Stats_101A_hw8_anna_piskun

Anna Piskun

2/25/2020

###Part A:

Load the dietstudy data into R (from "data used in class" folder). Create a new subset of data that includes only these variables: DIET, AGE, SEX, WEIGHT_0, DROPOUT2, WEIGHT_2, ADHER_2).

```
setwd("~/Desktop")
dietstudy <- read.csv("dietstudy.csv")
new.dietstudy <- dplyr::select(dietstudy, DIET, AGE, SEX, WEIGHT_0, DROPOUT2, WEIGHT_2, ADHER_2)
```

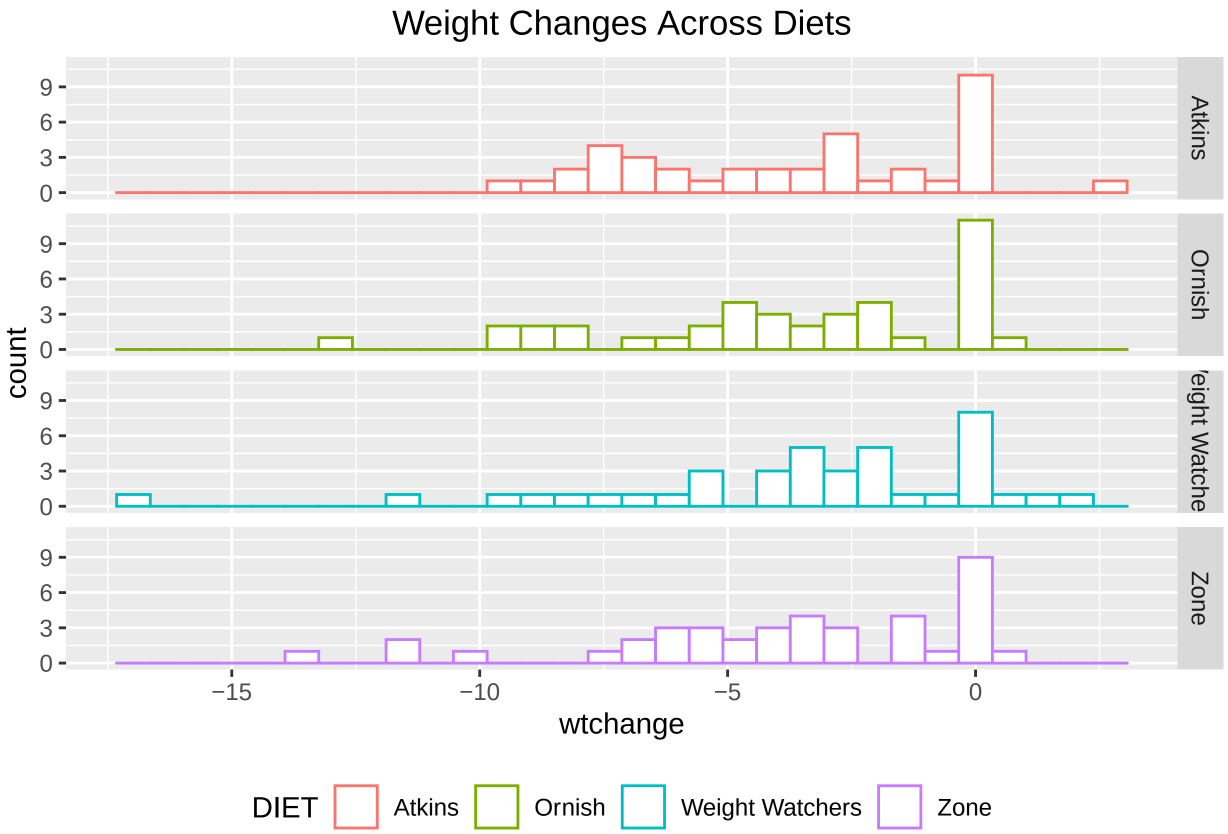
To this dataframe add a new variable that represents the change in weight after two months: wtchange=WEIGHT_2-WEIGHT_0

```
new.df <- dplyr::mutate(new.dietstudy, wtchange = WEIGHT_2-WEIGHT_0)
```

The data come from a randomized study to determine which data was best for losing weight at 2, 6 and 12 months. We'll examine only 2 month weight change. WEIGHT_0 is baseline weight, and WEIGHT_2 is weight after two months. DROPOUT2 is an indicator variable that indicates whether the subject dropped out of the study. ADHER_2 measures how well the subject adhered to the diet, with higher scores indicating higher adherence (self-reported).

a) Make a graphic to compare weight changes across diets. Based on this plot, which diet, if any, would you conclude was most effective?

```
library(ggplot2)
ggplot(new.df, aes(x=wtchange, color=DIET)) +
  geom_histogram(fill= "white", position="dodge", alpha=1, bins = 30) +
  facet_grid(DIET ~ .) + ggtitle("Weight Changes Across Diets") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "bottom")
```



Looking at the distribution of the four diets, no one diet stands out as being significantly more effective than the other. They all have similar ranges and sizes, and all four diets display a relatively equal proportion of negative values.

b) Note that some weight change values are exactly equal to 0? Why is this? Explain why, and then drop the 0 values from all subsequent analyses.

```
new.df$DROPOUT2[new.df$wtchange == 0]
```

```
##  [1] yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes
## [20] yes yes yes yes yes yes yes yes yes yes yes yes yes yes yes
## Levels: no yes
```

```
new.df$ADHER_2[new.df$wtchange == 0]
```

```
##  [1] 3 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Weight change values equal to 0 represents the subjects that experienced no weight loss on the specific diet they were assigned. For the wtchange values that equalled 0, they also all dropped out of the study (we see this by looking at the DROUPOUT2 variable) and generally had very low adherance levels (>3). By removing the zeros we remove individuals who did not adhere to the diet and then later dropped out. This will lead to a more representative and accurate model.

```
clean.df <- dplyr::filter(new.df, wtchange != 0)
```

c) Create a linear model that includes as predictors age, diet, sex, baseline weight and adherence. What does this model say about the effectiveness of the diets? Based on this model, what should a physician tell her patients about losing weight? (Don't worry, for now, about assessing model validity.)

