final_data_analysis

- install and load necessary packages

```
#install.packages("jsonlite")
library(jsonlite)
#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("purrr")
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:jsonlite':
##
##     flatten
```

```
#install.packages("stringr")
library(stringr)
#install.packages("readr")
library(readr)
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
#install.packages("ggplot2")
library(ggplot2)
#install.packages("MASS")
library("MASS")
```

```
##
## Attaching package: 'MASS'

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

```
#install.packages("car")
library("car")
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:purrr':
##
##     some

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

```
#install.packages("broom")
library(broom)
```

- This is where I'm loading the dataset in
- converting file to csv
- creating a us subset for only us races
- also creating a csv file for just all races

```
utmb_data <- fromJSON("utmb-race-data-raw.json", flatten=TRUE)
```

```
str(utmb_data[[1]])
```

```
## List of 11
##  $ Age           :List of 11
##   ..$ 20-34: int 89
##   ..$ 35-39: int 48
##   ..$ 40-44: int 49
##   ..$ 45-49: int 32
##   ..$ 50-54: int 18
##   ..$ 55-59: int 12
##   ..$ 60-64: int 6
##   ..$ 65-69: int 1
##   ..$ 70-74: int 1
##   ..$ U18  : int 30
##   ..$ U20  : int 6
##  $ City / Country: chr "Spain"
##  $ Country       :List of 1
```

```
##   ..$  Spain: int 292
##  $ Date          : chr "5th November 2017"
##  $ Distance      : chr "12 KM"
##  $ Elevation Gain: chr "400 M+"
##  $ N Results     : int 292
##  $ Race Category : chr "-"
##  $ Race Title    : chr "Caravaca Trail Experience 2017 - Promo"
##  $ Results       : num [1:292] 0.905 0.95 0.963 0.981 1.004 ...
##  $ Sex           :List of 2
##   ..$ Men  : int 150
##   ..$ Women: int 142
```

US races creating a json variable of us races

```
us_races <- utmb_data[
  sapply(utmb_data, function(x) {
    grepl("United States|USA|US$", x$`City / Country`, ignore.case = TRUE)
  })
]
```

turning us races ito a data frame for analysis also changing age categories to have larger bins: under 34, 35-59, 60+ as there isnt that much variation with only a few years

```
us_df <- imap_dfr(us_races, ~{
  results <- .x$Results
  total_age <- sum(unlist(.x$Age))
  tibble(
    race_id        = .y,
    city_country   = .x$`City / Country`,
    date           = .x$Date,
    distance       = .x$Distance,
    elevation_gain = .x$`Elevation Gain`,
    n_participants = length(results),
    mean_time_hrs      = mean(results, na.rm = TRUE),
    median_time_hrs    = median(results, na.rm = TRUE),
    sd_time            = sd(results, na.rm = TRUE),
    min_time_hrs       = min(results, na.rm = TRUE),
    max_time_hrs       = max(results, na.rm = TRUE),
    pct_women      = .x$Sex$Women / sum(unlist(.x$Sex)),
    pct_age_u35     = sum(unlist(.x$Age[c("U18", "U20", "20-34")]), na.rm = TRUE) / total_age,
    pct_age_35_59     = sum(unlist(.x$Age[c("35-39", "40-44", "45-49", "50-54", "55-59")]), na.rm = TRU
    pct_age_60_plus = sum(unlist(.x$Age[c("60-64", "65-69", "70-74", "75-79", "80+")]), na.rm = TRUE) /
  )
})
```

find a cutoff for race size (number of participants)

```
# min number of participants
min(us_df$n_participants, na.rm = TRUE)
```

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

```r
# 17

# max number of participants
max(us_df$n_participants, na.rm = TRUE)
```

```
## [1] 1842
```

```r
1842
```

```
## [1] 1842
```

```r
# distribution
summary(us_df$n_participants)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1      48      91     132     164    1842
```

will use the 1st quartile as the cutoff (min number of participants for a race to be included) (85 participants)

```r
us_df <- us_df %>% filter(n_participants >= 85)
```

convert distance and elevation columns to be numeric

```r
colnames(us_df)[4] <- "distance_km"
colnames(us_df)[5] <- "elevation_gain_m"

us_df <- us_df %>%
  mutate(
    distance_km = parse_number(distance_km),
    elevation_gain_m = parse_number(elevation_gain_m)
  )
```

will also only use distances greater than or equal to 50km

```r
us_df <- us_df %>% filter(distance_km >= 50)
```

center distance and elevation gain values

```r
mean_distance_km <- mean(us_df$distance_km)
mean_elevation_m <- mean(us_df$elevation_gain_m)

us_df$distance_km <- us_df$distance_km - mean_distance_km
us_df$elevation_gain_m <- us_df$elevation_gain_m - mean_elevation_m
```

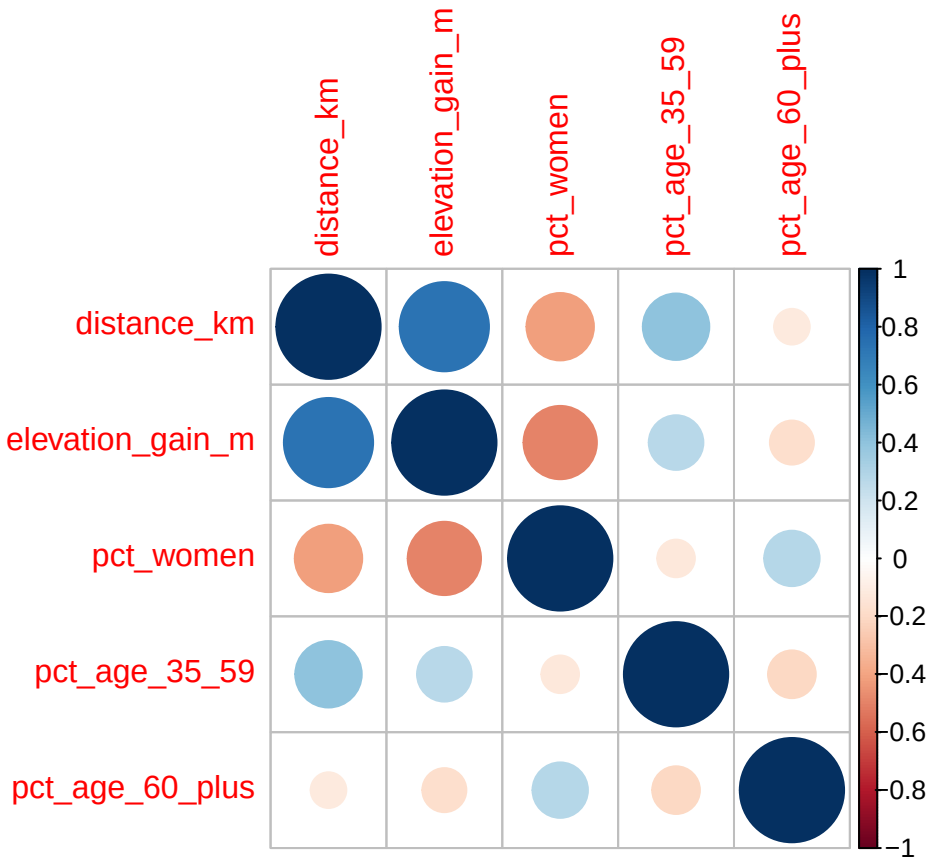
reminder that now all distance and elevation values are centered

distance baseline = 85.0581

elevation baseline = 2564.2234

EDA making a correlation matrix to see which factors should be included in regression remove pct_u35 to fix correlation issues (perfect multicollinearity)

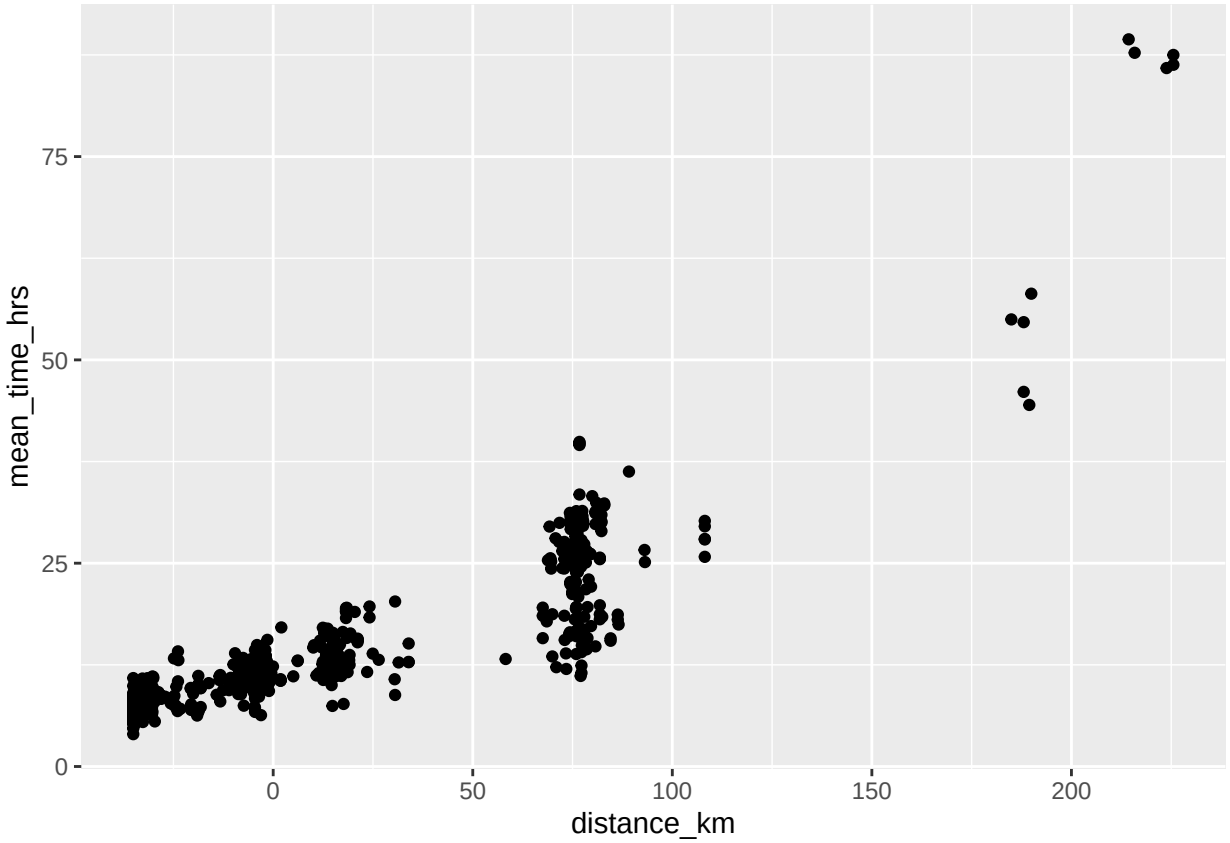
```
correlation_matrix <- us_df %>% dplyr::select(distance_km, elevation_gain_m, pct_women, pct_age_35_59, p

corrplot(cor(correlation_matrix))
```



