126 Final Project Step 3

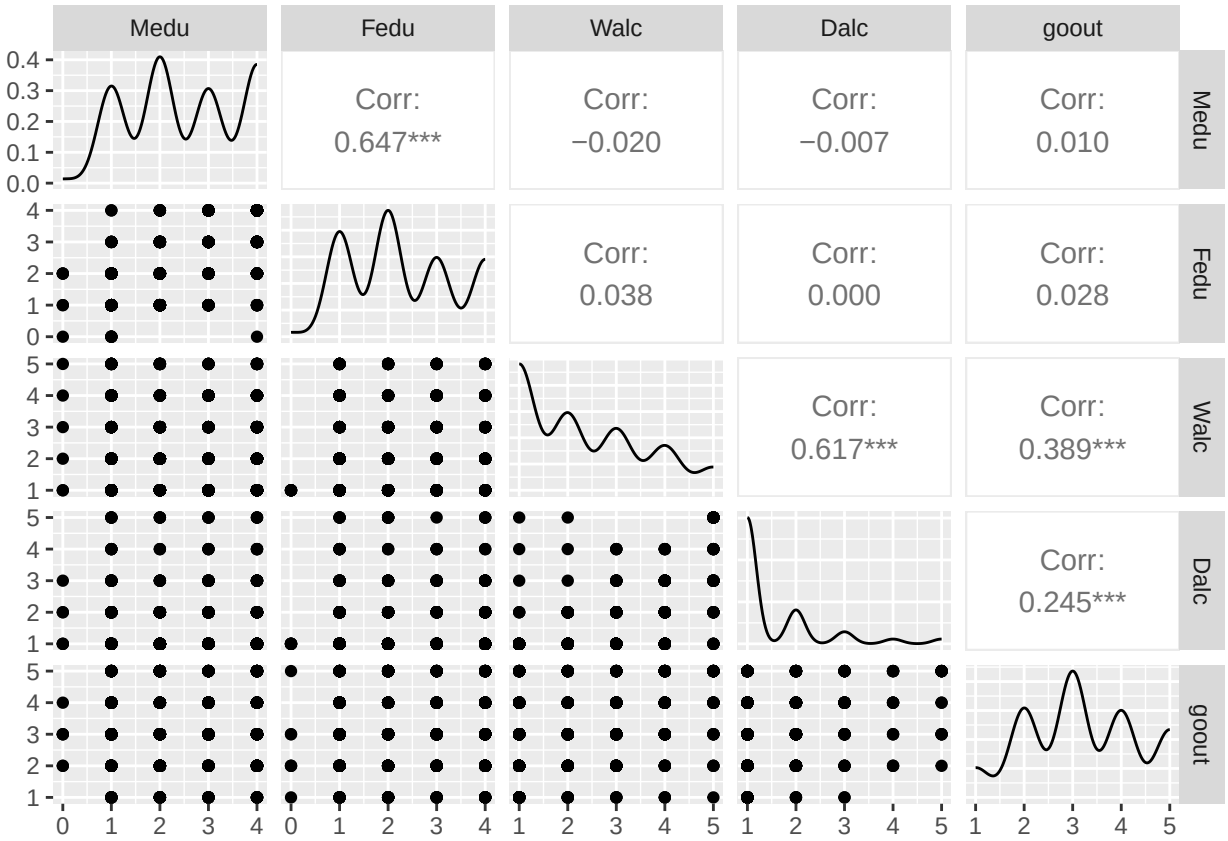
Angel Abdelnour, Andrew Hansen, Sammy Suliman

2023-05-09

Review: The UCI student performance dataset, collected from students at a secondary school in Portugal, has 16 numeric variables. Of those we have the three response variables, G1, G2, and G3, which represent the student's grades in each quarter respectively (for the sake of our project we will be using G3). There are 13 numeric predictor variables, and 17 categorical predictors.

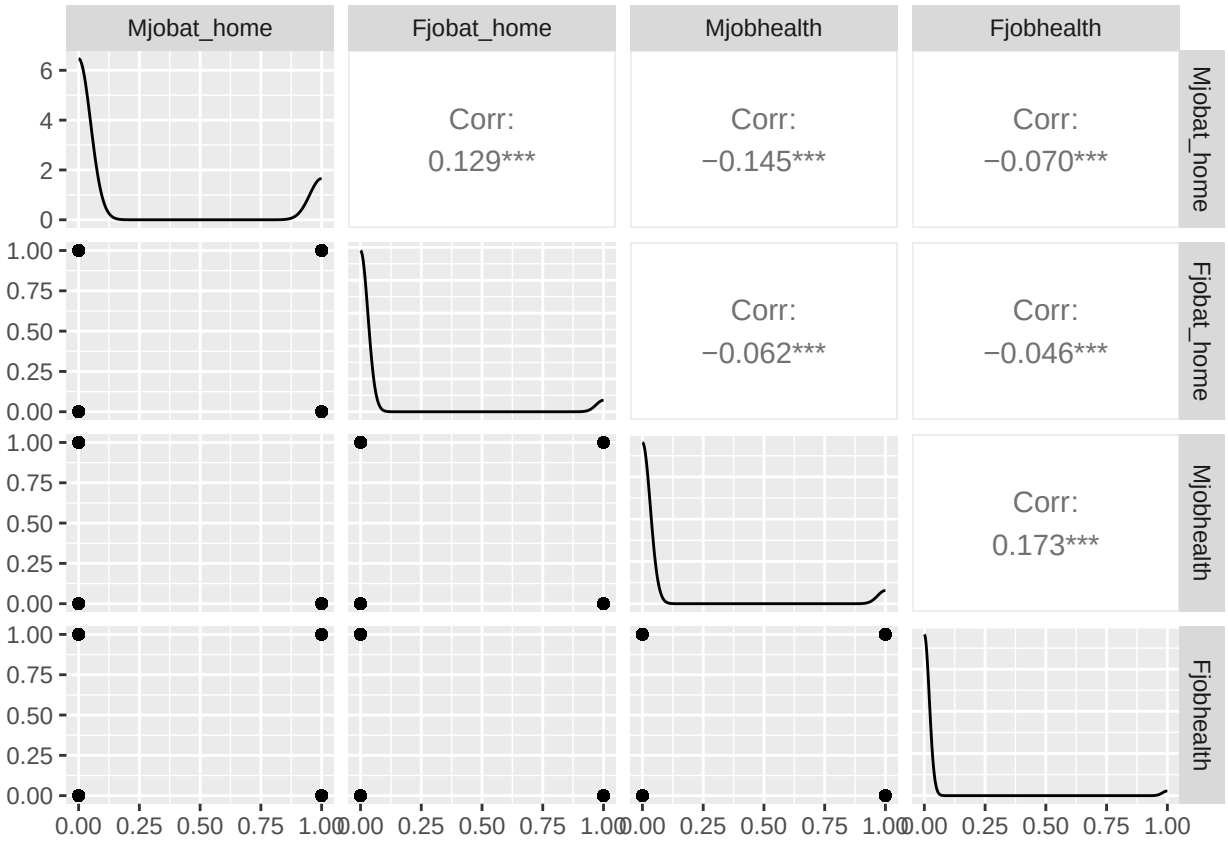
First, we will partition the data into a training and testing set for future validation purposes.

```
#make this example reproducible
set.seed(1)

#create ID column
performance$id <- 1:nrow(performance)

#use 70% of dataset as training set and 30% as test set
train <- performance %>% dplyr::sample_frac(0.70)
test  <- dplyr::anti_join(performance, train, by = 'id')
```



