Among several plausible explanations for a phenomenon, the simplest is the best.

Variable Selection

select the best subset of predictors to explain the data in the simplest way
unnecessary predictors add noise to the estimation of other quantities that interest us, wasting degrees of freedom
collinearity is caused by having too many variables provide the same information
save time any money by not measuring redundant predictors

Model Hierarchy
lower order terms should not be removed before higher order terms of the same variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

x is not significant but x^2 is
remove x term to get reduced model

$$y = \beta_0 + \beta_2 x^2 + \varepsilon$$

if you make a scale change from x to $(x + a)$, model becomes

$$y = \beta_0 + \beta_2 a^2 + 2\beta_2 a x + \beta_2 x^2 + \varepsilon$$

the first order x term reappears
scale changes shouldn't make any important changes to the model

Stepwise Testing

compares successive models
$\alpha_{criteria}$ doesn't have to be 5%, 15-20% cutoff may work best
possible to miss the optimal model because adding/dropping one variable at a time
no multiple testing correction so p-values should not be assumed to be the Type I Error
procedures are not directly linked to final objectives of prediction or explanation so may not help solve the problem of interest

Backward Elimination/Top Down

Step 1	start with all the predictors in the model
Step 2	remove the predictor with the highest p-value > $\alpha_{criteria}$
Step 3	refit the model
Step 4	repeat Steps 2-3 until all non-significant predictors are removed

Forward Selection/Bottom Up

Step 1	start with no predictors
Step 2	add the predictor with the lowest p-value < $\alpha_{criteria}$
Step 3	refit the model
Step 4	repeat Step 2-3 until no new predictors can be added

Stepwise Regression
combination of backward elimination and forward selection
at each stage, a variable may be added or removed

top down stepwise regression	alternate drop step and add step
bottom up stepwise regression	alternate add step with drop step

Among several plausible explanations for a phenomenon, the simplest is the best.

Criterion Testing

find the model that optimizes the measure of goodness of fit and reduces the model's complexity
prefer small values of AIC and BIC with smaller RSS and a small number of parameters
larger models fit better so will have smaller RSS but use more parameters
need to find a balance between RSS and number of parameters
BIC penalizes larger models more heavily and tend to prefer smaller models compared to AIC

Akaike Information Criterion (AIC) $\quad n \times log\left(\frac{RSS}{n}\right) + 2p'$

Bayes Information Criterion (BIC) $\quad n \times log\left(\frac{RSS}{n}\right) + log(n)p'$

Adjusted R^2

$$R^2 = \frac{SS_{reg}}{S_{YY}} = 1 - \frac{RSS}{S_{YY}}$$

adding variables can only decrease RSS and increase R^2
R^2 is not a good criterion because it always prefers the larger model
use significant changes of RSS and the criterion

$$R_a{}^2 = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{S_{YY}}{n-1}} = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2) = 1 - \frac{\hat{\sigma}_{Model}{}^2}{\hat{\sigma}_{Null}{}^2}$$

adding a predictor will only increase $R_a{}^2$ if it has some predictive value

Mallow's C_p Statistics

average mean square error of prediction = $\frac{1}{\sigma^2}\sum E\left(\hat{y}_i - E(y_i)\right)^2$

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

$\hat{\sigma}^2$ is from the model with all the predictors
RSS_p is from a model with p parameters

for the full model, $C_p = p$
if a p-predictor model fits, then $E[RSS_p] = (n-p)\sigma^2$ and $E(C_p) \approx p$
if a model has a bad fit, C_p will be much larger than p
models with a good fit will have small p and a C_p around or less than p