```
model <- lm(wtchange~AGE+SEX+WEIGHT_0+ADHER_2+DIET, data = clean.df)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: wtchange
##            Df Sum Sq Mean Sq F value    Pr(>F)
## AGE         1   9.08    9.08  1.3703    0.2441
## SEX         1 112.19  112.19 16.9307 7.210e-05 ***
## WEIGHT_0    1  49.75   49.75  7.5073    0.0071 **
## ADHER_2     1 422.55  422.55 63.7686 1.036e-12 ***
## DIET        3   3.11    1.04  0.1566    0.9252
## Residuals 118 781.91    6.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model)
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + SEX + WEIGHT_0 + ADHER_2 + DIET,
##     data = clean.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5178 -1.2538 -0.0252  1.6350  5.9320
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.142936   2.094564   2.455   0.0155 *
## AGE                 -0.003341   0.024284  -0.138   0.8908
## SEXMale             -0.957940   0.500626  -1.913   0.0581 .
## WEIGHT_0            -0.027415   0.016431  -1.668   0.0979 .
## ADHER_2             -0.871638   0.109861  -7.934 1.36e-12 ***
## DIETOrnish           0.154200   0.669211   0.230   0.8182
## DIETWeight Watchers -0.217142   0.660208  -0.329   0.7428
## DIETZone            -0.253694   0.661869  -0.383   0.7022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.574 on 118 degrees of freedom
## Multiple R-squared:  0.4328, Adjusted R-squared:  0.3992
## F-statistic: 12.86 on 7 and 118 DF,  p-value: 3.322e-12
```

Looking at the anova table of our model, the high p-value for the variable indicates that no matter which type is used, diet is not statistically significant nor useful for our model in predicting weight change given that we have already accounted for age, sex, initial weight, and adherence level. Given age, sex, and starting weight, adherence to diet is the only statistically significant variable. Therefore the physician can tell her patients that with regards to losing weight there is no evidence that one diet is better than the other, but what is most important is that they choose a diet and stick to it.

d) Interpret the "DIETOrnish" slope.

Since DIETOrnish is not statistically significant when compared to DIETAtkins due to the p-value being greater than 0.05, there is no statistically significant difference in weight change between the two diets.

e) It appears that those who adhere to the diet tend to lose more weight. However, some diets might be easier to adhere to than others. Add an interaction effect to test whether the effect of adherence is the same for the diets.