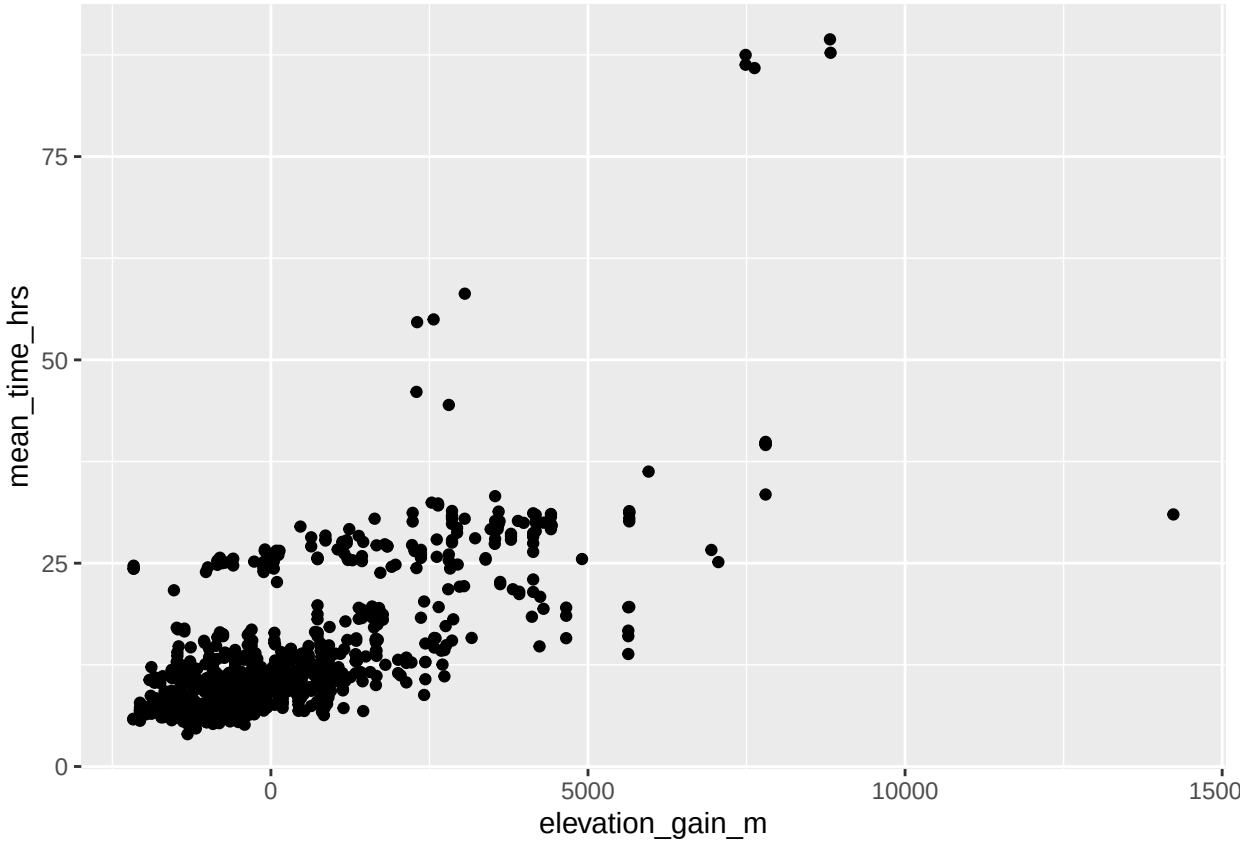
Preliminary scatterplots for analysis

- scatterplot of distance and elevation to see if linear relationship (maybe non-linear)
- scatterplot indicates potential quadratic relationship between distance and mean finish time

```
ggplot(us_df, aes(x = distance_km, y = mean_time_hrs)) +
      geom_point()
```




```
ggplot(us_df, aes(x = elevation_gain_m, y = mean_time_hrs)) +
      geom_point()
```



Regression

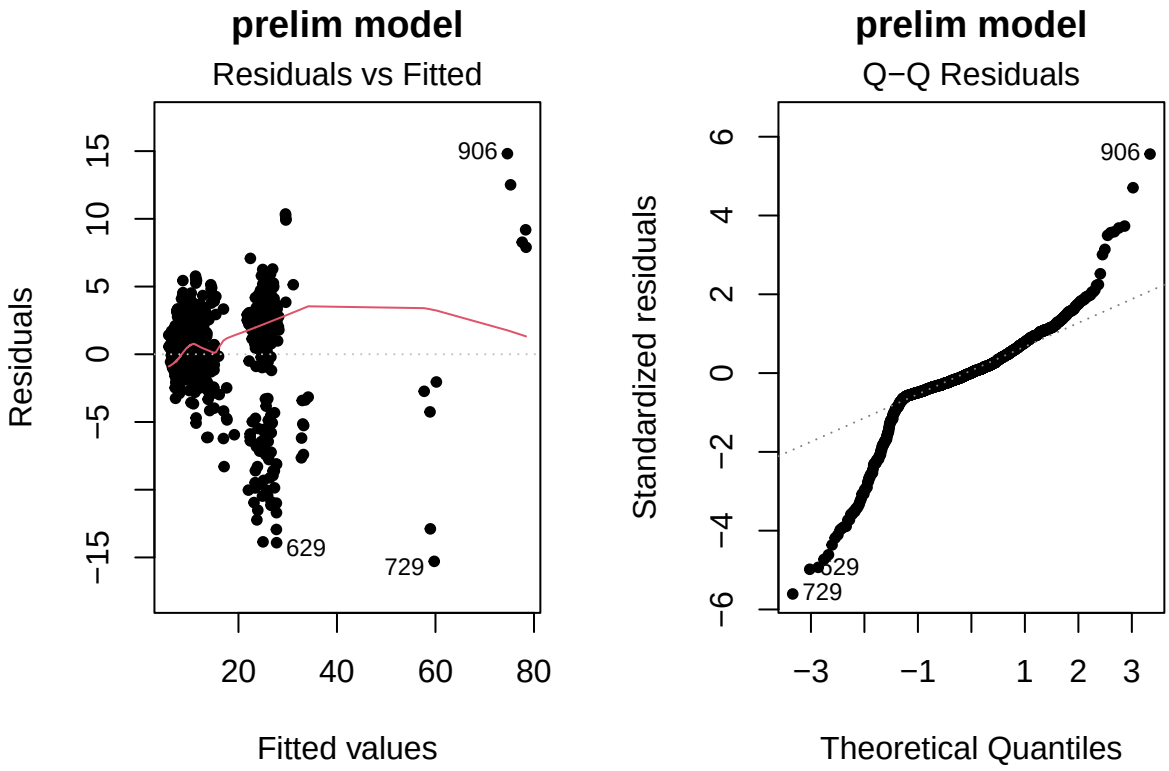
prelim regression with just main effects remember that we are using pct_u35 as the base so it is not included in models (to avoid perfect multicollinearity)

```
prelim_regression <- lm(mean_time_hrs ~ distance_km + I(distance_km^2) + elevation_gain_m +
                          pct_women + pct_age_35_59 + pct_age_60_plus, data = us_df)
summary(prelim_regression)
```

```
##
## Call:
## lm(formula = mean_time_hrs ~ distance_km + I(distance_km^2) +
##     elevation_gain_m + pct_women + pct_age_35_59 + pct_age_60_plus,
##     data = us_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2897  -0.9539   0.0285   1.3338  14.8097
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.057e+01  7.952e-01  13.291   <2e-16 ***
## distance_km       1.054e-01  3.585e-03  29.394   <2e-16 ***
## I(distance_km^2)  7.398e-04  2.845e-05  26.001   <2e-16 ***
## elevation_gain_m  7.546e-04  7.126e-05  10.590   <2e-16 ***
```

```
## pct_women        -2.082e-01  1.038e+00  -0.201    0.841
## pct_age_35_59     1.079e+00  1.075e+00   1.004    0.316
## pct_age_60_plus  -2.523e+00  1.752e+00  -1.440    0.150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.815 on 1188 degrees of freedom
## Multiple R-squared:  0.8983, Adjusted R-squared:  0.8978
## F-statistic:  1748 on 6 and 1188 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,2))
plot(prelim_regression, which=1, main="prelim model", pch=20)
plot(prelim_regression, which=2, main="prelim model", pch=20)
```



Using log regression as main model for robustness due to data not appearing to be normally distributing, going to attempt logistic regression mmodel

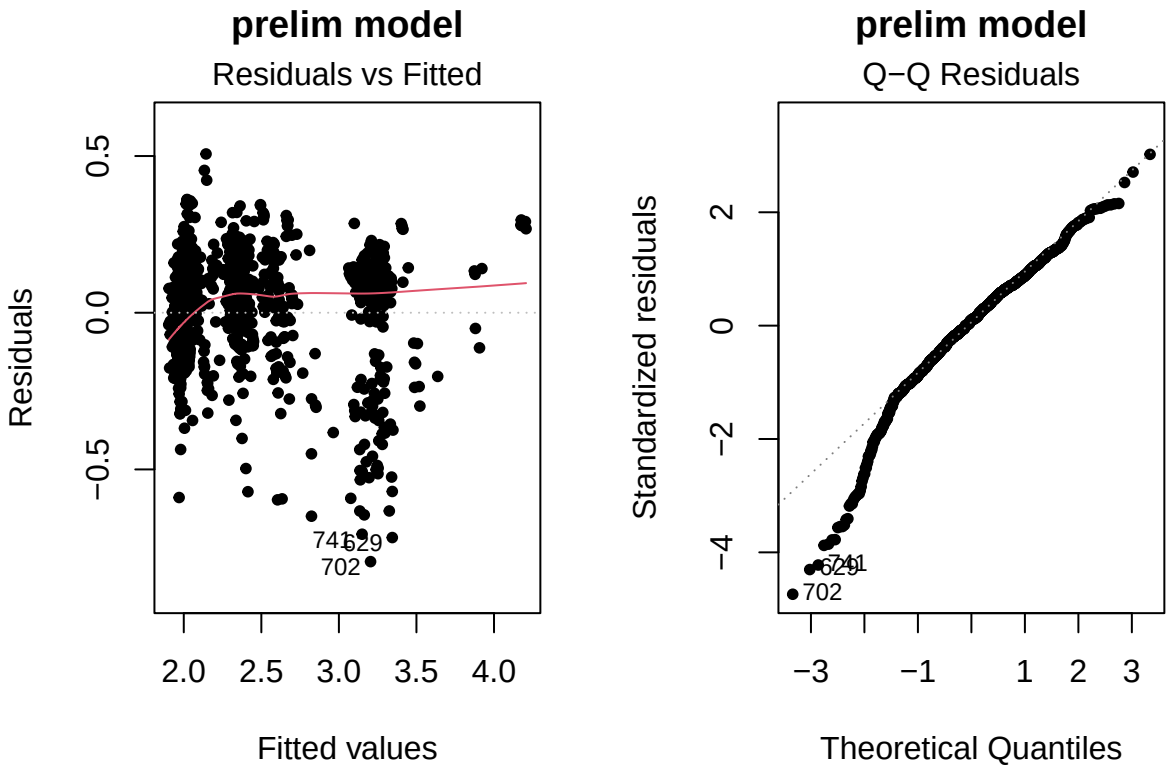
```
us_df <- us_df %>%
  mutate(log_mean_time = log(mean_time_hrs))
```

parsimonious model

```
log_regression <- lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
                       pct_women + pct_age_35_59 + pct_age_60_plus, data = us_df)

summary(log_regression)
```

```
##
## Call:
## lm(formula = log_mean_time ~ distance_km + I(distance_km^2) +
##     elevation_gain_m + pct_women + pct_age_35_59 + pct_age_60_plus,
##     data = us_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79419 -0.09046  0.01595  0.11109  0.50661
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.402e+00  4.744e-02  50.645   <2e-16 ***
## distance_km       1.101e-02  2.138e-04  51.466   <2e-16 ***
## I(distance_km^2) -1.932e-05  1.697e-06 -11.383   <2e-16 ***
## elevation_gain_m  3.543e-05  4.251e-06   8.335   <2e-16 ***
## pct_women         7.844e-02  6.192e-02   1.267    0.205
## pct_age_35_59    -1.224e-03  6.414e-02  -0.019    0.985
## pct_age_60_plus  -1.444e-01  1.045e-01  -1.382    0.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1679 on 1188 degrees of freedom
## Multiple R-squared:  0.889,  Adjusted R-squared:  0.8884
## F-statistic:  1585 on 6 and 1188 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(1,2))
plot(log_regression, which=1, main="prelim model", pch=20)
plot(log_regression, which=2, main="prelim model", pch=20)
```