We see from our pairsplot of a limited selection of numeric potential predictors, that the pairs of variables 'Dalc'/'Walc' (weekday / weekend consumption of alcohol) and 'Medu/Fedu' (mothers'/fathers' education level) are understandably highly correlated. So we will only consider 'Fedu' and/or 'Walc' as potential predictors in our model.

We will now perform one-hot encoding on all of the categorical variables as part of feature engineering.



Correlation between Mjob / Fjob factors is not as strong as we expected - we don't have to drop those columns We can now move on to fitting a model with all of our predictor variables.

```
model3 <- lm(G3 ~ ., data=train3)
summary(model3)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = train3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8497  -1.0253  -0.1091   0.9038   7.5648
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.7266747  0.2190994  44.394  < 2e-16 ***
## age               0.1332431  0.0105200  12.666  < 2e-16 ***
## Fedu              0.3123745  0.0114392  27.307  < 2e-16 ***
## traveltime       -0.1755126  0.0162139 -10.825  < 2e-16 ***
## studytime         0.2708783  0.0152343  17.781  < 2e-16 ***
## failures         -0.9888586  0.0246393 -40.133  < 2e-16 ***
## famrel            0.0776069  0.0133287   5.823 5.88e-09 ***
## freetime         -0.0880817  0.0124636  -7.067 1.63e-12 ***
## goout             0.0176277  0.0118384   1.489 0.136496
## Walc             -0.1125421  0.0108305 -10.391  < 2e-16 ***
```

```
## health           -0.0635395  0.0086616  -7.336 2.28e-13 ***
## absences         -0.0235735  0.0027909  -8.446  < 2e-16 ***
## schoolGP          0.4913750  0.0308332  15.937  < 2e-16 ***
## sexM             -0.5636908  0.0266429 -21.157  < 2e-16 ***
## addressR         -0.1342312  0.0293133  -4.579 4.69e-06 ***
## famsizeGT3        0.0885584  0.0282199   3.138 0.001702 **
## PstatusA          0.0722183  0.0403874   1.788 0.073767 .
## Mjobat_home      -0.7022898  0.0521160 -13.476  < 2e-16 ***
## Mjobhealth       -0.1850054  0.0604741  -3.059 0.002222 **
## Mjobother        -0.3828336  0.0470730  -8.133 4.42e-16 ***
## Mjobservices     -0.2971418  0.0495208  -6.000 2.00e-09 ***
## Fjobat_home      -0.8531634  0.0739372 -11.539  < 2e-16 ***
## Fjobhealth       -0.5487602  0.0909007  -6.037 1.60e-09 ***
## Fjobother        -0.9854546  0.0573489 -17.183  < 2e-16 ***
## Fjobservices     -1.1374398  0.0594535 -19.132  < 2e-16 ***
## reasoncourse     -0.1661850  0.0318747  -5.214 1.87e-07 ***
## reasonother      -0.1826162  0.0481316  -3.794 0.000149 ***
## reasonreputation  0.0008495  0.0371475   0.023 0.981755
## paidno            0.1534300  0.0497431   3.084 0.002042 **
## guardianfather    0.1285734  0.0292839   4.391 1.14e-05 ***
## guardianother     0.1908793  0.0553275   3.450 0.000562 ***
## schoolsupno       0.8648579  0.0385401  22.440  < 2e-16 ***
## famsupno          0.0507649  0.0260377   1.950 0.051228 .
## activitiesno     -0.1454399  0.0253802  -5.730 1.01e-08 ***
## nurseryno         0.0594305  0.0301636   1.970 0.048819 *
## higherno         -1.3717566  0.0458450 -29.922  < 2e-16 ***
## internetno       -0.3439550  0.0308328 -11.155  < 2e-16 ***
## romanticno        0.2137143  0.0261173   8.183 2.92e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.771 on 21938 degrees of freedom
## Multiple R-squared:  0.3268, Adjusted R-squared:  0.3256
## F-statistic: 287.8 on 37 and 21938 DF,  p-value: < 2.2e-16
```

