Homework 9

Kevin Linares and Jamila Sani

2024-11-22

```
library(jtools)
library(leaps)
library(car)
library(gratia)
library(dplyr)
library(faraway)
library(huxtable)
library(MASS)
library(mgcv)
library(tidyverse)

options(scipen=999)
```

1. Faraway Chapter 9. Exercise 7. Use the cheddar data for this question.

```
data("cheddar")
summary(cheddar)
```

```
     taste           Acetic           H2S             Lactic     
 Min.   : 0.70   Min.   :4.477   Min.   : 2.996   Min.   :0.860  
 1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978   1st Qu.:1.250  
 Median :20.95   Median :5.425   Median : 5.329   Median :1.450  
 Mean   :24.53   Mean   :5.498   Mean   : 5.942   Mean   :1.442  
 3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575   3rd Qu.:1.667  
 Max.   :57.20   Max.   :6.458   Max.   :10.199   Max.   :2.010  
```

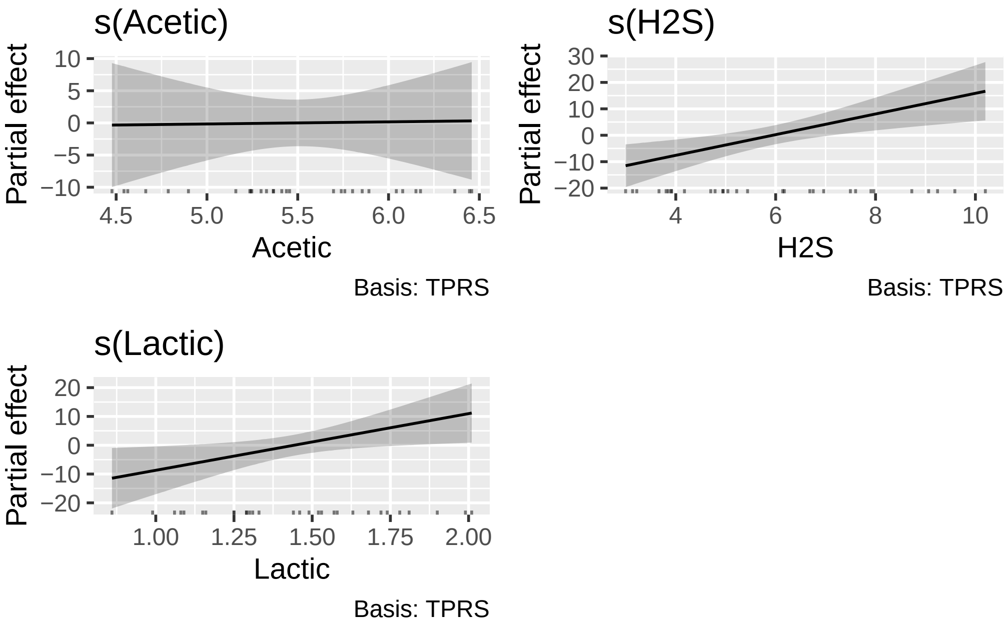
1.A. Fit a generalized additive model (GAM) for a response of taste with the other three variables as predictors. Do the predictors appear to have a non-linear relationship with the outcome?

- Based on our plots for the predictors in the GAM we do not see evidence of non-linear relationship with Taste as the outcome variable.

```
# fit a GAM to the data
summary(mod_taste_gam <- gam(taste ~ s(Acetic) + s(H2S) + s(Lactic), data=cheddar))
```

```
Family: gaussian
Link function: identity

Formula:
taste ~ s(Acetic) + s(H2S) + s(Lactic)

Parametric coefficients:
            Estimate Std. Error t value          Pr(>|t|)
(Intercept)    24.53       1.85   13.26 0.000000000000441 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(Acetic)   1      1 0.005 0.94198
s(H2S)      1      1 9.818 0.00425 **
s(Lactic)   1      1 5.196 0.03108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.612   Deviance explained = 65.2%
GCV = 118.42  Scale est. = 102.63    n = 30
```

