ADA2: Class 05, Ch 03 A Taste of Model Selection for Multiple Regression

Tim Farkas

February 04, 2022

CCHD birth weight

The California Child Health and Development Study involved women on the Kaiser Health plan who received prenatal care and later gave birth in the Kaiser clinics. Approximately 19,000 live-born children were delivered in the 20,500 pregnancies. We consider the subset of the 680 live-born white male infants in the study. Data were collected on a variety of features of the child, the mother, and the father.

The columns in the data set are, from left to right:

col	var name	description
1	id	ID
2	cheadcir	child's head circumference (inches)
3	clength	child's length (inches), y response
4	cbwt	child's birth weight (pounds)
5	gest	gestation (weeks)
6	mage	maternal age (years)
7	msmoke	maternal smoking (cigarettes/day)
8	mht	maternal height (inches)
9	mppwt	maternal pre-pregnancy weight (pounds)
10	page	paternal age (years)
11	ped	paternal education (years)
12	psmoke	paternal smoking (cigarettes/day)
13	pht	paternal height (inches)

```
library(erikmisc)
library(tidyverse)

# Leading 0s cause otherwise numeric columns to be class character.
# Thus, we add the column format "col_double()" for those columns with
#   leading 0s that we wish to be numeric.

dat_cchd <-
  read_csv(
    "~/Dropbox/3_Education/Courses/stat_528_ada2/ADA2_CL_05_cchd-birthwt.csv"
  , col_types =
    cols(
      msmoke  = col_double()
    , mppwt   = col_double()
    , ped     = col_double()
    , psmoke  = col_double()
```

```
    )
  ) %>%
# only keep the variables we're analyzing
  select(
    cbwt
  , mage, msmoke, mht, mppwt
  , page, psmoke, pht, ped
  )
  #   %>%
  # slice(
  #   -123  #  -123 excludes observation (row number) 123
  # )
str(dat_cchd)
```

```
tibble [680 x 9] (S3: tbl_df/tbl/data.frame)
 $ cbwt  : num [1:680] 7.3 8 7.5 7 5.3 8.6 9.1 6.5 3.3 8.1 ...
 $ mage  : num [1:680] 33 28 32 27 32 30 23 27 32 28 ...
 $ msmoke: num [1:680] 25 0 0 2 17 0 0 17 0 0 ...
 $ mht   : num [1:680] 66 63 61 68 67 63 65 64 64 66 ...
 $ mppwt : num [1:680] 140 130 126 150 112 131 134 125 142 113 ...
 $ page  : num [1:680] 37 35 38 30 28 34 26 29 32 41 ...
 $ psmoke: num [1:680] 25 7 17 7 17 17 0 7 0 0 ...
 $ pht   : num [1:680] 74 71 65 73 71 66 71 71 66 68 ...
 $ ped   : num [1:680] 12 10 12 16 10 12 12 12 14 16 ...
```

```
head(dat_cchd)
```

cbwt	mage	msmoke	mht	mppwt	page	psmoke	pht	ped
7.3	33	25	66	140	37	25	74	12
8.0	28	0	63	130	35	7	71	10
7.5	32	0	61	126	38	17	65	12
7.0	27	2	68	150	30	7	73	16
5.3	32	17	67	112	28	17	71	10
8.6	30	0	63	131	34	17	66	12

Rubric

A goal here is to build a multiple regression model to predict child's birth weight (column 4, `cbwt`) from the data on the mother and father (columns 6–13). A reasonable strategy would be to:

1. Examine the relationship between birth weight 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.

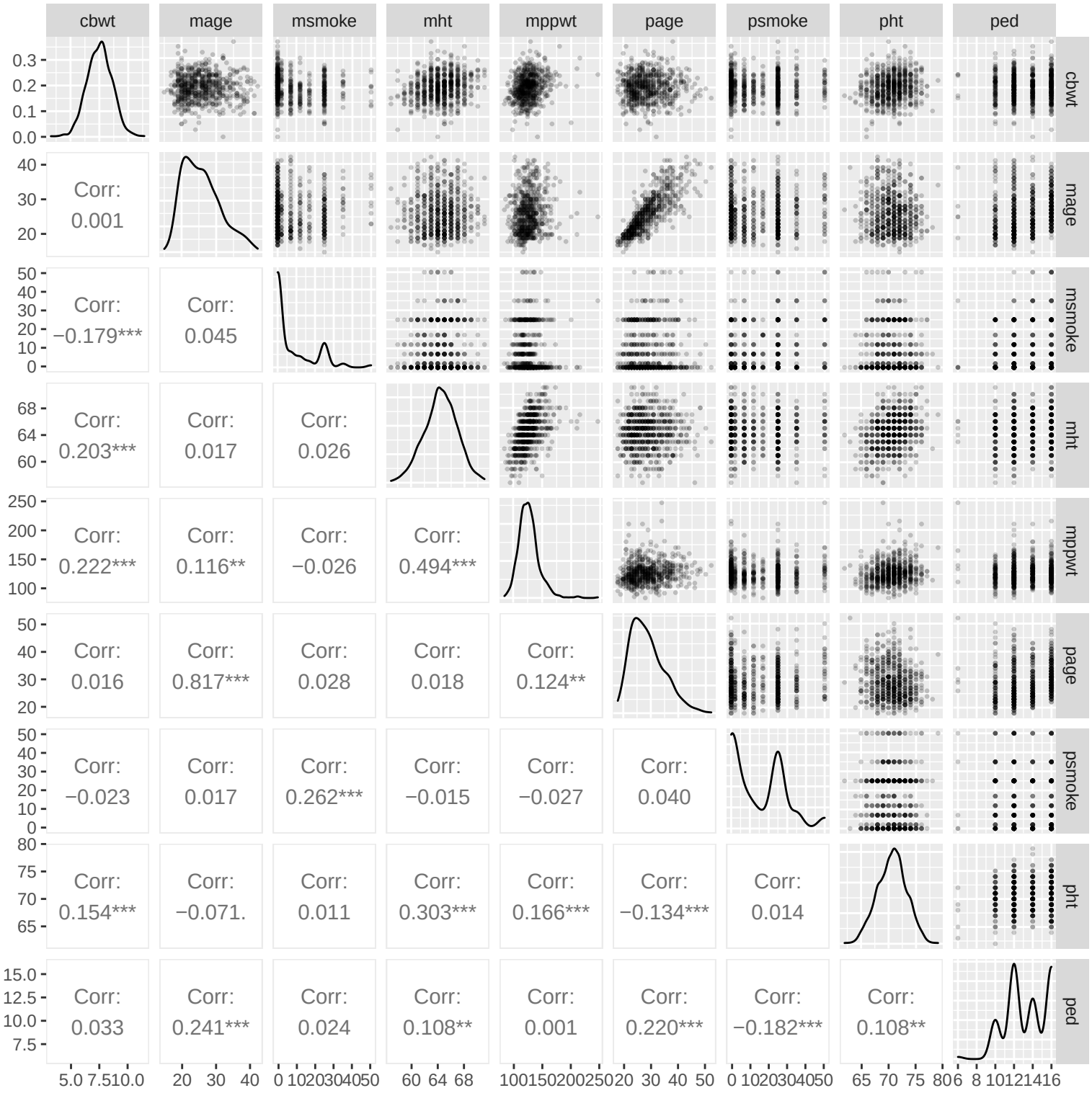
(1 p) Looking at the data

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?

```
summary(dat_cchd)
```

```
      cbwt             mage           msmoke            mht       
 Min.   : 3.300   Min.   :15.00   Min.   : 0.000   Min.   :57.00  
 1st Qu.: 6.800   1st Qu.:21.00   1st Qu.: 0.000   1st Qu.:63.00  
 Median : 7.600   Median :25.00   Median : 0.000   Median :64.00  
 Mean   : 7.516   Mean   :25.86   Mean   : 7.431   Mean   :64.43  
 3rd Qu.: 8.200   3rd Qu.:29.00   3rd Qu.:12.000   3rd Qu.:66.00  
 Max.   :11.400   Max.   :42.00   Max.   :50.000   Max.   :71.00  
     mppwt            page          psmoke           pht             ped        
 Min.   : 85.0   Min.   :18.0   Min.   : 0.00   Min.   :62.00   Min.   : 6.00  
 1st Qu.:115.0   1st Qu.:24.0   1st Qu.: 0.00   1st Qu.:69.00   1st Qu.:12.00  
 Median :125.0   Median :28.0   Median :12.00   Median :71.00   Median :14.00  
 Mean   :126.9   Mean   :28.8   Mean   :14.44   Mean   :70.62   Mean   :13.38  
 3rd Qu.:135.0   3rd Qu.:33.0   3rd Qu.:25.00   3rd Qu.:72.00   3rd Qu.:16.00  
 Max.   :246.0   Max.   :52.0   Max.   :50.00   Max.   :79.00   Max.   :16.00  
```

```
library(ggplot2)
library(GGally)
#p <- ggpairs(dat_cchd)
# put scatterplots on top so y axis is vertical
p <-
  ggpairs(
    dat_cchd
  , upper = list(continuous = wrap("points", alpha = 0.2, size = 0.5))
  , lower = list(continuous = "cor")
  )
print(p)
```




```
# correlation matrix and associated p-values testing "H0: rho == 0"
#library(Hmisc)
dat_cchd %>% as.matrix() %>% Hmisc::rcorr()
```

```
        cbwt  mage msmoke   mht mppwt  page psmoke   pht   ped
cbwt    1.00  0.00  -0.18  0.20  0.22  0.02  -0.02  0.15  0.03
mage    0.00  1.00   0.05  0.02  0.12  0.82   0.02 -0.07  0.24
msmoke -0.18  0.05   1.00  0.03 -0.03  0.03   0.26  0.01  0.02
mht     0.20  0.02   0.03  1.00  0.49  0.02  -0.01  0.30  0.11
mppwt   0.22  0.12  -0.03  0.49  1.00  0.12  -0.03  0.17  0.00
page    0.02  0.82   0.03  0.02  0.12  1.00   0.04 -0.13  0.22
psmoke -0.02  0.02   0.26 -0.01 -0.03  0.04   1.00  0.01 -0.18
pht     0.15 -0.07   0.01  0.30  0.17 -0.13   0.01  1.00  0.11
```

```
ped     0.03  0.24   0.02  0.11  0.00  0.22  -0.18  0.11  1.00

n= 680


P
       cbwt   mage   msmoke mht    mppwt  page   psmoke pht    ped
cbwt          0.9729 0.0000 0.0000 0.0000 0.6685 0.5430 0.0000 0.3899
mage   0.9729        0.2412 0.6490 0.0025 0.0000 0.6653 0.0639 0.0000
msmoke 0.0000 0.2412        0.4996 0.5024 0.4707 0.0000 0.7791 0.5370
mht    0.0000 0.6490 0.4996        0.0000 0.6396 0.7019 0.0000 0.0048
mppwt  0.0000 0.0025 0.5024 0.0000        0.0012 0.4745 0.0000 0.9736
page   0.6685 0.0000 0.4707 0.6396 0.0012        0.3015 0.0004 0.0000
psmoke 0.5430 0.6653 0.0000 0.7019 0.4745 0.3015        0.7224 0.0000
pht    0.0000 0.0639 0.7791 0.0000 0.0000 0.0004 0.7224        0.0049
ped    0.3899 0.0000 0.5370 0.0048 0.9736 0.0000 0.0000 0.0049
```

Solution

1. The ranges for these variables look reasonable. One concern is that the max for number of cigarettes is exactly 50 for both mothers and fathers, suggesting anything higher than 50 might be censored. The range for birthweight is very high, suggesting we might want to have included number of days pregnant at birth here, too.
2. There is one outlier in maternal heigh that is particularly concerning, but we can look at leverage to see how far out it really is. There's also rather few records in the lowest paternal education category, making them somewhat outlier like.
3. There aren't any very obvious non-linear trends, but we should more closely examine the relationship between maternal smoking and birth weight, which might have a concave, monotonically decreasing pattern.

(2 p) Backward selection, diagnostics of reduced model

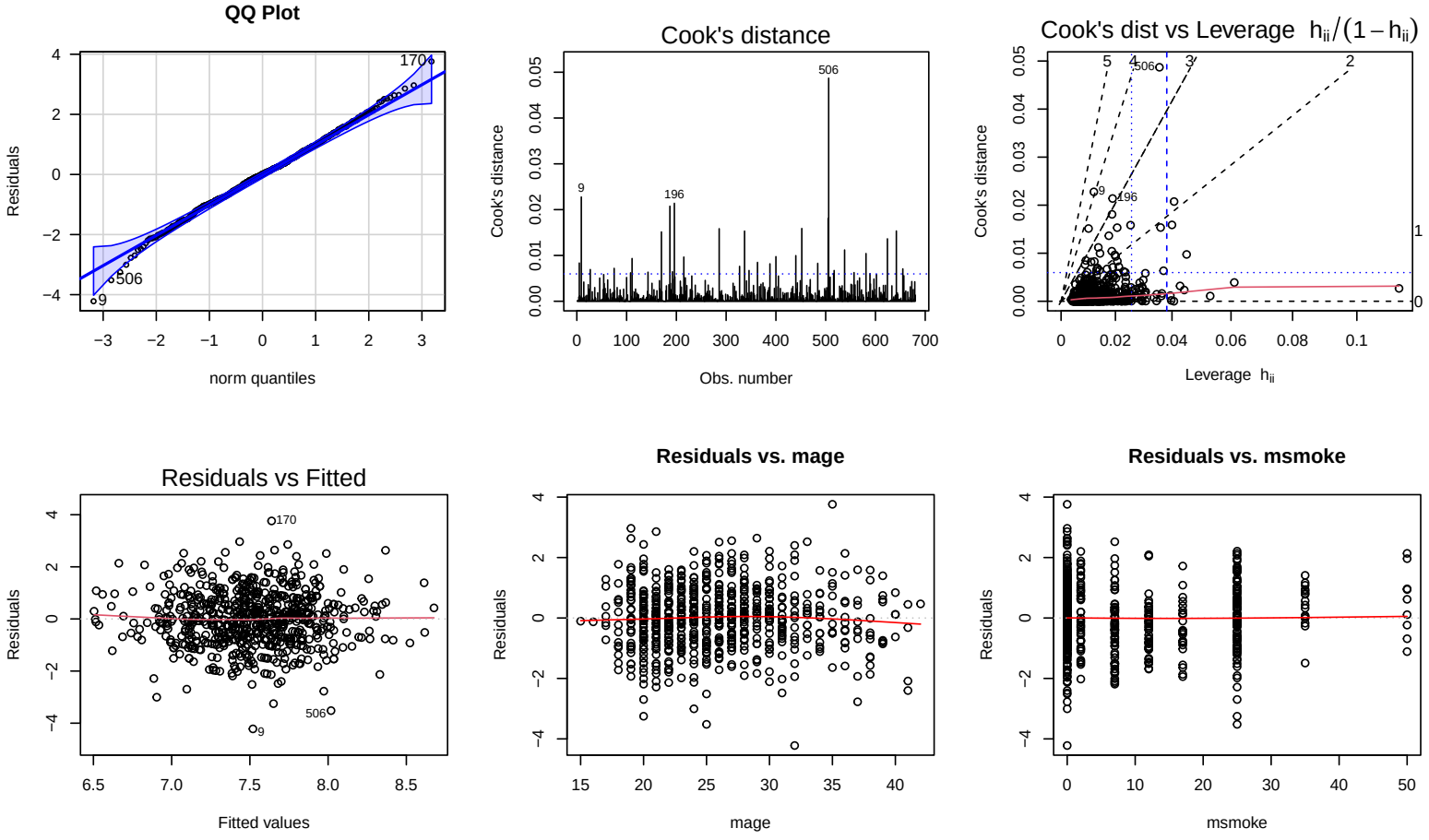
Below I fit the linear model with all the selected main effects.