```
model1 <- update(model, . ~ . + ADHER_2:DIET)
summary(model1)
```

```
##
## Call:
## lm(formula = wtchange ~ AGE + SEX + WEIGHT_0 + ADHER_2 + DIET +
##     ADHER_2:DIET, data = clean.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0759 -1.2948 -0.0646  1.5416  6.0222
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.857858   2.532731   1.918   0.0576 .
## AGE                          -0.004431   0.024607  -0.180   0.8574
## SEXMale                      -1.028814   0.525928  -1.956   0.0529 .
## WEIGHT_0                     -0.026165   0.017010  -1.538   0.1267
## ADHER_2                      -0.839098   0.191953  -4.371 2.72e-05 ***
## DIETOrnish                   -0.718937   2.520455  -0.285   0.7760
## DIETWeight Watchers           0.858656   2.077323   0.413   0.6801
## DIETZone                     -0.050935   2.224800  -0.023   0.9818
## ADHER_2:DIETOrnish            0.111664   0.318882   0.350   0.7268
## ADHER_2:DIETWeight Watchers -0.156166   0.278737  -0.560   0.5764
## ADHER_2:DIETZone             -0.025607   0.295947  -0.087   0.9312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.599 on 115 degrees of freedom
## Multiple R-squared:  0.4364, Adjusted R-squared:  0.3874
## F-statistic: 8.904 on 10 and 115 DF,  p-value: 1.029e-10
```

Since there is no statistically significant coefficients on the interactions, there is no statistical significance that adherence is different among the four diets.

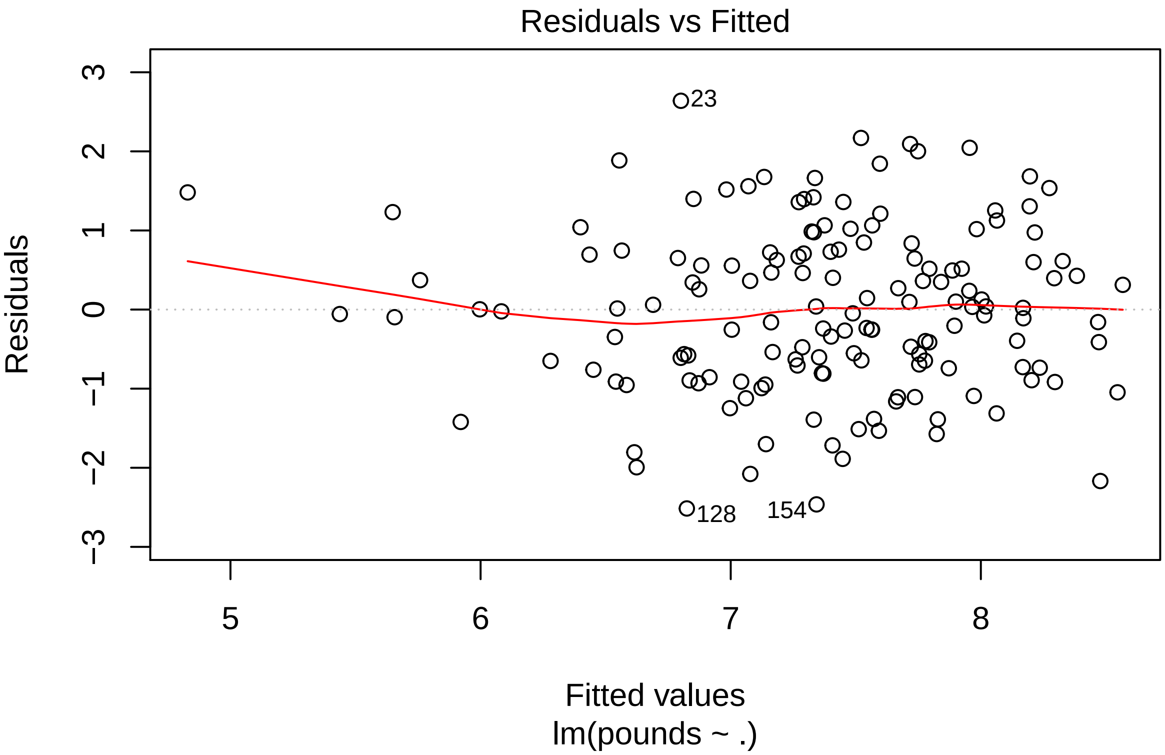
###Part B:

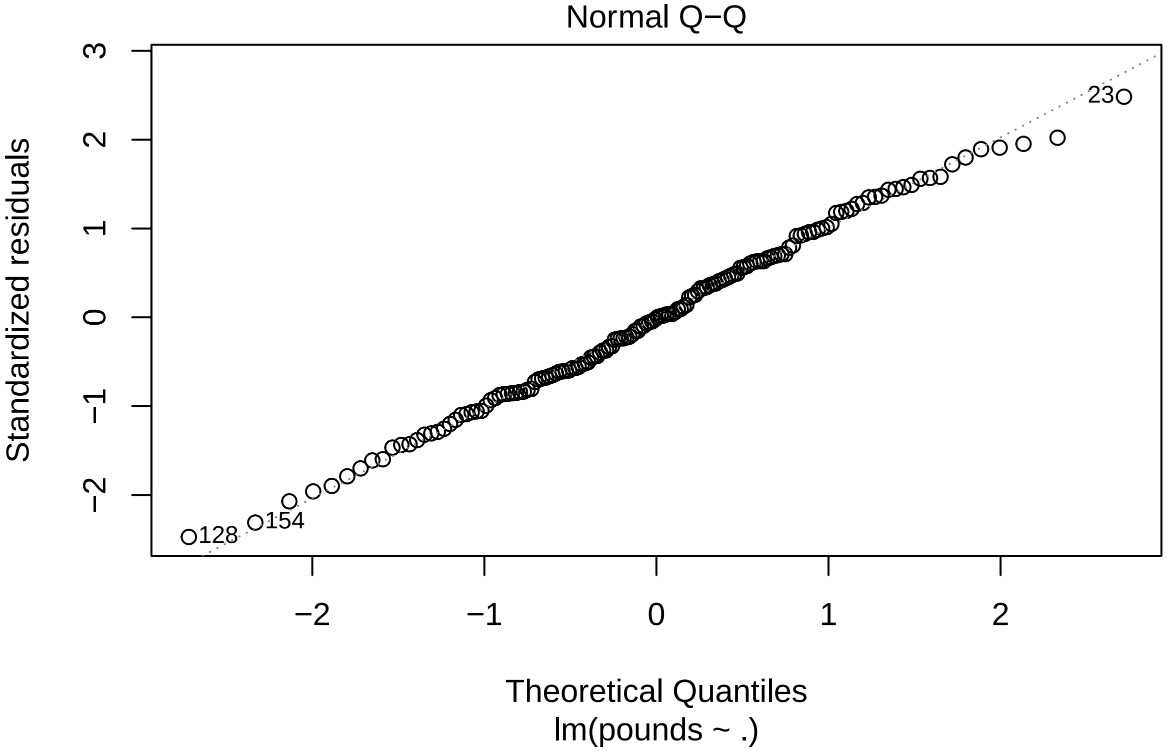
To answer these questions, use the NCbirths data file in the folder under Site Info. We'll return to these data in HW 9.

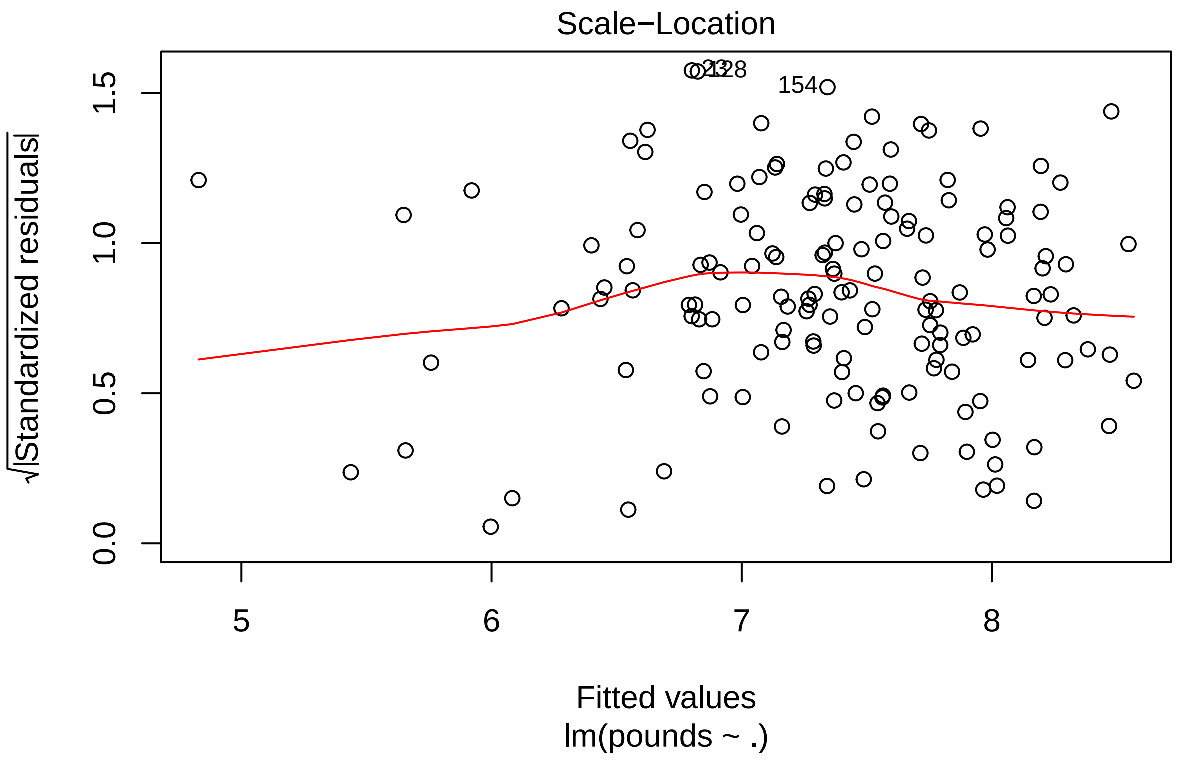
```
births <- read.csv("NCbirths.csv", row.names = 1)
```

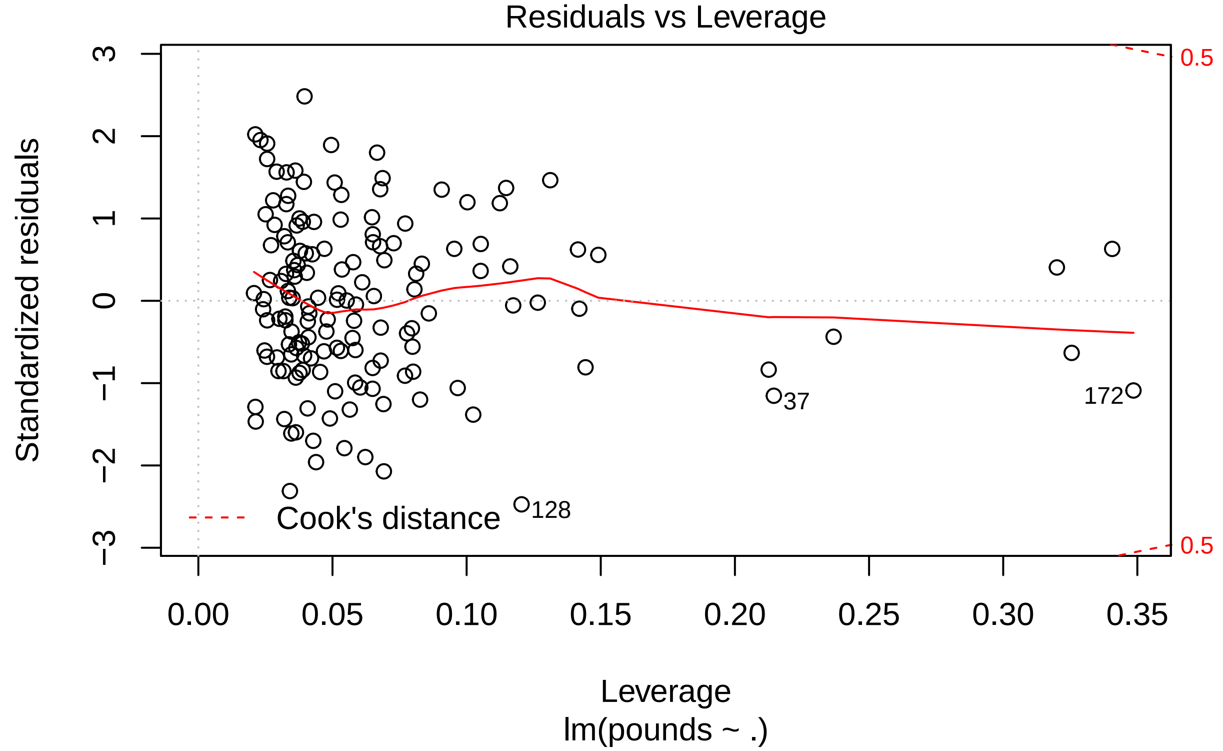
a) Fit a model using all of the available predictors to predict the weight of the baby (pounds). A description of the variables is below. Are the conditions of model validity satisfied? Explain which are and which aren't, and provide appropriate graphical support.

