Z_{\geq 1}$ $\mathrm{P(t)}$ is true.

WLOG, assume that the lastest derived principal component is in the last column of Z i.e. $z_i$ is orthogonal to $x_j^{(t-1)}$ for i=1,2,..,(t-1),

$$\begin{aligned}
\hat{\beta}^{(t-1)} &= (Z'Z)^{-1}Z'x_j^{(t-1)} \\
&= (\begin{bmatrix} Z'_{-t} \\ z'_t \end{bmatrix} [\, Z_{-t} \quad z_t \,])^{-1} \begin{bmatrix} Z'_{-t} \\ z'_t \end{bmatrix} x_j^{(t-1)} \\
&= (\begin{bmatrix} Z'_{-t} \\ z'_t \end{bmatrix} [\, Z_{-t} \quad z_t \,])^{-1} \begin{bmatrix} Z'_{-t}x_j^{(t-1)} \\ z'_t x_j^{(t-1)} \end{bmatrix} \\
&= \begin{bmatrix} Z'_{-t}Z_{-t} & Z'_{-t}z_t \\ z'_t Z_{-t} & z'_t z_t \end{bmatrix}^{-1} \begin{bmatrix} Z'_{-t}x_j^{(t-1)} \\ z'_t x_j^{(t-1)} \end{bmatrix} \\
&= \begin{bmatrix} Z'_{-t}Z_{-t} & 0 \\ 0 & z'_t z_t \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ z'_t x_j^{(t-1)} \end{bmatrix} \\
&= \begin{bmatrix} (Z'_{-t}Z_{-t})^{-1} & 0 \\ 0 & (z'_t z_t)^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ z'_t x_j^{(t-1)} \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ (z'_t z_t)^{-1} z'_t x_j^{(t-1)} \end{bmatrix}
\end{aligned}$$

4. **Theorem supporting SIR's assumption**

Theorem

If distribution of standardised X is elliptically symmetric and $p(y|x) = f(\eta_1' x, \eta_2' x, ..., \eta_K' x)$ for some function $f: R^K \to R$, then the centered inverse regression curve lies in the subspace spanned by $\{\eta_1, \eta_2, ..., \eta_K\}$.

Box Cox transformations can be used to transform X variables to become more normally-distributed.

5. **Some other optimisers for neural networks**

Nesterov accelerated gradient

Procedure

Step 1: Initialize $\theta^{(0)}$ randomly. Set $v^{(0)} = 0, t = 0$.

Step 2. While training loss not converged,

$$\begin{aligned}
&(\mathrm{a})\ t \leftarrow t+1 \\
&(\mathrm{b})\ v^{(t)} \leftarrow \gamma v^{(t-1)} + \eta \nabla_\theta L(\theta^{(t-1)} - \gamma v^{(t-1)}) \\
&(\mathrm{c})\ \theta^{(t)} \leftarrow \theta^{(t-1)} - v^{(t)}
\end{aligned}$$

**Gradient**

Adagrad

Step 1: Initialize $\theta^{(0)}$ randomly. Set $t = 0, G^{(0)} = 0$.
Step 2. While training loss not converged,

(a) $t \leftarrow t + 1$
(b) $g^{(t)} \leftarrow \nabla_\theta L(\theta^{(t-1)})$
(c) $G^{(t)} \leftarrow G^{(t-1)} + g^{(t)} \circ g^{(t)}$ where $\circ$ represent the pointwise (Hadamard) product
(d) $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta g^{(t)} \oslash (\sqrt{G^{(t)}} + \epsilon)$ where $\oslash$ represents the pointwise (Hadamard) division

The square root in 2(d) is applied elementwise.

The default is $\epsilon = 10^{-8}, \eta = 0.01$. Adagrad eliminates the need for $\eta$ to be tuned.

Adadelta

Step 1: Initialize $\theta^{(0)}$ randomly. Set $t = 0, h^{(0)} = 0$.
Step 2. While training loss not converged,

(a) $t \leftarrow t + 1$
(b) $g^{(t)} \leftarrow \nabla_\theta L(\theta^{(t-1)})$
(c) $h^{(t)} \leftarrow \gamma h^{(t-1)} + (1-\gamma) g^{(t)} \circ g^{(t)}$ where $\circ$ represent the pointwise (Hadamard) product
(c) $s^{(t)} \leftarrow -\eta g^{(t)} \oslash (\sqrt{h^{(t)}} + \epsilon)$ where $\oslash$ represents the pointwise (Hadamard) division
(c) $u^{(t)} \leftarrow \gamma u^{(t-1)} + (1-\gamma) s^{(t)}$
(d) $\theta^{(t)} \leftarrow \theta^{(t-1)} - g^{(t)} \circ u^{(t-1)} \oslash (\sqrt{h^{(t)} + \epsilon})$

where $h^{(t)}$ estimates $E(g^2)$, $u^{(t)}$ estimates $E(\Delta\theta^2)$ and $\sqrt{h^{(t)} + \epsilon}$ estimates root mean squared (RMS) error at iteration t
The default is $\gamma = 0.9$.

