Statistical Modeling Course

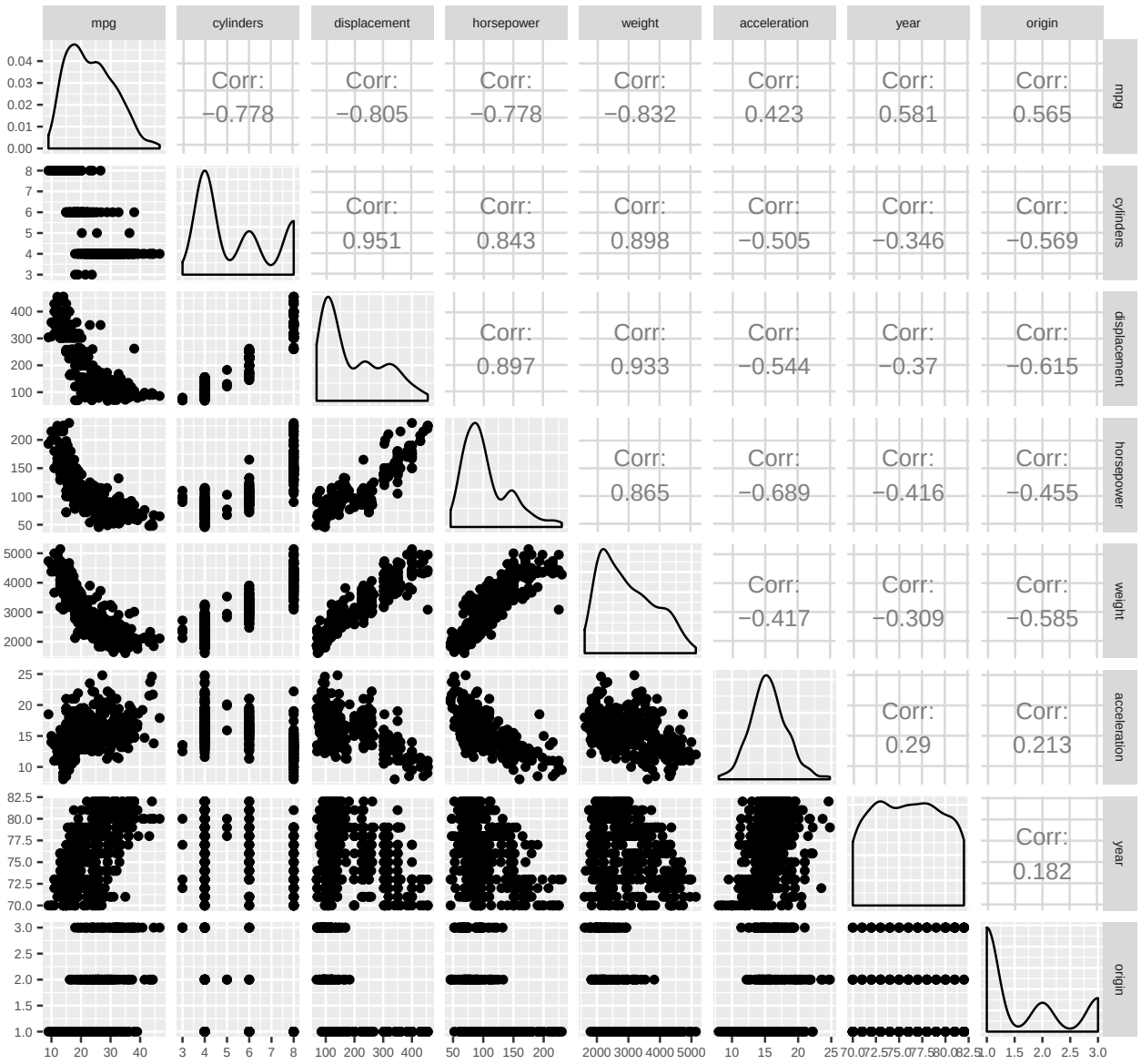
Multiple Linear Regression Assignment

The following problems involves the use of multiple linear regression on the `Auto` data set available in the `ISLR` package.

Problem 1

Use `GGpairs` in the `GGally` package to produce a scatterplot matrix which includes all of the variables in the data set and the pairwise corrlations. Set `progress = FALSE` so only the plot is printed. You will need to exclude the `name` variable, which is qualitative. Make sure to set the size to make it legible. You can change the font size by adding `theme(text=element_text(size=8))` to the plot.

```
Auto %>%
  select(-name) %>%
  ggpairs(data=., progress = FALSE) +
  theme(text=element_text(size=8))
```