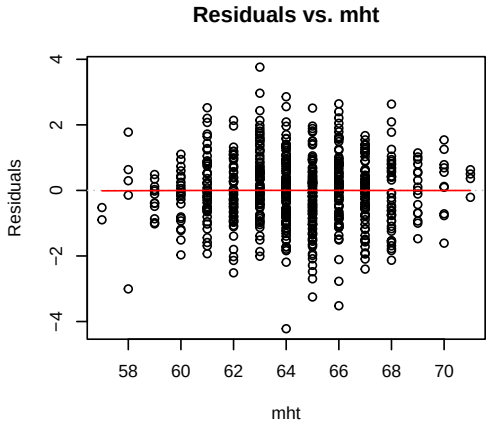
```r
# fit full model
lm_cchd_full <- lm(cbwt ~ mage + msmoke + mht + mppwt
                        + page + ped + psmoke + pht
                       , data = dat_cchd)

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

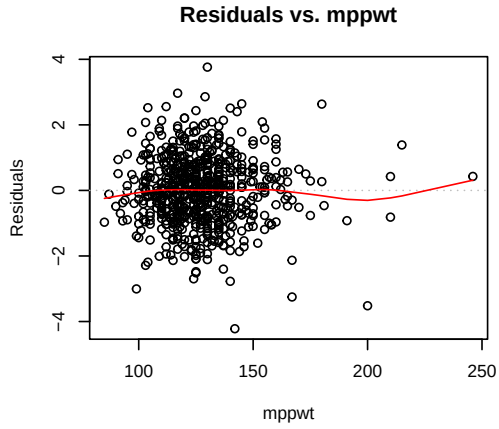
```
Call:
lm(formula = cbwt ~ mage + msmoke + mht + mppwt + page + ped +
    psmoke + pht, data = dat_cchd)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2194 -0.7005  0.0236  0.6527  3.7613

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

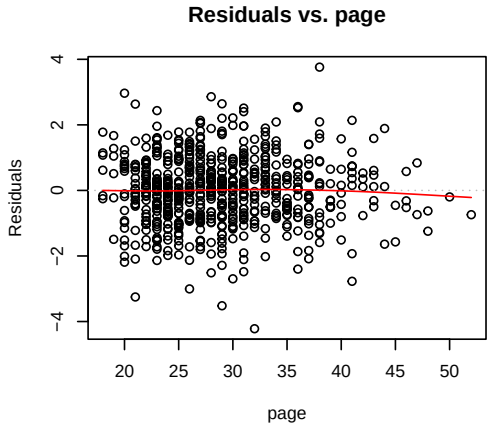
```
(Intercept)  0.510508   1.374821   0.371 0.710511
mage        -0.009105   0.012791  -0.712 0.476840
msmoke      -0.018180   0.003687  -4.931 1.03e-06 ***
mht          0.044131   0.019280   2.289 0.022389 *
mppwt        0.009221   0.002613   3.529 0.000445 ***
page         0.008121   0.011481   0.707 0.479591
ped          0.011172   0.019446   0.575 0.565820
psmoke       0.002546   0.002993   0.851 0.395230
pht          0.041670   0.016233   2.567 0.010473 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 671 degrees of freedom
Multiple R-squared:  0.1036,    Adjusted R-squared:  0.09291
F-statistic: 9.693 on 8 and 671 DF,  p-value: 8.952e-13
```

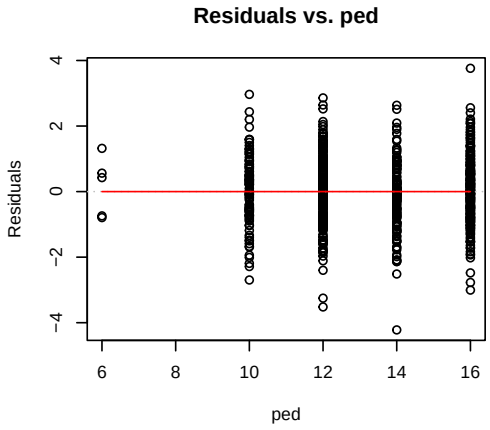
```
# plot diagnostics
e_plot_lm_diagostics(lm_cchd_full, sw_plot_set = "simpleAV")
```

