ADA2: Class 06, Ch 03 A Taste of Model Selection for Multiple Regression

Tim Farkas

February 07, 2022

Prostate-specific antigen (PSA)

A university medical center urology group (Stamey, *et al.*, 1989) was interested in the association between a prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies. We will look at a subset of this dataset for this assignment; later in the semester we'll be able to analyze the full complex dataset.

column	Variable name	Description
1	Identification number	1-97
2*	PSA level	Serum prostate-specific antigen level (ng/ml)
3*	Cancer volume	Estimate of prostate cancer volume (cc)
4*	Weight	Prostate weight (gm)
5*	Age	Age of patient (years)
6	Benign prostatic	Amount of benign prostatic hyperplasia (cm2) hyperplasia
7	Seminal vesicle invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 if no
8	Capsular penetration	Degree of capsular penetration (cm)
9	Gleason score	Pathologically determined grade of disease (6,7,8), higher indicates worse prognosis

Background: Until recently, PSA was commonly recommended as a screening mechanism for detecting prostate cancer. To be an efficient screening tool it is important that we understand how PSA levels relate to factors that may determine prognosis and outcome. The PSA test measures the blood level of prostate-specific antigen, an enzyme produced by the prostate. PSA levels under 4 ng/mL (nanograms per milliliter) are generally considered normal, while levels over 4 ng/mL are considered abnormal (although in men over 65 levels up to 6.5 ng/mL may be acceptable, depending upon each laboratorys reference ranges). PSA levels between 4 and 10 ng/mL indicate a risk of prostate cancer higher than normal, but the risk does not seem to rise within this six-point range. When the PSA level is above 10 ng/mL, the association with cancer becomes stronger. However, PSA is not a perfect test. Some men with prostate cancer do not have an elevated PSA, and most men with an elevated PSA do not have prostate cancer. PSA levels can change for many reasons other than cancer. Two common causes of high PSA levels are enlargement of the prostate (benign prostatic hypertrophy (BPH)) and infection in the prostate (prostatitis).

Rubric

A goal here is to build a multiple regression model to predict PSA level `PSA` from the cancer volume `V`, Prostate weight `Wt`, and Age `Age`. A reasonable strategy would be to:

1. Examine the relationship between the response and the potential predictors.
2. Decide whether any of the variables should be transformed.
3. Perform a backward elimination using the desired response and predictors.
4. Given the selected model, examine the residuals and check for influential cases.
5. Repeat the process, if necessary.
6. Interpret the model and discuss any model limitations.

(1 p) Looking at the data

Describe any patterns you see in the data. Are the ranges for each variable reasonable? Extreme/unusual observations? Strong nonlinear trends with the response suggesting a transformation?

```
library(erikmisc)
library(tidyverse)

# First, download the data to your computer,
#   save in the same folder as this Rmd file.

# read the data
  dat_psa <- read_csv("~/Dropbox/3_Education/Courses/stat_528_ada2/ADA2_CL_06_psa.csv")
str(dat_psa)
```

```
spec_tbl_df [97 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ID                          : num [1:97] 1 2 3 4 5 6 7 8 9 10 ...
 $ PSA                         : num [1:97] 0.651 0.852 0.852 0.852 1.448 ...
 $ cancer_volume               : num [1:97] 0.56 0.372 0.601 0.301 2.117 ...
 $ prostate_weight             : num [1:97] 16 27.7 14.7 26.6 30.9 ...
 $ patient_age                 : num [1:97] 50 58 74 58 62 50 64 58 47 63 ...
 $ benign_prostatic_hyperplasia: num [1:97] 0 0 0 0 0 ...
 $ seminal_vesicle_invasion    : num [1:97] 0 0 0 0 0 0 0 0 0 0 ...
 $ capsular_penetration        : num [1:97] 0 0 0 0 0 0 0 0 0 0 ...
 $ gleason_score               : num [1:97] 6 7 7 6 6 6 6 6 7 6 ...
 - attr(*, "spec")=
  .. cols(
  ..   ID = col_double(),
  ..   PSA = col_double(),
  ..   cancer_volume = col_double(),
  ..   prostate_weight = col_double(),
  ..   patient_age = col_double(),
  ..   benign_prostatic_hyperplasia = col_double(),
  ..   seminal_vesicle_invasion = col_double(),
  ..   capsular_penetration = col_double(),
  ..   gleason_score = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
#dat_psa

dat_psa <-
```

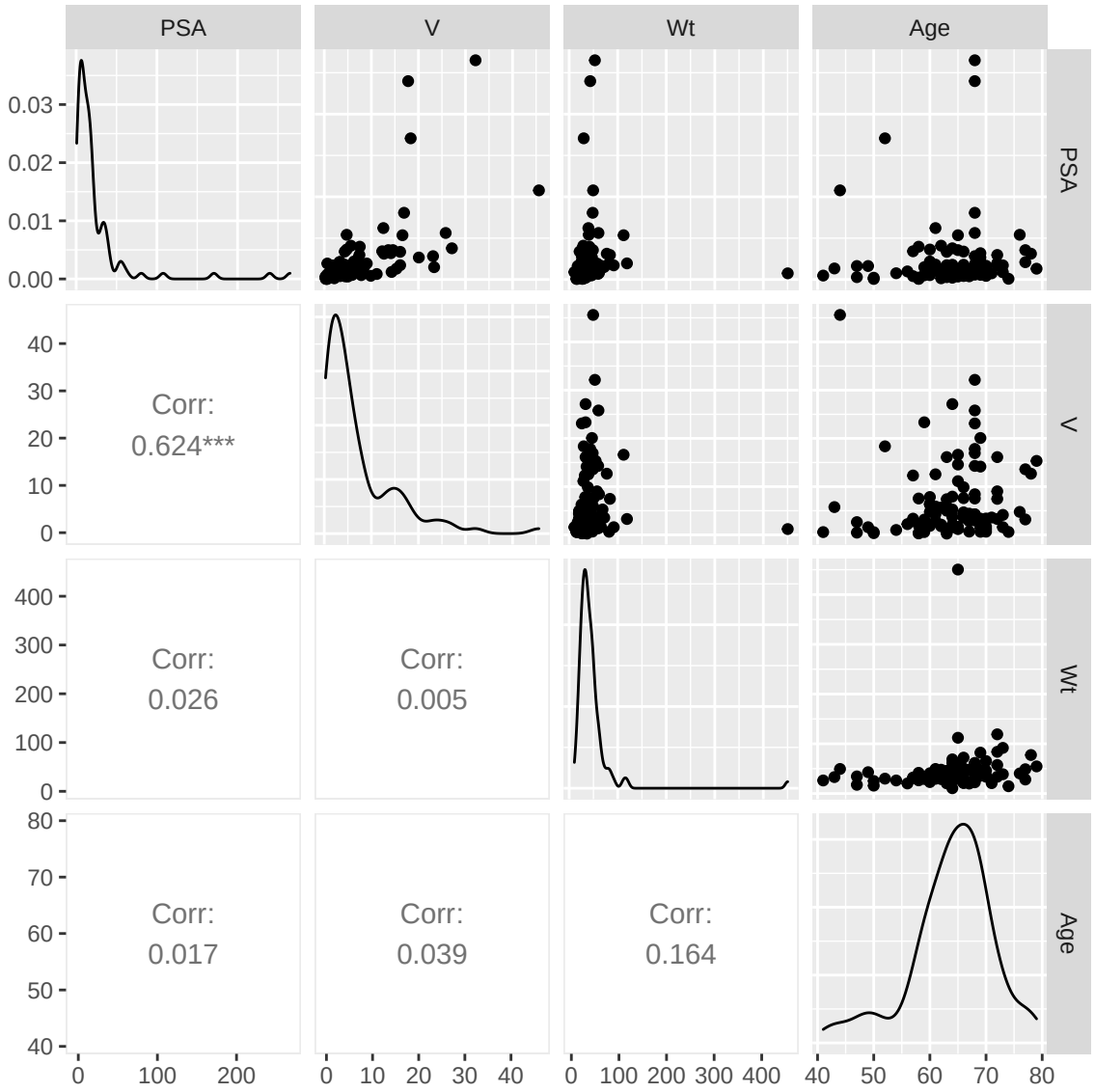
```
  dat_psa %>%
  # select variables we want to use for this assignment
  select(
    ID
  , PSA
  , cancer_volume
  , prostate_weight
  , patient_age
  ) %>%
  # simplify column names
  rename(
    PSA = PSA
  , V   = cancer_volume
  , Wt  = prostate_weight
  , Age = patient_age
  )

str(dat_psa)
```

```
tibble [97 x 5] (S3: tbl_df/tbl/data.frame)
 $ ID : num [1:97] 1 2 3 4 5 6 7 8 9 10 ...
 $ PSA: num [1:97] 0.651 0.852 0.852 0.852 1.448 ...
 $ V  : num [1:97] 0.56 0.372 0.601 0.301 2.117 ...
 $ Wt : num [1:97] 16 27.7 14.7 26.6 30.9 ...
 $ Age: num [1:97] 50 58 74 58 62 50 64 58 47 63 ...
```

```
summary(dat_psa)
```

```
       ID           PSA                V                  Wt        
 Min.   : 1   Min.   :  0.651   Min.   : 0.2592   Min.   : 10.70  
 1st Qu.:25   1st Qu.:  5.641   1st Qu.: 1.6653   1st Qu.: 29.37  
 Median :49   Median : 13.330   Median : 4.2631   Median : 37.34  
 Mean   :49   Mean   : 23.730   Mean   : 6.9987   Mean   : 45.49  
 3rd Qu.:73   3rd Qu.: 21.328   3rd Qu.: 8.4149   3rd Qu.: 48.42  
 Max.   :97   Max.   :265.072   Max.   :45.6042   Max.   :450.34  
      Age       
 Min.   :41.00  
 1st Qu.:60.00  
 Median :65.00  
 Mean   :63.87  
 3rd Qu.:68.00  
 Max.   :79.00  
```

