MLR.

Theory.

$$Y_i = \beta_0 + \beta_1 X + \beta_2 X + \dots + \beta_p X_p + \varepsilon_i$$

$$E[Y] = \beta_0 + \beta X + \beta_2 X + \dots + \beta_p X_p$$

Obs: linear model means that variables have an additive effect.

General Linear Regression Model can be.

1) → X are continous variables

2) → X are qualitative variables (gender, color)

3) → polynomial regression

→ some $X_i = X_{i-j}^{R}$ where $j \geq 1$

4) → transformed variables

→ $X_i' = f(X_i)$ and

$$Y_i = \beta_0 + \beta_1 X_1 + \dots + \beta_K f(X_2) + \dots + \beta_p X_p + \varepsilon_i$$

5) → interaction effects

→ $X'_i = X_i \cdot X_j$ with $i \neq j$

$$Y_i = \beta_0 + \beta_1 X_1 + \dots + \beta_i X_i + \beta_j X_j + \beta_{j+1} X_i X_j + \varepsilon_i$$

Matrix formulation:

$$Y = Xb + \varepsilon$$

$$\sigma^2\{e\} = \sigma^2 \cdot I$$

$$\left.\begin{array}{l} E[Y] = X\beta \\ \sigma^2\{Y\} = \sigma^2 I \end{array}\right\}$$

$$Y = Xb \Rightarrow X'Y = X'Xb$$

$$\Rightarrow (X'X)^{-1}(X'Y) = b$$

$$\hat{Y} = Xb$$

$$e = Y - \hat{Y} = Y - Xb = HY$$

$$H = X(X'X)^{-1}X'$$

$$e = (I - H)Y$$

$$\sigma^2[e] = \sigma^2(I - H)$$

$$s^2\{e\} = MSE\{I - H\}.$$

ANOVA for MLR

$$SST = \sum (Y_i - \bar{Y})^2 = Y'Y - \left(\frac{1}{n}\right)Y'JY$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = Y'Y - b'X'Y$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 = b'X'Y - \left(\frac{1}{n}\right)Y'JY$$

$$E\left[\frac{SSE}{n-p}\right] = E\{MSE\} = \sigma^2$$

$$E\left[\frac{SSR}{p-1}\right] = E\{MSR\} = \sigma^2 + \frac{1}{2}\left[\sum_{j=1}^{p} \beta_j \left(X_{ij} - \bar{X}_j\right)^2 + 2\sum_{j,k} \beta_j \beta_k \sum \left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)\right]$$

$$\left.\begin{array}{l} MSE \sim \chi^2(n-p) \\ MSR \sim \chi^2(p-1) \end{array}\right\} \Rightarrow F = \frac{MSR}{MSE} \sim F(p-1, n-p)$$

Coefficient of multiple Determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$0 \leq R^2 \leq 1$$

$$\text{Adjusted } R^2 = 1 - \frac{MSE}{MST} = 1 - \frac{n-1}{n-p} \frac{SSE}{SST}$$

Inference about Regression parameters

LSE & MLE estimators are unbiased.

$$E[b] = \beta$$

$$\sigma^2[b] = \begin{bmatrix} \sigma^2[b_0] & \sigma(b_0 b_1) & \cdots & \sigma(b_0, b_{p-1}) \\ \sigma(b_{p-1}, b_0) & \sigma(b_{p-1}, b_1) & \cdots & \sigma^2(p-1) \end{bmatrix}$$

$$\sigma^2\{b\} = \sigma^2 (X^t X)^{-1}$$

$$s^2(b) = MSE (X^t X)^{-1}$$

$$\frac{b_k - \beta_k}{s(b_k)} \sim t(n-p).$$

Assumption of linear regression.

1. The relanship is linear
2. There is no multicolinearity (predictor are independent algebraicaly $\Rightarrow \exists$ inv
   → the inverse of $X^t X$ exists.
3. the residuals are normaly distributed
4. residuals are not correlated (no auto-correlation) sequential
5. Variance is constant (homoskedasticity).

steps in developing linear regression models.

① plot.
② fit regression using information from plots.
③ check significance
④ check computational issues.
   → outliers for LSE
   → difference in magnitude → standardized model.
⑤ check assumptions.
⑥ fix or mitigate violation of assumptions or computational issues.
⑦ redo step 2-6 with a different regression equation.
⑧ select the final model from the models that have a solution in step 6.
⑨. Validate the model on independent data.

Heter et. al presents a simplified version in Fig 8.1
① Data collection and preparation
② Reduction of explanatory or predictor variables
③ Model refinement and selection
④ Model validation.

In this class we focus on model refinement & selection.

Plot is offering you a first insight on the relationship between predicted and predictor, usualy a one-to-one relationship (very rare predicted vs. 2 predictor).

Brute force assessment = all possible regressions.

$y = Xb \Rightarrow$

$y = x_1$

$y = x_2$

$y = x_p$

$y = x_1\ x_2$

$y = x_1\ x_3$

$y = x_1 \cdots x_p$

$y = x_2\ x_3$

$y = x_{p-1}\ x_p$.