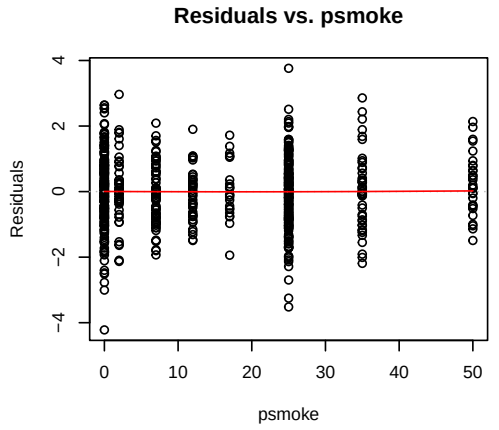


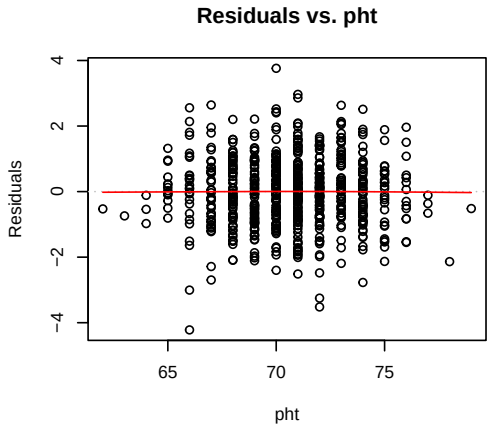

Residuals vs. mht
Residuals
mht

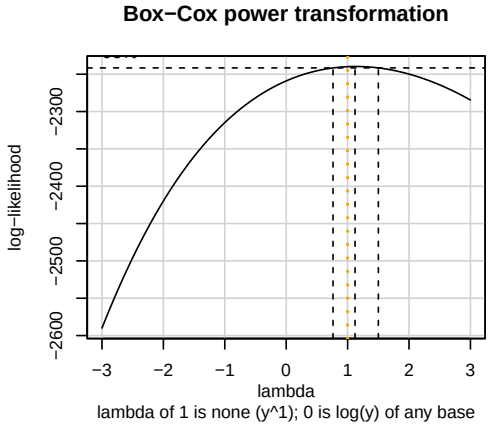

Residuals vs. mppwt
Residuals
mppwt

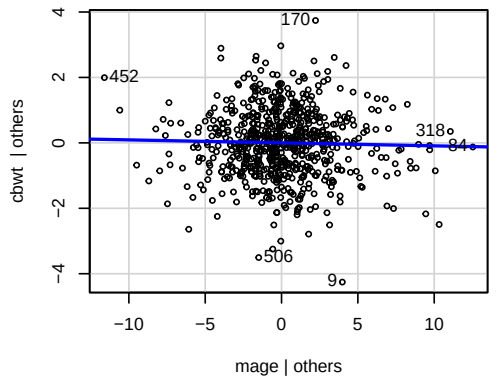

Residuals vs. page
Residuals
page

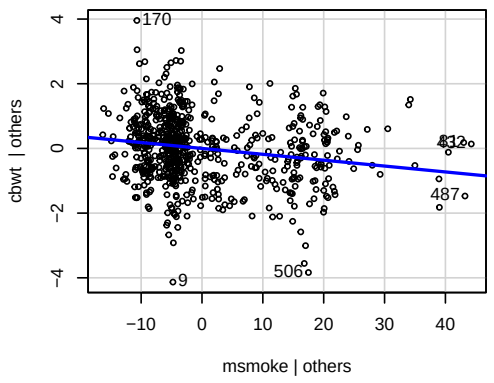

Residuals vs. ped
Residuals
ped

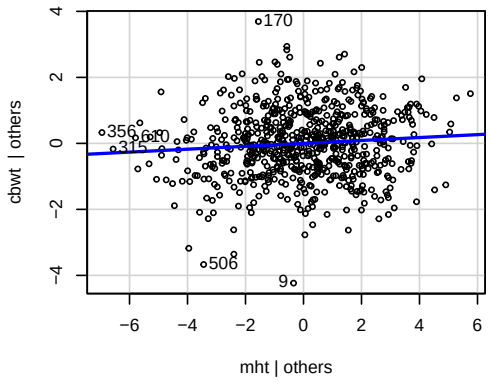

Residuals vs. psmoke
Residuals
psmoke

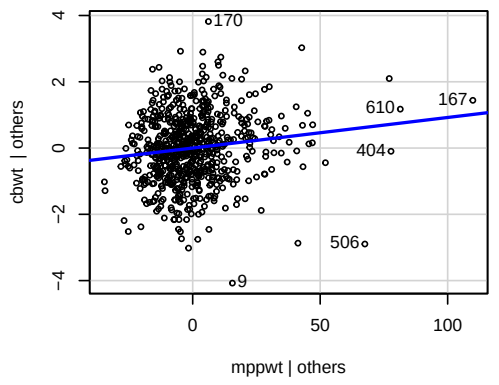

Residuals vs. pht
Residuals
pht

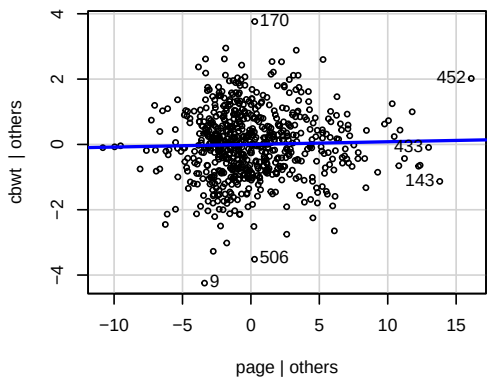

Box−Cox power transformation
log−likelihood
lambda
lambda of 1 is none (y^1); 0 is log(y) of any base

