SOLUTION Lab 6 - Added Variable Plots, Unusual Points, Choosing Predictors (4.1, 4.2 & 4.4)

P.B. Matheson adopted from A.S. Wagaman

This lab will lead you through associated topics (4.1-4.5) in Chapter 4 with example code. You should not need to write your own code unless you choose to remove the unusual points after identifying them. The lab is one long example. Work with those around you to address the questions throughout.

For this lab, we will be investigating a data set on bulls sold at auction, trying to understand the relationships between certain variables. In particular, we want to see if we can predict the sale price of the bulls.

```
Bulls <- read.table("https://pmatheson.people.amherst.edu/stat230/bulls.txt", header = T)
glimpse(Bulls)
```

```
## Rows: 76
## Columns: 9
## $ breed    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ salepr   <int> 2200, 2250, 1625, 4600, 2150, 1225, 2250, 4000, 1600, 1525, 1~
## $ yrhgt    <dbl> 51.0, 51.9, 49.9, 53.1, 51.2, 49.2, 51.0, 51.5, 50.1, 49.6, 5~
## $ ftfrbody <int> 1128, 1108, 1011, 993, 996, 985, 959, 1060, 979, 1083, 1036, ~
## $ prctffb  <dbl> 70.9, 72.1, 71.6, 68.9, 68.6, 71.4, 72.1, 69.3, 71.2, 75.8, 6~
## $ frame    <int> 7, 7, 6, 8, 7, 6, 7, 7, 6, 6, 6, 7, 7, 6, 6, 6, 6, 5, 6, 6, 6~
## $ bkfat    <dbl> 0.25, 0.25, 0.15, 0.35, 0.25, 0.15, 0.20, 0.30, 0.25, 0.30, 0~
## $ saleht   <dbl> 54.8, 55.3, 53.1, 56.4, 55.0, 51.4, 54.0, 55.6, 51.5, 54.6, 5~
## $ salewt   <int> 1720, 1575, 1410, 1595, 1488, 1500, 1522, 1765, 1365, 1640, 1~
```

The summary reveals the following variables are present: **breed** where 1- Angus, 5- Hereford, 8 - Simental, **salepr** - sale price, **yrhgt** = yearling height at shoulder (inches), **ftfrbody** = fat free body (pounds), **prctffb** = percent fat-free body, **frame** = scaled from 1-8 (1 = small, 8 = large), **bkfat** = back fat (inches), **saleht** and **salewt** = sale height at shoulder (inches) and weight (pounds).

You may have noticed that R is currently providing a NUMERIC summary for breed. That is because it doesn't know it should be a factor variable (categorical). We can adjust that as follows:

```
# Bulls <- mutate(Bulls, breed = as.factor(breed)) #quick easy but no labels (to show you how this can
Bulls <- mutate(Bulls, breed = cut(breed, breaks = c(0, 4, 6, 10), labels = c("Angus", "Hereford", "Sime
glimpse(Bulls)
```

```
## Rows: 76
## Columns: 9
## $ breed    <fct> Angus, Angus, Angus, Angus, Angus, Angus, Angus, Angus, Angus~
## $ salepr   <int> 2200, 2250, 1625, 4600, 2150, 1225, 2250, 4000, 1600, 1525, 1~
## $ yrhgt    <dbl> 51.0, 51.9, 49.9, 53.1, 51.2, 49.2, 51.0, 51.5, 50.1, 49.6, 5~
## $ ftfrbody <int> 1128, 1108, 1011, 993, 996, 985, 959, 1060, 979, 1083, 1036, ~
## $ prctffb  <dbl> 70.9, 72.1, 71.6, 68.9, 68.6, 71.4, 72.1, 69.3, 71.2, 75.8, 6~
## $ frame    <int> 7, 7, 6, 8, 7, 6, 7, 7, 6, 6, 6, 7, 7, 6, 6, 6, 6, 5, 6, 6, 6~
## $ bkfat    <dbl> 0.25, 0.25, 0.15, 0.35, 0.25, 0.15, 0.20, 0.30, 0.25, 0.30, 0~
## $ saleht   <dbl> 54.8, 55.3, 53.1, 56.4, 55.0, 51.4, 54.0, 55.6, 51.5, 54.6, 5~
## $ salewt   <int> 1720, 1575, 1410, 1595, 1488, 1500, 1522, 1765, 1365, 1640, 1~
```

You might wonder why the breaks were selected as those values. The breed values were 1, 5, and 8, so these choices for cuts separate those 3 numbers. Other values would also have worked.