Problem 2

Perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Print the results (including R^2 and p-values).

```
Auto_sub <- Auto %>% select(-name)
fit <- lm(mpg~., data=Auto_sub)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto_sub)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Problem 3

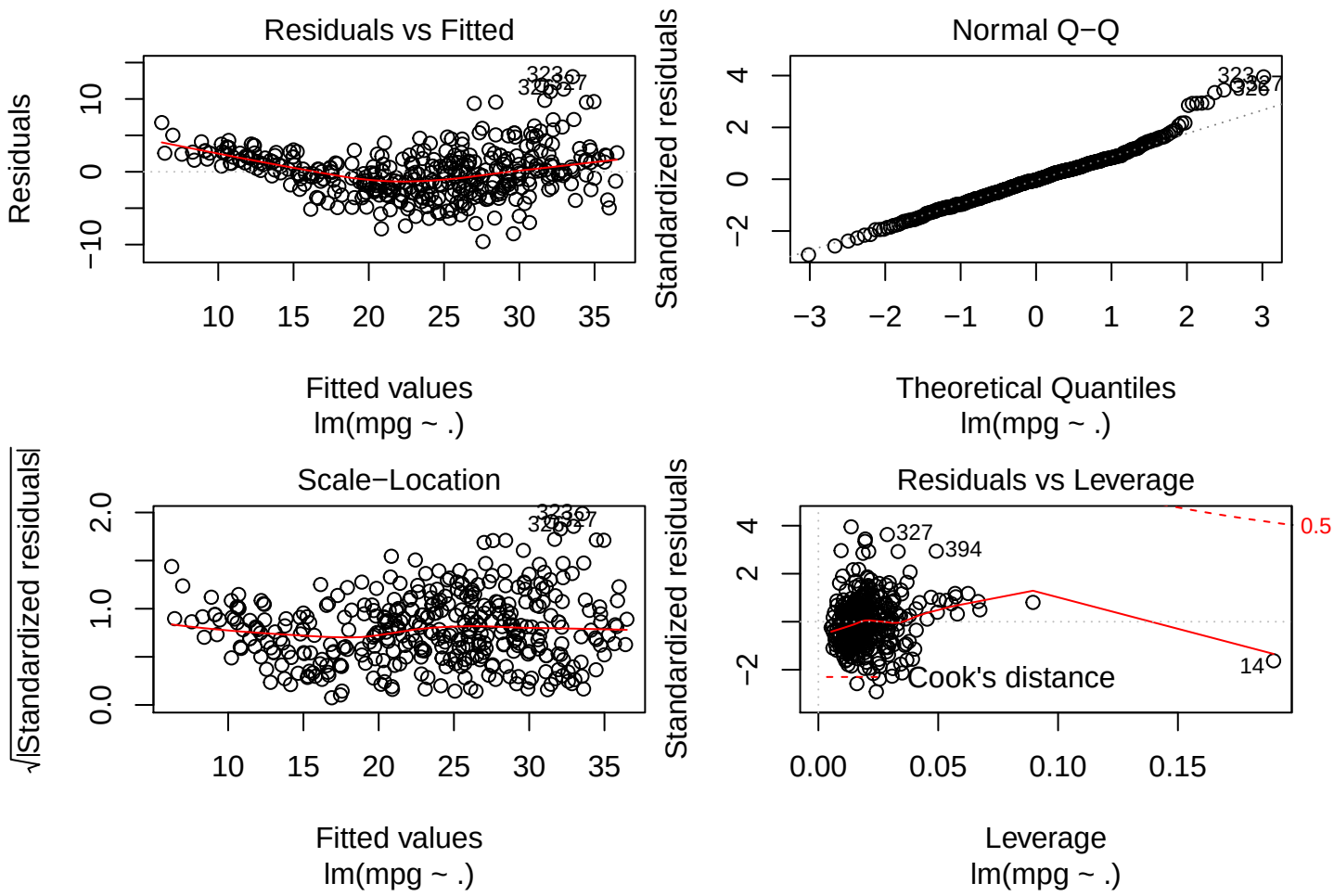
Comment on the output. Include the answers to the following questions: What fraction of the variance of `mpg` is explained by the model? Is there a relationship between the predictors and the response? Which predictors appear to have statistically significant relationship to the response? What does the coefficient for the `year` variable suggest?

Answer: The model explains 81.8% of the variance in mpg which indicates there is a relationship between the predictors and the response. Engine displacement, weight, year, and origin are all significantly correlated with mpg. The coefficient of year is 0.75 which suggest that in this dataset each one year increase corresponds to an increase in mpg of 0.75.

Problem 4

Use the `plot()` function to produce diagnostic plots of the linear regression fit. Make sure your plots are visible in your pdf. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```r
plot(fit)
```



Answer: There is a slight curve in the Residuals vs Fitted indicating there may be a non-linear effect we are not properly accounting for. The residuals appear approximately normally distributed except for the highest quantiles. The constant variance assumption appears to be met. There is one point with higher leverage but it is inside the Cook's distance intervals.