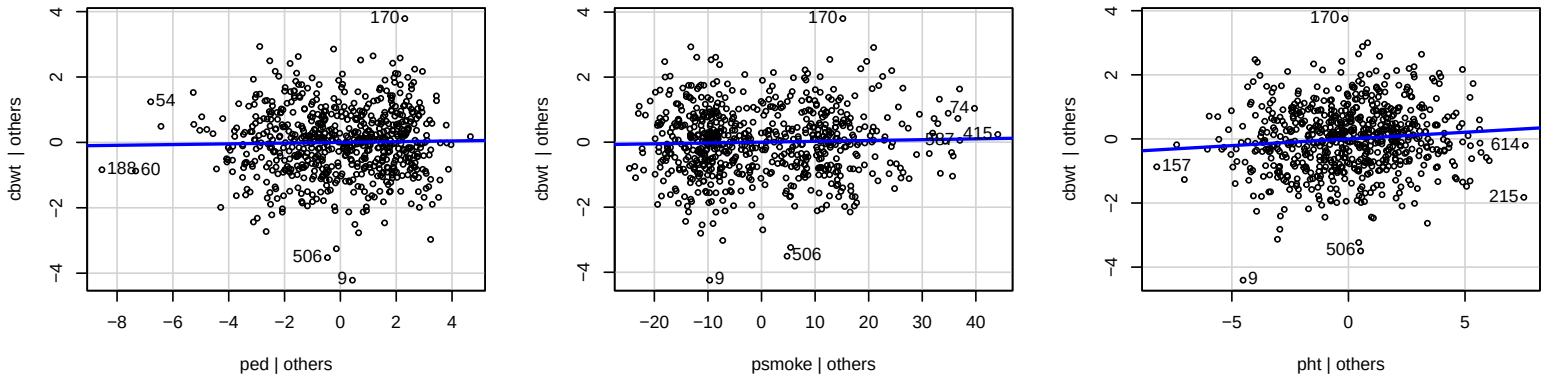

cbwt | others
mage | others


cbwt | others
msmoke | others


cbwt | others
mht | others


cbwt | others
mppwt | others


cbwt | others
page | others



Model selection starts here.

```
## AIC
# option: test="F" includes additional information
#           for parameter estimate tests that we're familiar with
# option: for BIC, include k=log(nrow( [data.frame name] ))
lm_cchd_red_AIC <- step(lm_cchd_full, direction="backward", test="F")
```

```
Start:  AIC=62.76
cbwt ~ mage + msmoke + mht + mppwt + page + ped + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- ped     1    0.3572 726.62 61.093  0.3301 0.5658197
- page    1    0.5416 726.81 61.265  0.5004 0.4795906
- mage    1    0.5484 726.81 61.272  0.5067 0.4768396
- psmoke  1    0.7833 727.05 61.491  0.7237 0.3952296
<none>                726.26 62.758
- mht     1    5.6711 731.94 66.048  5.2396 0.0223886 *
- pht     1    7.1324 733.40 67.404  6.5896 0.0104731 *
- mppwt   1   13.4808 739.74 73.265 12.4550 0.0004454 ***
- msmoke  1   26.3137 752.58 84.960 24.3114 1.034e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=61.09
cbwt ~ mage + msmoke + mht + mppwt + page + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- mage    1    0.4705 727.09 59.533  0.4351 0.5097278
- psmoke  1    0.6029 727.22 59.657  0.5576 0.4555018
- page    1    0.6150 727.24 59.668  0.5688 0.4510024
<none>                726.62 61.093
- mht     1    6.0455 732.67 64.727  5.5911 0.0183357 *
- pht     1    7.6382 734.26 66.204  7.0641 0.0080516 **
- mppwt   1   13.1673 739.79 71.305 12.1775 0.0005153 ***
- msmoke  1   26.0343 752.66 83.030 24.0772 1.162e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=59.53
cbwt ~ msmoke + mht + mppwt + page + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
```

```
- page    1    0.1513 727.24 57.674  0.1401  0.708332
- psmoke  1    0.6475 727.74 58.138  0.5994  0.439092
<none>                727.09 59.533
- mht     1    6.1466 733.24 63.257  5.6893  0.017344 *
- pht     1    7.4130 734.50 64.431  6.8615  0.009006 **