Variable Selection

Now that we have fitted the full model with all of our predictors, we want to try backwards selection to reduce the model so it only has statistically significant predictors. We will use AIC (Akaike information criterion) as our metric to evaluate this to maximize predictive power.

```
backward <- stepAIC(model3, direction='backward', trace=0)
summary(backward)
```

```
##
## Call:
## lm(formula = G3 ~ age + Fedu + traveltime + studytime + failures +
##     famrel + freetime + goout + Walc + health + absences + schoolGP +
##     sexM + addressR + famsizeGT3 + PstatusA + Mjobat_home + Mjobhealth +
##     Mjobother + Mjobservices + Fjobat_home + Fjobhealth + Fjobother +
##     Fjobservices + reasoncourse + reasonother + paidno + guardianfather +
##     guardianother + schoolsupno + famsupno + activitiesno + nurseryno +
##     higherno + internetno + romanticno, data = train3)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8493  -1.0251  -0.1089   0.9039   7.5648
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.726989   0.218664  44.484  < 2e-16 ***
## age             0.133243   0.010520  12.666  < 2e-16 ***
## Fedu            0.312373   0.011439  27.308  < 2e-16 ***
## traveltime     -0.175511   0.016213 -10.825  < 2e-16 ***
## studytime       0.270877   0.015234  17.781  < 2e-16 ***
## failures       -0.988855   0.024638 -40.135  < 2e-16 ***
## famrel          0.077605   0.013328   5.823 5.87e-09 ***
## freetime       -0.088081   0.012463  -7.067 1.63e-12 ***
## goout           0.017630   0.011838   1.489 0.136436
## Walc           -0.112542   0.010830 -10.391  < 2e-16 ***
## health         -0.063538   0.008661  -7.336 2.28e-13 ***
## absences       -0.023573   0.002791  -8.447  < 2e-16 ***
## schoolGP        0.491405   0.030805  15.952  < 2e-16 ***
## sexM           -0.563732   0.026580 -21.209  < 2e-16 ***
## addressR       -0.134113   0.028852  -4.648 3.37e-06 ***
## famsizeGT3      0.088534   0.028200   3.140 0.001694 **
## PstatusA        0.072162   0.040312   1.790 0.073455 .
## Mjobat_home    -0.702228   0.052045 -13.493  < 2e-16 ***
## Mjobhealth     -0.184887   0.060250  -3.069 0.002153 **
## Mjobother      -0.382828   0.047071  -8.133 4.41e-16 ***
## Mjobservices   -0.297069   0.049416  -6.012 1.87e-09 ***
## Fjobat_home    -0.853181   0.073932 -11.540  < 2e-16 ***
## Fjobhealth     -0.548739   0.090894  -6.037 1.59e-09 ***
## Fjobother      -0.985493   0.057323 -17.192  < 2e-16 ***
## Fjobservices   -1.137473   0.059435 -19.138  < 2e-16 ***
## reasoncourse   -0.166584   0.026680  -6.244 4.35e-10 ***
## reasonother    -0.183029   0.044616  -4.102 4.11e-05 ***
## paidno          0.153493   0.049666   3.090 0.002001 **
## guardianfather  0.128628   0.029185   4.407 1.05e-05 ***
## guardianother   0.190903   0.055317   3.451 0.000559 ***
## schoolsupno     0.864874   0.038533  22.445  < 2e-16 ***
## famsupno        0.050801   0.025990   1.955 0.050642 .
## activitiesno   -0.145520   0.025136  -5.789 7.16e-09 ***
## nurseryno       0.059416   0.030157   1.970 0.048821 *
## higherno       -1.371781   0.045831 -29.931  < 2e-16 ***
## internetno     -0.343990   0.030794 -11.171  < 2e-16 ***
## romanticno      0.213743   0.026087   8.194 2.67e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.771 on 21939 degrees of freedom
## Multiple R-squared:  0.3268, Adjusted R-squared:  0.3256
## F-statistic: 295.8 on 36 and 21939 DF,  p-value: < 2.2e-16
```

The only variable that was dropped by the backward selection algorithm was reasonreputation. It's large p-value in the full model indicated that it was not a statistically significant predictor.

Interaction effects

We have decided not to use interaction terms for the following reasons: Firstly, it is not clear that the numeric variables will have a different association with the response at different levels of the categorical factors, for example, why the same amount of studytime would have a different association with the response based on whether or not the subject is in a romantic relationship, etc. Additionally, we already have a large amount of terms in our model, some of them with relatively high p-values (for example, goout). So it is not clear that interaction terms involving these predictors will be statistically significant.

Computation Model

```
new_model <- lm(G3 ~ studytime + age + romanticno, data=train3)
summary(new_model)
```

```
##
## Call:
## lm(formula = G3 ~ studytime + age + romanticno, data = train3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9317  -1.3411  -0.3411   1.1938   7.1938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79772    0.20022  58.923  < 2e-16 ***
## studytime    0.46505    0.01734  26.822  < 2e-16 ***
## age         -0.06278    0.01167  -5.380 7.54e-08 ***
## romanticno   0.20838    0.02975   7.004 2.56e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 21972 degrees of freedom
## Multiple R-squared:  0.0358, Adjusted R-squared:  0.03567
## F-statistic:   272 on 3 and 21972 DF,  p-value: < 2.2e-16
```

```
new_model2 <- lm(G3 ~ studytime + age + romanticno + age:romanticno, data=train3)
summary(new_model2)
```

```
##
## Call:
## lm(formula = G3 ~ studytime + age + romanticno + age:romanticno,
##     data = train3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9074  -1.3928  -0.3928   1.1418   7.1418
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.86837    0.32185  39.983  < 2e-16 ***
## studytime       0.46534    0.01733  26.849  < 2e-16 ***
```

```
## age            -0.12696    0.01909  -6.651 2.98e-11 ***
## romanticno     -1.49815    0.40285  -3.719 0.000201 ***
## age:romanticno  0.10237    0.02410   4.248 2.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.117 on 21971 degrees of freedom
## Multiple R-squared:  0.03659,    Adjusted R-squared:  0.03642
## F-statistic: 208.6 on 4 and 21971 DF,  p-value: < 2.2e-16
```