Our goal is to predict **salepr** so let's begin by fitting a full model. (We are skipping checking linearity with scatterplots so we can get to the concepts here.)

```
fmfull <- lm(salepr ~ ., data = Bulls)
msummary(fmfull)
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9075.8225  4245.8207  -2.138 0.036259 *
## breedHereford  -159.3594   144.3514  -1.104 0.273615
## breedSimental  -746.6135   190.3182  -3.923 0.000211 ***
## yrhgt            95.8523   103.3010   0.928 0.356844
## ftfrbody         -1.9431     0.9952  -1.953 0.055119 .
## prctffb         -20.4921    25.7312  -0.796 0.428662
## frame           290.9162   158.7596   1.832 0.071399 .
## bkfat           940.0862   829.4844   1.133 0.261173
## saleht          130.5066    65.2020   2.002 0.049444 *
## salewt            0.3717     0.5573   0.667 0.507139
##
## Residual standard error: 419.6 on 66 degrees of freedom
## Multiple R-squared:  0.6009, Adjusted R-squared:  0.5465
## F-statistic: 11.04 on 9 and 66 DF,  p-value: 2.755e-10
```

```
car::vif(fmfull)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## breed     4.684208  2        1.471157
## yrhgt    13.631307  1        3.692060
## ftfrbody  3.626672  1        1.904382
## prctffb   3.016194  1        1.736719
## frame     9.224431  1        3.037175
## bkfat     2.351862  1        1.533578
## saleht    7.278147  1        2.697804
## salewt    2.229945  1        1.493300
```

Does this output resemble the VIF output we had previously? If not, what is different? Why might that be?

> SOLUTION: Breed is categorical, so this is a modified version of the VIF output that can deal with categorical variables. The basic premise is the same. You don't want large VIFs - same cutoff, so values > 5 are indicative of issues (given the context). You use the GVIF column.

There appear to be some multicollinearity issues, so before we continue, let's remove yearling height and frame. These two variables appear to be highly correlated with saleht. Does multicollinearity have to be only pairwise between predictor variables?

> SOLUTION: No. There can be a linear relationship between the predictors. It doesn't need to be just pairwise correlations.

```
cor(select(Bulls, yrhgt, frame, saleht))
```

```
##            yrhgt     frame    saleht
## yrhgt  1.0000000 0.9402488 0.8595129
## frame  0.9402488 1.0000000 0.8007440
## saleht 0.8595129 0.8007440 1.0000000
```

We refit the *full* model to proceed.

```
fmfull <- lm(salepr ~ breed + ftfrbody + prctffb + bkfat + saleht + salewt, data = Bulls)
msummary(fmfull)
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.180e+04  3.181e+03  -3.710  0.00042 ***
## breedHereford -1.322e+02  1.626e+02  -0.813  0.41920
## breedSimental -6.754e+02  2.137e+02  -3.160  0.00235 **
## ftfrbody      -1.494e+00  1.125e+00  -1.328  0.18849
## prctffb       -1.699e+01  2.923e+01  -0.581  0.56293
## bkfat          1.163e+03  9.394e+02   1.238  0.22003
## saleht         3.063e+02  5.556e+01   5.512 5.93e-07 ***
## salewt        -1.937e-01  6.151e-01  -0.315  0.75382
##
## Residual standard error: 476.9 on 68 degrees of freedom
## Multiple R-squared:  0.4687, Adjusted R-squared:  0.414
## F-statistic:  8.57 on 7 and 68 DF,  p-value: 1.718e-07
```

```
car::vif(fmfull)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## breed    4.500401  2        1.456508
## ftfrbody 3.585504  1        1.893543
## prctffb  3.012479  1        1.735649
## bkfat    2.334573  1        1.527931
## saleht   4.090701  1        2.022548
## salewt   2.102185  1        1.449891
```

There may still be some issues, but this is better, no VIF (GVIF) is over 5.