- mppwt   1   13.0524 740.14 69.632 12.0813  0.000542 ***
- msmoke  1   26.4382 753.53 81.820 24.4713 9.538e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=57.67
cbwt ~ msmoke + mht + mppwt + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- psmoke  1    0.6734 727.92 56.304  0.6241 0.4298016
<none>                727.24 57.674
- mht     1    6.1270 733.37 61.379  5.6784 0.0174507 *
- pht     1    7.2623 734.51 62.431  6.7306 0.0096834 **
- mppwt   1   13.7100 740.95 68.374 12.7062 0.0003902 ***
- msmoke  1   26.3602 753.60 79.886 24.4303 9.733e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=56.3
cbwt ~ msmoke + mht + mppwt + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                727.92 56.304
- mht     1    6.0566 733.97 59.938  5.6163 0.0180745 *
- pht     1    7.3497 735.27 61.135  6.8154 0.0092379 **
- mppwt   1   13.6367 741.55 66.925 12.6454 0.0004028 ***
- msmoke  1   25.9791 753.90 78.150 24.0905 1.154e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm_cchd_final <- lm_cchd_red_AIC
summary(lm_cchd_final)
```

```
Call:
lm(formula = cbwt ~ msmoke + mht + mppwt + pht, data = dat_cchd)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2726 -0.6965  0.0145  0.6661  3.8167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.643876   1.331579   0.484 0.628867
msmoke      -0.017376   0.003540  -4.908 1.15e-06 ***
mht          0.045320   0.019124   2.370 0.018074 *
mppwt        0.009129   0.002567   3.556 0.000403 ***
pht          0.041392   0.015855   2.611 0.009238 **
---
```

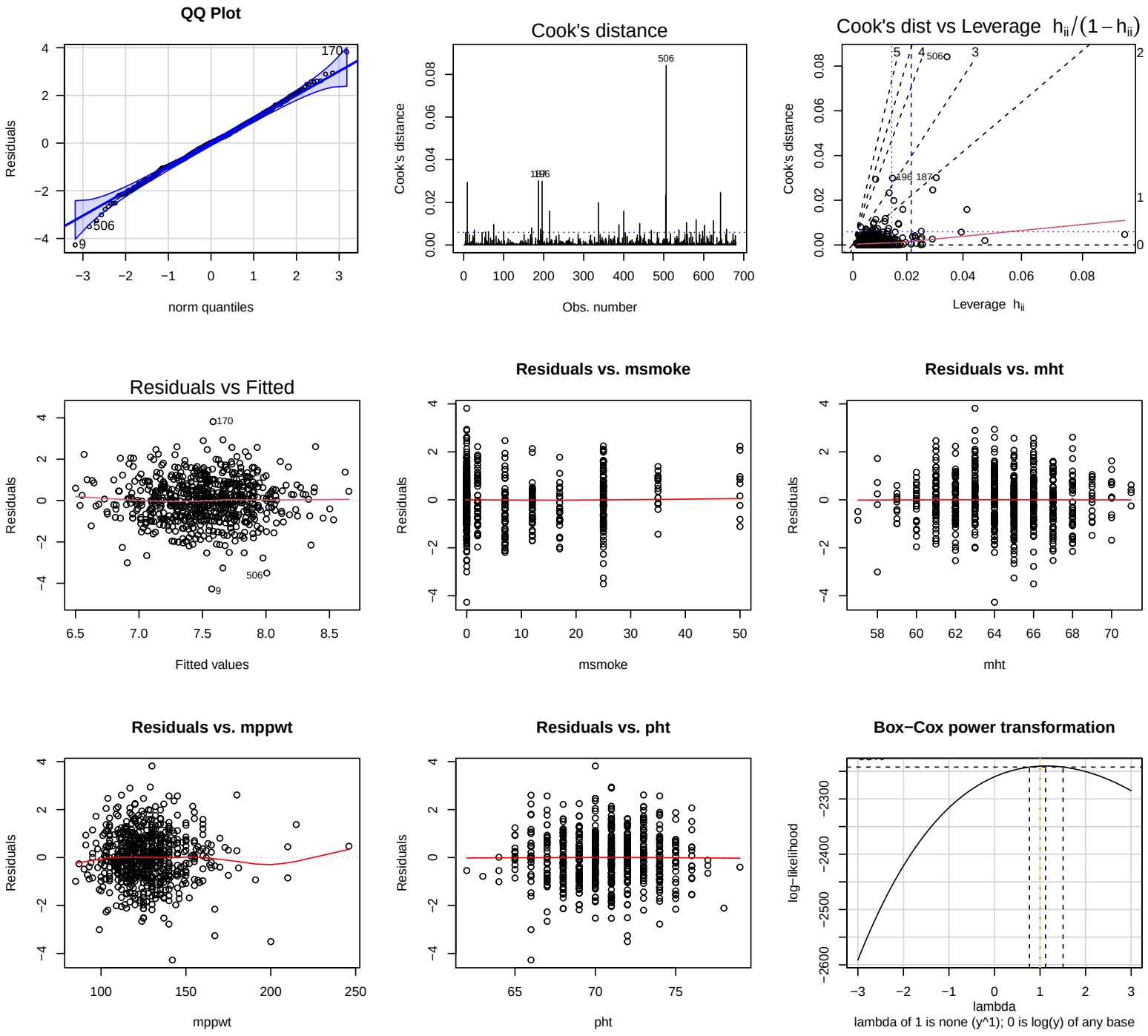
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.038 on 675 degrees of freedom
Multiple R-squared:  0.1016,    Adjusted R-squared:  0.09623
F-statistic: 19.07 on 4 and 675 DF,  p-value: 7.095e-15
```

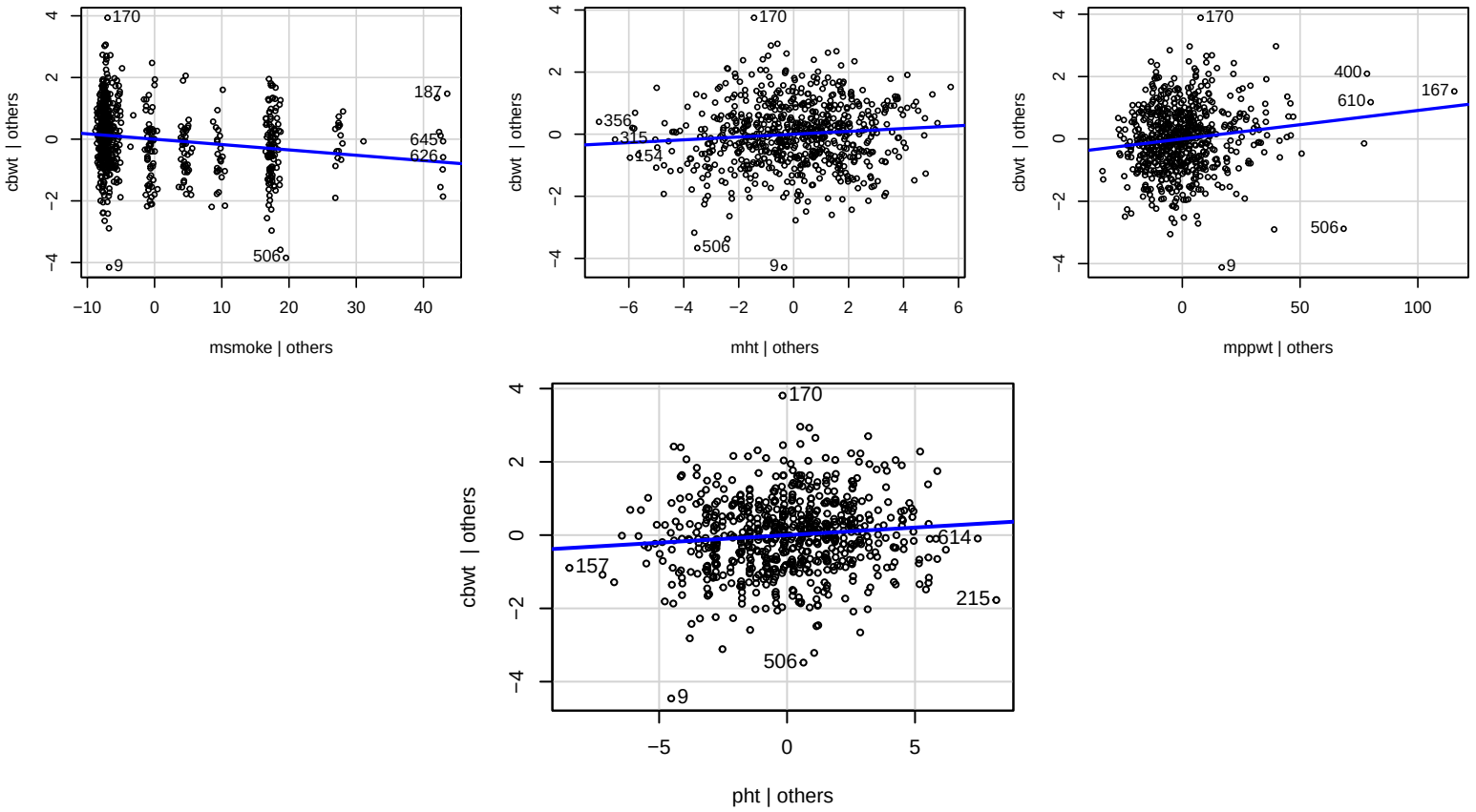
```r
# BIC (not shown)
# step(lm_cchd_full, direction="backward", test="F", k=log(nrow(dat_cchd)))
```

Backward selection results in a model with `msmoke`, `mht`, `mppwt`, and `pht` all significant at a 0.05 level.

Diagnostics