The 2 models we will be using are a multiple linear regression model, 'new_model' containing 2 numeric predictors, age and studytime, as well as 1 categorical predictor 'romanticno', and a second model 'new_model2', containing the same predictors as new_model, but with an additional interaction term between romanticno and age. We chose these 2 new models to test whether our assertion that the interaction terms were not needed was justified.

```
predictions <- predict(new_model, test3)

data.frame(
  R2 = R2(predictions, test3$G3),
  RMSE = rmse(predictions, test3$G3),
  MAE = mae(predictions, test3$G3)
)
```

```
##           R2     RMSE      MAE
## 1 0.06620296 2.141269 1.654857
```

```
predictions2 <- predict(new_model2, test3)

data.frame(
  R2 = R2(predictions2, test3$G3),
  RMSE = rmse(predictions2, test3$G3),
  MAE = mae(predictions2, test3$G3)
)
```

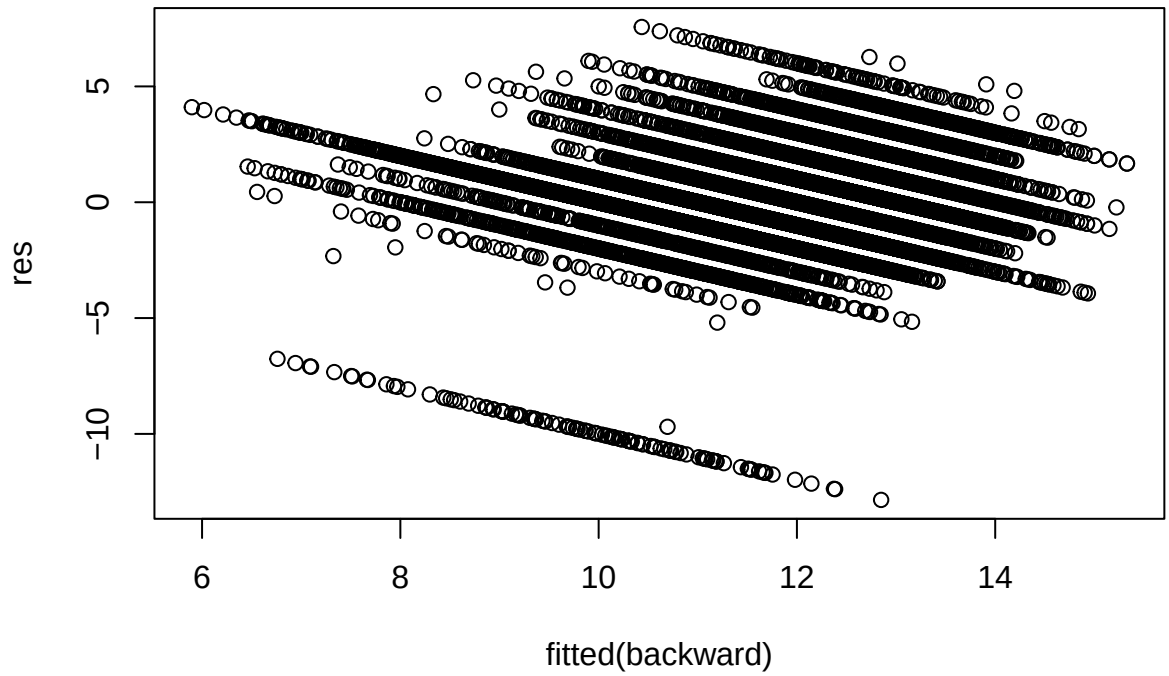
```
##           R2     RMSE      MAE
## 1 0.06447631 2.142224 1.655882
```

To evaluate the performance of our 2 models, we chose the metrics R^2, RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error). R^2 is a measure of goodness of fit, whereas RMSE and MAE are measures of total error. In all 3 metrics, the extended model including the interaction term performed nearly identically to the model without the interaction term. This indicates that including an interaction term between the numeric and categorical variables in our dataset may not have a large effect on model fit in some cases.

Statistical model

We already computed the model 'backward', through the backward selection algorithm, starting with the full model involving all predictors in the training data, and testing p-values to drop the least significant predictors (in this case, only 1). To justify our model, we will plot a graph of residuals against fit to check whether the core assumptions of the linear model (linearity and constant variance) are satisfied.

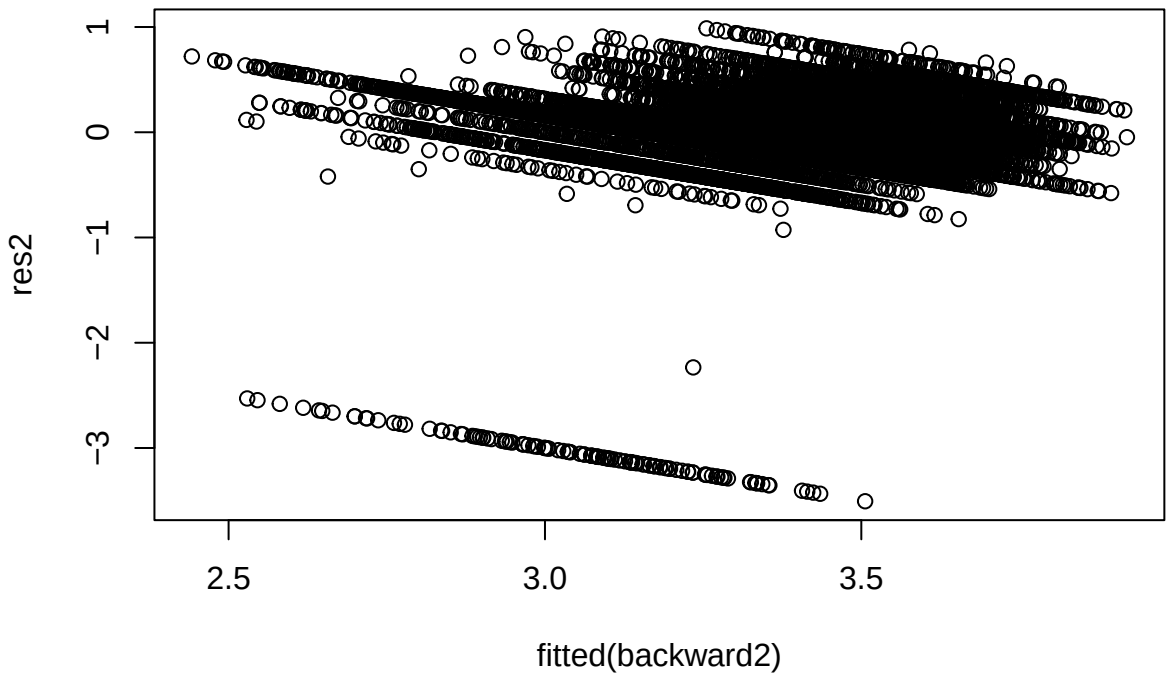
```
res <- resid(backward)
plot(fitted(backward), res)
```