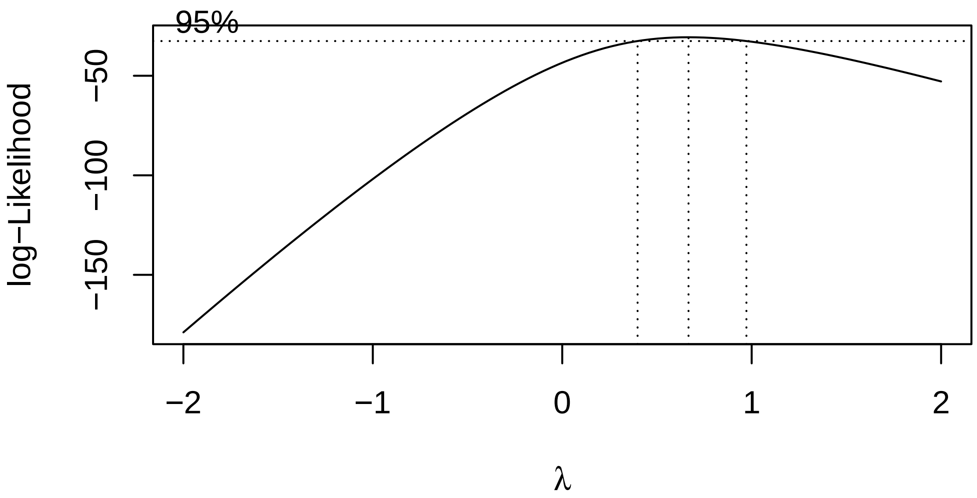
```
gratia::draw(mod_taste_gam)
```



1.B. Use the Box-Cox method to determine an optimal transformation of the response. Would it be reasonable to leave the response as is (i.e., no transformation)?

- The Lambda value is between .50 and 1. We can either leave the outcome variable un-transformed or try a squared root transformation.

```
mod_taste_lm <- lm(taste ~., data = cheddar)
boxcox(mod_taste_lm)
```



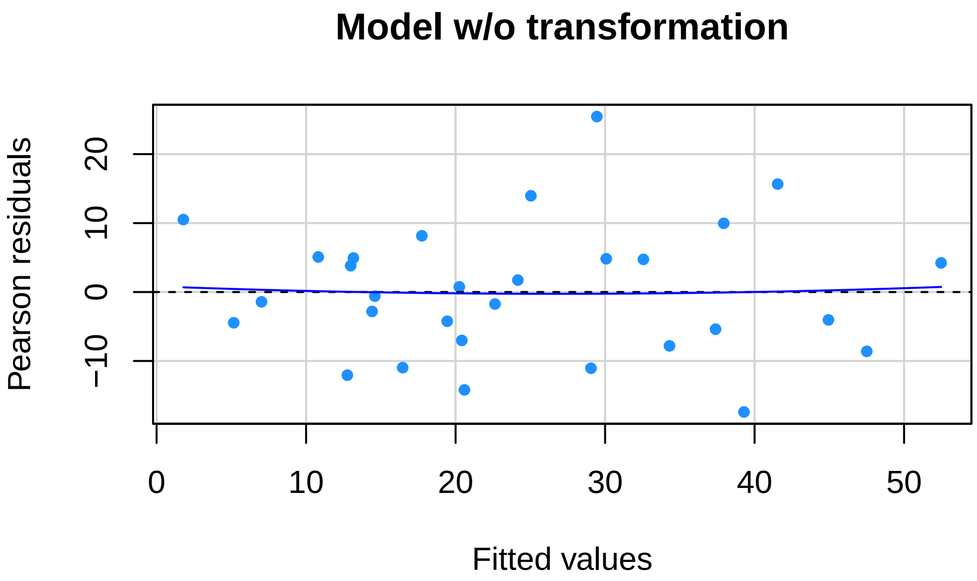
- If we do decide to squared root transform the outcome response, our residual plot appears to show more evidence on non-constant variance than the residual plot before the transformation.

```
mod_taste_lm_t <- lm(I(sqrt(taste)) ~ ., data = cheddar)

export_summs(mod_taste_lm, mod_taste_lm_t)
```