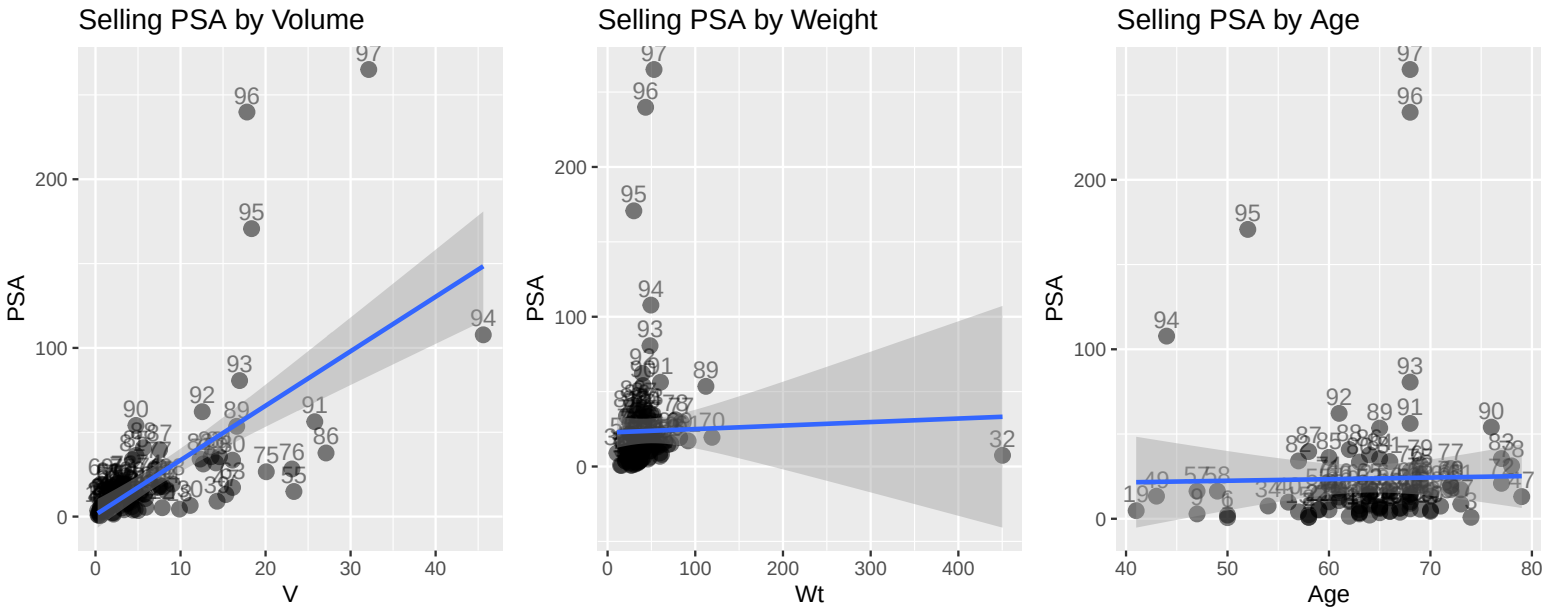
```
library(ggplot2)
library(GGally)
#p <- ggpairs(dat_psa)
# put scatterplots on top so y axis is vertical
p <- ggpairs(dat_psa %>% select(PSA, V, Wt, Age)
             #, upper = list(continuous = wrap("points", alpha = 0.2, size = 0.5))
             , upper = list(continuous = "points")
             , lower = list(continuous = "cor")
             )
print(p)
```




```
# correlation matrix and associated p-values testing "H0: rho == 0"
#library(Hmisc)
dat_psa %>%
  select(PSA, V, Wt, Age) %>%
  as.matrix() %>%
  Hmisc::rcorr()
```

```
     PSA    V   Wt  Age
PSA 1.00 0.62 0.03 0.02
V   0.62 1.00 0.01 0.04
Wt  0.03 0.01 1.00 0.16
Age 0.02 0.04 0.16 1.00

n= 97


P
    PSA    V      Wt     Age
PSA        0.0000 0.7988 0.8672
V   0.0000        0.9604 0.7038
Wt  0.7988 0.9604        0.1078
Age 0.8672 0.7038 0.1078
```

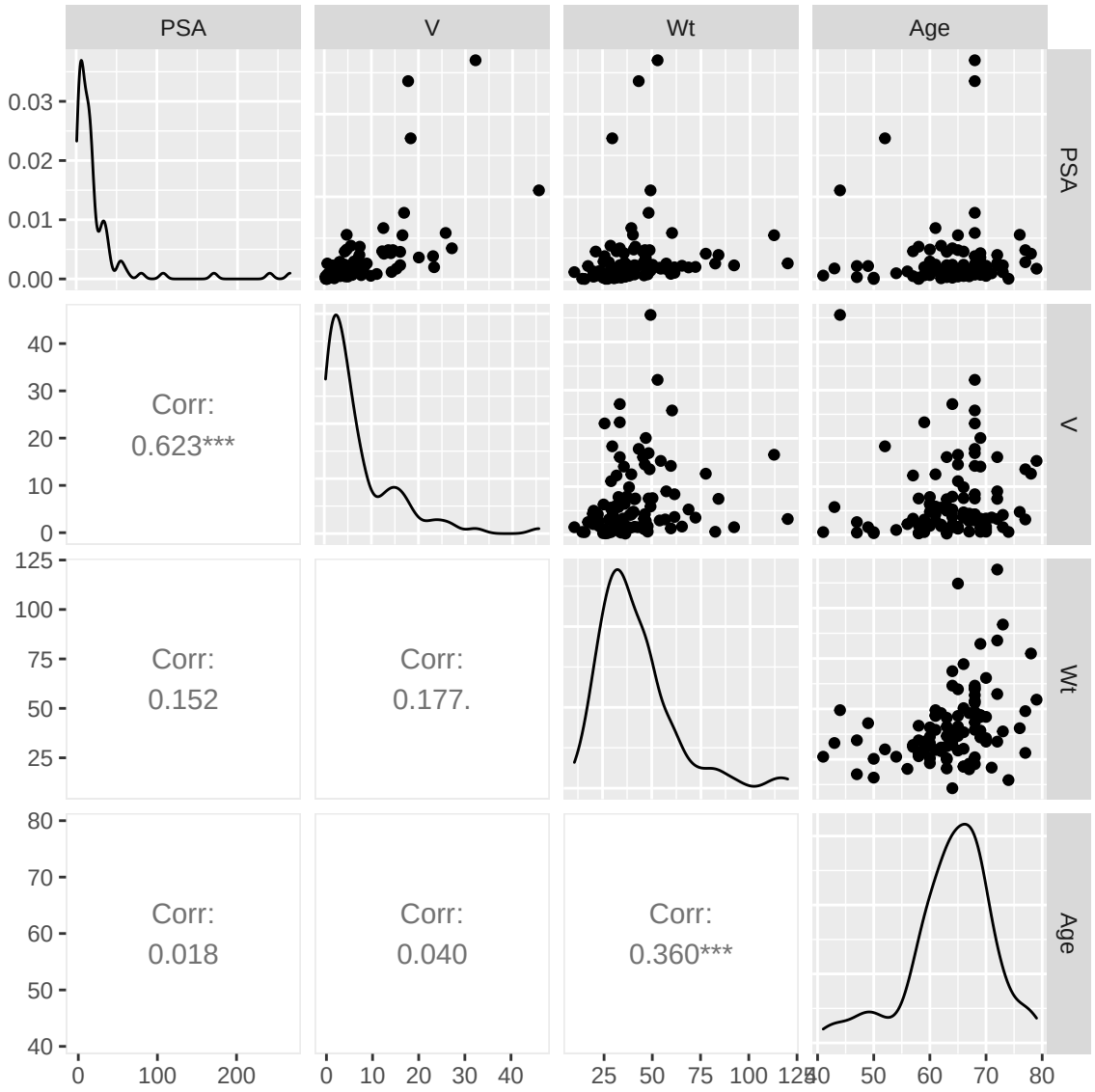
Bivariate scatterplots with points labeled by ID number to help characterize relationships and identify outliers.