```r
# plot diagnostics
e_plot_lm_diagostics(lm_cchd_final, sw_plot_set = "simpleAV")
```





Discuss the diagnostics in terms of influential observations or problematic structure in the residuals. In particular, if an observation is influential, describe *how* it is influential; does it change the slope, intercept, or both for the regression surface?

Solution

Outliers with especially high Cook's distance, like observation 506, will change both the slope and intercept of the regression. Looking at the added-variable plots, I don't see this particular point obviously influencing things very much. The one case that looks problematic to me is 167, which may be major source of the positive relationship between maternal weight and birth weight. A DFBETAs analysis would help determine that.

(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.

Solution

The diagnostics all look pretty great. I'd be comfortable moving forward with this analysis. But let's see what happens if we remove observation 167.

```r
dat_cchd_rm <- dat_cchd %>%
  slice(-167)
lm_cchd_full_rm <- lm(cbwt ~ mage + msmoke + mht + mppwt
                        + page + ped + psmoke + pht
                      , data = dat_cchd_rm)
lm_cchd_red_AIC_rm <- step(lm_cchd_full_rm, direction="backward", test="F")
```

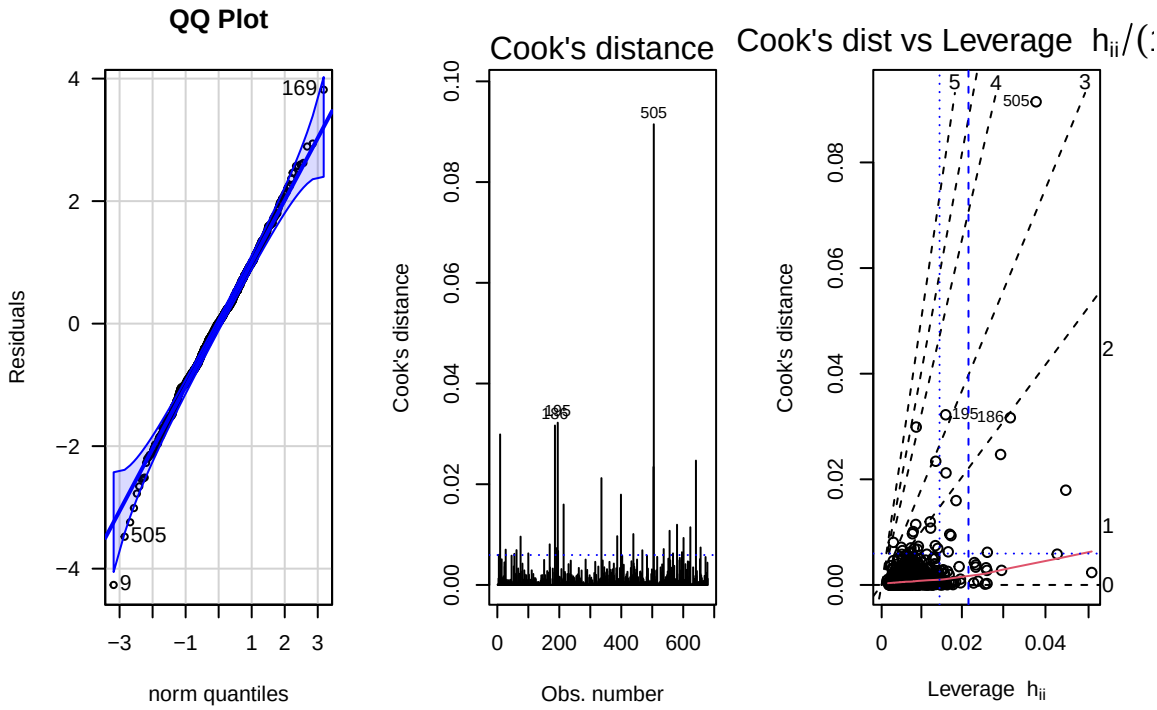
```
Start:  AIC=63.5
cbwt ~ mage + msmoke + mht + mppwt + page + ped + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- ped     1    0.4093 726.47 61.883  0.3777  0.539052
- page    1    0.5070 726.57 61.974  0.4679  0.494213
- mage    1    0.5096 726.57 61.976  0.4702  0.493129
- psmoke  1    0.7343 726.79 62.186  0.6776  0.410718
<none>                726.06 63.500
- mht     1    5.8346 731.89 66.935  5.3841  0.020621 *
- pht     1    7.1938 733.25 68.194  6.6384  0.010193 *
- mppwt   1   11.5341 737.59 72.202 10.6436  0.001161 **
- msmoke  1   26.5172 752.58 85.856 24.4698 9.555e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=61.88
cbwt ~ mage + msmoke + mht + mppwt + page + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- mage    1    0.4338 726.90 60.288  0.4006  0.526976
- psmoke  1    0.5508 727.02 60.397  0.5087  0.475937
- page    1    0.5879 727.06 60.432  0.5430  0.461448
<none>                726.47 61.883
- mht     1    6.1868 732.66 65.641  5.7144  0.017101 *
- pht     1    7.7124 734.18 67.053  7.1235  0.007792 **
- mppwt   1   11.2781 737.75 70.343 10.4170  0.001309 **
- msmoke  1   26.1830 752.65 83.924 24.1838 1.102e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=60.29
cbwt ~ msmoke + mht + mppwt + page + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- page    1    0.1577 727.06 58.435  0.1458  0.702751
- psmoke  1    0.5859 727.49 58.835  0.5416  0.462015
<none>                726.90 60.288
- mht     1    6.3122 733.21 64.159  5.8355  0.015972 *
- pht     1    7.5040 734.41 65.261  6.9372  0.008636 **
- mppwt   1   11.1100 738.01 68.587 10.2708  0.001415 **
- msmoke  1   26.6279 753.53 82.716 24.6167 8.872e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

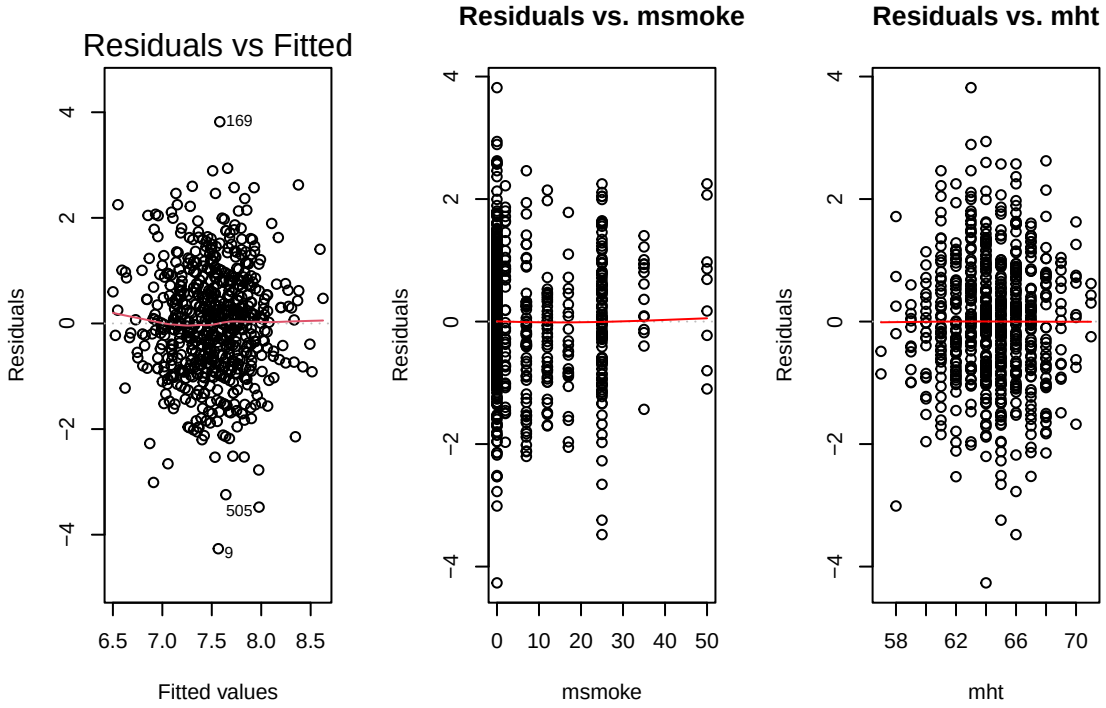
```
Step:  AIC=58.44
cbwt ~ msmoke + mht + mppwt + psmoke + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
- psmoke  1    0.6119 727.67 57.006  0.5664  0.451955
<none>                727.06 58.435
- mht     1    6.2882 733.35 62.282  5.8207  0.016105 *
- pht     1    7.3468 734.41 63.262  6.8005  0.009315 **
- mppwt   1   11.7029 738.76 67.277 10.8328  0.001049 **
- msmoke  1   26.5435 753.60 80.782 24.5699 9.079e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=57.01
cbwt ~ msmoke + mht + mppwt + pht

         Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                727.67 57.006
- mht     1    6.2557 733.93 60.819  5.7942   0.01635 *
- pht     1    7.4463 735.12 61.919  6.8971   0.00883 **
- mppwt   1   11.5267 739.20 65.678 10.6766   0.00114 **
- msmoke  1   26.2189 753.89 79.041 24.2851 1.047e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

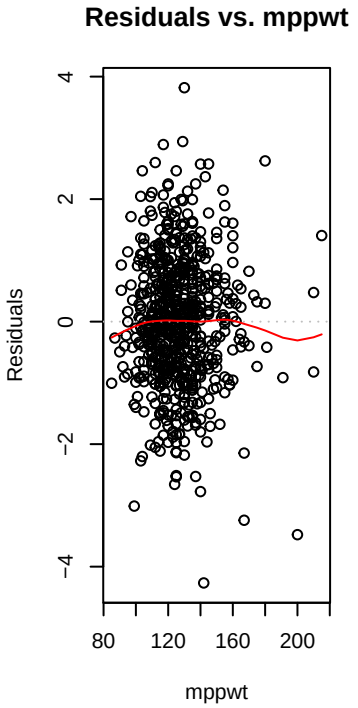
```
lm_cchd_final_rm <- lm_cchd_red_AIC_rm
summary(lm_cchd_final_rm)
```

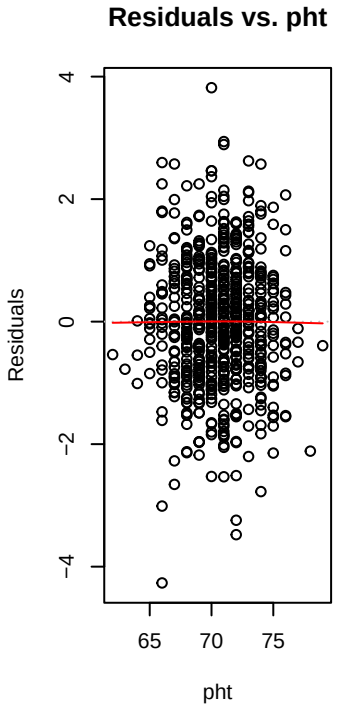
```
Call:
lm(formula = cbwt ~ msmoke + mht + mppwt + pht, data = dat_cchd_rm)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2659 -0.6988  0.0122  0.6698  3.8189

