HW 7

1.Model with *lpsa* as the response and the *other variables* as predictors from *prostate* data.

The full model contains both significant (lcavol,lweight,svi) & insignificant (age,lbph,lcp,gleason, pgg45) predicotrs (alpha=0.05). The R^2 is farily high 65.48% with a residual standard error of 0.7084. The coefficients for *age* & *lcp* are negative, whereas the coefficients for *lcavol*,*lweight*, *lbph*,*svi*,*gleason*, and *pgg45* are positive.

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

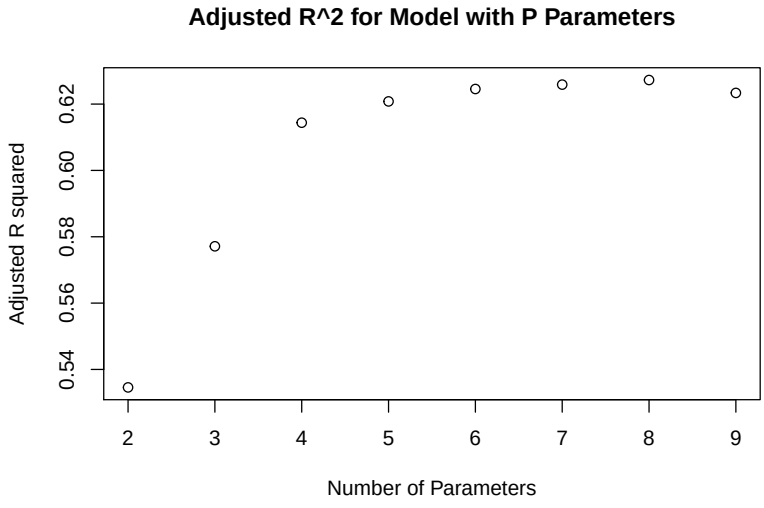
A. Backward Elimination- The final model given by backward substitution has as significant predicotrs (in order of importance): **lcavol**,**lweight**, and **svi**. All coefficients for the beta^ estimates are positive, but the intercept is negative. The R^2 is fairly high 62.64 and the residual standard error is 0.7168.

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

B. Adjusted R^2- The model with the highest Adjusted R^2 is the model with a total of 8 parameters. This means that the final model will have 7 predictors, and such ones being (in order of importance)- **lcavol**, **svi**,**lweight**,**lbph**, **age**, **pgg45**, and **lcp**. Out of the 7 predictors only the first three are indeed significant (lcavol,svi,lweight); the rest are insignificant. The coefficients for the beta^ estimates are all positive with the exception of age & lcp (negative). The R^2 is fairly high 65.44 and the residual standard error is 0.7048.

```
## Subset selection object
## Call: regsubsets.formula(lpsa ~ ., data = prostate)
## 8 Variables  (and intercept)
##         Forced in Forced out
## lcavol      FALSE      FALSE
## lweight     FALSE      FALSE
## age         FALSE      FALSE
## lbph        FALSE      FALSE
## svi         FALSE      FALSE
## lcp         FALSE      FALSE
## gleason     FALSE      FALSE
## pgg45       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          lcavol lweight age lbph svi lcp gleason pgg45
## 1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "
## 2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "
## 3  ( 1 ) "*"    "*"     " " " "  "*" " " " "     " "
## 4  ( 1 ) "*"    "*"     " " "*"  "*" " " " "     " "
## 5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "
## 6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     "*"
## 7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"
## 8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"
```