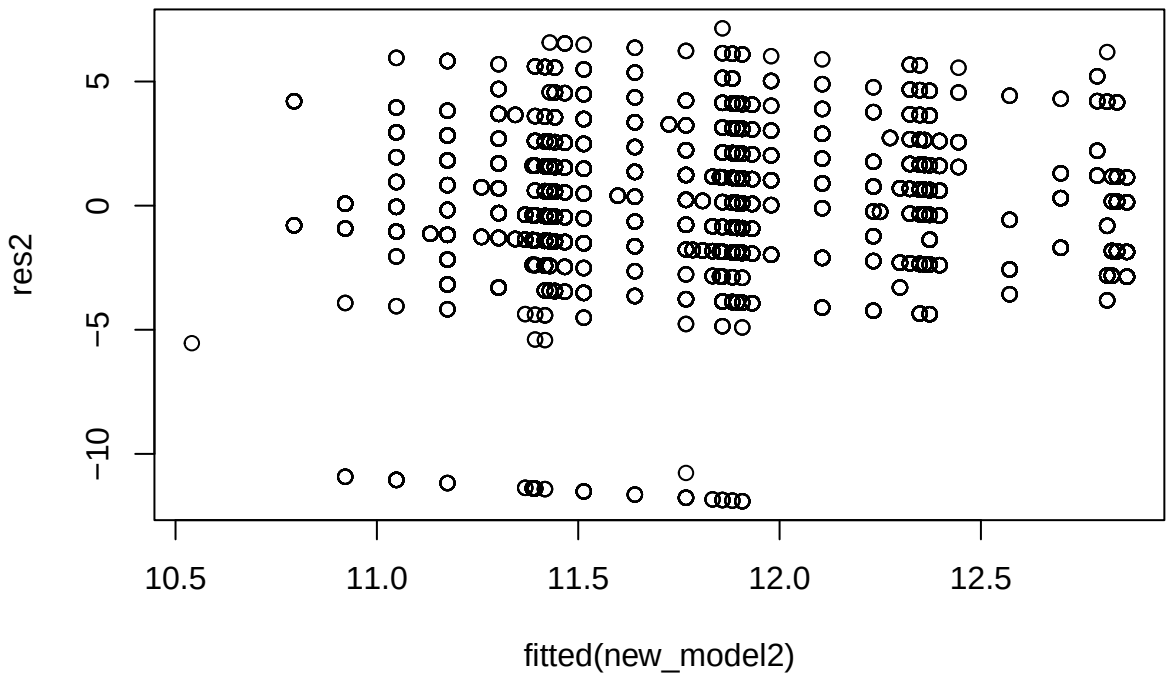
In the visual above we can see that variance is not constant so we try this procedure again, this time we transform the response variable by taking the square root of G3.

```
model4 <- lm(sqrt(G3) ~ ., data=(train3))
summary(model4)
backward2 <- stepAIC(model4, direction='backward', trace=0)
summary(backward2)
```

This time more predictors are dropped by the backward selection algorithm, goout, famsizeGT3, PstatusA, in addition to reasonreputation.



However, the residuals vs fit plot still suggests the linearity and variance assumptions are violated. For curiosity, lets plot the residuals vs fit graph for our previously fitted smaller-sized model.



With the exception of some outlying points which we will examine later, we see the assumptions of the linear model appear to be satisfied. So it seems that perhaps the issue with the 'backward' model is too many predictors. Our smaller model only contains the most significant predictors of our model (as judged by the p-values of these variables in our 'backward' summary), plus an interaction term that intuitively feels reasonable (being in a romantic relationship may affect performance even after controlling for age).

Since the residuals vs fit plot appears to satisfy the most important assumptions of the linear model, we will adopt this model going forward.

```
overall_p <- function(model) {
    f <- summary(model)$fstatistic
    p <- pf(f[1],f[2],f[3],lower.tail=F)
    attributes(p) <- NULL
    return(p)
}

print(paste("P-value:", overall_p(new_model2)))
```

```
## [1] "P-value: 5.54057405924183e-176"
```

The very low p-value confirms the significance of regression for our model.