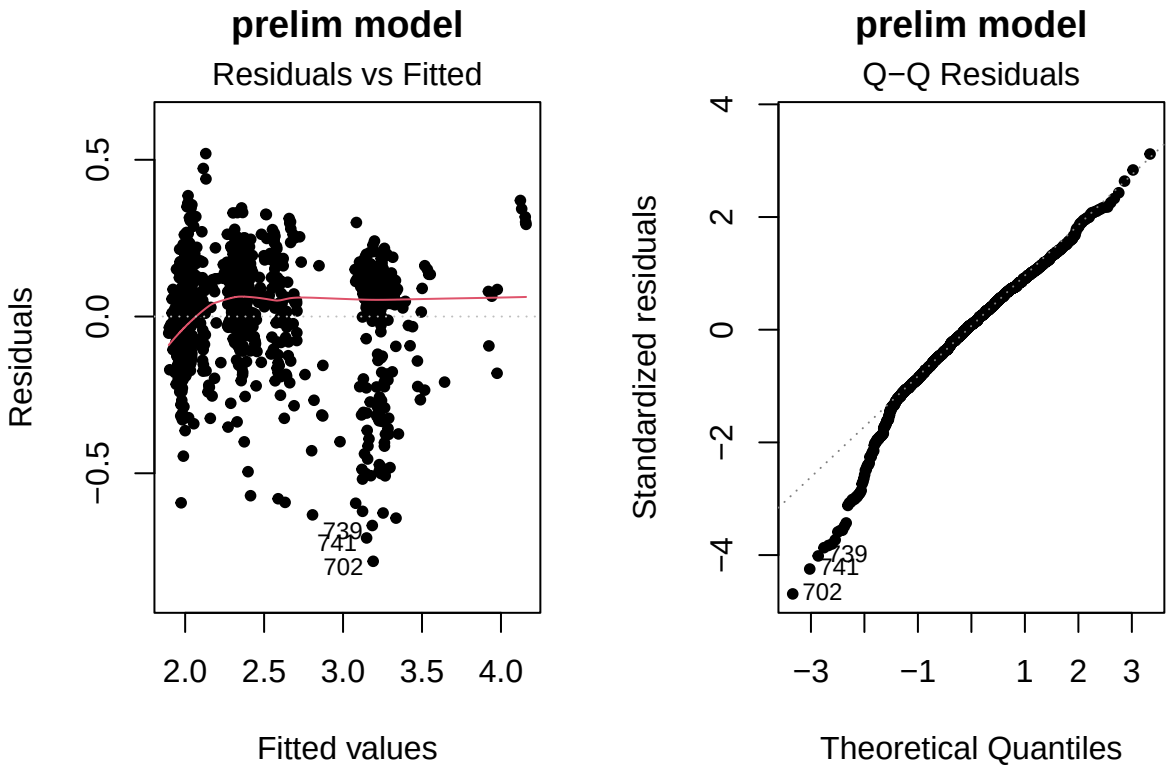
regressions with full model (interaction and higher order) adding only terms that are hypothesized to be significant and relevant

```
full_log_regression <- lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
                          pct_women + pct_age_35_59 + pct_age_60_plus +
                          distance_km*elevation_gain_m + distance_km*pct_women +
                          elevation_gain_m*pct_women, data = us_df)
summary(full_log_regression)
```

```
##
## Call:
## lm(formula = log_mean_time ~ distance_km + I(distance_km^2) +
##     elevation_gain_m + pct_women + pct_age_35_59 + pct_age_60_plus +
##     distance_km * elevation_gain_m + distance_km * pct_women +
##     elevation_gain_m * pct_women, data = us_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78075 -0.08917  0.01458  0.11163  0.51964
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.423e+00  4.805e-02  50.416  < 2e-16 ***
## distance_km                   9.389e-03  6.351e-04  14.785  < 2e-16 ***
## I(distance_km^2)             -1.490e-05  2.924e-06  -5.096 4.04e-07 ***
## elevation_gain_m              9.313e-05  1.518e-05   6.134 1.17e-09 ***
```

```
## pct_women                     4.085e-02  6.409e-02   0.637   0.5240
## pct_age_35_59                -2.428e-02  6.440e-02  -0.377   0.7062
## pct_age_60_plus              -1.748e-01  1.048e-01  -1.667   0.0957 .
## distance_km:elevation_gain_m -1.365e-07  8.572e-08  -1.592   0.1117
## distance_km:pct_women         5.189e-03  2.046e-03   2.537   0.0113 *
## elevation_gain_m:pct_women   -2.043e-04  5.119e-05  -3.992 6.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.167 on 1185 degrees of freedom
## Multiple R-squared:  0.8905, Adjusted R-squared:  0.8896
## F-statistic:  1070 on 9 and 1185 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,2))
plot(full_log_regression, which=1, main="prelim model", pch=20)
plot(full_log_regression, which=2, main="prelim model", pch=20)
```



regression with reduced model after stepwise backward elimination

```
log_backward_step_model <- stepAIC(full_log_regression, direction = "backward")
```

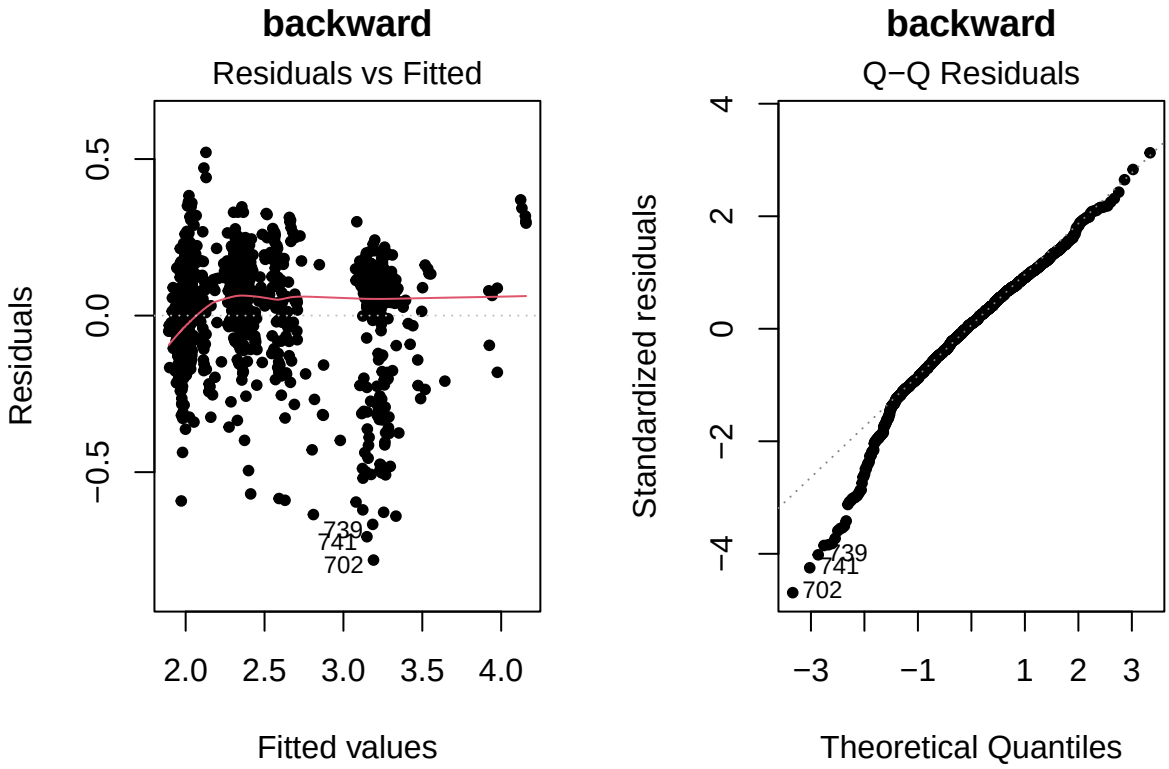
```
## Start:  AIC=-4267.96
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + pct_age_35_59 + pct_age_60_plus + distance_km *
##     elevation_gain_m + distance_km * pct_women + elevation_gain_m *
```

```
##     pct_women
##
##                                Df Sum of Sq    RSS     AIC
## - pct_age_35_59                 1   0.00396 33.042 -4269.8
## <none>                                      33.038 -4268.0
## - distance_km:elevation_gain_m  1   0.07065 33.108 -4267.4
## - pct_age_60_plus               1   0.07750 33.115 -4267.2
## - distance_km:pct_women         1   0.17939 33.217 -4263.5
## - elevation_gain_m:pct_women    1   0.44421 33.482 -4254.0
## - I(distance_km^2)              1   0.72389 33.762 -4244.1
##
## Step:  AIC=-4269.82
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + pct_age_60_plus + distance_km:elevation_gain_m +
##     distance_km:pct_women + elevation_gain_m:pct_women
##
##                                Df Sum of Sq    RSS     AIC
## <none>                                      33.042 -4269.8
## - distance_km:elevation_gain_m  1   0.07254 33.114 -4269.2
## - pct_age_60_plus               1   0.07357 33.115 -4269.2
## - distance_km:pct_women         1   0.17742 33.219 -4265.4
## - elevation_gain_m:pct_women    1   0.44025 33.482 -4256.0
## - I(distance_km^2)              1   0.71999 33.762 -4246.1
```

```r
summary(log_backward_step_model)
```

```
##
## Call:
## lm(formula = log_mean_time ~ distance_km + I(distance_km^2) +
##     elevation_gain_m + pct_women + pct_age_60_plus + distance_km:elevation_gain_m +
##     distance_km:pct_women + elevation_gain_m:pct_women, data = us_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78083 -0.09029  0.01419  0.11212  0.52128
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.406e+00  1.873e-02 128.432  < 2e-16 ***
## distance_km                    9.372e-03  6.331e-04  14.804  < 2e-16 ***
## I(distance_km^2)              -1.481e-05  2.913e-06  -5.084 4.30e-07 ***
## elevation_gain_m               9.281e-05  1.515e-05   6.124 1.24e-09 ***
## pct_women                      3.932e-02  6.394e-02   0.615   0.5387
## pct_age_60_plus               -1.667e-01  1.026e-01  -1.625   0.1044
## distance_km:elevation_gain_m -1.381e-07  8.558e-08  -1.614   0.1069
## distance_km:pct_women          5.156e-03  2.043e-03   2.524   0.0117 *
## elevation_gain_m:pct_women    -2.025e-04  5.094e-05  -3.975 7.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1669 on 1186 degrees of freedom
## Multiple R-squared:  0.8905, Adjusted R-squared:  0.8897
## F-statistic:  1205 on 8 and 1186 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(1,2))
plot(log_backward_step_model, which=1, main="backward", pch=20)
plot(log_backward_step_model, which=2, main="backward", pch=20)
```



regression using stepwise elimination (both directions)

```r
log_both_step_model <- stepAIC(full_log_regression, direction = "both")
```