Solution

1. The ranges look reasonable, but I'm not a SME.
2. Observation 32 has an extreme prostate weight value. I remove it below and rerun these exploratory analyses before answering #3.

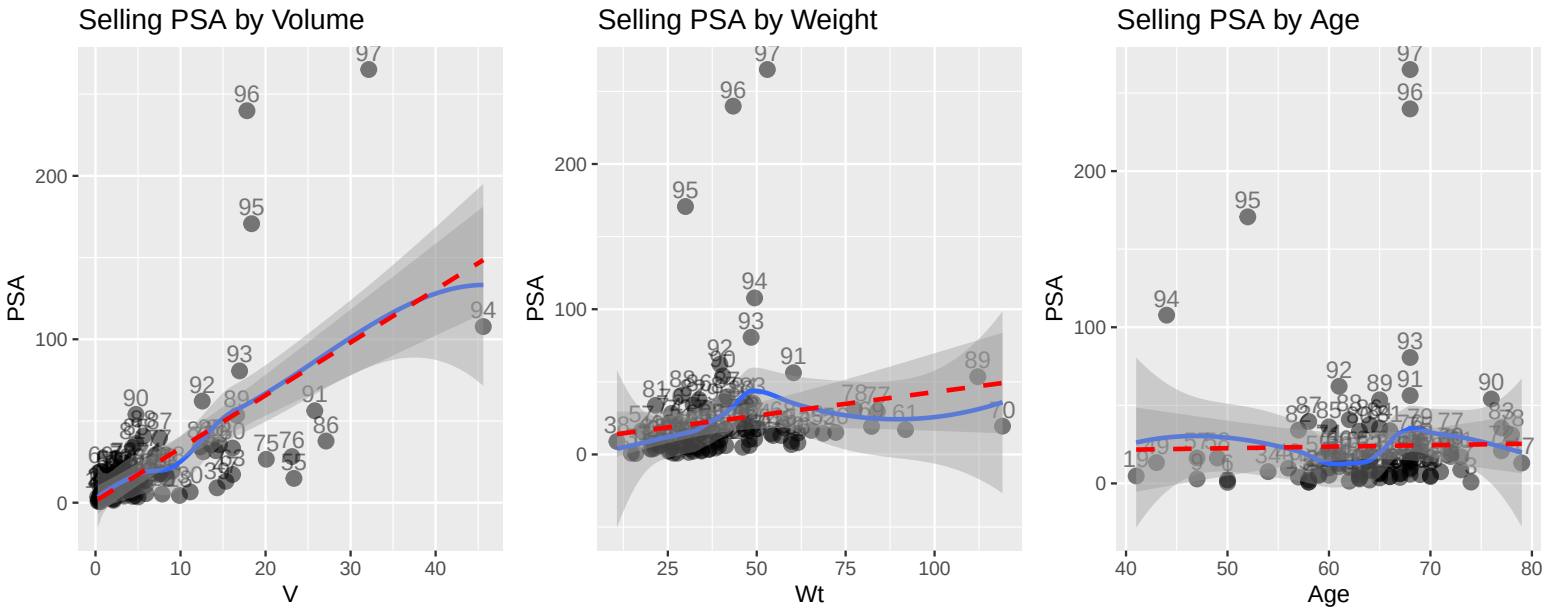
```
library(ggplot2)
library(GGally)

# remove #32
dat_psa <- dat_psa %>% filter(ID != "32")
#p <- ggpairs(dat_psa)
# put scatterplots on top so y axis is vertical
p <- ggpairs(dat_psa %>% select(PSA, V, Wt, Age)
             #, upper = list(continuous = wrap("points", alpha = 0.2, size = 0.5))
             , upper = list(continuous = "points")
             , lower = list(continuous = "cor")
             )
print(p)
```




```
# correlation matrix and associated p-values testing "H0: rho == 0"
#library(Hmisc)
dat_psa %>%
  select(PSA, V, Wt, Age) %>%
  as.matrix() %>%
  Hmisc::rcorr()
```

```
     PSA    V   Wt  Age
PSA 1.00 0.62 0.15 0.02
V   0.62 1.00 0.18 0.04
Wt  0.15 0.18 1.00 0.36
Age 0.02 0.04 0.36 1.00

n= 96


P
    PSA    V      Wt     Age
PSA        0.0000 0.1392 0.8629
V   0.0000        0.0842 0.6960
Wt  0.1392 0.0842        0.0003
Age 0.8629 0.6960 0.0003
```

Bivariate scatterplots with points labeled by ID number to help characterize relationships and identify outliers.



Solution: Post-Outlier Removal

1. The ranges still look good.
2. There are a few rather large values of PSA (observations 95-97) that may be exerting undue influence on the results. They're not yet obviouly outliers to me, so we'll leave them in for now.
3. I decided to add a LOESS smoother to evaluate the shape of relationships. The relationship between PSA and cancer volume most conforms to the linear fit. The other two relationships show complex patterns, but would be a challenge to model with a transformation, so we'll stick with linear for now.

(2 p) Backward selection, diagnostics of reduced model

Fit an appropriate full model and perform backward selection using **BIC**. Discuss the diagnostics in terms of influential observations or problematic structure in the residuals.

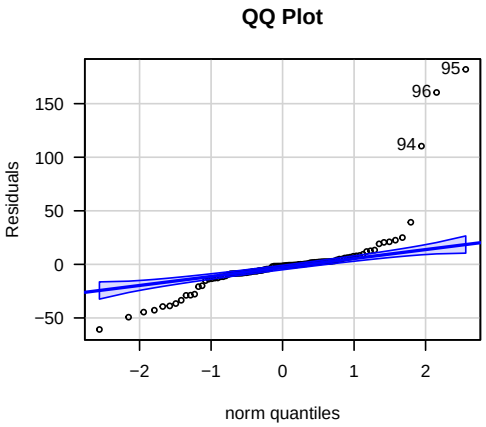
Solution

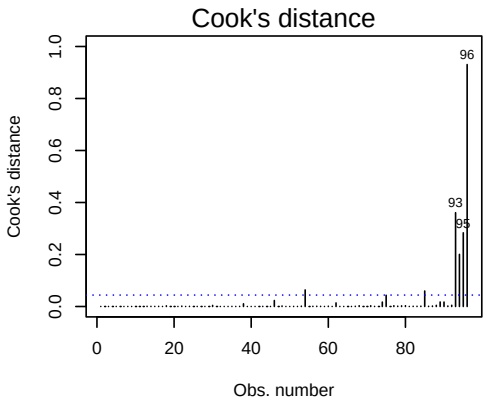
Below I'll get you started with the linear model with all the selected main effects and a set of diagnostic plots.

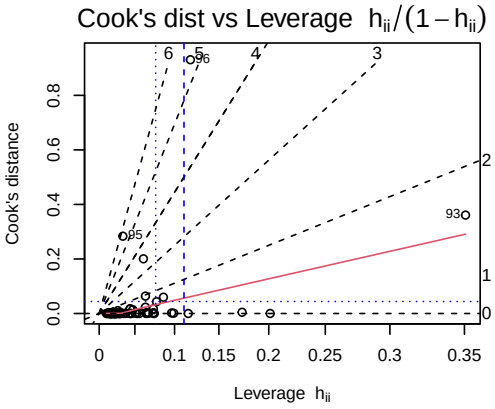
```
# fit full model
lm_psa_full <-
  lm(
    PSA ~ V + Wt + Age
  , data = dat_psa
  )

#library(car)
#Anova(aov(lm_psa_full), type=3)
summary(lm_psa_full)
```

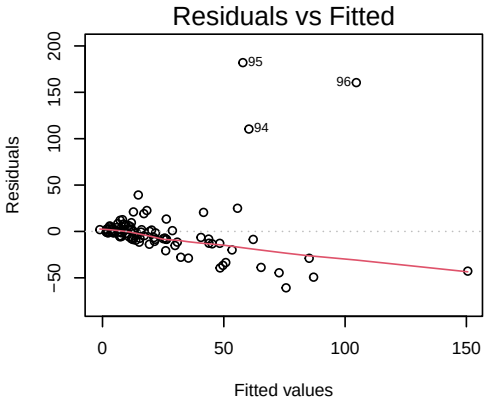
```
Call:
lm(formula = PSA ~ V + Wt + Age, data = dat_psa)

Residuals:
```

