1	2	$\sum$

Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

Marina Dittschar & Clarissa Auckenthaler

Assignment 3

(Submitted 23.05.2022)

Task 1

Prove the bias-variance tradeoff with irreducible error:

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{1}$$

During the proof, we refer to $f(x_0)$ as f and $\hat{f}(x_0)$ as $\hat{f}$ for simplicity of depiction. Every data point in a data set D is defined by [1]:

$$y_0 = f + \epsilon \tag{2}$$

with underlying "true" function f and noise ϵ. We input this into the left side of (1) and multiply out [3]:

$$\begin{aligned} & E[(f + \epsilon - \hat{f})^2] && (3) \\ &= E[\epsilon^2] + E[(f - \hat{f})^2] + E[2\epsilon(f - \hat{f})] && (4) \\ &= E[\epsilon^2] + E[(f - \hat{f})^2] + 2 \cdot E[\epsilon] \cdot E[(f - \hat{f})] && (5) \\ &= E[\epsilon^2] + E[(f - \hat{f})^2] && (6) \end{aligned}$$

$E[\epsilon] = 0$ because it is by definition randomly distributed around zero. This means the last term becomes zero.

We subtract and add $E[\hat{f}]$ inside the second term of (5) [1]:

$$\begin{aligned} & E[\epsilon^2] + E[(f - E[\hat{f}] + E[\hat{f}] - \hat{f})^2] + E[\epsilon^2] && (7) \\ &= E[\epsilon^2] + E[(f - E[\hat{f}])^2] + E[E[\hat{f}] - \hat{f})^2] + E[2(f - E[\hat{f}])(E[\hat{f}] - \hat{f})] && (8) \end{aligned}$$

We look closer at the last term [3]:

$$\begin{aligned} & E[2(f - E[\hat{f}])(E[\hat{f}] - \hat{f})] && (9) \\ &= E[E[\hat{f}] - \hat{f}]] \cdot E[2(f - E[\hat{f}])] && (10) \\ &= E[E[\hat{f}] - \hat{f}]] \cdot 2(f - E[\hat{f}]) && (11) \\ &= 0 && (12) \end{aligned}$$

The step from (10) to (11) is possible because the function of the expected value is dependent on the data set D, but the true function f is not [3], and $E[\hat{f}]$ is just a constant. The step from (11) to (12) is true because

$$E[E[\hat{f}] - \hat{f}] = E[E[\hat{f}]] - E[\hat{f}] = E[\hat{f}] - E[\hat{f}] = 0 \tag{13}$$

From (8) we are now left with:

$$E[\epsilon^2] + E[(f - E[\hat{f}])^2] + E[(E[\hat{f}] - \hat{f})^2] = E[\epsilon^2] + (f - E[\hat{f}])^2 + E[(E[\hat{f}] - \hat{f})^2] \tag{14}$$

The second term was transformed as in the step (10)-(11). From this, the first term can directly be derived from the formula for $Var(\epsilon)$:

$$Var(\epsilon) = E[(E[\epsilon] - \epsilon)^2] \tag{15}$$

$$= E[(0 - \epsilon)^2] \tag{16}$$

$$= E[\epsilon^2] \tag{17}$$

The Bias of a function $\hat{f}$ can also be written as:

$$Bias(\hat{f}) = E[\hat{f}] - f \tag{18}$$

Consequently, we see that:

$$(f - E[\hat{f}])^2 = Bias(\hat{f})^2 \tag{19}$$

Finally, the third term corresponds to the definition of the variance of $\hat{f}$:

$$E[(E[\hat{f}] - \hat{f})^2] = Var(\hat{f}) \tag{20}$$

All together, we have shown that

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{21}$$

q.e.d.

Task 2

a) **Create scatterplots between all the variables. Is the relationship between those variables linear? Describe the connection between the variables. (Exclude the name variable, which is qualitative.)**