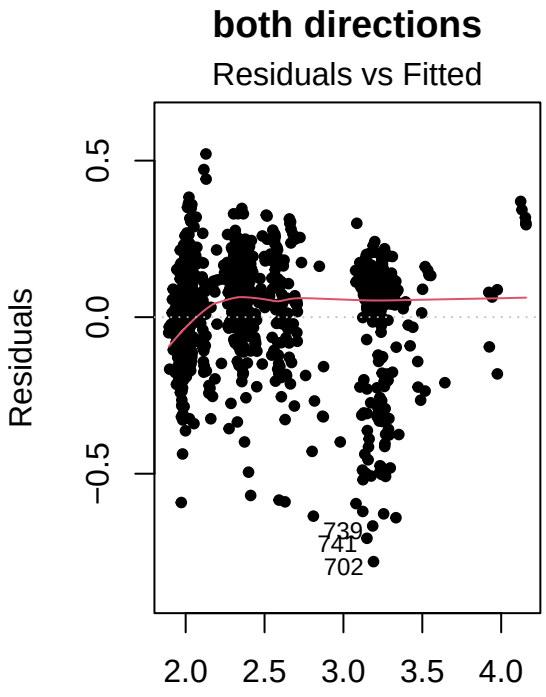
```
## Start:  AIC=-4267.96
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + pct_age_35_59 + pct_age_60_plus + distance_km *
##     elevation_gain_m + distance_km * pct_women + elevation_gain_m *
##     pct_women
##
##                                Df Sum of Sq    RSS     AIC
## - pct_age_35_59                 1   0.00396 33.042 -4269.8
## <none>                                      33.038 -4268.0
## - distance_km:elevation_gain_m  1   0.07065 33.108 -4267.4
## - pct_age_60_plus               1   0.07750 33.115 -4267.2
## - distance_km:pct_women         1   0.17939 33.217 -4263.5
## - elevation_gain_m:pct_women    1   0.44421 33.482 -4254.0
## - I(distance_km^2)              1   0.72389 33.762 -4244.1
##
## Step:  AIC=-4269.82
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
```

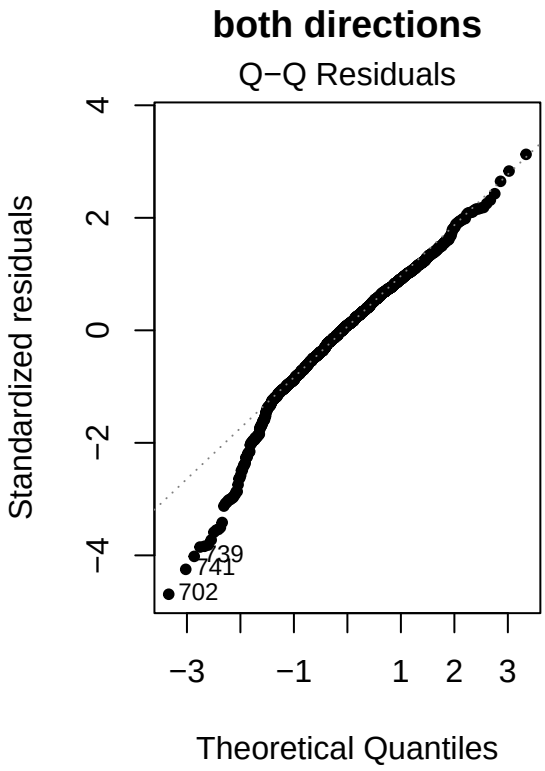
```
##     pct_women + pct_age_60_plus + distance_km:elevation_gain_m +
##     distance_km:pct_women + elevation_gain_m:pct_women
##
##                                Df Sum of Sq    RSS     AIC
## <none>                                      33.042 -4269.8
## - distance_km:elevation_gain_m  1   0.07254 33.114 -4269.2
## - pct_age_60_plus               1   0.07357 33.115 -4269.2
## + pct_age_35_59                 1   0.00396 33.038 -4268.0
## - distance_km:pct_women         1   0.17742 33.219 -4265.4
## - elevation_gain_m:pct_women    1   0.44025 33.482 -4256.0
## - I(distance_km^2)              1   0.71999 33.762 -4246.1
```

```
summary(log_both_step_model)
```

```
##
## Call:
## lm(formula = log_mean_time ~ distance_km + I(distance_km^2) +
##     elevation_gain_m + pct_women + pct_age_60_plus + distance_km:elevation_gain_m +
##     distance_km:pct_women + elevation_gain_m:pct_women, data = us_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78083 -0.09029  0.01419  0.11212  0.52128
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.406e+00  1.873e-02 128.432  < 2e-16 ***
## distance_km                   9.372e-03  6.331e-04  14.804  < 2e-16 ***
## I(distance_km^2)             -1.481e-05  2.913e-06  -5.084 4.30e-07 ***
## elevation_gain_m              9.281e-05  1.515e-05   6.124 1.24e-09 ***
## pct_women                     3.932e-02  6.394e-02   0.615   0.5387
## pct_age_60_plus              -1.667e-01  1.026e-01  -1.625   0.1044
## distance_km:elevation_gain_m -1.381e-07  8.558e-08  -1.614   0.1069
## distance_km:pct_women         5.156e-03  2.043e-03   2.524   0.0117 *
## elevation_gain_m:pct_women   -2.025e-04  5.094e-05  -3.975 7.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1669 on 1186 degrees of freedom
## Multiple R-squared:  0.8905, Adjusted R-squared:  0.8897
## F-statistic:  1205 on 8 and 1186 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,2))
plot(log_both_step_model, which=1, main="both directions", pch=20)
plot(log_both_step_model, which=2, main="both directions", pch=20)
```





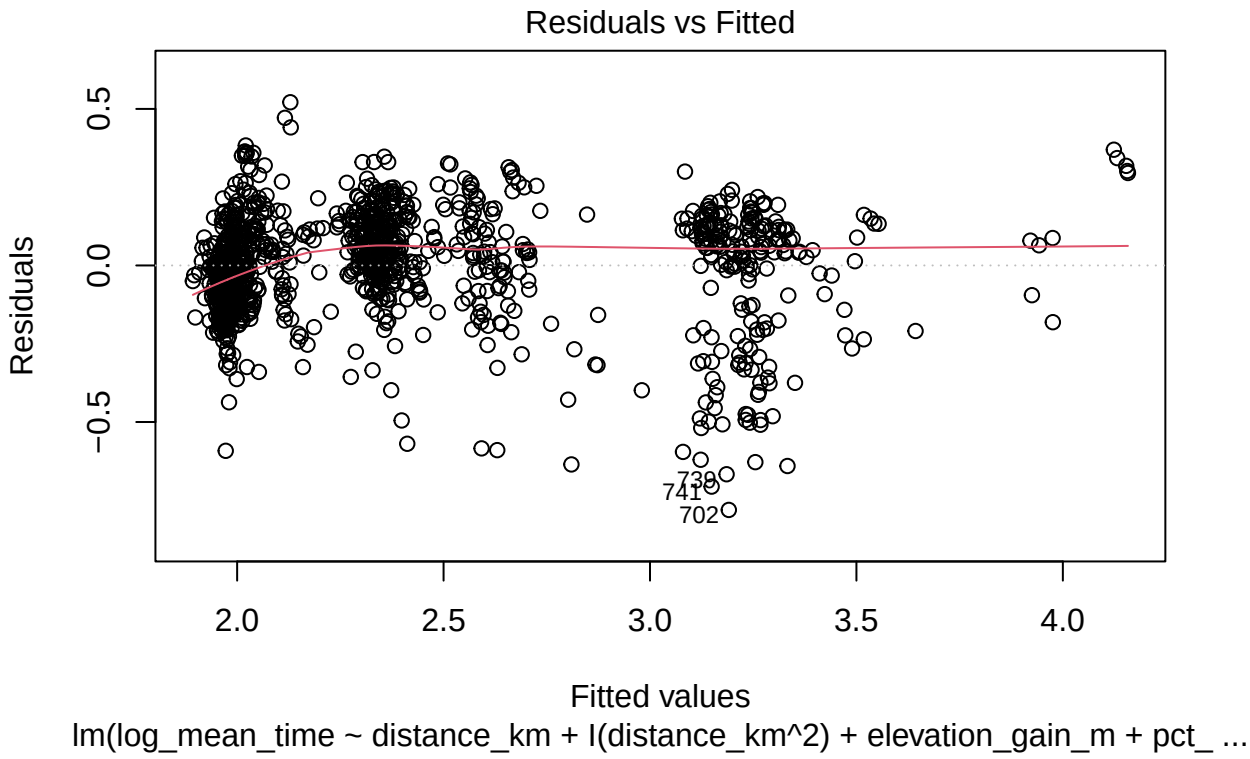

```
vif(log_backward_step_model, type = "predictor")
```

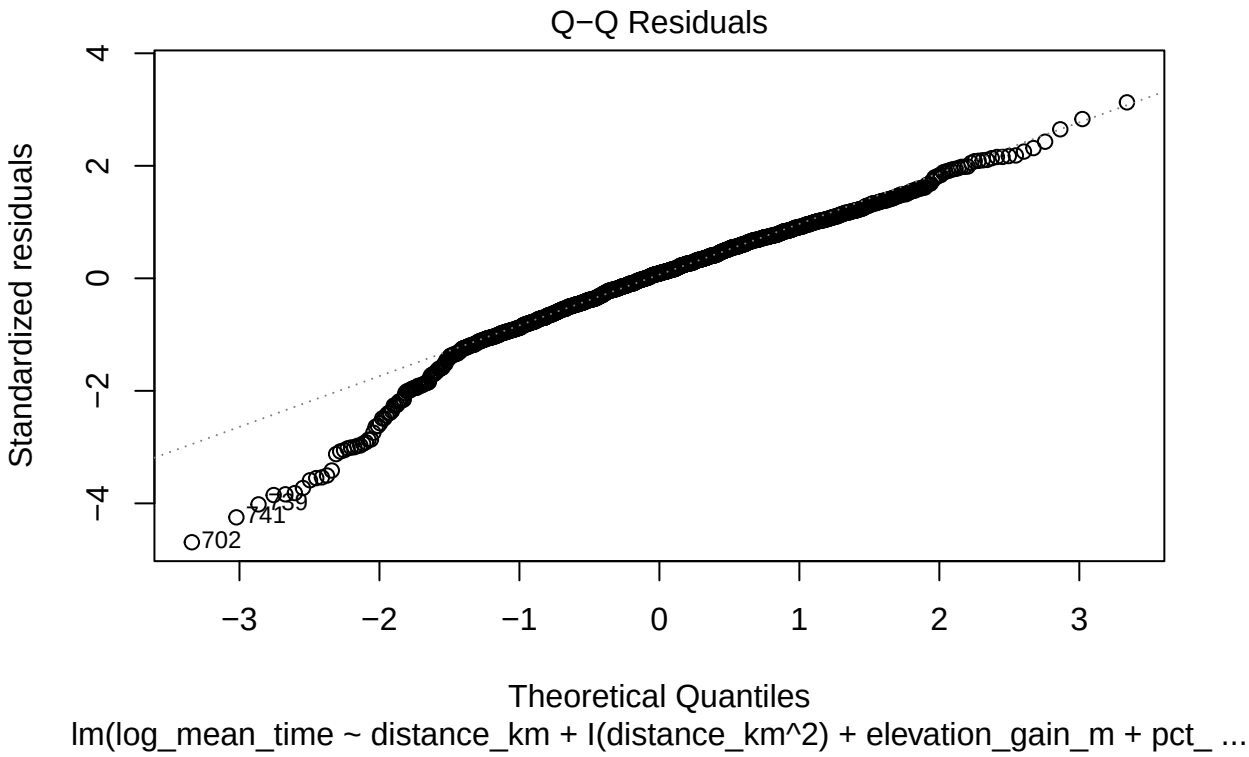
checking VIF of log stepwise regression models