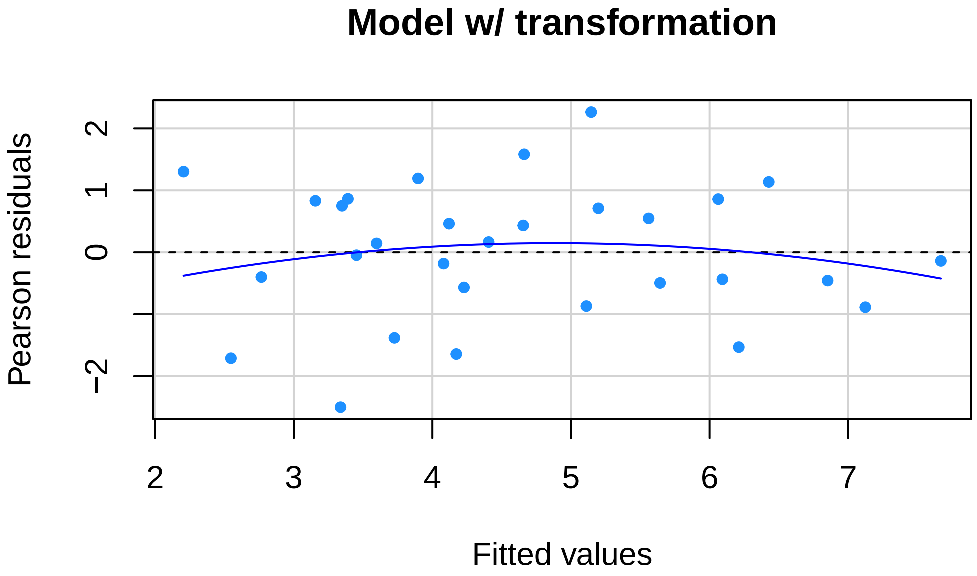
```
residualPlot(mod_taste_lm, pch=20, col= "dodgerblue",
             main="Model w/o transformation")
```

	Model 1	Model 2
(Intercept)	-28.88	-0.92
	(19.74)	(2.26)
Acetic	0.33	-0.00
	(4.46)	(0.51)
H2S	3.91 **	0.44 **
	(1.25)	(0.14)
Lactic	19.67 *	2.02
	(8.63)	(0.99)
N	30	30
R2	0.65	0.63

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.



```r
residualPlot(mod_taste_lm_t, pch=20, col= "dodgerblue",
             main="Model w/ transformation")
```



2. Faraway Chapter 10. Exercise 2. Using the teengamb dataset with gamble as the response and the other variables. Implement the following variable selection methods to determine the "best" model.

```r
data("teengamb")
glimpse(teengamb)
```

```
Rows: 47
Columns: 5
$ sex    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ~
$ status <int> 51, 28, 37, 28, 65, 61, 28, 27, 43, 18, 18, 43, 30, 28, 38, 38,~
$ income <dbl> 2.00, 2.50, 2.00, 7.00, 2.00, 3.47, 5.50, 6.42, 2.00, 6.00, 3.0~
$ verbal <int> 8, 8, 6, 4, 8, 6, 7, 5, 6, 7, 6, 6, 4, 6, 6, 8, 8, 5, 8, 9, 8, ~
$ gamble <dbl> 0.00, 0.00, 0.00, 7.30, 19.60, 0.10, 1.45, 6.60, 1.70, 0.10, 0.~
```

2.A. Backward elimination (based on the significance of predictors)

- The full model contains gamble as a function of sex, income, status, and verbal. Given our significance level, we remove status, and finally status and verbal together to refit the models.

```r
summ(mod_back <- lm(gamble ~., teengamb))
```

```
MODEL INFO:
Observations: 47
Dependent Variable: gamble
Type: OLS linear regression

MODEL FIT:
F(4,42) = 11.69, p = 0.00
R² = 0.53
Adj. R² = 0.48

Standard errors:OLS
--------------------------------------------------
                      Est.    S.E.   t val.      p
----------------- --------- -------- --------- ------
(Intercept)          22.56   17.20     1.31   0.20
sex                 -22.12    8.21    -2.69   0.01
status                0.05    0.28     0.19   0.85
income                4.96    1.03     4.84   0.00
verbal               -2.96    2.17    -1.36   0.18
--------------------------------------------------
```

```r
# drop status
summ(mod_back_status <- update(mod_back,.~. -status))
```

```
MODEL INFO:
Observations: 47
Dependent Variable: gamble
Type: OLS linear regression

MODEL FIT:
F(3,43) = 15.93, p = 0.00
R² = 0.53
Adj. R² = 0.49
```

```
Standard errors:OLS
--------------------------------------------------
                      Est.    S.E.   t val.      p
----------------- -------- ------- -------- ------
(Intercept)          24.14   14.77     1.63   0.11
sex                 -22.96    6.77    -3.39   0.00
income                4.90    0.96     5.13   0.00
verbal               -2.75    1.83    -1.50   0.14
--------------------------------------------------
```

```r
# drop status and verbal
summ(mod_back_status_verbal <- update(mod_back,.~. -status-verbal))
```

```
MODEL INFO:
Observations: 47
Dependent Variable: gamble
Type: OLS linear regression

MODEL FIT:
F(2,44) = 22.12, p = 0.00
R² = 0.50
Adj. R² = 0.48

Standard errors:OLS
-------------------------------------------------
                      Est.   S.E.   t val.      p
----------------- -------- ------ -------- ------
(Intercept)           4.04   6.39     0.63   0.53
sex                 -21.63   6.81    -3.18   0.00
income                5.17   0.95     5.44   0.00
-------------------------------------------------
```

2.B. Now use AIC. Which is the "best" model?