For this model, does it appear that all predictors are significant with the rest in the model?

> SOLUTION: Only two predictors appear to be significant with the others in the model - saleht and breedSimental (which means keep breed). Hopefully we can reduce the model size substantially (eliminate predictors).

What percentage of variability in the response does this model explain?

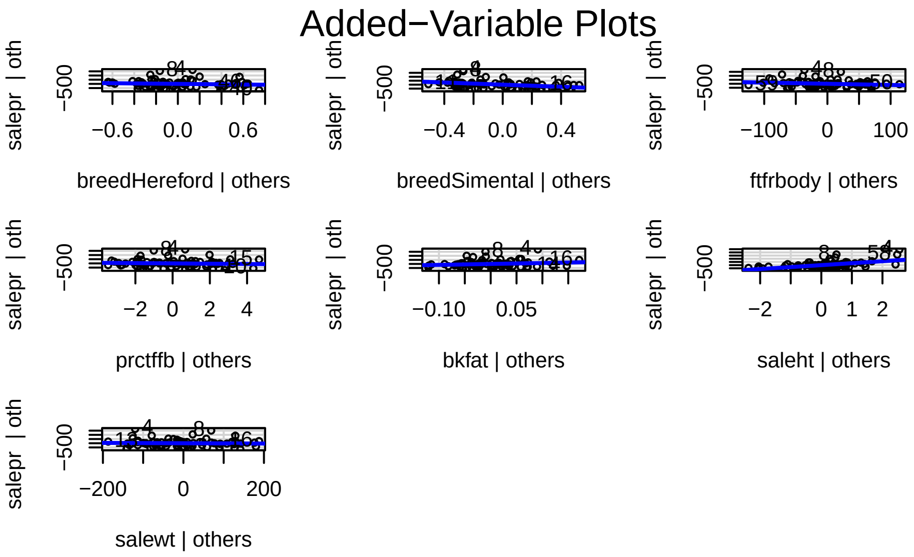
> SOLUTION: The R-squared is 0.4687. Overall this suggests predictions wouldn't be terribly accurate.

Added Variable Plots (4.1)

We can use added variable plots to see what the relationships are between predictors and the response with other variables in the model.

Investigate the added variable plots for *prctffb* and *saleht* and explain what they show you. In particular, which of these two variables would you want to keep in the model?

```
car::avPlots(fmfull)
```



SOLUTION: For prctffb, the added variable plot shows a tiny negative association between the two sets of residuals, but it isn't likely strong enough to warrant keeping prctffb in the model with the other variables included. For saleht, we see a clear positive association between the sets of residuals, which suggests saleht is important (it adds something) to the model when the other variables are included. We see a few points that may be unusual in both of these plots. We would keep saleht but probably not prctffb in the model. (And may remove other variables as well.)