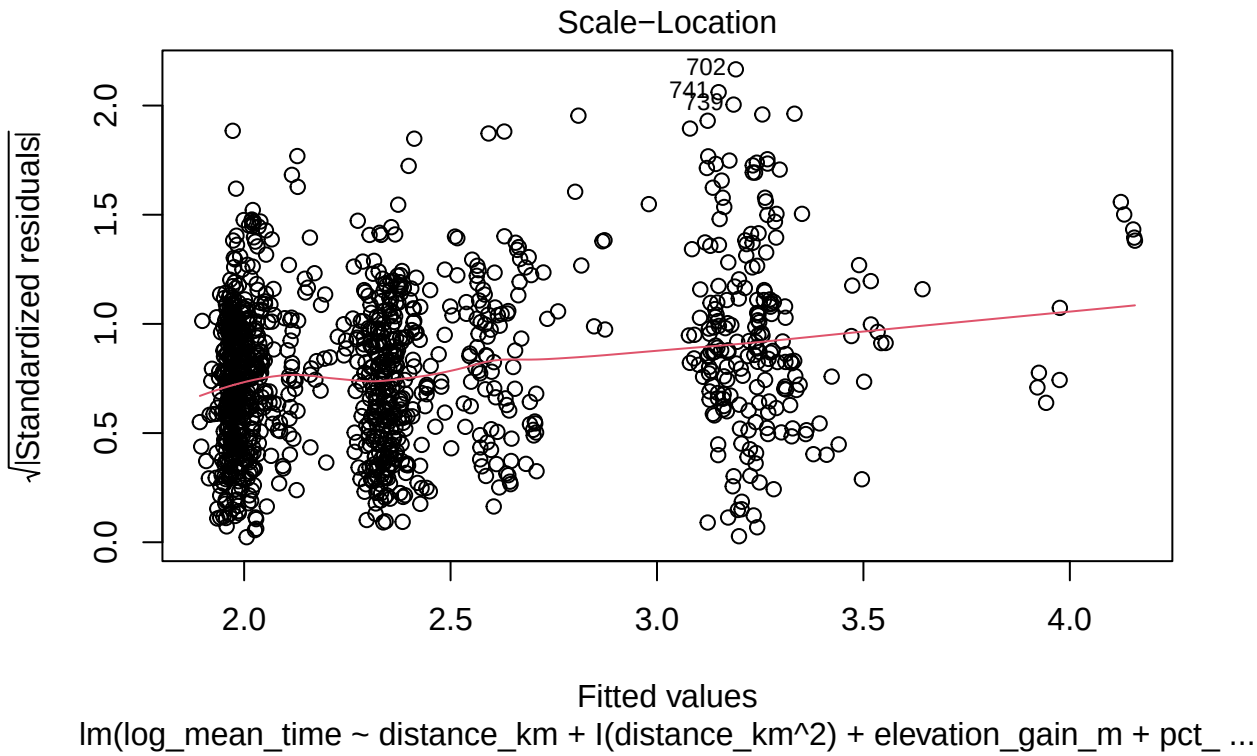
```
## GVIFs computed for predictors

##                      GVIF Df GVIF^(1/(2*Df))
## distance_km      1.110801  7        1.007534
## elevation_gain_m 1.110801  7        1.007534
## pct_women        1.110801  7        1.007534
## pct_age_60_plus  1.110801  1        1.053946
##                                              Interacts With
## distance_km      I(distance_km^2), elevation_gain_m, pct_women
## elevation_gain_m                        distance_km, pct_women
## pct_women                          distance_km, elevation_gain_m
## pct_age_60_plus                                             --
##                                           Other Predictors
## distance_km                                pct_age_60_plus
## elevation_gain_m                           pct_age_60_plus
## pct_women                                  pct_age_60_plus
## pct_age_60_plus  distance_km, elevation_gain_m, pct_women
```

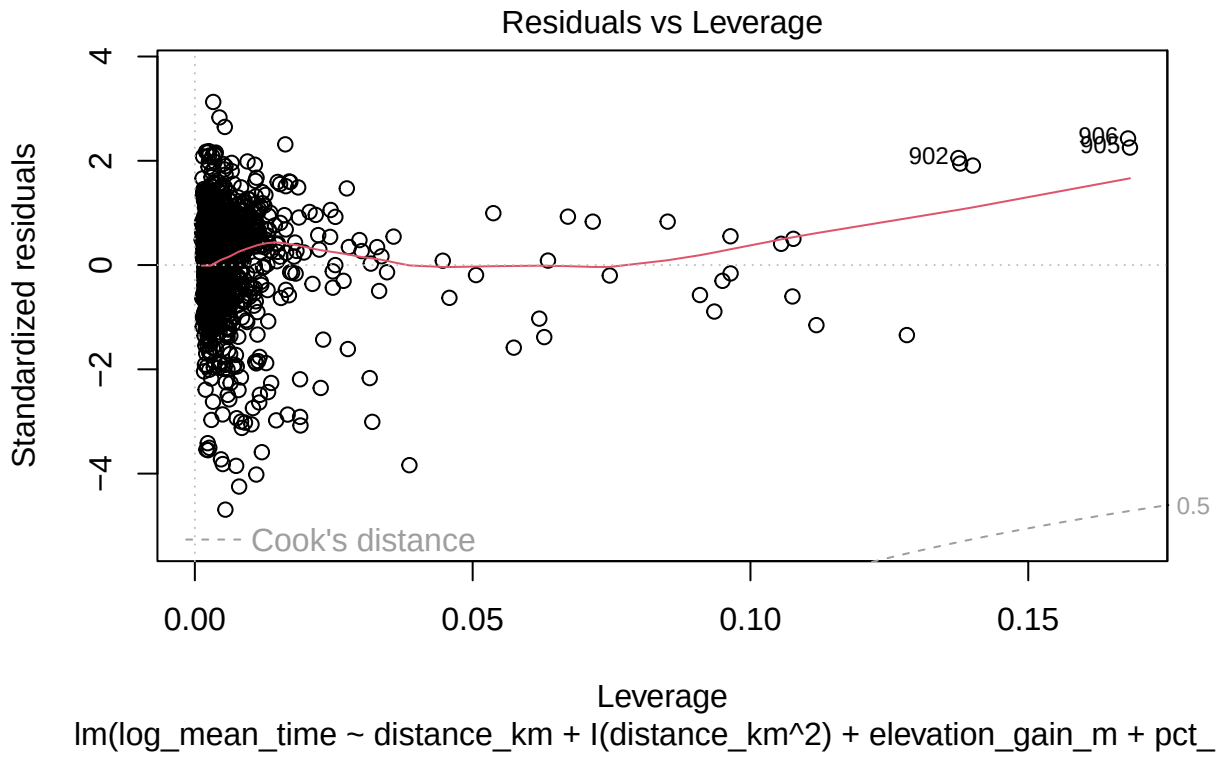
look for outliers using cooks distance

```r
cooks_distance <- cooks.distance(log_both_step_model)
threshold <- 4 / nrow(us_df)
influential_observations <- which(cooks_distance > threshold)

plot(log_both_step_model)
```




Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m + pct_ ...
2239
741
702


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m + pct_ ...
702
741
739



rerun regression without outliers to see if conclusions change (sensitivity analysis) remove outliers

```
keep_rows <- which(cooks_distance <= threshold)
us_df_no_outliers <- us_df[keep_rows, ]
```

stepwise regression using data with outliers removes to check sensitivity and robustness

```
no_outlier_log_regression <- lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
                          pct_women + pct_age_35_59 + pct_age_60_plus +
                          distance_km*elevation_gain_m + distance_km*pct_women +
                          elevation_gain_m*pct_women, data = us_df_no_outliers)

outlierless_log_both_step_model <- stepAIC(no_outlier_log_regression, direction = "both")
```

```
## Start:  AIC=-4535.9
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + pct_age_35_59 + pct_age_60_plus + distance_km *
##     elevation_gain_m + distance_km * pct_women + elevation_gain_m *
##     pct_women
##
##                               Df Sum of Sq    RSS     AIC
## - pct_age_60_plus              1   0.00024 19.344 -4537.9
## - pct_age_35_59                1   0.00255 19.346 -4537.7
```