$y = x_1\ x_2 \cdots x_p$.

Total # regressions = ? $= 0 + C_p^1 + C_p^2 + \cdots + C_p^{p-1} + C_p^p = 2^p$

selection among candidates using various statistics.

(1) R_p^2 or SSE_p.

(2) MSE_p or R_a^2

(3) C_p criterion

(4) $PRESS_p$ criterion.

① $R^2 = 1 - \frac{SSE}{SST}$

$R^2_p = 1 - \frac{SSE_p}{SST}$ stands for the p variable(s).

R^2_p is not intended to identify which subset is "the best" (maximizes or minimizes a criterion)

R^2 increases with the # variables

R^2_p is intended for finding the point where adding more explanatory variables is not worthy.

② MSE_p or R_a criterion.

$R^2_a = 1 - \frac{n-1}{n-p}\frac{SSE}{SST}$

$MSE_p = \frac{SSE_p}{n-p-1}$ where p is the # variables

Example.

$MSE_1 = \frac{SSE(X_1)}{n-2}$

$MSE_{1,2} = \frac{SSE(X_1, X_2)}{n-3}$

Interpretation with scree plot.

③ C_p - criterion.

C_p is concerned with total mean square error of the n fitted values for each subset regression model.

error in fitted value is: $\hat{Y}_i - \mu_i = \hat{Y}_i - E[\hat{Y}_i]$

$MSE_{\hat{Y}_i - \mu_i} = E\left[(\hat{Y}_i - \mu_i)\right]^2 = E\left[(E[Y_i] - \mu_i) + (\hat{Y}_i - E[\hat{Y}_i])\right]^2 =$

$= (E[\hat{Y}_i] - \mu_i)^2 + \sigma^2[\hat{Y}_i]$

$$\sum\left[\left(E[\hat{Y}_i] - \mu_i\right)^2 + V(\hat{Y}_i)\right] = \sum\left(E[Y_i] - \mu_i\right)^2 + \sum V(\hat{Y}_i)$$

$$\Gamma_p = \frac{MSE_{\hat{Y}_i} - \mu_i}{\sigma^2}$$

an unbiased estimator of Γ_p is C_p

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots X_{p-1})} - (n - 2p)$$

if there is no bias then $E\{\hat{Y}_i\} = \mu_i$ and

$E\{C_p\} \cong p$ (Note) C_p identifies subset of X for predicting Y

(4) PRESS p criterion.

PRESS = prediction of sum of squares

PRESS measures how well the use of fitted values for a subset model can predict the observed Y.

Steps: (1) In PRESS the ith observation is deleted and the regression is estimated from the n−1 observations.

(2) Predict the ith observation with the fitted regression

(3) compute PRESS statistics:

$$PRESS_p = \sum (Y_i - \hat{Y}_{i(i)})^2$$

where $\hat{Y}_{i(i)}$ is the predicted Y_i from the model with no ith case.

Decision: Small PRESS indicates good model.

AUTOMATIC SEARCH PROCEDURES for variable reduction

Identify the "best" subset of variables from a large # of predictors.

① Forward selection procedure.

- is an algorithm that is based on F test.

Step 1. Fit a SLR for each p variable

$$y = b_0 + b_1 x$$

$$F^* = \frac{MSR(X_k)}{MSE(X_k)}$$

where for SLR $MSR(X_k) = SSR(X_k)$

The variable with LARGEST F^* is selected.

Step 2. Fit all regression models with three selected variable in Step 1

let say X_{s1} was selected

$$y = b_0 + b_1 X_{s1} + X_k$$

compute $$F^* = \frac{MSR(X_k | X_{s1})}{MSE(X_{s1}, X_k)}$$

where $MSR(X_k | X_{s1})$ is MSR with the variable X_k and X_{s1} (X_k given X_{s1})

$MSE(X_{s1}, X_k)$ is MSE for the regression with 2 variables X_k and X_{s1}

$$F^* = \left(\frac{b_k}{s(b_k)}\right)^2.$$

Among all models select the one with the largest F^*, let say X_{s2}

Step 3. Asses if any variable should be dropped from the model.

This is done to eliminate variable whose participation in the model is "covered" by other variables or combination of variables.

If no variable is dropped then both X_{S_1} & X_{S_2} are kept in the model

Step 4. Add a new variable, and execute the regression $Y = b_0 + b_1 X_{S_1} + b_2 X_{S_2} + b_3 X_K$ among all $F = \frac{MSR(X_K | X_{S_1}, X_{S_2})}{MSE(X_{S_1}, X_{S_2}, X_K)}$

select the largest (if significant).

Step 5. → Go to Step 3.
→ Stop if no new variable is added or dropped

Step 6. → Go to Step 4.
→ stop if no new variable is added.

Forward selection procedure

Same as stepwise but only add variable, do not drop them (looks only forward)

Backward selection procedure

Start with all variables in and drop one variable at a time.

there is no addition of variables.