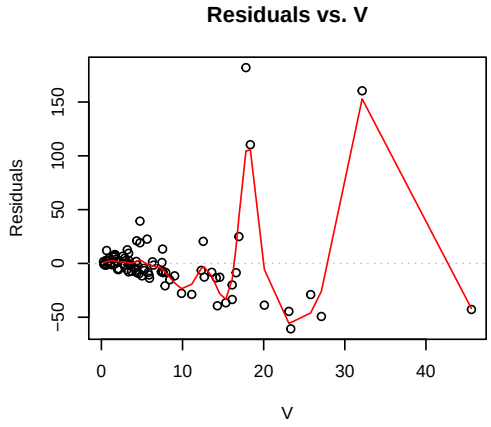
```
    Min      1Q  Median      3Q     Max
-60.787  -8.582  -1.389   2.710 181.992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.7933    28.9219   0.200    0.842
V             3.1897     0.4288   7.438 5.22e-11 ***
Wt            0.1118     0.1892   0.591    0.556
Age          -0.1413     0.4775  -0.296    0.768
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.49 on 92 degrees of freedom
Multiple R-squared:  0.3909,    Adjusted R-squared:  0.3711
F-statistic: 19.68 on 3 and 92 DF,  p-value: 6.091e-10
```

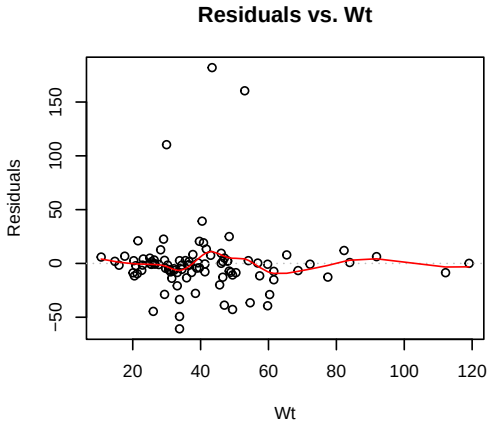


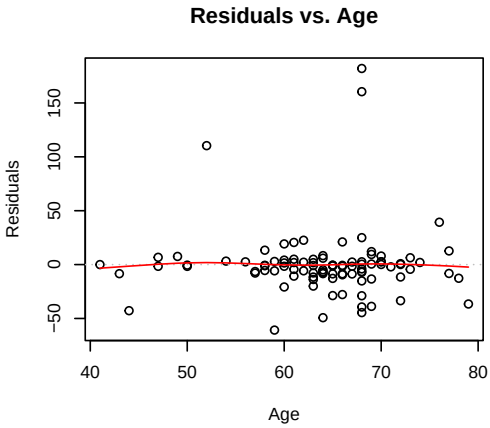


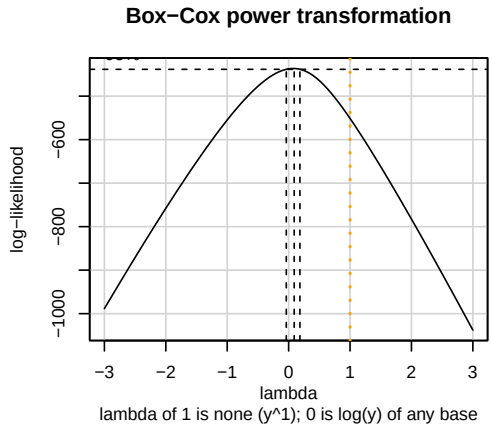


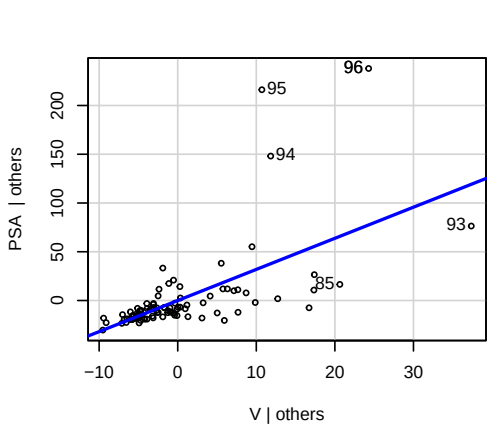


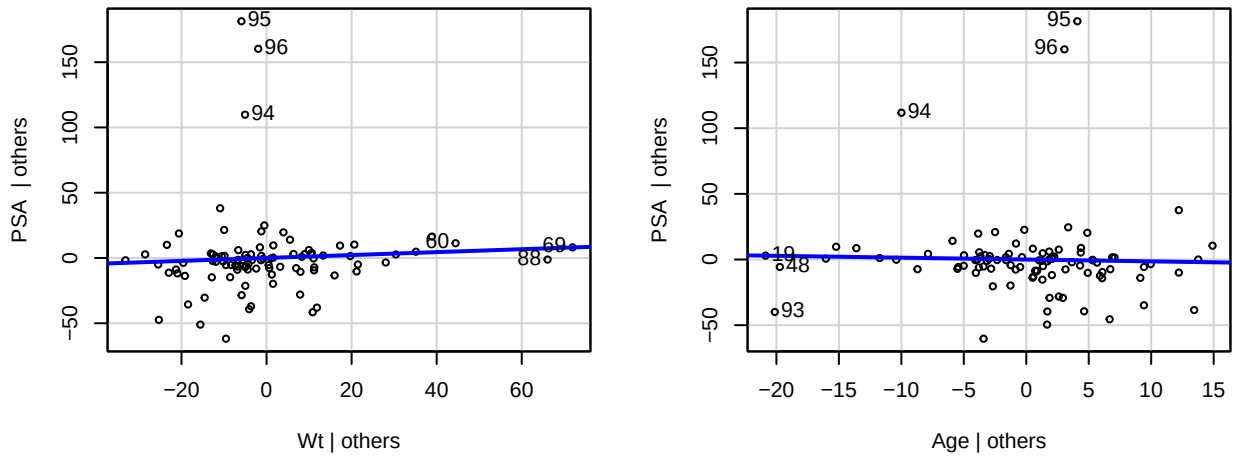






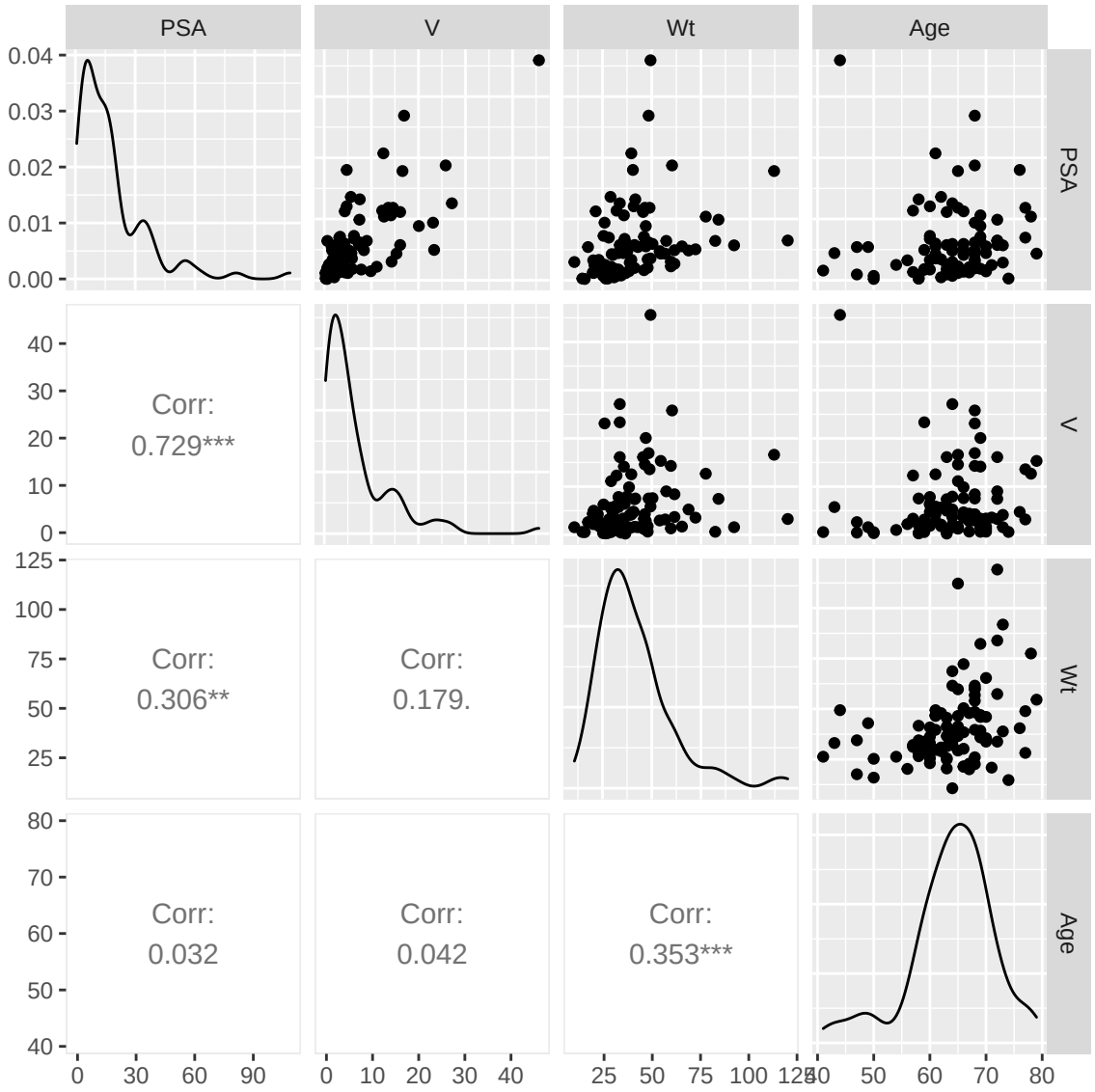






Yeesh, no those three values are definitely exerting extreme influence. Let's remove them and start over.