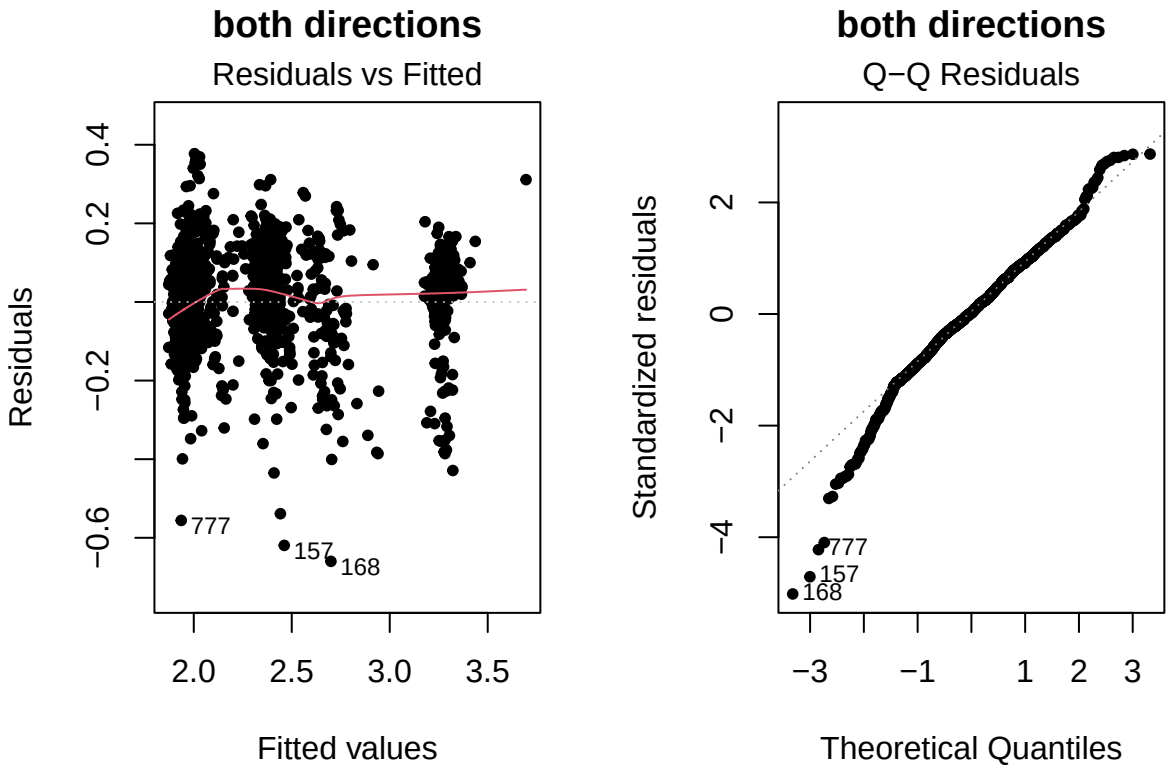
```
## - distance_km:pct_women         1   0.00850 19.352 -4537.4
## <none>                                  19.343 -4535.9
## - elevation_gain_m:pct_women    1   0.13276 19.476 -4530.2
## - distance_km:elevation_gain_m  1   0.51286 19.856 -4508.5
## - I(distance_km^2)              1   0.92339 20.267 -4485.6
##
## Step:  AIC=-4537.88
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + pct_age_35_59 + distance_km:elevation_gain_m +
##     distance_km:pct_women + elevation_gain_m:pct_women
##
##                                 Df Sum of Sq    RSS     AIC
## - pct_age_35_59                  1   0.00293 19.347 -4539.7
## - distance_km:pct_women          1   0.00830 19.352 -4539.4
## <none>                                  19.344 -4537.9
## + pct_age_60_plus                1   0.00024 19.343 -4535.9
## - elevation_gain_m:pct_women     1   0.13253 19.476 -4532.2
## - distance_km:elevation_gain_m   1   0.52069 19.864 -4510.1
## - I(distance_km^2)               1   0.92786 20.272 -4487.3
##
## Step:  AIC=-4539.71
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + distance_km:elevation_gain_m + distance_km:pct_women +
##     elevation_gain_m:pct_women
##
##                                 Df Sum of Sq    RSS     AIC
## - distance_km:pct_women          1   0.00858 19.355 -4541.2
## <none>                                  19.347 -4539.7
## + pct_age_35_59                  1   0.00293 19.344 -4537.9
## + pct_age_60_plus                1   0.00062 19.346 -4537.7
## - elevation_gain_m:pct_women     1   0.13564 19.482 -4533.9
## - distance_km:elevation_gain_m   1   0.51896 19.866 -4512.0
## - I(distance_km^2)               1   0.93827 20.285 -4488.6
##
## Step:  AIC=-4541.21
## log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m +
##     pct_women + distance_km:elevation_gain_m + elevation_gain_m:pct_women
##
##                                 Df Sum of Sq    RSS     AIC
## <none>                                  19.355 -4541.2
## + distance_km:pct_women          1   0.00858 19.347 -4539.7
## + pct_age_35_59                  1   0.00321 19.352 -4539.4
## + pct_age_60_plus                1   0.00026 19.355 -4539.2
## - elevation_gain_m:pct_women     1   0.16544 19.521 -4533.7
## - distance_km:elevation_gain_m   1   0.53072 19.886 -4512.9
## - I(distance_km^2)               1   0.95482 20.310 -4489.2
```

```
summary(outlierless_log_both_step_model)
```

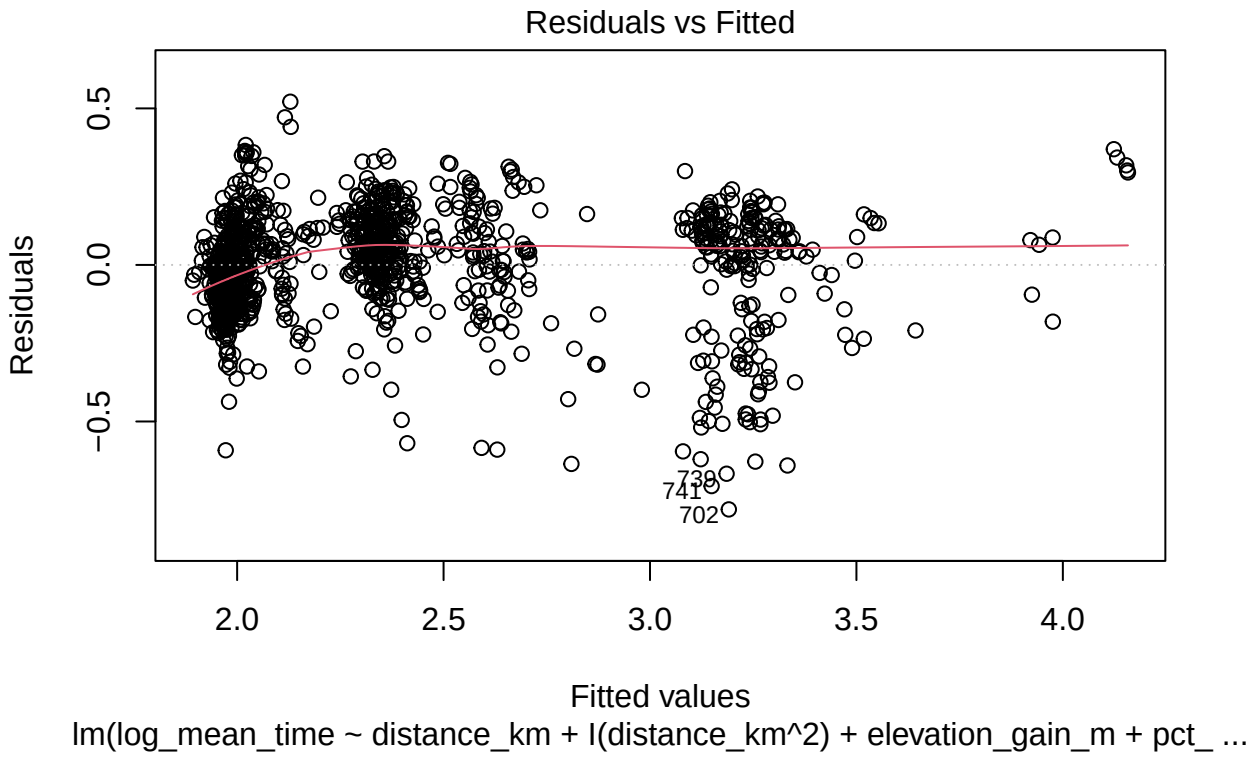
```
##
## Call:
## lm(formula = log_mean_time ~ distance_km + I(distance_km^2) +
##     elevation_gain_m + pct_women + distance_km:elevation_gain_m +
##     elevation_gain_m:pct_women, data = us_df_no_outliers)
```