```
model2 <- lm(pounds~., data = births)
plot(model2)
```


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
−3
23
128
154
5
6
7
8
Fitted values
lm(pounds ~ .)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
23
128
154
−2
−1
0
1
2
Theoretical Quantiles
lm(pounds ~ .)





Looking at the standardized residual plot there is no clear trend or pattern therefore the linearity condition is satisfied. Likewise, there is no fan-shape pattern in the residuals so the constant variance condition is also satisfied. The normal QQ plot has very little to no deviation from the straight line, therefore the errors are normally distributed. Since there is no increasing or decreasing trend in the scale-location plot, this again confirms that the constant variance condition is satisfied.

b) Create marginal model plots for each predictor. Which predictors are not fit well by the model?

```r
library(alr3)
```

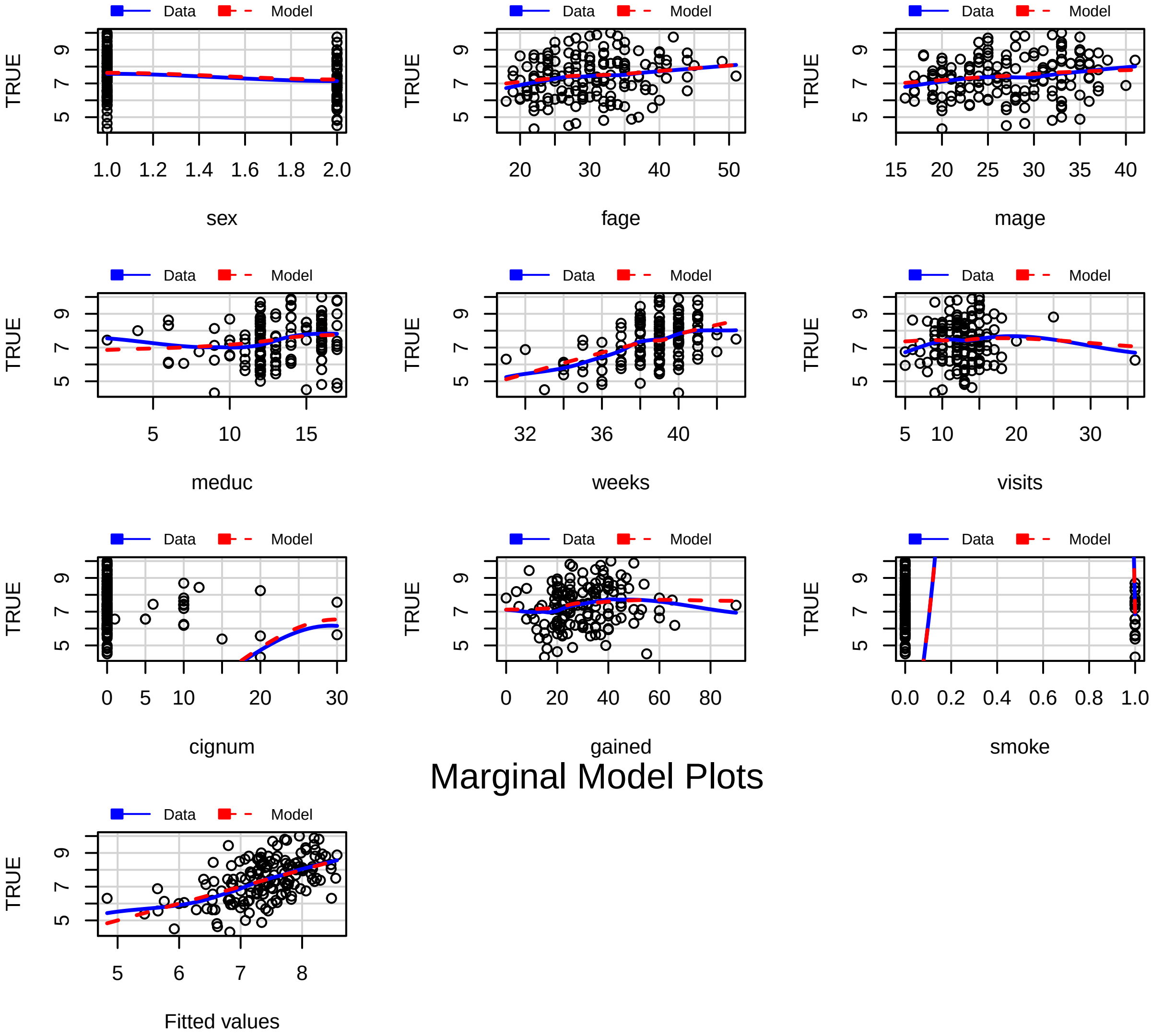
```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