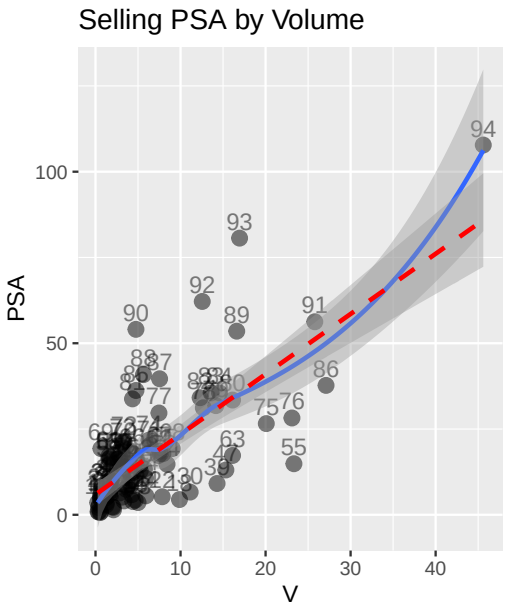
```
library(ggplot2)
library(GGally)

# remove #32
dat_psa <- dat_psa %>% filter(!ID %in% c(95, 96, 97))
#p <- ggpairs(dat_psa)
# put scatterplots on top so y axis is vertical
p <- ggpairs(dat_psa %>% select(PSA, V, Wt, Age)
             #, upper = list(continuous = wrap("points", alpha = 0.2, size = 0.5))
             , upper = list(continuous = "points")
             , lower = list(continuous = "cor")
             )
print(p)
```

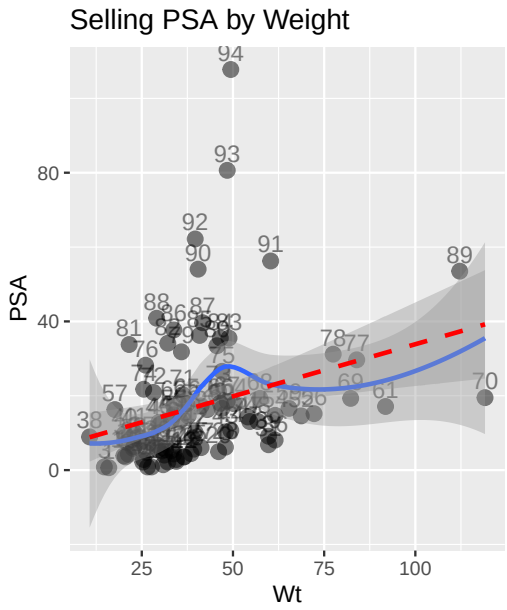


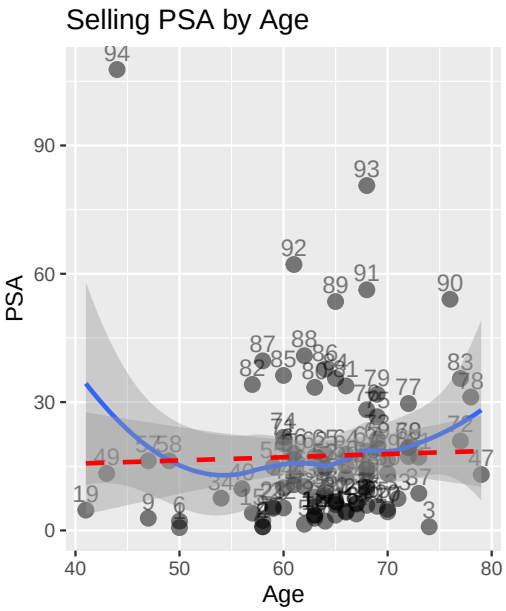

```
# correlation matrix and associated p-values testing "H0: rho == 0"
#library(Hmisc)
dat_psa %>%
  select(PSA, V, Wt, Age) %>%
  as.matrix() %>%
  Hmisc::rcorr()
```

```
     PSA    V   Wt  Age
PSA 1.00 0.73 0.31 0.03
V   0.73 1.00 0.18 0.04
Wt  0.31 0.18 1.00 0.35
Age 0.03 0.04 0.35 1.00

n= 93


P
    PSA    V      Wt     Age
PSA        0.0000 0.0029 0.7628
V   0.0000        0.0866 0.6889
Wt  0.0029 0.0866        0.0005
Age 0.7628 0.6889 0.0005
```

Bivariate scatterplots with points labeled by ID number to help characterize relationships and identify outliers.

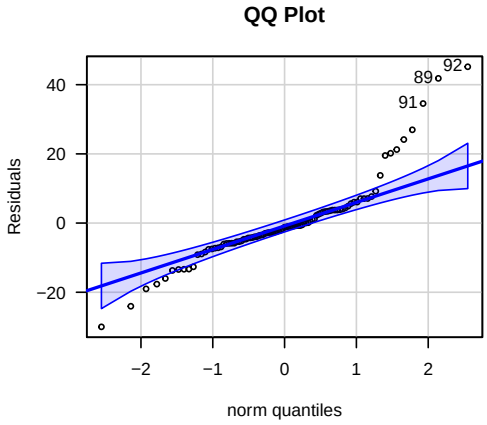


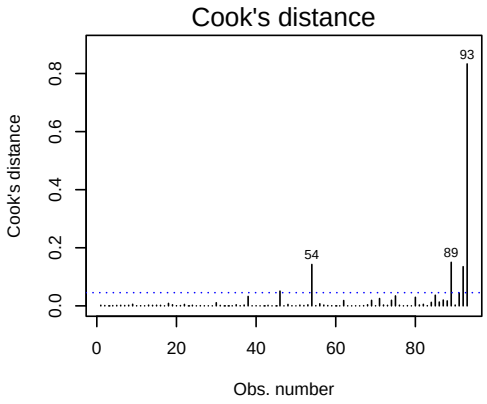


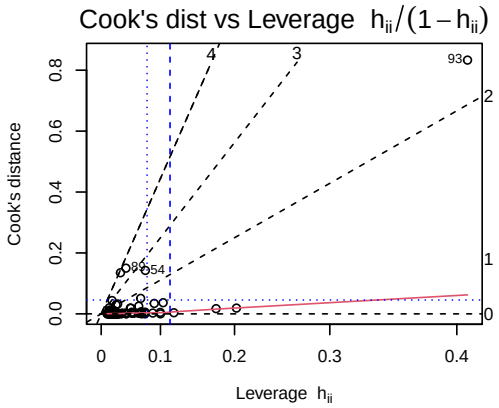


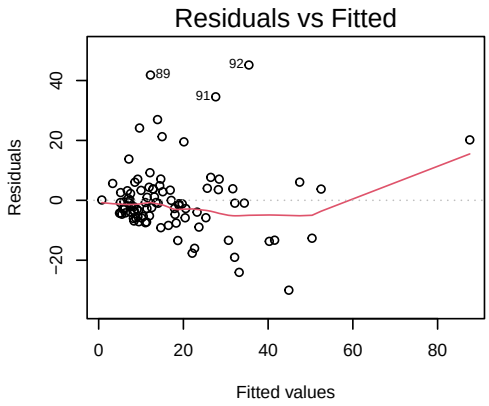

```
# fit full model
lm_psa_full <-
  lm(
    PSA ~ V + Wt + Age
  , data = dat_psa
  )

#library(car)
#Anova(aov(lm_psa_full), type=3)
summary(lm_psa_full)
```

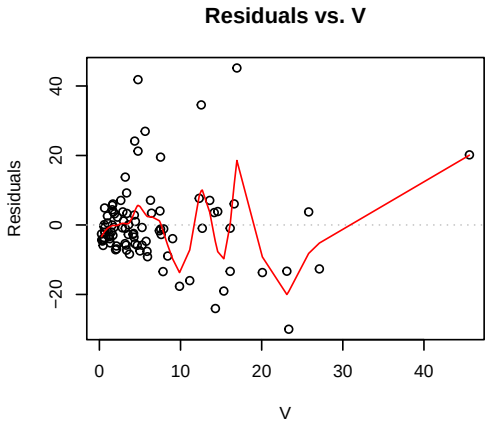
```
Call:
lm(formula = PSA ~ V + Wt + Age, data = dat_psa)

Residuals:
    Min      1Q  Median      3Q     Max
-30.006  -5.418  -1.183   3.756  45.181

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.33455   10.76854   0.867  0.38836
V            1.67263    0.17048   9.811 7.86e-16 ***
Wt           0.18900    0.06936   2.725  0.00775 **
Age         -0.16722    0.17730  -0.943  0.34815
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.9 on 89 degrees of freedom
Multiple R-squared:  0.5676,    Adjusted R-squared:  0.5531
F-statistic: 38.95 on 3 and 89 DF,  p-value: 3.598e-16
```