Variable Selection Automated Techniques (4.2)

```
fmfull <- lm(salepr ~ breed*saleht + ftfrbody + prctffb + bkfat + salewt, data = Bulls) #note an intera
msummary(fmfull)
```

```
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -1.609e+04  4.366e+03  -3.684 0.000465 ***
## breedHereford         6.481e+03  5.877e+03   1.103 0.274086    
## breedSimental         6.474e+03  5.648e+03   1.146 0.255847    
## saleht                3.742e+02  7.313e+01   5.116 2.89e-06 ***
## ftfrbody             -1.544e+00  1.134e+00  -1.361 0.178013    
## prctffb              -9.168e+00  2.980e+01  -0.308 0.759301    
## bkfat                 8.629e+02  9.631e+02   0.896 0.373533    
## salewt               -4.711e-02  6.241e-01  -0.075 0.940065    
## breedHereford:saleht -1.248e+02  1.115e+02  -1.119 0.267104    
## breedSimental:saleht -1.319e+02  1.034e+02  -1.275 0.206753    
## 
## Residual standard error: 476.7 on 66 degrees of freedom
## Multiple R-squared:  0.4847, Adjusted R-squared:  0.4145 
## F-statistic: 6.899 on 9 and 66 DF,  p-value: 6.214e-07
```

Best Subsets and Mallow's Cp Code

```
best <- regsubsets(salepr ~ breed*saleht + ftfrbody + prctffb + bkfat + salewt, data = Bulls, nbest = 1
with(summary(best), data.frame(rsq, adjr2, cp, outmat))
```

```
##                 rsq     adjr2        cp breedHereford breedSimental saleht
## 1  ( 1 ) 0.1520597 0.1406011 36.614154                                  *
## 2  ( 1 ) 0.4084108 0.3922029  5.777695                                  *
## 3  ( 1 ) 0.4451100 0.4219896  3.076830                                  *
## 4  ( 1 ) 0.4638033 0.4335950  2.682379                                  *
## 5  ( 1 ) 0.4714634 0.4337108  3.701179                           *      *
## 6  ( 1 ) 0.4747761 0.4291045  5.276845             *             *      *
## 7  ( 1 ) 0.4839929 0.4308745  6.096252             *             *      *
## 8  ( 1 ) 0.4846999 0.4231715  8.005696             *             *      *
##          ftfrbody prctffb bkfat salewt breedHereford.saleht
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )        *
## 4  ( 1 )        *             *
## 5  ( 1 )        *             *
## 6  ( 1 )        *                                         *
## 7  ( 1 )        *             *                           *
## 8  ( 1 )        *       *     *                           *
##          breedSimental.saleht
## 1  ( 1 )
## 2  ( 1 )                    *
## 3  ( 1 )                    *
## 4  ( 1 )                    *
## 5  ( 1 )                    *
## 6  ( 1 )                    *
## 7  ( 1 )                    *
## 8  ( 1 )                    *
```

Using Cp as our criteria for the best model, we want the fourth row, which says the model with the lowest Cp contains: saleht, ftfrbody, bkfat, and then the interaction between breedSimental and saleht. In order to include that interaction, we have to keep breed in the model as well (and will have Hereford terms as well). Thus, (by convention of keeping lower order terms and other levels of categorical variables) the model this is effectively telling us to use is:

```
Cpmod <- lm(salepr ~ breed*saleht + ftfrbody + bkfat, data = Bulls)
msummary(Cpmod)
```

```
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -16876.242   3456.971  -4.882 6.67e-06 ***
## breedHereford          6711.238   5696.678   1.178   0.2429    
## breedSimental          6824.983   5439.812   1.255   0.2139    
## saleht                  379.261     67.983   5.579 4.55e-07 ***
## ftfrbody                 -1.764      0.859  -2.054   0.0439 *  
## bkfat                   931.030    844.788   1.102   0.2743    
## breedHereford:saleht   -128.926    108.178  -1.192   0.2375    
## breedSimental:saleht   -138.479     99.474  -1.392   0.1684    
## 
## Residual standard error: 470 on 68 degrees of freedom
## Multiple R-squared:  0.484,  Adjusted R-squared:  0.4309 
## F-statistic: 9.112 on 7 and 68 DF,  p-value: 6.872e-08
```

(Note that if we really think there is no difference between Angus and Hereford, we could redefine the breed levels and make the variable binary.)

Are all the predictors significant in this model?

SOLUTION: Not all the predictors are significant. In fact, only saleht and ftfrbody are significant.

Let's see what model(s) the other three methods of automated variable selection suggest.