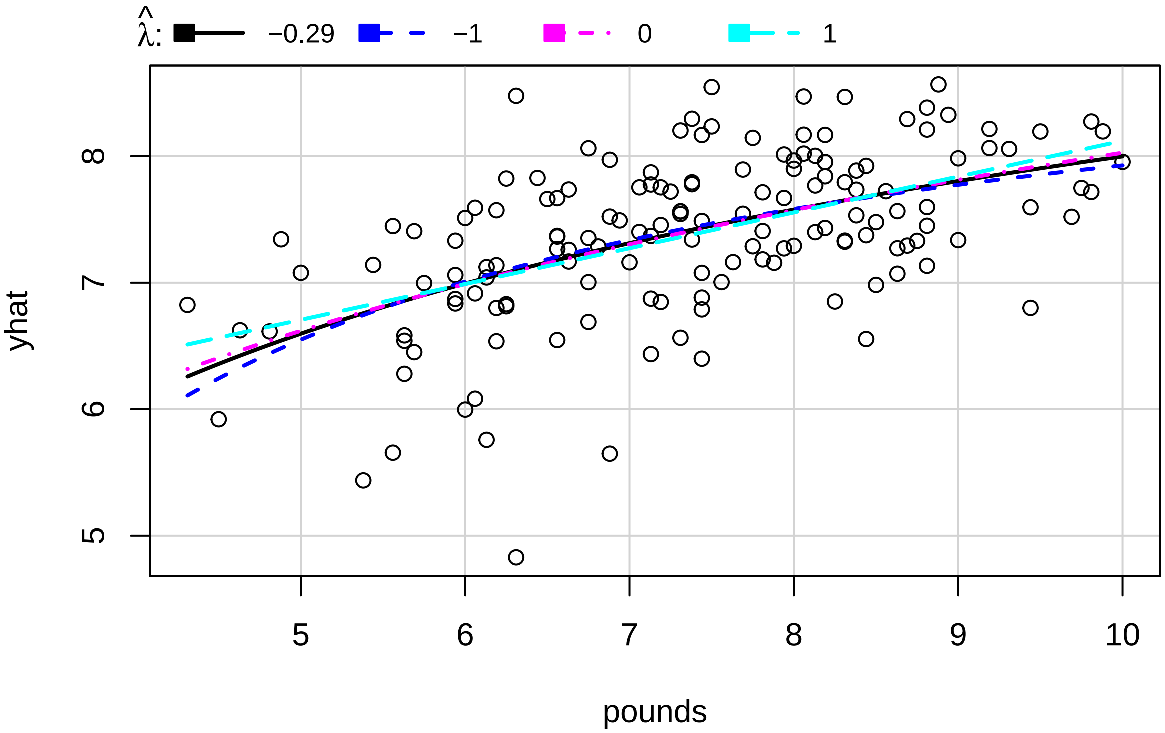
```r
mmps(model2)
```



While all of the predictors are fit relatively well by the model (the regression line and loess line are about the same), the *meduc*, *gained*, and *visits* variables can be improved upon as their loess line has slightly more deviation from the regression line.

c) Use both an inverse response plot and a box-cox approach to consider a transform for the response variable. What do they recommend? (Hint: apply summary to the powerTransform() function.)

```
invResPlot(model2)
```



```
##       lambda      RSS
## 1 -0.2892921 46.77429
## 2 -1.0000000 46.94127
## 3  0.0000000 46.80061
## 4  1.0000000 47.25936
```

```
summary(powerTransform(model2))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored

## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1     0.969           1       0.1761        1.762
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df     pval
## LR test, lambda = (0) 6.007746  1 0.014243
##
## Likelihood ratio test that no transformation is needed
##                                 LRT df    pval
## LR test, lambda = (1) 0.005859047  1 0.93899
```

The inverse response plot suggests raising the response variable weight to the power of -0.2892921, but since the RSS is nearly the same for this transform as the RSS for the original model, and the lines are almost the same, it is best to stick with the original, more simplified model. Looking at the boxcox method, since there is a small p-value (<0.05) for lambda = 0, we reject the null hypothesis that we should do a log transform. Looking at the pvalue for lambda = 1 it is greater than 0.05, therefore we fail to reject the null hypothesis and the untransformed model is best. Thus, the BoxCox method does not recommend a model transform.

d) Use the Box-Cox approach to consider transforms for the predictor variables. Do NOT include any categorical variables. Which variables are suggested for transformations and what are these transformations?

```r
summary(powerTransform(cbind(births$fage, births$mage, births$meduc, births$weeks, births$visits + 0.1,
```