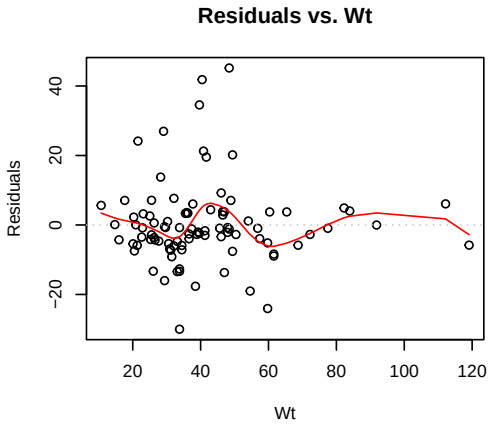


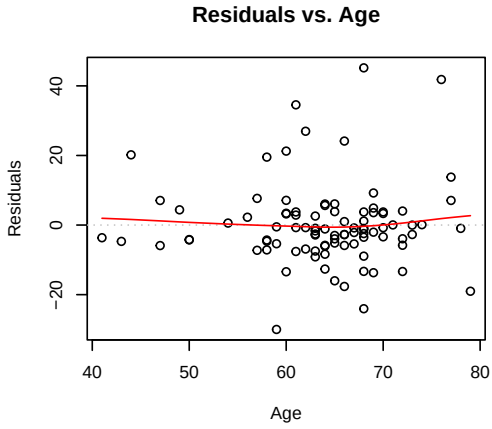


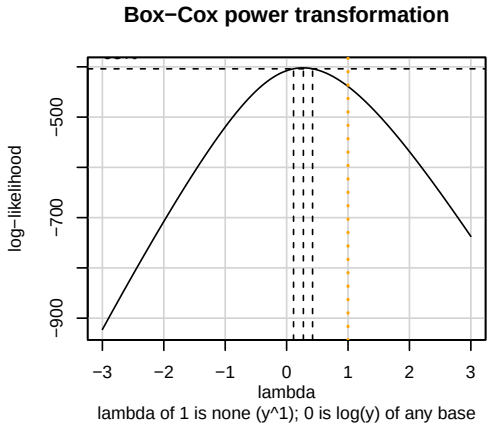


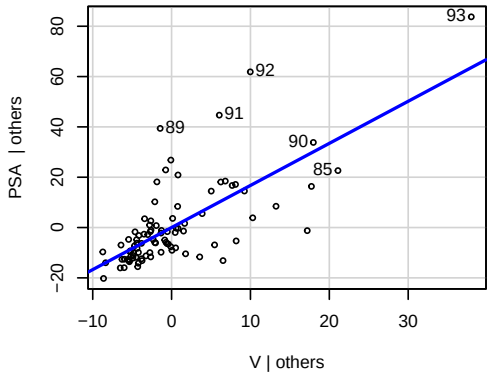


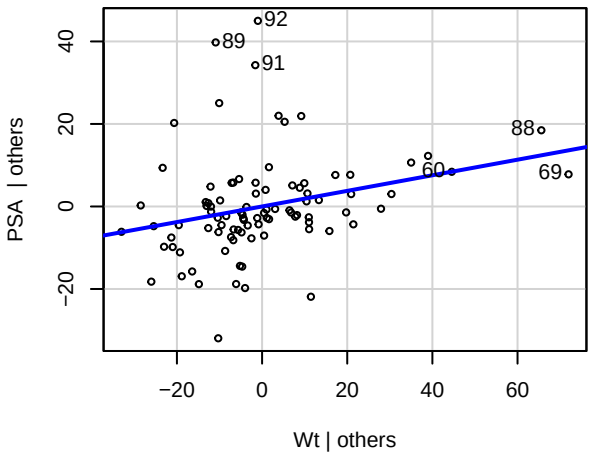


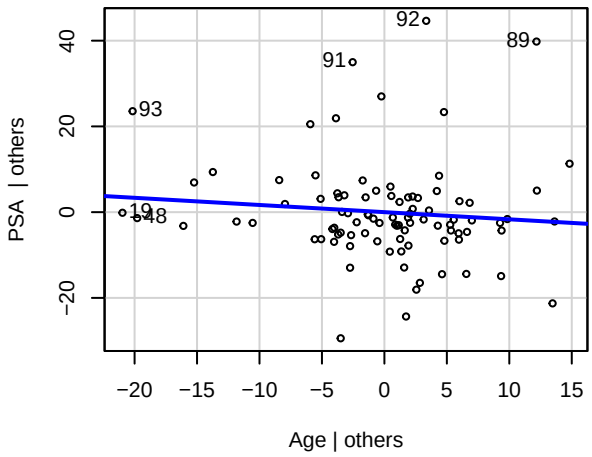












OK, that's better. I'm not totally happy with it, but I don't want to remove outliers forever and ever.

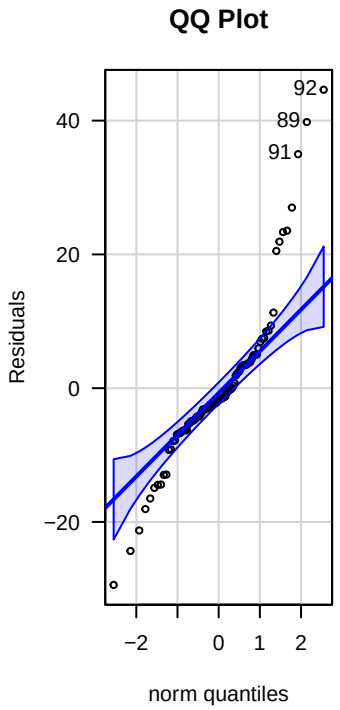
```
lm_psa_rm <- step(lm_psa_full, direction="backward", test="F", k = log(nrow(dat_psa)))
```

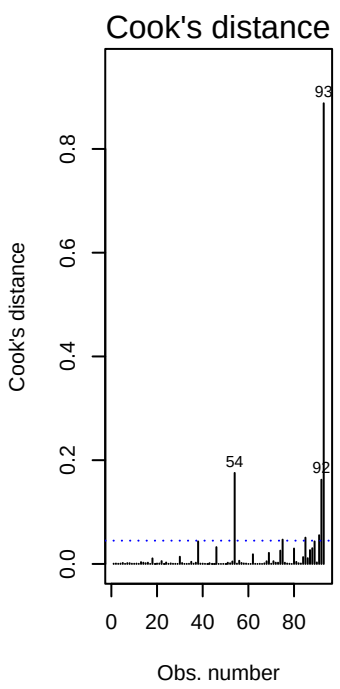
```
Start:  AIC=474.61
PSA ~ V + Wt + Age

       Df Sum of Sq   RSS    AIC F value    Pr(>F)
- Age   1     125.9 12720 471.01  0.8896  0.348146
<none>              12595 474.61
- Wt    1    1050.6 13645 477.53  7.4242  0.007747 **
- V     1   13622.2 26217 538.26 96.2609 7.859e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=471.01
PSA ~ V + Wt

       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              12720 471.01
- Wt    1     924.8 13645 473.00  6.5429    0.0122 *
- V     1   13689.1 26410 534.41 96.8527 6.102e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

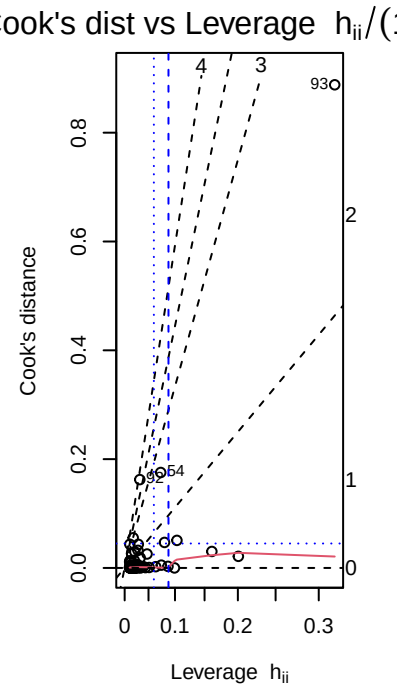
```
summary(lm_psa_rm)
```

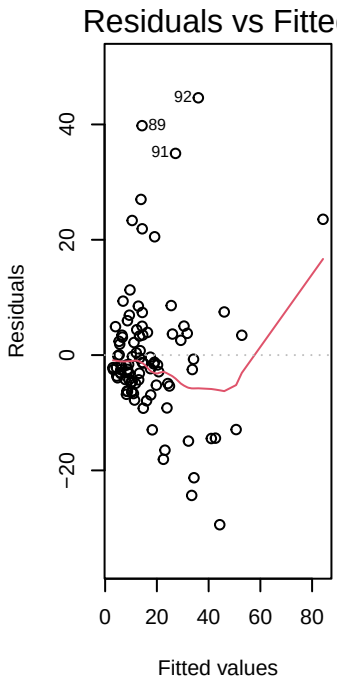
```
Call:
lm(formula = PSA ~ V + Wt, data = dat_psa)