- The model with sex + income + verbal has the lower AIC, and we determine it is the "best" model for these data based on this criterion.

```
back.mod <- stepAIC(mod_back, direction = "backward")
```

```
Start:  AIC=298.18
gamble ~ sex + status + income + verbal

         Df Sum of Sq   RSS    AIC
- status  1      17.8 21642 296.21
<none>                21624 298.18
- verbal  1     955.7 22580 298.21
- sex     1    3735.8 25360 303.67
- income  1   12056.2 33680 317.00

Step:  AIC=296.21
gamble ~ sex + income + verbal

         Df Sum of Sq   RSS    AIC
<none>                21642 296.21
- verbal  1    1139.8 22781 296.63
- sex     1    5787.9 27429 305.35
- income  1   13236.1 34878 316.64
```

```
back.mod$anova
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
			42	2.16e+04	298
- status	1	17.8	43	2.16e+04	296

2.C. Now use adjusted R2. Which is the "best" model?

- "Best" model is gamble ~ sex + income + verbal, the adjusted r-squared is close to the full model while dropping status or one less parameter to estimate.

```r
huxtable::huxreg(
  mod_back, mod_back_status, mod_back_status_verbal,
  statistics=c( "adj.r.squared"))
```

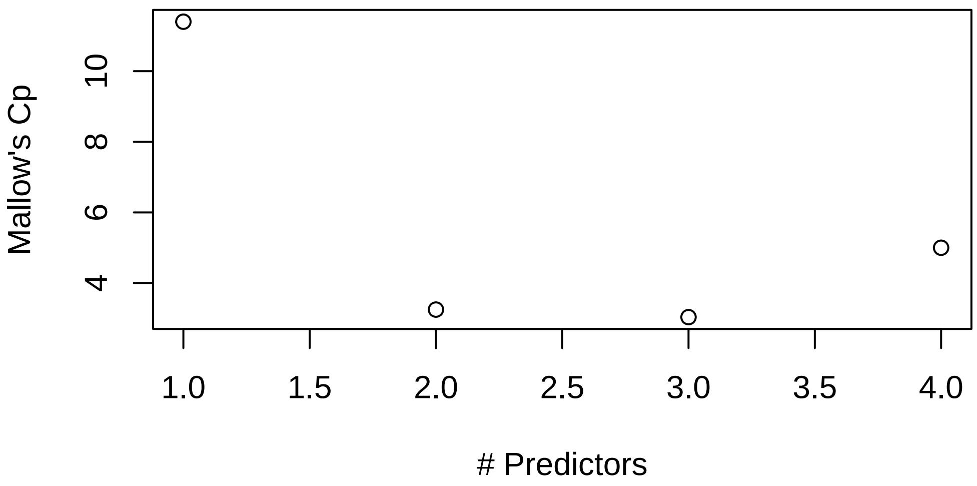
	(1)	(2)	(3)
(Intercept)	22.556	24.139	4.041
	(17.197)	(14.769)	(6.394)
sex	-22.118 *	-22.960 **	-21.634 **
	(8.211)	(6.771)	(6.809)
status	0.052		
	(0.281)		
income	4.962 ***	4.898 ***	5.172 ***
	(1.025)	(0.955)	(0.951)
verbal	-2.959	-2.747	
	(2.172)	(1.825)	
adj.r.squared	0.482	0.493	0.479

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

2.D. Now use Mallows Cp. Which is the "best" model?

- "Best" model is gamble ~ sex + income + verbal based on having the lowest Mllow c_p value.

```r
sub1<-regsubsets(gamble ~.,teengamb)
rsub1<-summary(sub1)
plot(I(1:4), rsub1$cp, ylab="Mallow's Cp", xlab="# Predictors")
```



```r
rsub1$which
```

```
  (Intercept)   sex status income verbal
1        TRUE FALSE  FALSE   TRUE  FALSE
2        TRUE  TRUE  FALSE   TRUE  FALSE
3        TRUE  TRUE  FALSE   TRUE   TRUE
4        TRUE  TRUE   TRUE   TRUE   TRUE
```