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.8482       -1.00      -1.4134      -0.2830
## Y2    0.1779        0.00      -0.4361       0.7920
## Y3    2.1376        2.00       1.6248       2.6504
## Y4    7.5044        7.50       4.9497      10.0590
## Y5    0.2392        0.00      -0.0883       0.5666
## Y6   -1.7078       -1.71      -1.9843      -1.4313
## Y7    0.6249        0.50       0.4349       0.8148
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                        LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0) 590.9812  7 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                        LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1) 1345.603  7 < 2.22e-16
```

All of the variables ($fage$, *mage*, *meduc*, *weeks*, *visits*, *cignum*, *gained*) are suggested for transformations. The transforms are as follows: raise $fage$ to the -1.00, log transform of *mage*, square *meduc*, raise *weeks* to the 7.50, log transform of *visits*, raise *cignum* to the -1.71, and lastly raise *gained* to the 0.5.

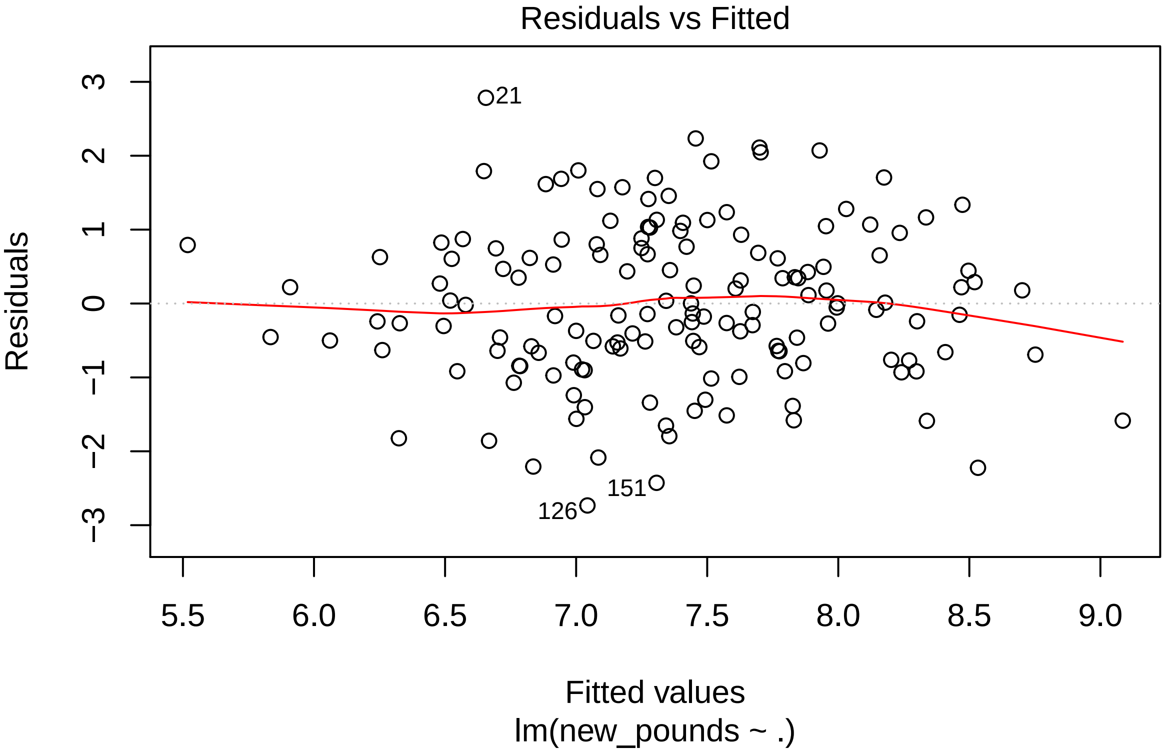
e) In question d, you were told to remove categorical predictors when seeking the box-cox transformation. Why is this?

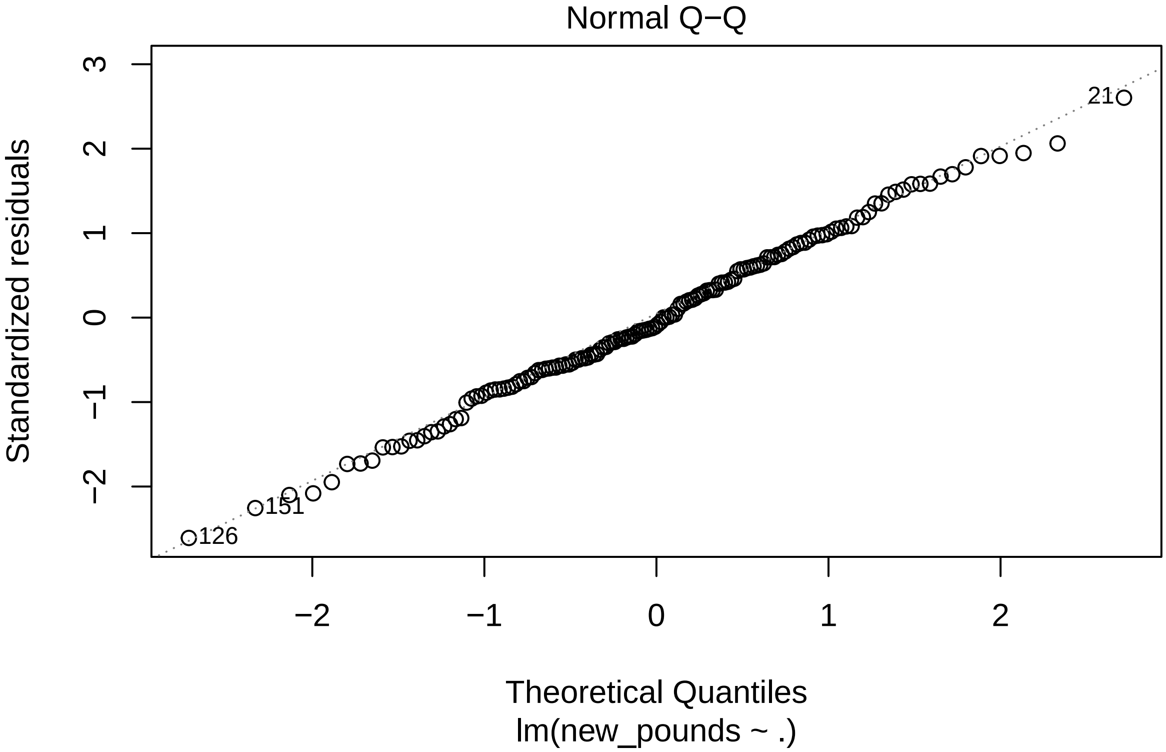
We need to remove categorical predictors because they only consist of values equalling either 0 or 1 so in this case they serve as dummy variables that will never have a transformation that will turn them into normal distributions.

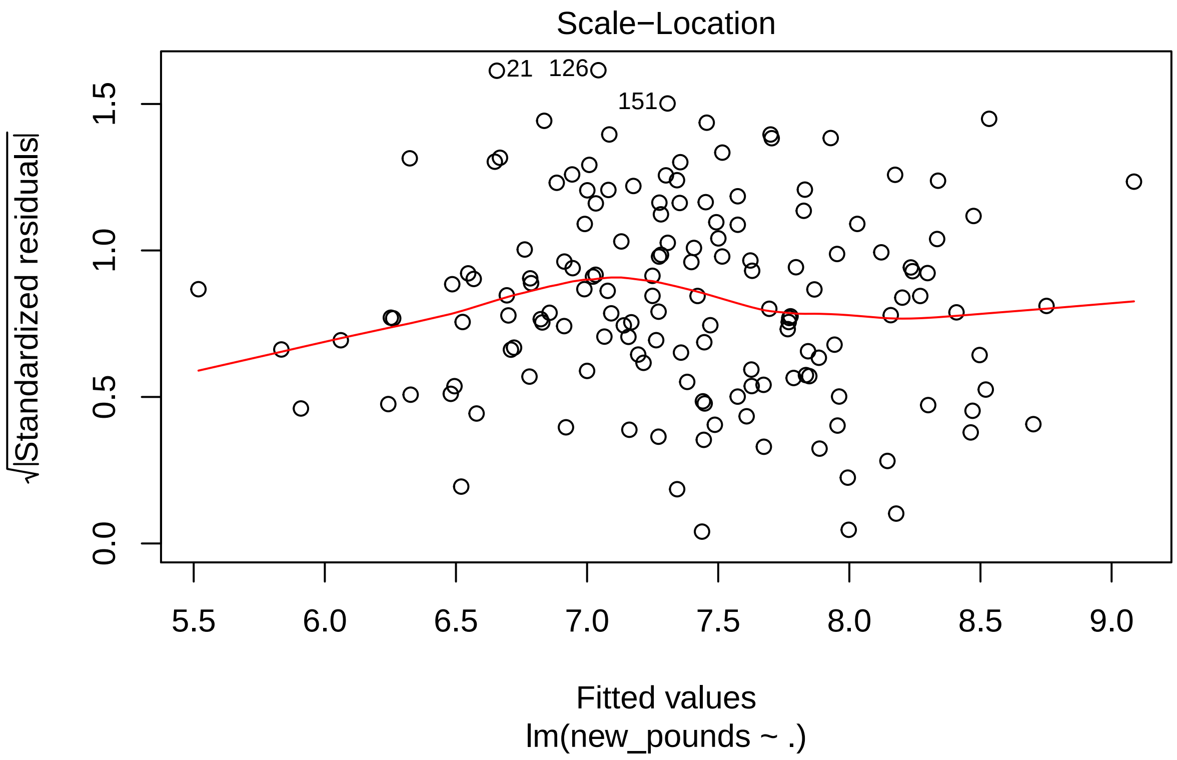
f) Create a new dataframe with the transformed variables as well as sex, smokes, and cignum. Fit a model using all of these predictors. Compare the validity of this model to the original untransformed model. Which do you think is best?