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.602857   1.335122   0.452  0.65175
msmoke      -0.017564   0.003564  -4.928 1.05e-06 ***
mht          0.046355   0.019257   2.407  0.01635 *
mppwt        0.008762   0.002682   3.268  0.00114 **
pht          0.041698   0.015877   2.626  0.00883 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.039 on 674 degrees of freedom
Multiple R-squared:  0.1006,    Adjusted R-squared:  0.09521
F-statistic: 18.84 on 4 and 674 DF,  p-value: 1.078e-14
```

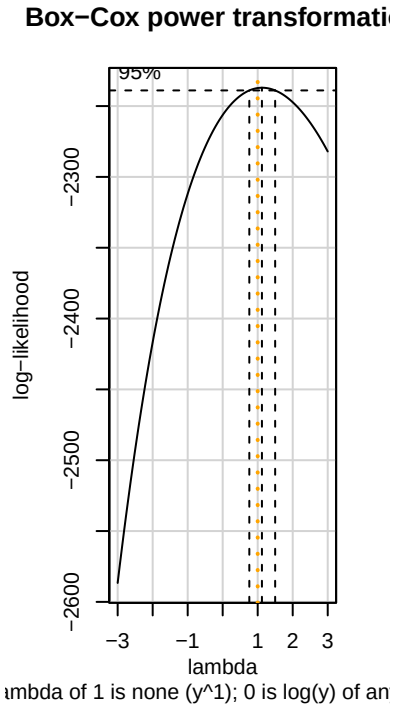
```
e_plot_lm_diagostics(lm_cchd_final_rm, sw_plot_set = "simpleAV")
```

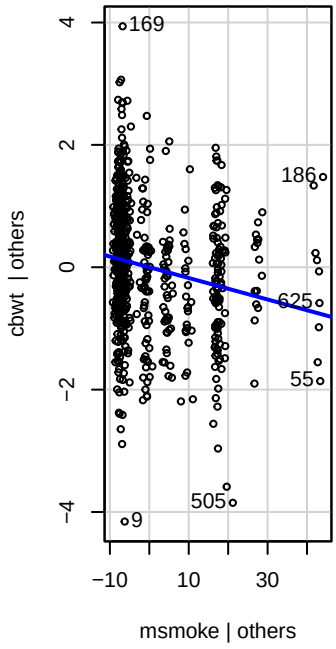

QQ Plot
Residuals
norm quantiles
169
505
9
Cook's distance
Cook's distance
Obs. number
505
Cook's dist vs Leverage $h_{ii}/(1$
Cook's distance
Leverage h_{ii}
5
4
3
505
195
186

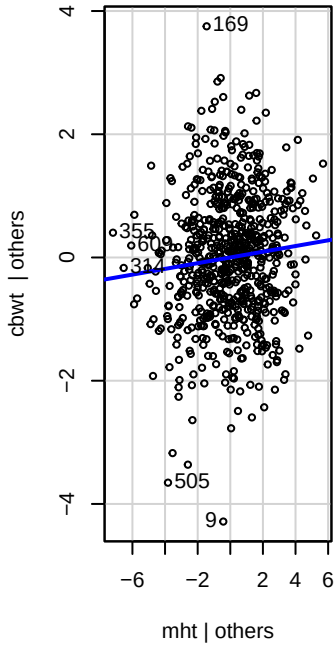

Residuals vs Fitted
Residuals
Fitted values
169
505
9
Residuals vs. msmoke
Residuals
msmoke
Residuals vs. mht
Residuals
mht

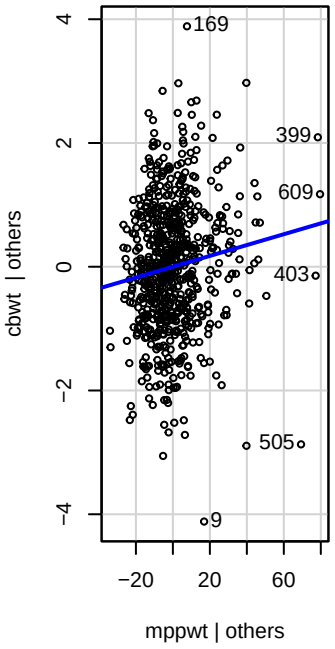

Residuals vs. mppwt
Residuals
mppwt

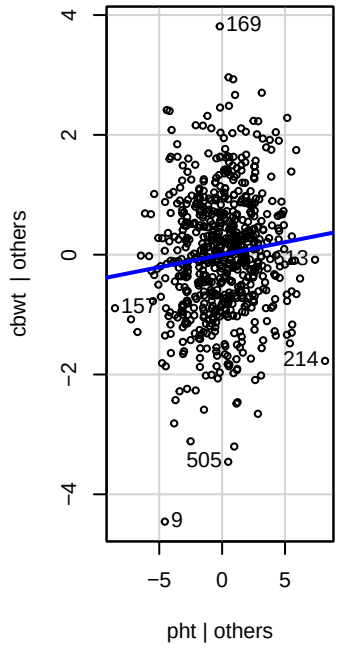

Residuals vs. pht
Residuals
pht


Box−Cox power transformati
95%
log−likelihood
lambda
ambda of 1 is none (y^1); 0 is log(y) of an


cbwt | others
msmoke | others


cbwt | others
mht | others


cbwt | others
mppwt | others



No difference. Let's stick with the original model.

(3 p) Interpret 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.)

Finally, write the equation for the final model and interpret each model coefficient. Do these quantities make sense?

Solution

The R^2 for this model is 0.10, indicating the model explains 10% of the variation in the response.

The equation is:

$BW_i = 0.6 - 0.018 * MSMOKE_i + 0.46 * MHT_i + 0.009 * MPPWT_i + 0.042 * PHT_i$. This all makes good sense. Smoking is BAD, so it reduces birth weight. That metrics of maternal and paternal body size positively relate to child size is not surprising.

(1 p) Inference to whom

To which population of people does this model make inference to? Does this generalize to all humans?

Sometimes this is call the "limitations" section. By carefully specifying what the population is that inference applies to, often that accounts for the limitations.

Solution

Oh, goodness no, this does not necessarily generalize to the entire population, since we've restricted attention to only white, male babies (and white parents by extension). It might generalize to the broader population, and (frankly), we might expect it to, given our particular findings, but it doesn't have to. We ought to do more investigation of these patterns for other demographics to see.