Significance of Coefficients

```
## [1] "Coefficient (Intercept) is significant."
## [1] "Coefficient studytime is significant."
```

```
## [1] "Coefficient age is significant."
## [1] "Coefficient romanticno is significant."
## [1] "Coefficient age:romanticno is significant."
```

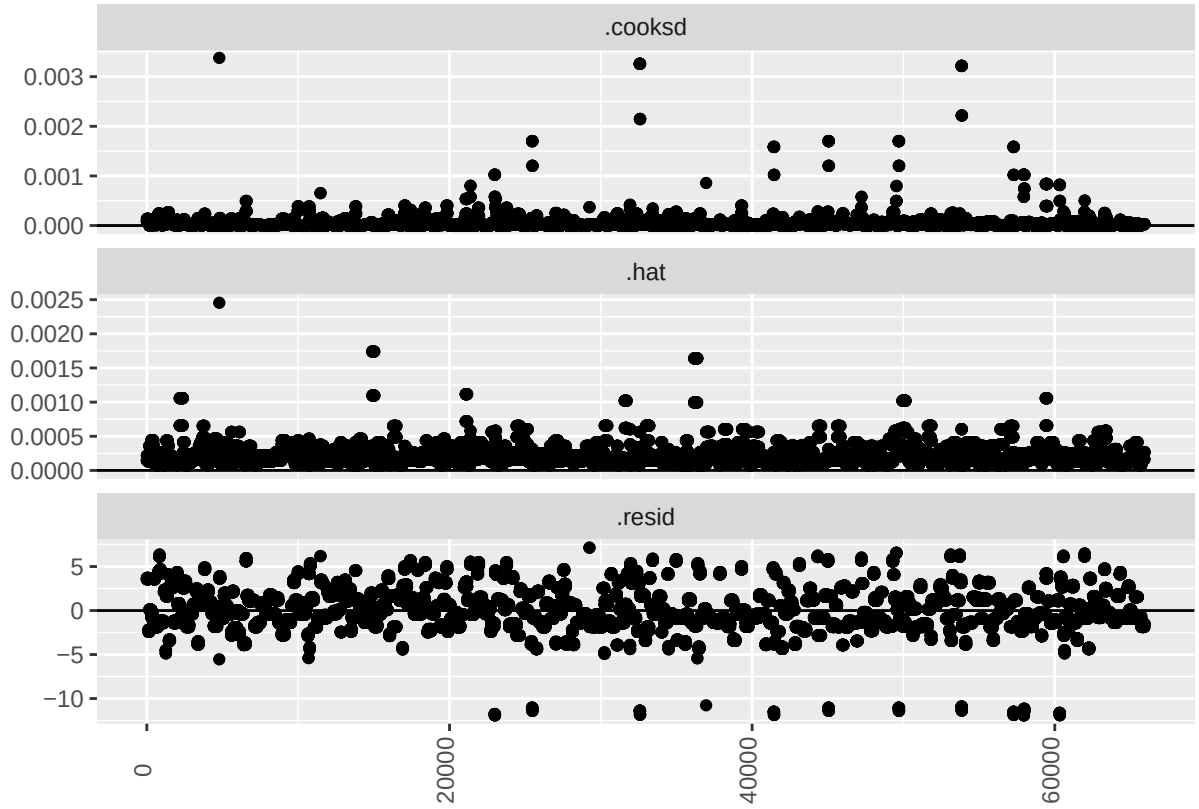
Even after applying the Bonferroni correction to control for the family-wise error rate, we see that all the coefficents on our model are statistically significant.

Goodness of Fit

```
##           R2     Adj_R2
## 1 0.06447631 0.06599044
```

Both the unadjusted and adjusted R^2 values are rather low. This means that relatively little of the variance of the data can be explained by our model. However, R^2 is not necessarily the best metric for model fit as it increases with number of predictors, regardless of whether they are significant.

Influence



Let's identify the biggest outliers, leverage points, and points of greatest influence.

```
# most influential point
high_influence <- augment(new_model2, train3) %>%
  slice_max(order_by = .cooksd, n = 5) %>%
  mutate(row_index = row_number())
high_influence
```

```
## # A tibble: 8 x 45
##     age  Fedu traveltime studytime failures famrel freetime goout  Walc health
##   <int> <int>      <int>     <int>    <int>  <int>    <int> <int> <int>  <int>
## 1    22     1          1         1        3      5        4     5     5      1
## 2    19     1          2         2        3      3        5     4     4      1
## 3    19     1          2         2        3      3        5     4     4      1
## 4    19     1          2         2        3      3        5     4     4      1
## 5    19     1          2         2        3      3        5     4     4      1
## 6    19     1          2         2        3      3        5     4     4      1
## 7    19     1          2         2        3      3        5     4     4      1
## 8    19     1          2         2        3      3        5     4     4      1
## # i 35 more variables: absences <int>, G3 <int>, schoolGP <dbl>, sexM <dbl>,
## #   addressR <dbl>, famsizeGT3 <dbl>, PstatusA <dbl>, Mjobat_home <dbl>,
## #   Mjobhealth <dbl>, Mjobother <dbl>, Mjobservices <dbl>, Fjobat_home <dbl>,
## #   Fjobhealth <dbl>, Fjobother <dbl>, Fjobservices <dbl>, reasoncourse <dbl>,
## #   reasonother <dbl>, reasonreputation <dbl>, paidno <dbl>,
## #   guardianfather <dbl>, guardianother <dbl>, schoolsupno <dbl>,
## #   famsupno <dbl>, activitiesno <dbl>, nurseryno <dbl>, higherno <dbl>, ...
```