Residuals:
    Min      1Q  Median      3Q     Max
-29.421  -4.934  -1.754   3.477  44.621

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4256     2.9773  -0.143   0.8866
V             1.6763     0.1703   9.841  6.1e-16 ***
Wt            0.1660     0.0649   2.558   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.89 on 90 degrees of freedom
Multiple R-squared:  0.5633,    Adjusted R-squared:  0.5536
F-statistic: 58.05 on 2 and 90 DF,  p-value: < 2.2e-16
```

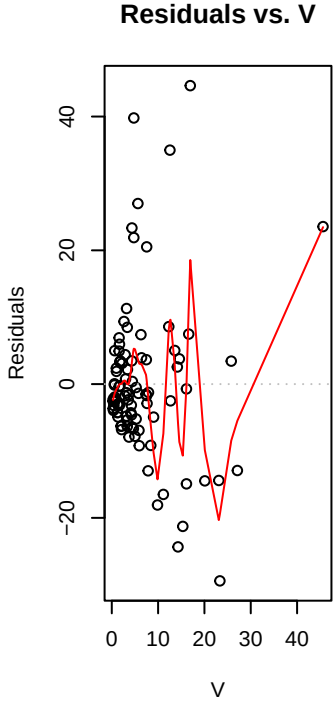
```
e_plot_lm_diagostics(lm_psa_rm, sw_plot_set = "simpleAV")
```

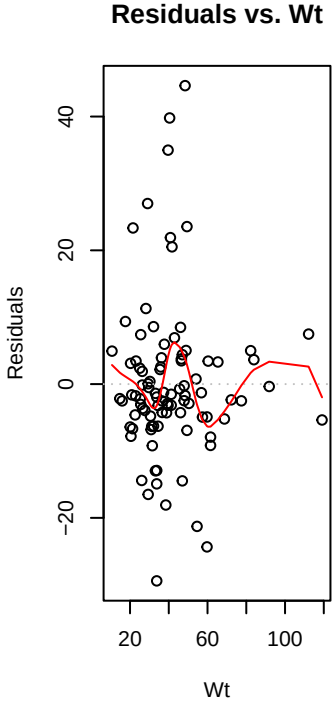

QQ Plot
Residuals
norm quantiles
92
89
91

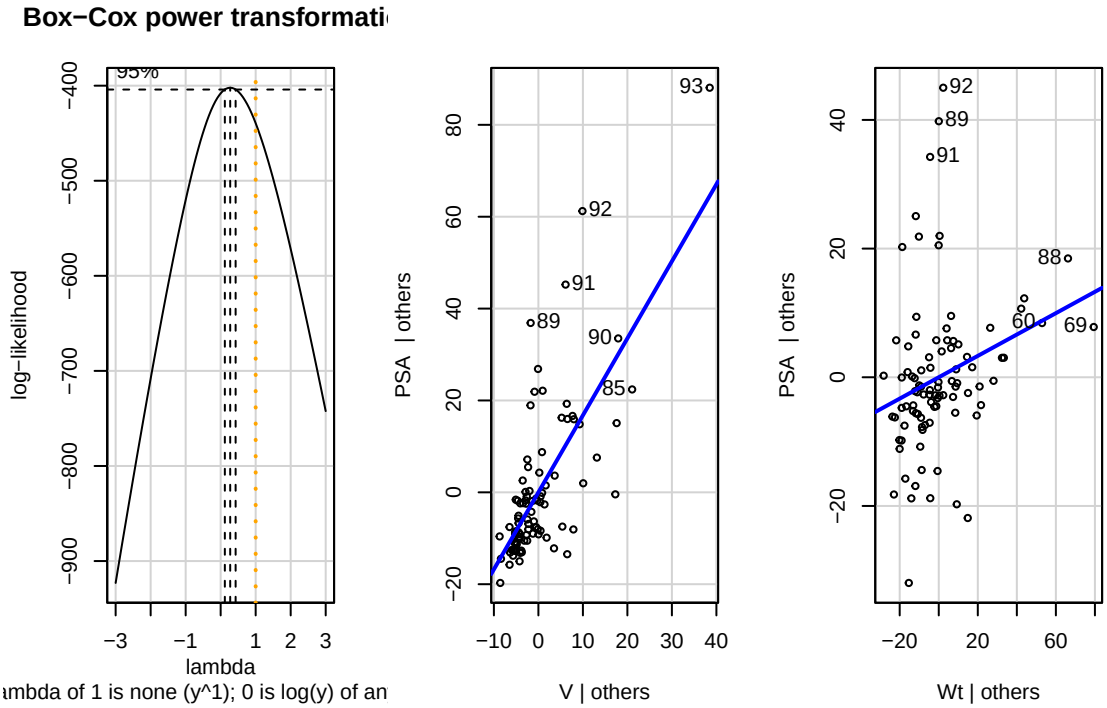

Cook's distance
Cook's distance
Obs. number
93
54
92


Cook's dist vs Leverage hii/(
Cook's distance
Leverage hii
4
3
93
2
1
0
92
54


Residuals vs Fitted
Residuals
Fitted values
92
89
91


Residuals vs. V
Residuals
V


Residuals vs. Wt
Residuals
Wt



These diagnostics show a strong need for a better model. The assumption of normality is severely violated, there are at least a couple extreme outliers (affecting normality), the error variance increases with the fitted values, this pattern appears to be due to variance in cancer volume, and the Box-Cox profile shows a log transformation would be better. Let's do it.

(3 p) Address model fit

If the model doesn't fit well (diagnostics tell you this, not R^2 or significance tests), then address the lack of model fit. Transformations and removing influential points are two strategies. The decisions you make should be based on what you observed in the residual plots. If there's an influential observation, remove it and see how that affects the backward selection (whether the same predictors are retained), the model fit (diagnostics), and regression coefficient estimates (betas). If there's a pattern in the residuals that can be addressed by a transformation, guess at the appropriate transformation and try it.

Repeat until you are satisfied with the diagnostics meeting the model assumptions. Below, briefly outline what you did (no need to show all the output) by (1) identifying what you observed in the diagostics and (2) the strategy you took to address that issue. Finally, show the final model and the diagnostics for that. Describe how the final model is different from the original; in particular discuss whether variables retained are different from backward selection and whether the sign and magnitude of the regression coefficients are much different.

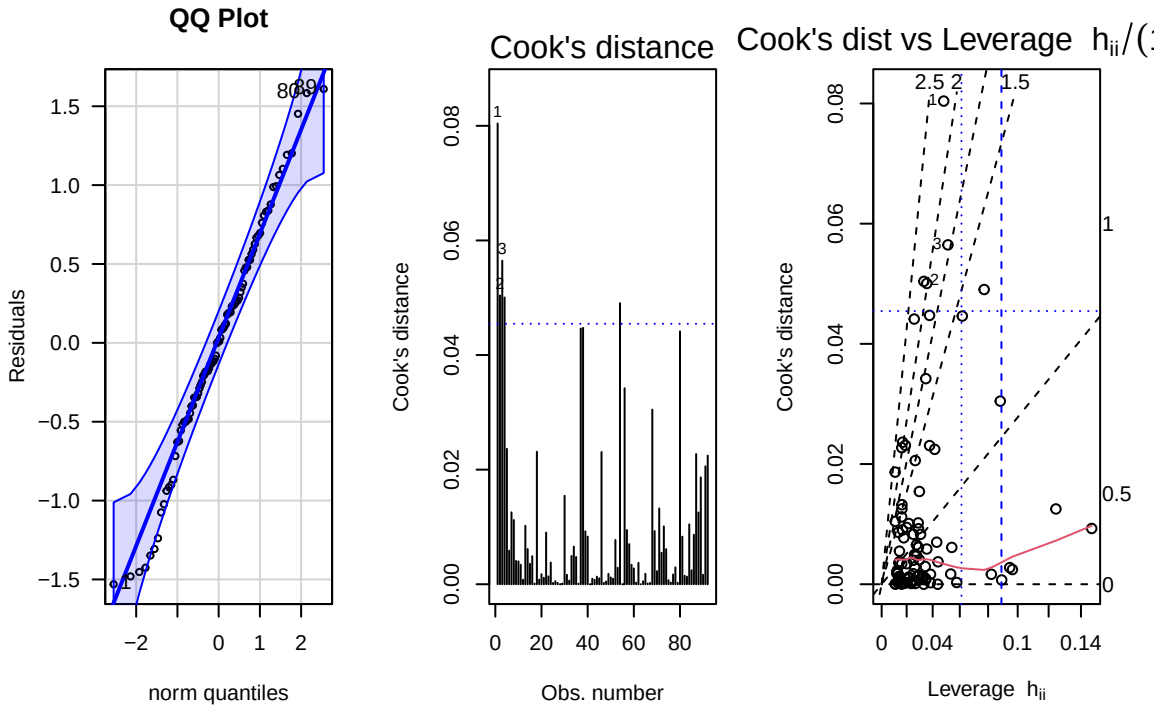
```
# In the diagnostic plots, R uses the row label as the "Obs. number"
# Thus, you need to remove observations by their ID number.
#   (An ID doesn't change when a lower row is removed, but the Obs. number will.)
# Below is an example of how to do that.