Backward Elimination

```r
backward <- regsubsets(salepr ~ breed*saleht + ftfrbody + prctffb + bkfat + salewt, data = Bulls, metho
with(summary(backward), data.frame(cp, outmat))
```

```
##                 cp breedHereford breedSimental saleht ftfrbody prctffb bkfat
## 1  ( 1 ) 36.614154                                  *
## 2  ( 1 )  5.777695                                  *
## 3  ( 1 )  3.076830                                  *        *
## 4  ( 1 )  3.600329                           *      *        *
## 5  ( 1 )  5.232608                           *      *        *
## 6  ( 1 )  5.276845             *             *      *        *
## 7  ( 1 )  6.096252             *             *      *        *             *
## 8  ( 1 )  8.005696             *             *      *        *       *     *
##          salewt breedHereford.saleht breedSimental.saleht
## 1  ( 1 )
## 2  ( 1 )                                                *
## 3  ( 1 )                                                *
## 4  ( 1 )                                                *
## 5  ( 1 )                           *                    *
## 6  ( 1 )                           *                    *
## 7  ( 1 )                           *                    *
## 8  ( 1 )                           *                    *
```

Forward selection

```r
forward <- regsubsets(salepr ~ breed*saleht + ftfrbody + prctffb + bkfat + salewt, data = Bulls, method
with(summary(forward), data.frame(cp, outmat))
```

```
##                 cp breedHereford breedSimental saleht ftfrbody prctffb bkfat
## 1  ( 1 ) 36.614154                                  *
## 2  ( 1 )  5.777695                                  *
## 3  ( 1 )  3.076830                                  *        *
## 4  ( 1 )  2.682379                                  *        *             *
## 5  ( 1 )  3.701179                           *      *        *             *
## 6  ( 1 )  5.445308                           *      *        *             *
## 7  ( 1 )  6.096252             *             *      *        *             *
## 8  ( 1 )  8.005696             *             *      *        *       *     *
##          salewt breedHereford.saleht breedSimental.saleht
## 1  ( 1 )
## 2  ( 1 )                                                *
## 3  ( 1 )                                                *
## 4  ( 1 )                                                *
## 5  ( 1 )                                                *
## 6  ( 1 )                           *                    *
## 7  ( 1 )                           *                    *
## 8  ( 1 )                           *                    *
```

Stepwise Regression

```r
stepwise <- regsubsets(salepr ~ breed*saleht + ftfrbody + prctffb + bkfat + salewt, data = Bulls, metho
with(summary(stepwise), data.frame(cp, outmat))
```

```
##                 cp breedHereford breedSimental saleht ftfrbody prctffb bkfat
## 1  ( 1 ) 36.614154                                  *
## 2  ( 1 )  5.777695                                  *
## 3  ( 1 )  3.076830                                  *        *
## 4  ( 1 )  2.682379                                  *        *             *
## 5  ( 1 )  3.701179                           *      *        *             *
## 6  ( 1 )  5.276845             *             *      *        *
## 7  ( 1 )  8.052480             *             *      *        *       *     *
## 8  ( 1 )  9.625813             *             *      *        *       *     *
##          salewt breedHereford.saleht breedSimental.saleht
## 1  ( 1 )
## 2  ( 1 )                                                *
## 3  ( 1 )                                                *
## 4  ( 1 )                                                *
## 5  ( 1 )                                                *
## 6  ( 1 )                           *                    *
## 7  ( 1 )      *
## 8  ( 1 )      *                    *
```

What final model(s) do all three of these methods propose using Cp as the criterion? Is it the same as the model suggested by the best subsets method (though we had to adjust that one to include breed to get the desired interaction)?

> SOLUTION: Backward elimination would not include bkfat. Forward selection and stepwise regression agree with best subsets, at least if your criterion is lowest Cp.

Remember, automated techniques can only help so much! The computer can't "think" about what makes sense in a model. You should use your own knowledge and expert opinion (from experts!) to assist whenever possible.