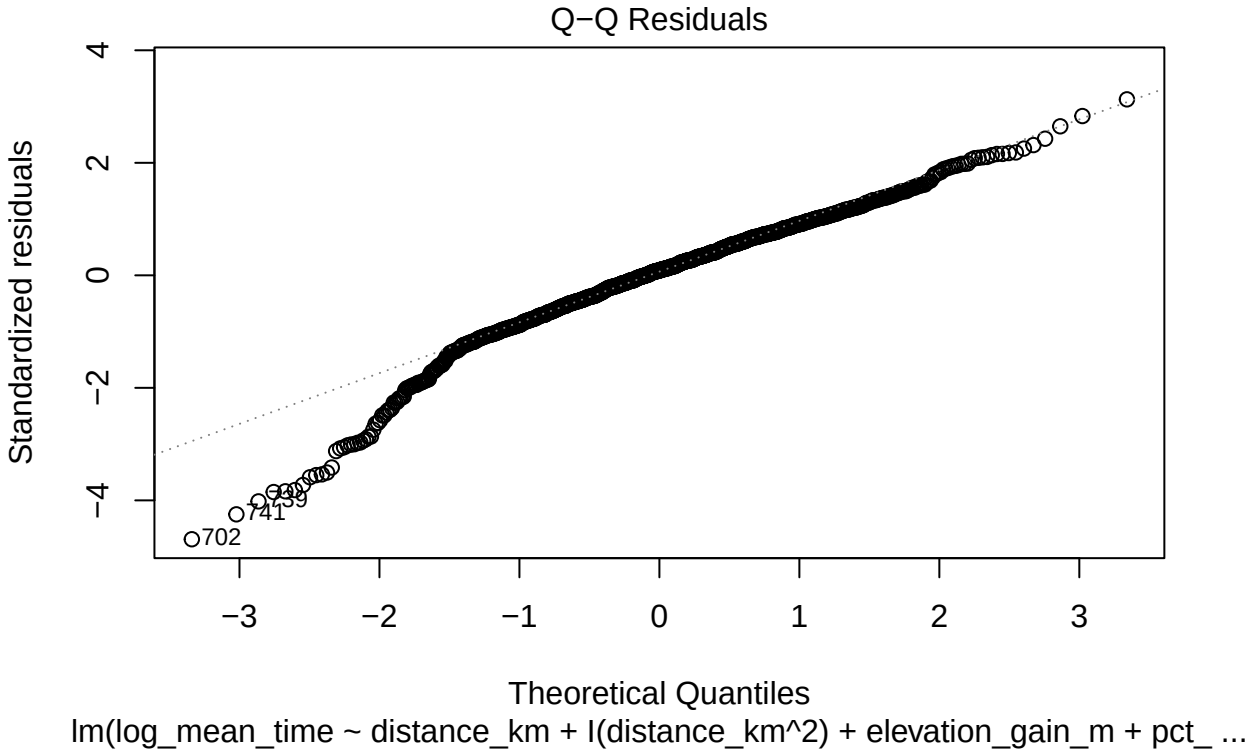
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65988 -0.07322  0.00263  0.08543  0.37731
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.445e+00  1.634e-02 149.564  < 2e-16 ***
## distance_km                     1.210e-02  2.015e-04  60.056  < 2e-16 ***
## I(distance_km^2)               -2.503e-05  3.375e-06  -7.417 2.38e-13 ***
## elevation_gain_m                8.128e-05  1.181e-05   6.882 9.80e-12 ***
## pct_women                       4.296e-02  5.676e-02   0.757  0.44929
## distance_km:elevation_gain_m   -5.275e-07  9.540e-08  -5.529 4.00e-08 ***
## elevation_gain_m:pct_women     -1.187e-04  3.846e-05  -3.087  0.00207 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1318 on 1115 degrees of freedom
## Multiple R-squared:  0.9225, Adjusted R-squared:  0.9221
## F-statistic:  2211 on 6 and 1115 DF,  p-value: < 2.2e-16
```

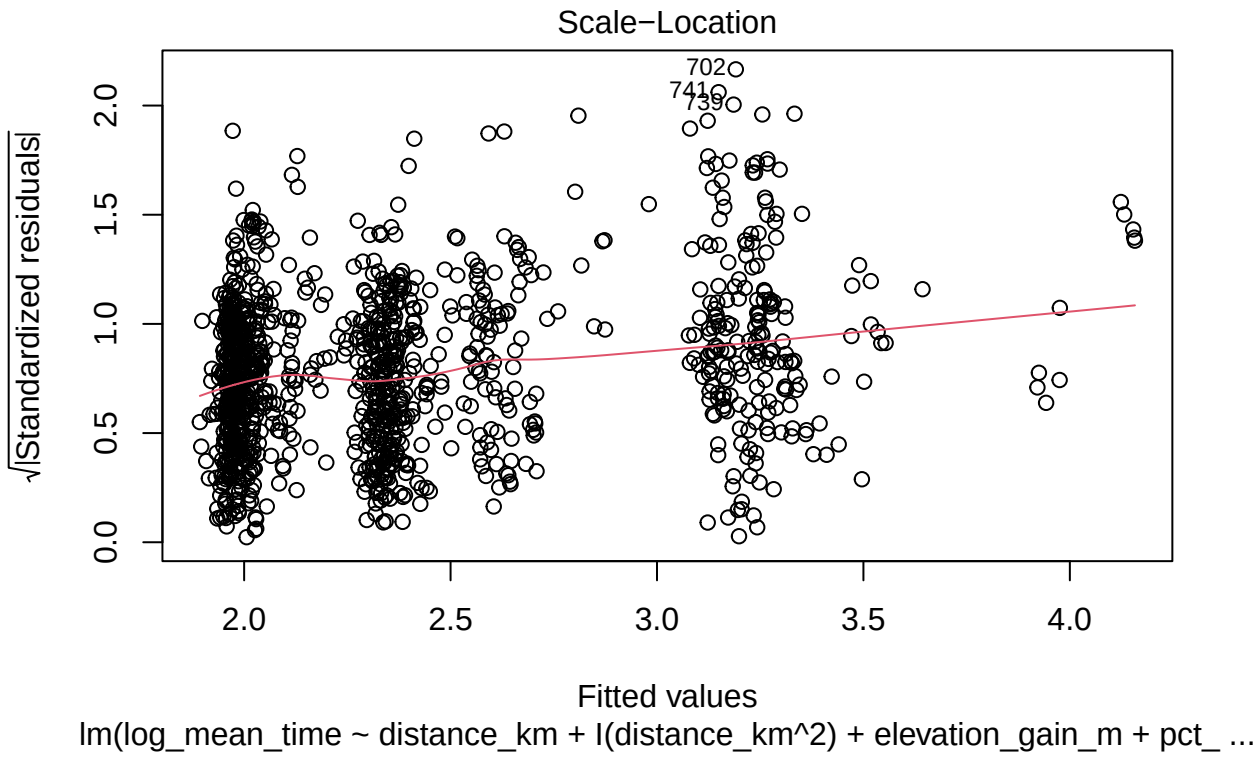
```r
par(mfrow = c(1,2))
plot(outlierless_log_both_step_model, which=1, main="both directions", pch=20)
plot(outlierless_log_both_step_model, which=2, main="both directions", pch=20)
```

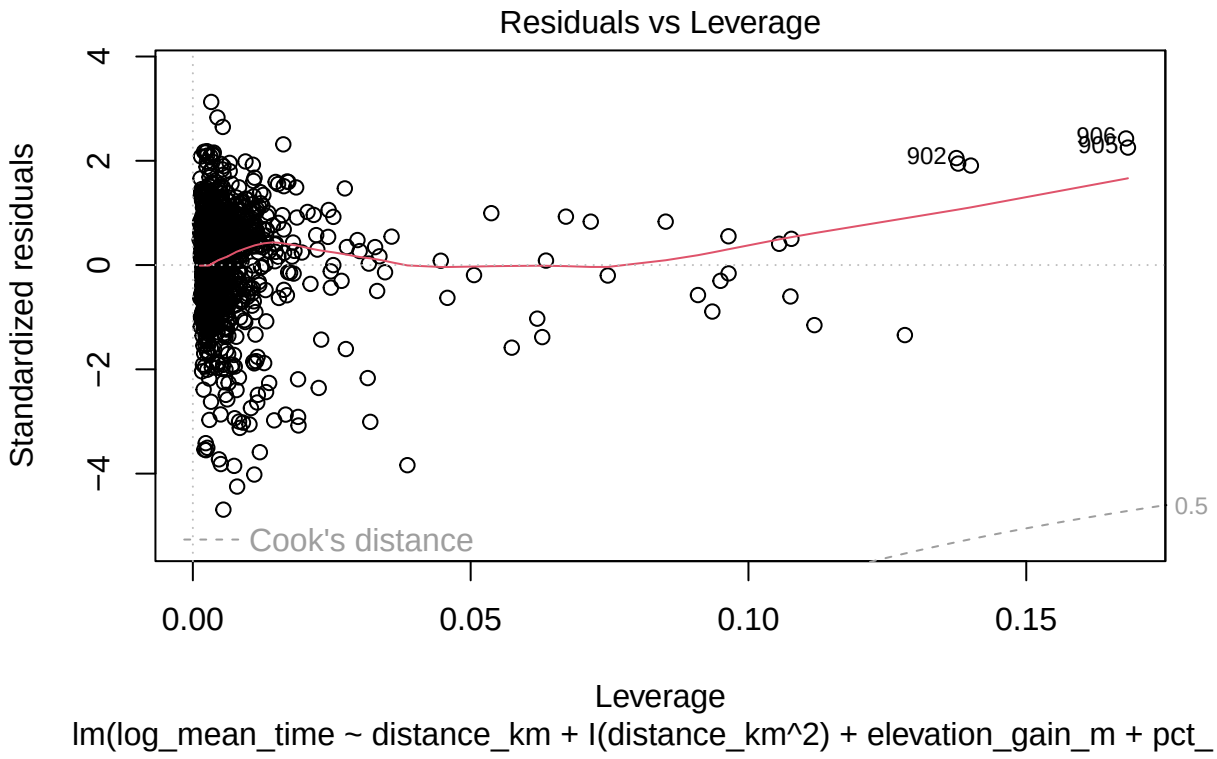


```
plot(log_both_step_model)
```




Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
2239
741
702
Theoretical Quantiles
lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m + pct_ ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(log_mean_time ~ distance_km + I(distance_km^2) + elevation_gain_m + pct_ ...
702
741
739



these clusters are due to the most common distances 50k, 50mi, 100k, and 100mi

deciding to use the outlier-free model

vif analysis

```
vif(outlierless_log_both_step_model, type = "predictor")
```

```
## GVIFs computed for predictors
```

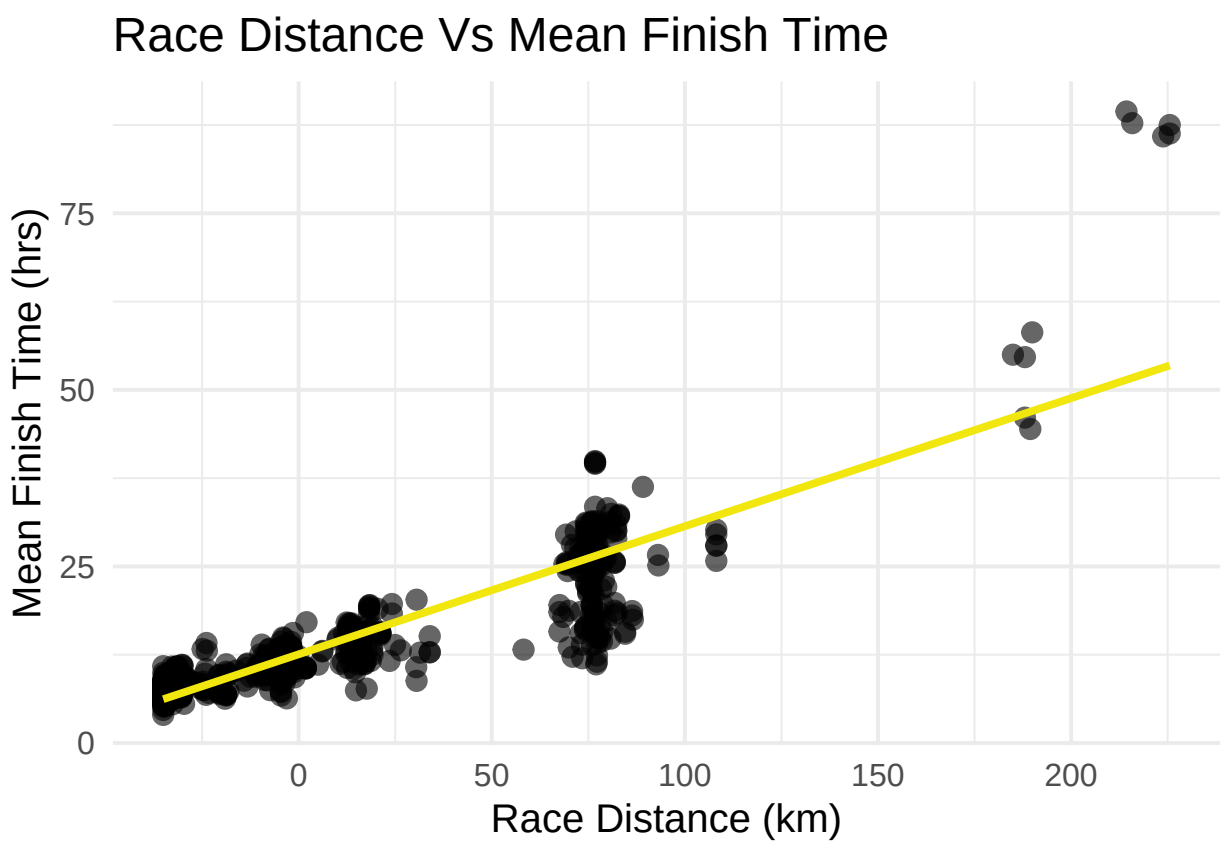
```
##                       GVIF Df GVIF^(1/(2*Df))
## distance_km      14.117138  4        1.392254
## elevation_gain_m  1.000000  6        1.000000
## pct_women         4.406172  3        1.280394
##                                  Interacts With Other Predictors
## distance_km      I(distance_km^2), elevation_gain_m        pct_women
## elevation_gain_m             distance_km, pct_women               --
## pct_women                          elevation_gain_m      distance_km
```

making scatterplots to put on the poster to show distributions scatterplot of distance and mean time (with outliers)

```
ggplot(us_df, aes(x = distance_km, y = mean_time_hrs)) +
  geom_point(alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", se = FALSE, col = "#f1e60e") +
  labs(
```