```r
new.births <- dplyr::transmute(births, new_fage = fage^(-1), new_mage = log(mage), new_meduc = meduc^2,

new.model <- lm(new_pounds~., data = new.births)
plot(new.model)
```

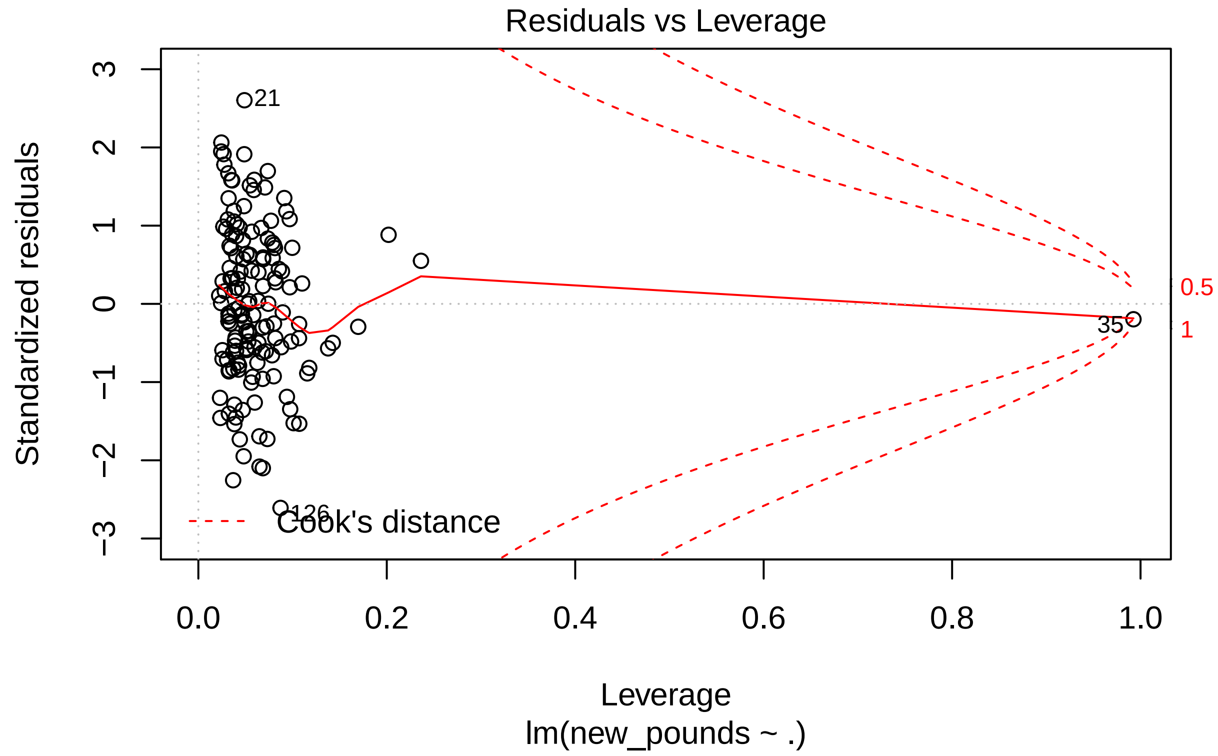

Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
−3
21
151
126
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Fitted values
lm(new_pounds ~ .)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
21
151
126
−2
−1
0
1
2
Theoretical Quantiles
lm(new_pounds ~ .)