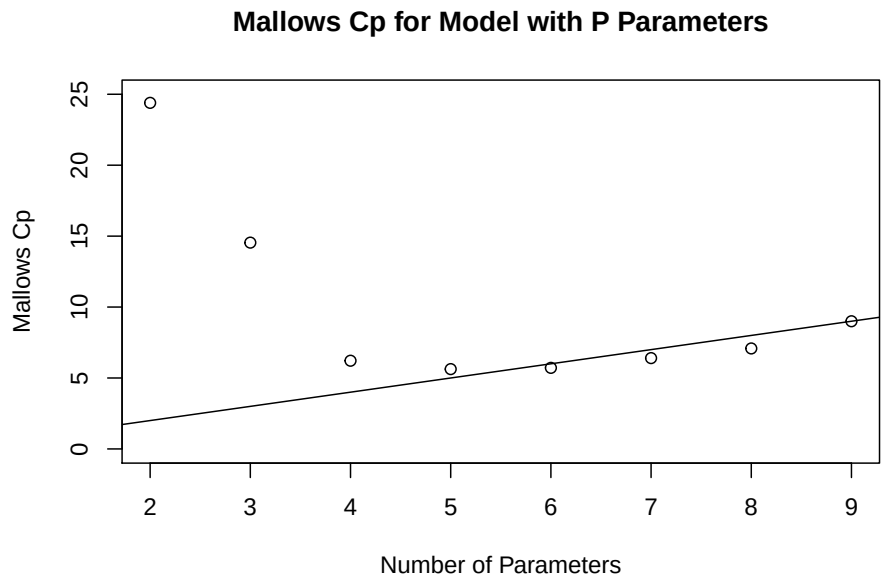


```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + pgg45 +
##     lcp, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## svi          0.757734   0.241282   3.140  0.00229 **
## lbph         0.107671   0.058108   1.853  0.06720 .
## age         -0.019336   0.011066  -1.747  0.08402 .
## pgg45        0.005318   0.003433   1.549  0.12488
## lcp         -0.104482   0.090478  -1.155  0.25127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

C. Mallow's Cp - The model with the minimum Mallow's Cp is the model with a total of 6 parameters. This means that the final model will have 5 predicotrs, and such ones being (in order of importance)- **lcavol**, **lweight**,**svi**,**lbph**, and **age**. Out of the 5 predictors only the first three are indeed significant (lcavol,svi,lweight); the rest are insignificant. The coefficients for the beta estimates are all positive with the exception of age(negative). The R^2 is fairly high 64.41 and the residual standard error is 0.7073.

```
## Subset selection object
## Call: regsubsets.formula(lpsa ~ ., data = prostate)
## 8 Variables  (and intercept)
##         Forced in Forced out
## lcavol      FALSE      FALSE
## lweight     FALSE      FALSE
## age         FALSE      FALSE
## lbph        FALSE      FALSE
## svi         FALSE      FALSE
## lcp         FALSE      FALSE
## gleason     FALSE      FALSE
## pgg45       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          lcavol lweight age lbph svi lcp gleason pgg45
## 1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "
## 2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "
## 3  ( 1 ) "*"    "*"     " " " "  "*" " " " "     " "
## 4  ( 1 ) "*"    "*"     " " "*"  "*" " " " "     " "
## 5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "
## 6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     "*"
## 7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"
```

```
## 8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"
```




```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## svi          0.72095    0.20902   3.449 0.000854 ***
## lbph         0.11184    0.05805   1.927 0.057160 .
## age         -0.01489    0.01075  -1.385 0.169528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

Model Comparison

Selected Models

- All best models from backwards substitution, adjusted R^2, and Mallow's Cp tend to have similar estimates & signs for the coefficients that they do share - with the exception of the **intercept** for backward sub (negative).

- The order of significance for predictors (**lcavol**, **lweight**, and **svi**) seems to be simlar across all modles- with the excpetion of backward subsitution for the pair (*lweight*, *svi*).
- All models, consider **gleason** the most insignificant predictor.
- Backward substituion only has significant predictors in its final suggested model, whereas both adjusted R^2 & Mallow's Cp have a mix of significant & insignifcant predictors.

Original fit

- **vs backward substitution**, boht models are similar in the coefficients, standard errors, and p-values of their shared predictors. The major difference is the amount of predictors- in the original model there are a total of 8 different predictors, whereas in the b.s. model there are only 3 predictors. Yet, the R^2 & RSE differ only slightly from each other. All of the predicors in the b.s. model are significant in the orignal model. Also, the intercept is negative for b.s., but positive for the original model. This method tends to favor smaller models.
- **vs adjusted r^2**, both models have almost the same number of predictors (minus gleason). Hence, the coefficients, standard errors, p-values, are almost identical. The R^2 did slightly drop, but the RSE is the same. The adjusted R^2 model does contain unsignificant predictors.
- **vs Mallow's Cp**, both models have similar coefficients, standard errors, p-values. The R^2 & RSE did slightly drop. This model contains unsignificant predictors. Mallow's Cp tends to pick larger models given similar penalty as AIC.

Final Conclusion

- The suggestion would be to use the model suggested by *backward subsitution* with predictors **lcavol**,**lweight**, and **svi** if the scope for the analysis is *inference*.
- Instead, if the scope for the analysis is *prediction*, the suggestion would be to use the model suggested by *Mallow's Cp* with predictors **lcavol**, **lweight**,**svi**,**lbph**, and **age**.