Root mean squared propagation (RMSprop)
Procedure

Step 1: Initialize $\theta^{(0)}$ randomly. Set $v^{(0)} = 0, t = 0$.
Step 2. While training loss not converged,

(a) $t \leftarrow t + 1$
(b) $g^{(t)} \leftarrow \nabla_\theta L(\theta^{(t-1)})$
(c) $v^{(t)} \leftarrow \beta_2 v^{(t-1)} + (1-\beta_2) g^{(t)} \circ g^{(t)}$ where $\circ$ represent the pointwise (Hadamard) product
(d) $\hat{v}^{(t)} \leftarrow \dfrac{v^{(t)}}{(1-\beta_2^t)}$
(e) $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta g^{(t)} \oslash (\sqrt{\hat{v}^{(t)}} + \epsilon)$ where $\oslash$ represents the pointwise (Hadamard) division

The default is $\beta_2 = 0.9, \epsilon = 10^{-7}$. Only $\eta$ needs to be tuned.

**Appendix C**
**Predictors**

| | Abbreviation | Name |
|---|---|---|
| 1 | absacc | Absolute accruals |
| 2 | acc | Working capital accruals |
| 3 | age | No. of years since first Compustat coverage |
| 4 | agr | Asset growth |
| 5 | b/m | Book-to-market (macro) |
| 6 | baspread | Bid-ask spread |
| 7 | bm | Book-to-market |
| 8 | cash | Cash holdings |
| 9 | cashdebt | Cash flow to debt |
| 10 | cashpr | Cash productivity |
| 11 | cfp | Cash flow to price ratio |
| 12 | chcsho | Change in shares outstanding |
| 13 | chinv | Change in inventory |
| 14 | chmom | Change in 6-month momentum |
| 15 | chtx | Change in tax expense |
| 16 | cinvest | Corporate investment |
| 17 | convind | Convertible debt indicator |
| 18 | currat | Current ratio |

| | | |
|---|---|---|
| 19 | depr | Depreciation/PP&E |
| 20 | dfy | Default spread (macro) |
| 21 | divi | Dividend initiation |
| 22 | divo | Dividend omission |
| 23 | dolvol | Dollar trading volume |
| 24 | dp | Dividend-price ratio (macro) |
| 25 | dy | Dividend to price |
| 26 | e/p | Earnings to price (macro) |
| 27 | egr | Growth in common shareholder equity |
| 28 | ep | Earnings to price |
| 29 | gma | Gross profitability |
| 30 | grcapx | Growth in capital expenditures |
| 31 | herf | Industry sales |
| 32 | hire | concentration |
| 33 | ill | Illiquidity |
| 34 | indmom | Industry momentum |
| 35 | invest | Capital expenditures and inventory |
| 36 | lev | Leverage |
| 37 | lgr | Growth in long-term debt |
| 38 | maxret | Maximum daily return |

| | | |
|---|---|---|
| 39 | mom12m | 12-month momentum |
| 40 | mom1m | 1-month momentum |
| 41 | mom36m | 36-month momentum |
| 42 | mom6m | 6-month momentum |
| 43 | ms | Financial statement score |
| 44 | mve_ia | Industry adjusted size |
| 45 | mvel1 | Size/Market capitalization |
| 46 | nincr | Number of earnings increases |
| 47 | ntis | Net equity expansion (macro) |
| 48 | operprof | Operating profitability |
| 49 | orgcap | Organizational capital |
| 50 | pchcurrat | % change in current ratio |
| 51 | pchdepr | % change in depreciation |
| 52 | pchgm_pchsale | % change in gross margin - % change in sale |
| 53 | pchquick | % change in quick ratio |
| 54 | pchsale_pchrect | % change in sale - % change in A/R |

| | | |
|---|---|---|
| 55 | pchsale_pchxsga | % change in sale - % change in SG&A |
| 56 | pctacc | Percent accruals |
| 57 | ps | Financial statements score |
| 58 | quick | Quick ratio |
| 59 | rd | R&D increase |
| 60 | ret | Monthly return |
| 61 | retvol | Return volatility |
| 62 | roaq | Return on assets |
| 63 | roavol | Earnings volatility |
| 64 | roeq | Return on equity |
| 65 | roic | Return on invested capital |
| 66 | rsup | Revenue surprise |
| 67 | salecash | Sales to cash |
| 68 | saleinv | Sales to inventory |
| 69 | salerec | Sales to receivables |
| 70 | securedind | Secured debt indicator |
| 71 | sgr | Sales growth |
| 72 | sin | Sin stocks |
| 73 | sp | Sales to price |
| 74 | std_dolvol | Volatility of liquidity (dollar trading volume) |
| 75 | std_turn | Volatility of liquidity (share turnover) |

| | | |
|---|---|---|
| 76 | svar | Stock variance (macro) |
| 77 | tang | Debt capacity/firm tangibility |
| 78 | tb | Tax income to book income |
| 79 | tbl | 3-months US treasury bill-rate (macro) |
| 80 | tms | Term spread (macro) |
| 81 | turn | Share turnover |
| 82 | zerotrade | Zero trading days |