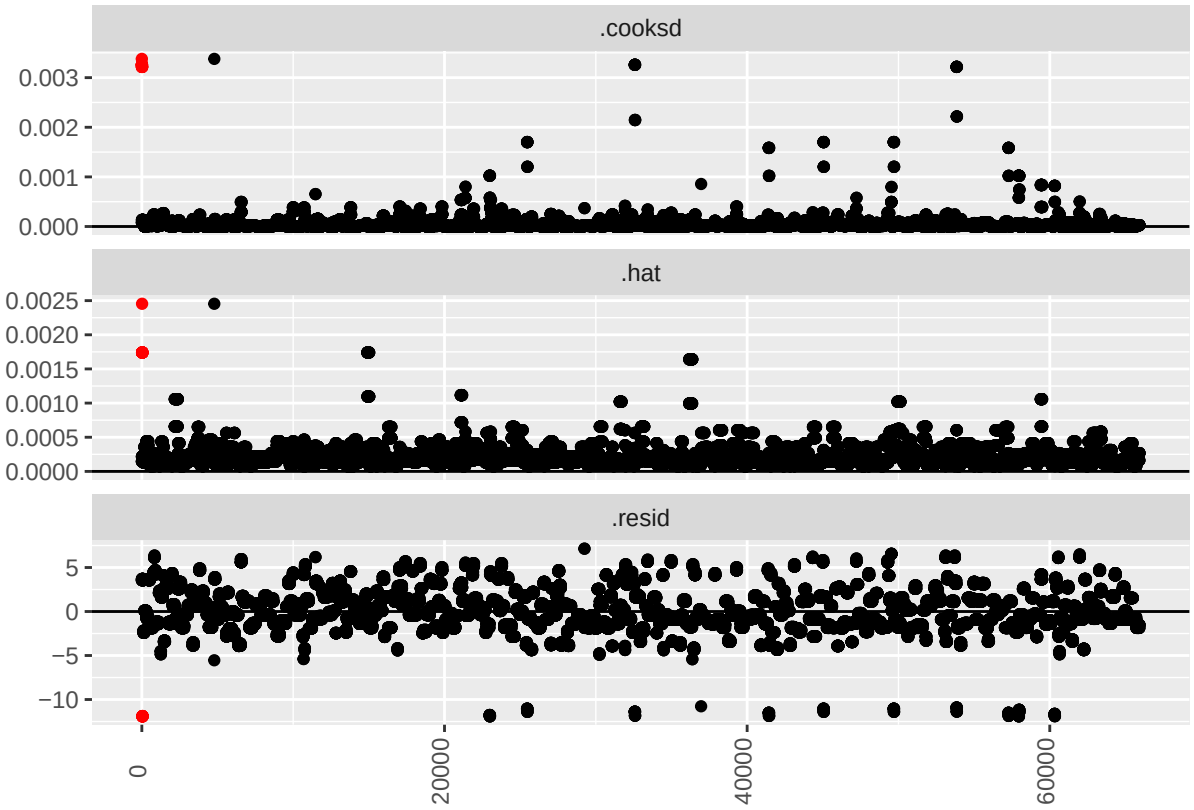
```
# most outlying points
high_residuals <- augment(new_model2, train3) %>%
  slice_max(order_by = .resid, n = 5) %>%
  mutate(row_index = row_number())
high_residuals
```

```
## # A tibble: 12 x 45
##      age  Fedu traveltime studytime failures famrel freetime goout  Walc health
##    <int> <int>      <int>     <int>    <int>  <int>    <int> <int> <int>  <int>
##  1    18     4          1         2        0      3        2     4     4      2
##  2    18     4          1         2        0      3        2     4     4      2
##  3    15     2          1         1        0      3        5     2     1      3
##  4    15     2          1         1        0      3        5     2     1      3
##  5    15     2          1         1        0      3        5     2     1      3
##  6    15     2          1         1        0      3        5     2     1      3
##  7    15     2          1         1        0      3        5     2     1      3
##  8    15     2          1         1        0      3        5     2     1      3
##  9    15     2          1         1        0      3        5     2     1      3
## 10    15     2          1         1        0      3        5     2     1      3
## 11    15     2          1         1        0      3        5     2     1      3
## 12    15     2          1         1        0      3        5     2     1      3
## # i 35 more variables: absences <int>, G3 <int>, schoolGP <dbl>, sexM <dbl>,
## #   addressR <dbl>, famsizeGT3 <dbl>, PstatusA <dbl>, Mjobat_home <dbl>,
## #   Mjobhealth <dbl>, Mjobother <dbl>, Mjobservices <dbl>, Fjobat_home <dbl>,
## #   Fjobhealth <dbl>, Fjobother <dbl>, Fjobservices <dbl>, reasoncourse <dbl>,
## #   reasonother <dbl>, reasonreputation <dbl>, paidno <dbl>,
## #   guardianfather <dbl>, guardianother <dbl>, schoolsupno <dbl>,
## #   famsupno <dbl>, activitiesno <dbl>, nurseryno <dbl>, higherno <dbl>, ...
```

```
# highest leverage point
high_leverage <- augment(new_model2, train3) %>%
  slice_max(order_by = .hat, n = 5) %>%
  mutate(row_index = row_number())
high_leverage
```

```
## # A tibble: 20 x 45
##      age  Fedu traveltime studytime failures famrel freetime goout  Walc health
##    <int> <int>      <int>     <int>    <int>  <int>    <int> <int> <int>  <int>
##  1    22     1          1         1        3      5        4     5     5      1
##  2    21     4          1         3        2      3        3     2     1      5
##  3    21     4          1         3        2      3        3     2     1      5
##  4    21     4          1         3        2      3        3     2     1      5
##  5    21     4          1         3        2      3        3     2     1      5
##  6    21     4          1         3        2      3        3     2     1      5
##  7    21     4          1         3        2      3        3     2     1      5
##  8    21     4          1         3        2      3        3     2     1      5
##  9    21     4          1         3        2      3        3     2     1      5
## 10    21     4          1         3        2      3        3     2     1      5
## 11    21     4          1         3        2      3        3     2     1      5
## 12    21     4          1         3        2      3        3     2     1      5
## 13    21     4          1         3        2      3        3     2     1      5
## 14    21     4          1         3        2      3        3     2     1      5
## 15    21     4          1         3        2      3        3     2     1      5
## 16    21     4          1         3        2      3        3     2     1      5
## 17    21     4          1         3        2      3        3     2     1      5
## 18    21     4          1         3        2      3        3     2     1      5
## 19    21     4          1         3        2      3        3     2     1      5
## 20    21     4          1         3        2      3        3     2     1      5
## # i 35 more variables: absences <int>, G3 <int>, schoolGP <dbl>, sexM <dbl>,
## #   addressR <dbl>, famsizeGT3 <dbl>, PstatusA <dbl>, Mjobat_home <dbl>,
## #   Mjobhealth <dbl>, Mjobother <dbl>, Mjobservices <dbl>, Fjobat_home <dbl>,
## #   Fjobhealth <dbl>, Fjobother <dbl>, Fjobservices <dbl>, reasoncourse <dbl>,
## #   reasonother <dbl>, reasonreputation <dbl>, paidno <dbl>,
## #   guardianfather <dbl>, guardianother <dbl>, schoolsupno <dbl>,
## #   famsupno <dbl>, activitiesno <dbl>, nurseryno <dbl>, higherno <dbl>, ...
```

```
b <- seq(1, 47)
unusual_obs <- broom::augment(new_model2, train3) %>%
  pivot_longer(cols = c(.resid, .hat, .cooksd)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n = 10) %>%
  ungroup()
p_caseinf + geom_point(data = unusual_obs, color = 'red', aes(x=b, y=value))
```