```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



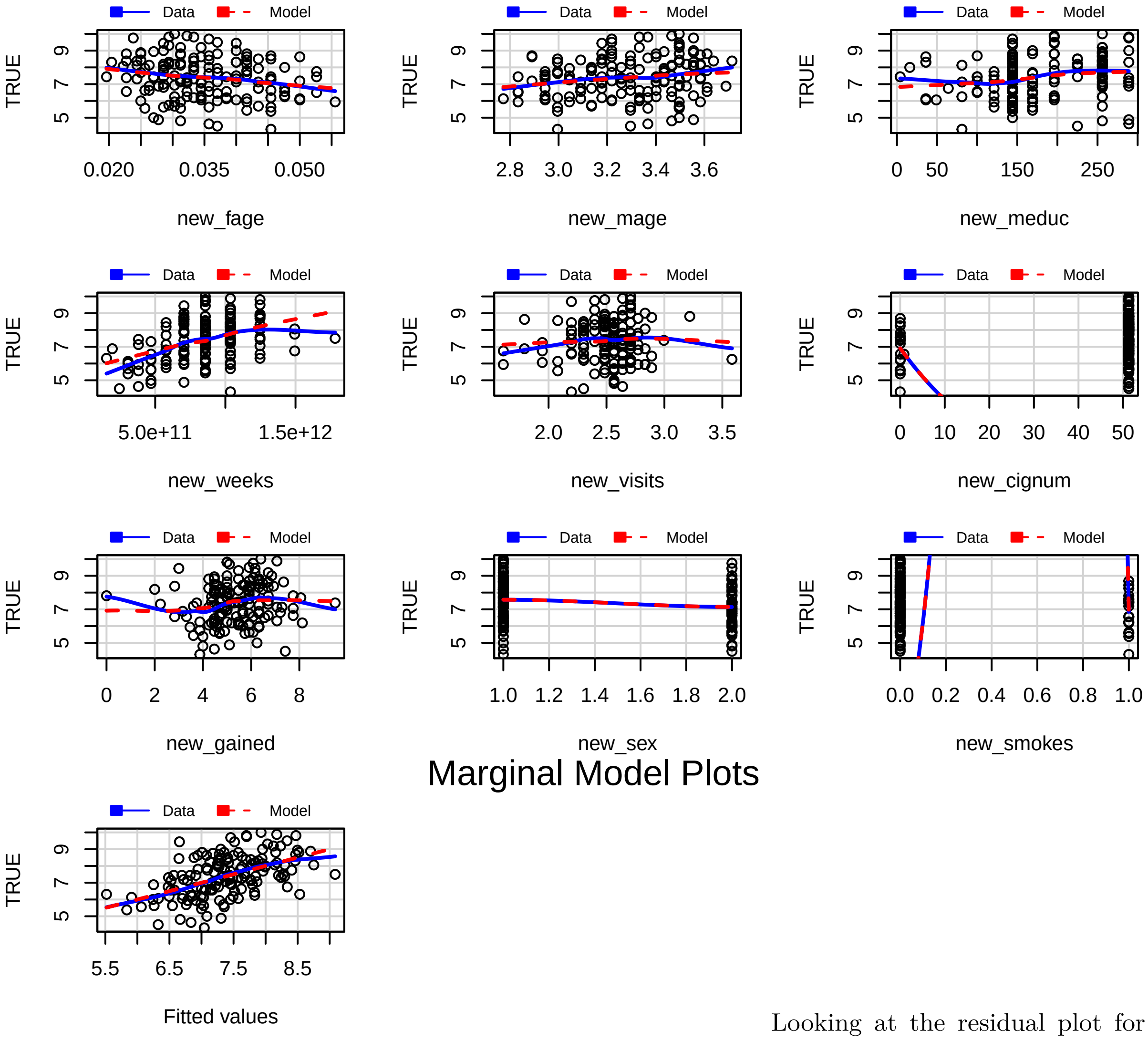
```
summary(new.model)
```

```
##
## Call:
## lm(formula = new_pounds ~ ., data = new.births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73311 -0.66067 -0.06971  0.75508  2.78410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.583e+01  7.130e+01   0.923   0.3575
## new_fage    -1.514e+01  1.914e+01  -0.791   0.4305
## new_mage     2.677e-01  7.141e-01   0.375   0.7083
## new_meduc    1.470e-03  1.687e-03   0.872   0.3849
## new_weeks    1.817e-12  3.232e-13   5.623 9.64e-08 ***
## new_visits   9.610e-02  3.662e-01   0.262   0.7934
## new_cignum  -1.184e+00  1.385e+00  -0.855   0.3938
## new_gained   7.883e-02  7.278e-02   1.083   0.2806
## new_sex     -3.324e-01  1.842e-01  -1.804   0.0733 .
## new_smokes  -6.108e+01  7.092e+01  -0.861   0.3906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 142 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.2668, Adjusted R-squared:  0.2204
## F-statistic: 5.743 on 9 and 142 DF,  p-value: 8.737e-07
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = pounds ~ ., data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51447 -0.73704 -0.00984  0.71179  2.63919
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.855408   1.848560  -2.086   0.0388 *
## sex         -0.312481   0.182321  -1.714   0.0887 .
## fage         0.011539   0.019377   0.595   0.5525
## mage         0.016641   0.023717   0.702   0.4840
## meduc        0.023626   0.038487   0.614   0.5403
## weeks        0.267083   0.044620   5.986 1.67e-08 ***
## visits       0.006242   0.027020   0.231   0.8176
## cignum      -0.024871   0.032537  -0.764   0.4459
## gained       0.007445   0.006701   1.111   0.2684
## smoke       -0.160720   0.506459  -0.317   0.7515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.084 on 142 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.283,  Adjusted R-squared:  0.2375
## F-statistic: 6.227 on 9 and 142 DF,  p-value: 2.184e-07
```

```r
mmps(new.model)
```



Looking at the residual plot for the new model there is no pattern so the linearity condition is satisfied. The normall QQ plot has little to no deviation from the straight line so the errors are normally distributed. There is no increasing or decreasing trend in the scale-location plot, or fanshape in the residual plot thus confirming that the constant variance condition is satisfied. Therefore both models are valid, and looking more closely at their individual summaries both models had almost the same RSS and R-squared values (with adjusted R-squared decreasing in the transformed model). Looking at the marginal model plots, all the variables are fit by the model pretty well with the loess lines being almost the same as the regression lines, except with this model now the *weeks*, *gained*, and *visits* have slightly more deviation. Thus, since there was no significant improvement in the transformed model it is better to keep the simpler, untransformed model. Therefore, in this case, the original model is best.