```
    x = "Race Distance (km)",
    y = "Mean Finish Time (hrs)",
    title = "Race Distance Vs Mean Finish Time"
  ) +
  theme_minimal(base_size = 15)
```

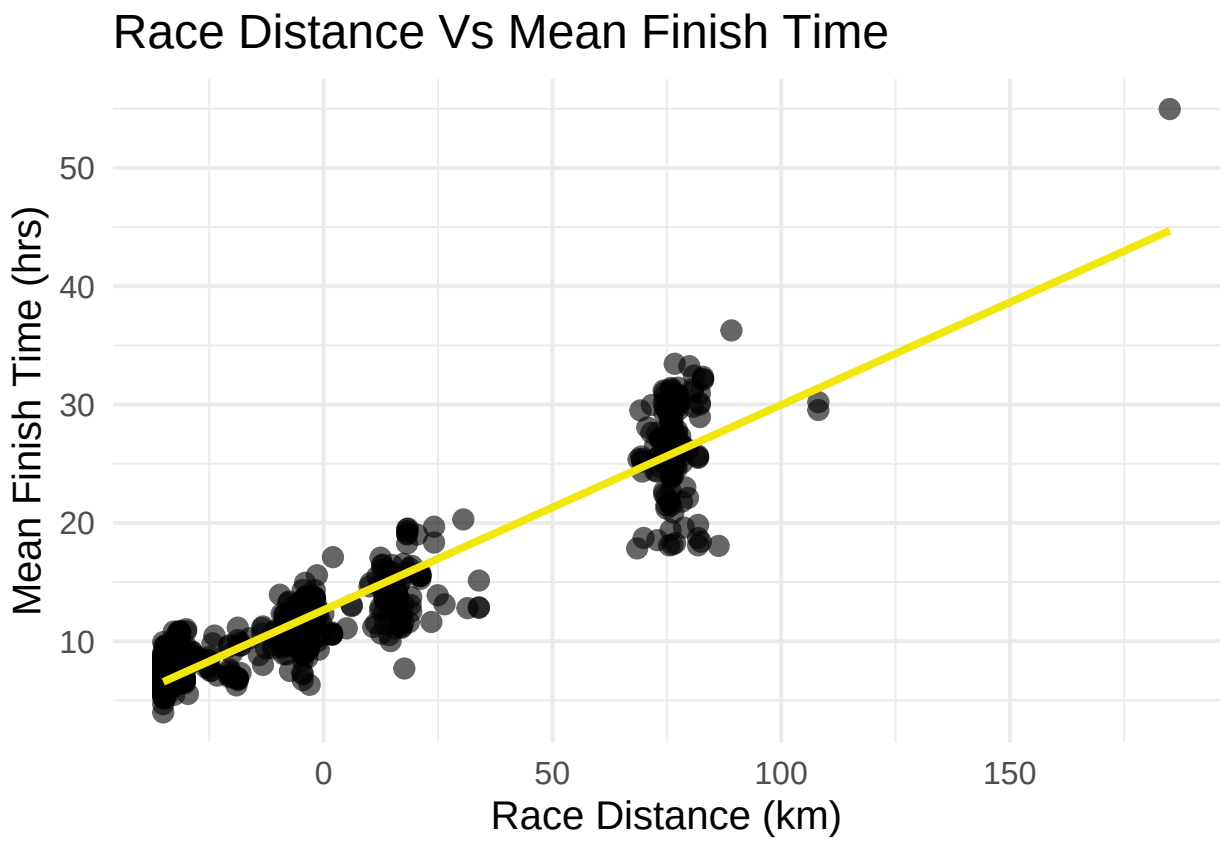
```
## `geom_smooth()` using formula = 'y ~ x'
```



first scatterplot of mean time and distance with outliers removed

```
ggplot(us_df_no_outliers, aes(x = distance_km, y = mean_time_hrs)) +
  geom_point(alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", se = FALSE, col = "#f1e60e") +
  labs(
    x = "Race Distance (km)",
    y = "Mean Finish Time (hrs)",
    title = "Race Distance Vs Mean Finish Time"
  ) +
  theme_minimal(base_size = 15)
```

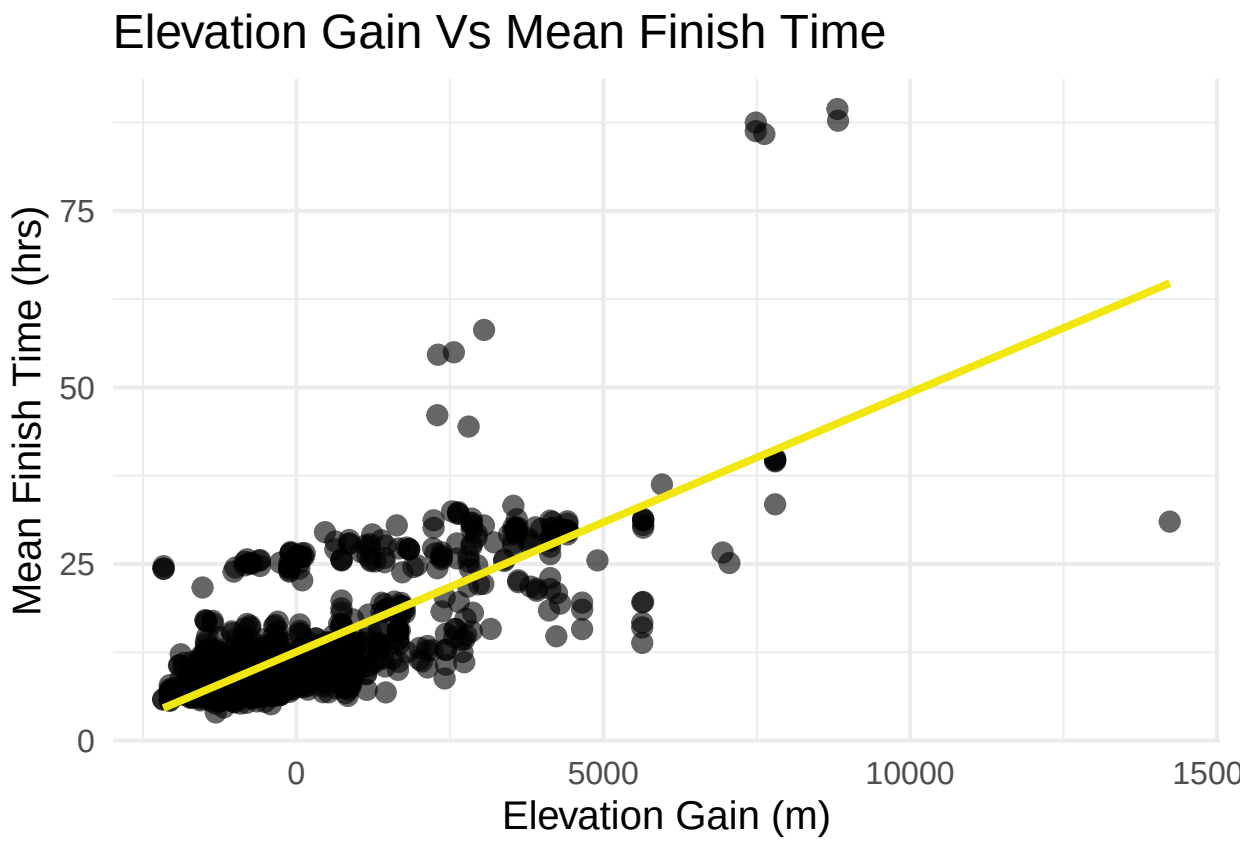
```
## `geom_smooth()` using formula = 'y ~ x'
```



second scatterplot of mean time and elevation gain (with outliers)

```
ggplot(us_df, aes(x = elevation_gain_m, y = mean_time_hrs)) +
  geom_point(alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", se = FALSE, col = "#f1e60e") +
  labs(
    x = "Elevation Gain (m)",
    y = "Mean Finish Time (hrs)",
    title = "Elevation Gain Vs Mean Finish Time"
  ) +
  theme_minimal(base_size = 15)
```

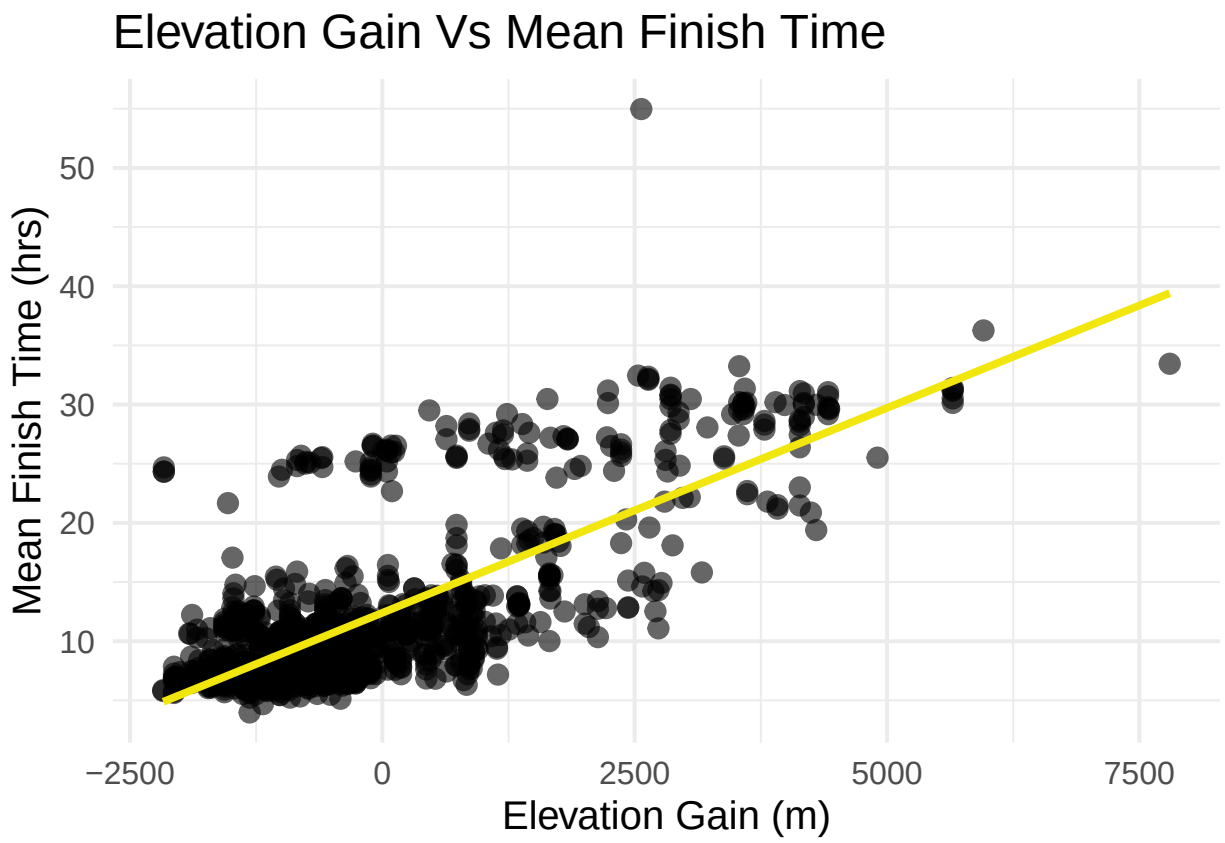
```
## `geom_smooth()` using formula = 'y ~ x'
```



second scatterplot of mean time and elevation gain with outliers removed

```
ggplot(us_df_no_outliers, aes(x = elevation_gain_m, y = mean_time_hrs)) +
  geom_point(alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", se = FALSE, col = "#f1e60e") +
  labs(
    x = "Elevation Gain (m)",
    y = "Mean Finish Time (hrs)",
    title = "Elevation Gain Vs Mean Finish Time"
  ) +
  theme_minimal(base_size = 15)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Elevation Gain Vs Mean Finish Time
Mean Finish Time (hrs)
50
40
30
20
10
−2500
0
2500
5000
7500
Elevation Gain (m)