Now lets try to exclude observations with high leverage and high residuals from the training set and refit our model.

```
high_leverage['.resid']
```

```
## # A tibble: 20 x 1
##    .resid
##     <dbl>
##  1 -5.54
##  2  0.402
##  3  0.402
##  4  0.402
##  5  0.402
##  6  0.402
##  7  0.402
##  8  0.402
##  9  0.402
## 10  0.402
## 11  0.402
## 12  0.402
## 13  0.402
## 14  0.402
## 15  0.402
## 16  0.402
## 17  0.402
## 18  0.402
```

```
## 19  0.402
## 20  0.402
```

```
(high_residuals['.hat'])
```

```
## # A tibble: 12 x 1
##        .hat
##       <dbl>
##  1 0.000163
##  2 0.000163
##  3 0.000413
##  4 0.000413
##  5 0.000259
##  6 0.000259
##  7 0.000259
##  8 0.000259
##  9 0.000259
## 10 0.000259
## 11 0.000259
## 12 0.000259
```

We notice there is a single 'high leverage' observation with higher than average residuals. We will test the effect of retraining our model without this observation.

```
high_resid_high_leverage <- high_leverage[1,]
print(high_resid_high_leverage$.cooksd)
```

```
## [1] 0.003378172
```

This observation has a relatively large Cook's distance, which indicates it is relatively influential to the model fit.

```
# we see absences = 12, age = 22 ... use this to find index of observation in original training set
train3[train3$absences == 12 & train3$age == 22,]
```

```
##      age Fedu traveltime studytime failures famrel freetime goout Walc health
## 1596  22    1          1         1        3      5        4     5    5      1
##      absences G3 schoolGP sexM addressR famsizeGT3 PstatusA Mjobat_home
## 1596       12  5        1    1        0          1        0           0
##      Mjobhealth Mjobother Mjobservices Fjobat_home Fjobhealth Fjobother
## 1596          0         0            1           0          0         0
##      Fjobservices reasoncourse reasonother reasonreputation paidno
## 1596            1            0           1                0      1
##      guardianfather guardianother schoolsupno famsupno activitiesno nurseryno
## 1596              0             0           1        1            1         1
##      higherno internetno romanticno
## 1596        1          0          0
```

```
# row index = 1596
train4 <- train3[-c(1596),]
```

```
new_model3 <- lm(G3 ~ studytime + age + romanticno + age:romanticno, data=train4)
summary(new_model3)
```

```
##
## Call:
## lm(formula = G3 ~ studytime + age + romanticno + age:romanticno,
##     data = train4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9073  -1.3931  -0.3931   1.1419   7.1419
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.82969    0.32214  39.826  < 2e-16 ***
## studytime       0.46507    0.01733  26.836  < 2e-16 ***
## age            -0.12457    0.01911  -6.519 7.21e-11 ***
## romanticno     -1.45882    0.40308  -3.619 0.000296 ***
## age:romanticno  0.09997    0.02411   4.146 3.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.117 on 21970 degrees of freedom
## Multiple R-squared:  0.03646,    Adjusted R-squared:  0.03629
## F-statistic: 207.9 on 4 and 21970 DF,  p-value: < 2.2e-16
```

```
## [1] "Coefficient (Intercept) is significant."
## [1] "Coefficient studytime is significant."
## [1] "Coefficient age is significant."
## [1] "Coefficient romanticno is significant."
## [1] "Coefficient age:romanticno is significant."
```

Interpretation

It appears that dropping this single observation was enough to change the values of the coefficients on the fitted model, but not enough to shift whether they were statistically significant.

Confidence / Prediction Intervals for $\hat{Y}$

Next, we will examine a confidence interval for the mean of the response (G3) at the mean value of the numeric predictor variables and at the most common value of the categorical predictor values:

```
freq_value <- function(x) {
as.numeric(tail(names(sort(table(x))), 1))
}
```

```
x_bar <- c(freq_value(test3$studytime), mean(test3$age), freq_value(test3$romanticno))
new_data <- data.frame(studytime = x_bar[1], age = x_bar[2], romanticno = x_bar[3])
```

```r
predict(new_model2, newdata = new_data, interval = 'confidence', level = 0.95)
```

```
##        fit      lwr     upr
## 1 11.88888 11.85346 11.9243
```

With 95% confidence, the final grade for a student with age equal to the mean and with the most common relationship status (single), and in the most common studytime bracket (2-5 hours) is between 11.85346 and 11.9243.

```r
new_data2 <- data.frame(age = 22, studytime = 3, romanticno = 0)
predict(new_model2, newdata = new_data2, interval = 'prediction', level = 0.95)
```

```
##        fit      lwr     upr
## 1 11.47116 7.316517 15.6258
```

With 95% confidence, the final grade for a student aged 22, in a relationship, and who studies between 5-10 hours a week is between 7.316517 and 15.6258.

Conclusion