# remove influential observations
dat_psa_sub <-
  dat_psa %>%
  filter(
    !(ID %in% c( [ID numbers here] )
  )
```

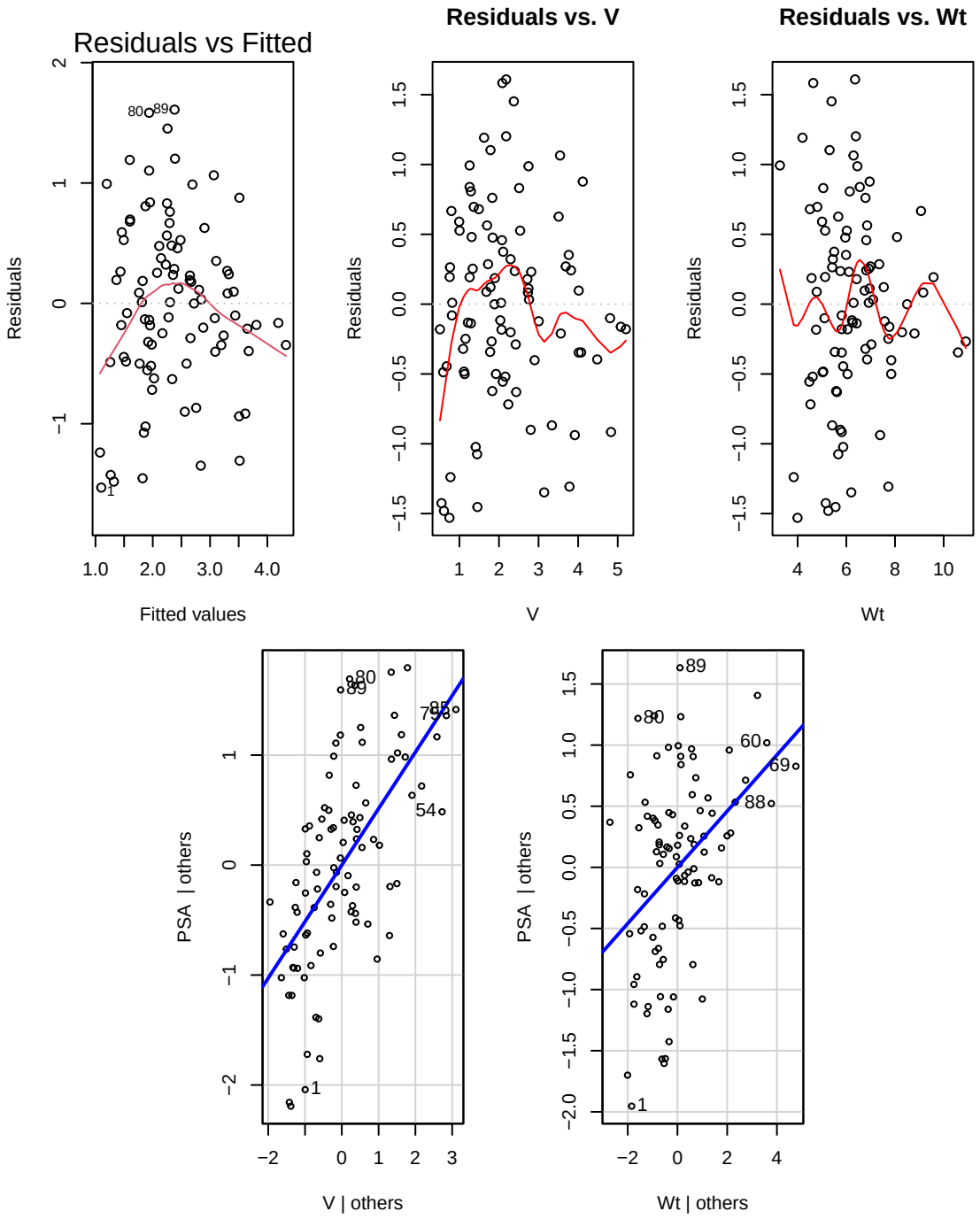
Solution

```
dat_psa_new <- dat_psa %>%
  mutate(across(PSA, ~log(.x)),
         across(c(V, Wt), ~sqrt(.x))) %>%
  slice(-93)

lm_psa_full <- lm(PSA ~ V + Wt + Age, data = dat_psa_new)
lm_psa_rm <- step(lm_psa_full, direction="backward", test="F", k = log(nrow(dat_psa_new)))
```

```
Start:  AIC=-45.63
PSA ~ V + Wt + Age

       Df Sum of Sq    RSS     AIC F value    Pr(>F)
- Age   1    0.0006 46.026 -50.153  0.0012 0.9720221
<none>              46.025 -45.632
- Wt    1    7.9366 53.962 -35.518 15.1748 0.0001906 ***
- V     1   28.6476 74.673  -5.633 54.7744 7.615e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-50.15
PSA ~ V + Wt

       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              46.026 -50.153
- Wt    1    8.9134 54.939 -38.388  17.236 7.548e-05 ***
- V     1   30.0565 76.082  -8.434  58.120 2.553e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
e_plot_lm_diagostics(lm_psa_rm, sw_plot_set = "simpleAV")
```





The Box-Cox profile showed a log transformation of PSA to be appropriate, and that observation 93 (not ID 93!) was an outlier. I removed #93 and performed a log tranformation. The resulting plots showed attenuating error variance across fitted values, and the AV plots showed a similar non-linearity, suggesting a square-root transformation might be appropriate, which I applied. I'm mostly happy with the result. The normality of error terms is the best we've seen yet, the couple outliers flagged by Cook's distance aren't that extreme, and the residuals appear to have stabilized. The AV plots are showing more linear relationships with PSA, though their not perfect. I'm going leave things here though, given the improvement in the other diagnostics.

(0 p) 3D plot

You may want to use the 3D visualization for your original final model and after making some decisions based on diagnostics. Are you convinced it's a reasonable summary of the general relationship?

```
## visualize in 3D with surface
# library(rgl)
# library(car)
```

```
# scatter3d(y ~ x1 + x2, data = dat)
```

(1 p) Predictive ability of the final model

What proportion of variation in the response does the model explain over the mean of the response? (This quantity indicates how precisely this model will predict new observations.)

Solution

```
summary(lm_psa_rm)
```

```
Call:
lm(formula = PSA ~ V + Wt, data = dat_psa_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5304 -0.4126  0.0094  0.4775  1.6093 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.20103    0.35045  -0.574    0.568    
V            0.51371    0.06738   7.624 2.55e-11 ***
Wt           0.22974    0.05534   4.152 7.55e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7191 on 89 degrees of freedom
Multiple R-squared:  0.5195,    Adjusted R-squared:  0.5088 
F-statistic: 48.12 on 2 and 89 DF,  p-value: 6.814e-15
```

The R^2 is 0.5195, indicating that about 52% of variance in PSA is explained by the combined effects of cancer volume and prostate weight.

(2 p) Interpret the final model

Write the equation for the final model and interpret each model coefficient.

Solution

The equation for this final model is

$$\log Y = -0.20 + 0.51 * \sqrt{Vol} + 0.22 * \sqrt{Wt}$$

indicating significant positive relationships between cancer volume and PSA, controlling for prostate weight, and prostate weight and PSA, controlling for cancer volume. A one unit increase in the square root of volume leads to a 0.51 unit increase in the natural log of PSA, and a one unit increase in the square root of prostate weight leads to a .22 unit increase in the natural log of PSA. The intercept indicates that the natural log of PSA is -.20 when both cancer volume and prostate weight are 0, a non-sensical concept, since prostate weight must be non-zero.

(1 p) Inference to whom

To which population of people does this model make inference to? Go back to the abstract of the original study (first sentence) to see which population this sample of men was drawn from.

Solution

This only applies to men who have already been diagnosed with advanced prostate cancer, limiting the ability of this study to help identify cancer in undiagnosed men.