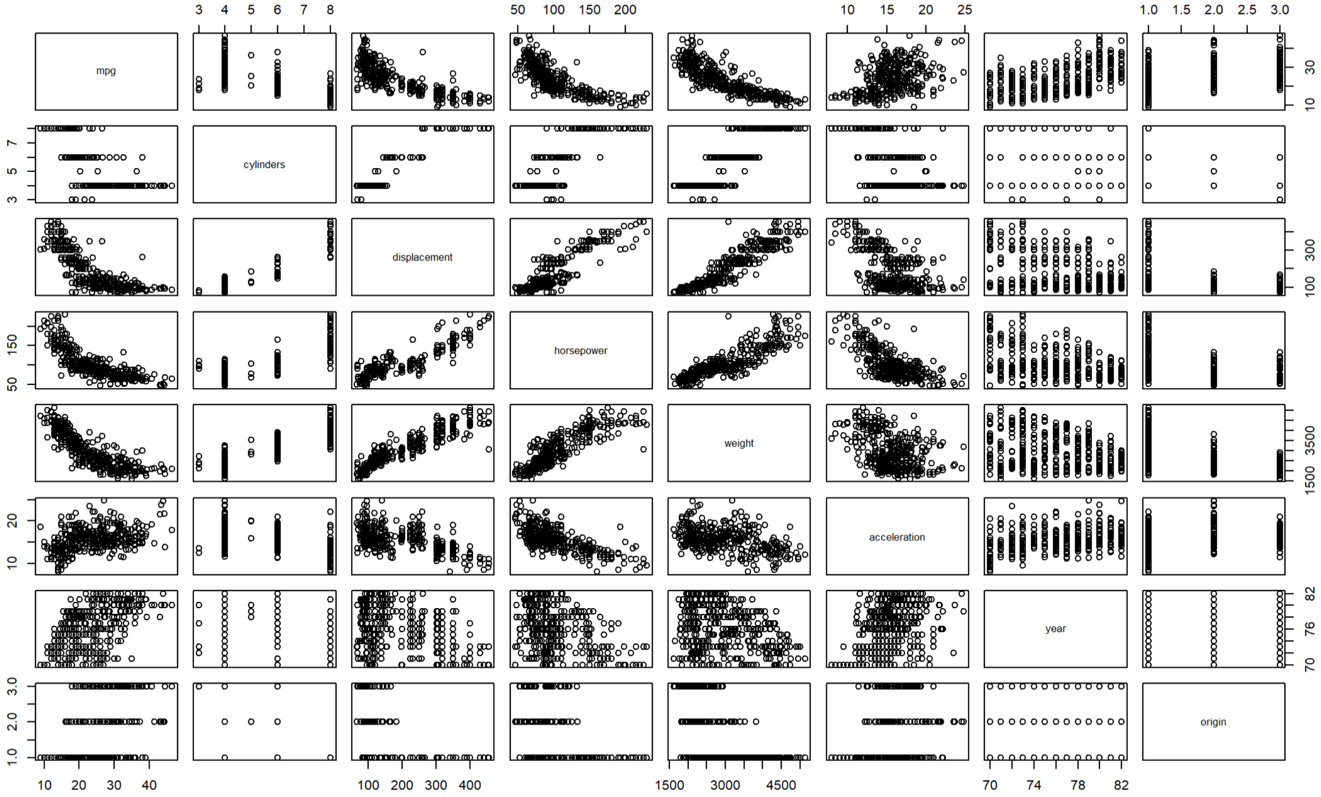


Figure 1: Scatterplot for every pair of variables in the dataset except *name*

As we see in Figure 1, the relationships between the different variables differ from each other. There are variables which seem to have a positive linear relationship, such as "horsepower" and "displacement", "horsepower" and "weight", "displacement" and "weight", "cylinders" and "displacement", "cylinders" and "horsepower" as well as "mpg" and "year". There also seem to be variables with a negative linear relationship such as "displacement" and "acceleration" or "horsepower" and "acceleration". For many other variables however, the nature of the relationships seem to be nonlinear, e.g. between "mpg" and "displacement", "mpg" and "horsepower", "mpg" and "weight" and "mpg" and "acceleration". For some cases, the samples differ to widely as that relationships could be assumed, e.g. this is the case for all variables' relationship with "year" except "mpg". In other cases, the possible value range is so narrow that it becomes hard to judge relationships between variables, especially by eye (see the relationships with "origin").

b) Detect the variables in the scatter plots that appear to be most highly correlated and anti-correlated, respectively. Justify your choice numerically.

The figure 2 displays the correlation matrix of all columns except name. The values which are close to 1 are highly correlated, the negative ones close to -1 are anti-correlated.

Highly correlated are:

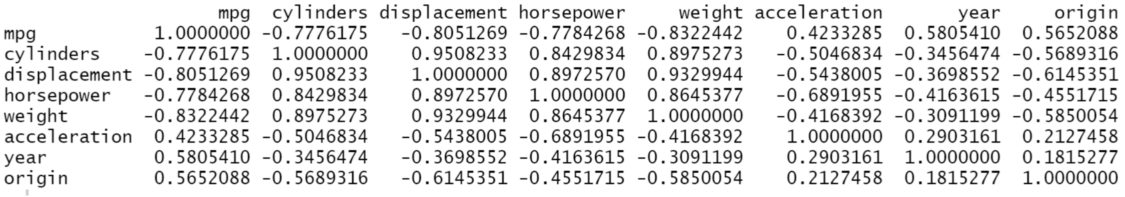
- cylinders - displacement = 0.951
- cylinders - horsepower = 0.843
- cylinders - weight = 0.898
- displacement - horsepower = 0.897
- displacement - weight = 0.933

Anti-correlated are:

- mpg - cylinders = -0.776
- mpg - displacement = -0.85
- mpg - horsepower = -0.778
- mpg - weight = -0.832

Not highly correlated or anti-correlated are:

- mpg - acceleration = 0.423
- cylinders - year = -0.346
- displacement - year = -0.370
- horsepower - year = -0.412
- weight - year = -0.309
- acceleration - year = 0.290
- acceleration - origin = 0.213
- year - origin = 0.182



```
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

Figure 2: Correlation Matrix of dataset Auto

We see that an especially high correlation exists between the variables "cylinders", "weight", "displacement" and "horsepower" and that the correlation between "mpg" and these four variables seem to be especially negative. Other relationships show more inconclusive values.

c) Perform simple linear regression with *mpg* as the response using the variables *cylinders, displacement, horsepower* and *year* as features. Which predictors appear to have a statistically significant relationship to the outcome and how good are the resulting models (measured using R^2)?

Residuals	Min	-14.2413	
	1Q	-3.1832	
	Median	-0.6332	
	3Q	2.5491	
	Max	17.9168	
Coefficients:		Intercept	cylinders
	Estimate	42.9155	-3.5581
	Std. Error	0.8349	0.1457
	t-value	51.40	-24.43
	Pr(>|t|)	$<2 * 10^{-16}$***	$<2.2 * 10^{-16}$***
Residual standard error		4.914 on 390 degrees of freedom	
Multiple R-squared		0.6047	
Adjusted R-squared		0.6037	
F-statistic		596.6 on 1 and 390 DF	
p-value		$<2.2 * 10^{-16}$	

Table 1: SLR- mpg and cylinders (*Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1*)

Residuals	Min	-12.9170	
	1Q	3.0243	
	Median	-0.5021	
	3Q	2.3512	
	Max	18.6128	
Coefficients:		Intercept	displacement
	Estimate	35.12064	-0.06005
	Std. Error	0.49442	0.00224
	t-value	71.03	-26.81
	Pr(>|t|)	$<2 * 10^{-16}$***	$<2 * 10^{-16}$***
Residual standard error		4.635 on 390 degrees of freedom	
Multiple R-squared		0.6482	
Adjusted R-squared		0.6473	
F-statistic		718.7 on 1 and 390 DF	
p-value		$<2.2 * 10^{-16}$	

Table 2: SLR- mpg and displacement (*Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1*)

Residuals	Min	-13.571	
	1Q	-3.25	
	Median	-0.3435	
	3Q	2.7630	
	Max	16.9240	
Coefficients:		Intercept	horsepower
	Estimate	39.935861	-0.157845
	Std. Error	0.717499	0.006446
	t-value	55.66	-24.49
	Pr(>|t|)	$<2 * 10^{-16}$***	$<2 * 10^{-16}$***
Residual standard error		4.906 on 390 degrees of freedom	
Multiple R-squared		0.6059	
Adjusted R-squared		0.6049	
F-statistic		599.7 on 1 and 390 DF	
p-value		$<2.2 * 10^{-16}$	

Table 3: SLR- mpg and horsepower (*Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1*)

Residuals	Min	-12.0212	
	1Q	-5.4411	
	Median	-0.4412	
	3Q	4.9739	
	Max	18.2088	
Coefficients:		Intercept	year
	Estimate	-70.01167	1.23004
	Std. Error	6.64516	0.08736
	t-value	-10.54	14.08
	Pr(>|t|)	$<2.2 * 10^{-16}$***	$<2 * 10^{-16}$***
Residual standard error		6.363 on 390 degrees of freedom	
Multiple R-squared		0.337	
Adjusted R-squared		0.3353	
F-statistic		198.3 on 1 and 390 DF	
p-value		$<2.2 * 10^{-16}$	

Table 4: SLR- mpg and year (*Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1*)

As depicted in Table 1, there is a negative relationship between cylinders and mpg, as can be seen in the coefficient estimate. The relationship seems to be highly significant, as the p-value is $< 2.2 \cdot 10^{-16}$. The determined intercept is also highly significant, with a value of 42.916 and a p-value of $< 2.2 \cdot 10^{-16}$. The adjusted R^2 is 0.605, which means that 60.5% of the variance in the data can be explained by the determined coefficient.

The linear coefficient between displacement and is also negative with a value of -0.06 as can be seen in Table 2. The coefficient might seem quite small in comparison to the coefficient of "cylinder", but this is because displacement has very large values, which means small changes have big impact on predicted values. To prevent this, the data could have been scaled beforehand. The relationship is highly significant, as the p-value is also $< 2.2 \cdot 10^{-16}$. With a value of 35.121 and a p-value of $< 2.2 \cdot 10^{-16}$, the intercept is also highly significant. The adjusted R^2 is 0.648, which means that 64.8% of the variance in the data can be explained by the determined coefficient.

The coefficient of horsepower and mpg is -0.157 (see Table 3, which also implies a negative relationship, which is highly significant with a p-value of $< 2.2 \cdot 10^{-16}$. The intercept has a value of 39.936 and is highly significant with a p-value $< 2.2 \cdot 10^{-16}$. The adjusted R^2 is 0.606, which means that 60.6% of the variance in the data can be explained by the determined coefficient. We can see that the R^2 is fairly similar for the first three coefficients.

The coefficient of "year" and "mpg" is 1.23. Thee seems to be a positive relationship, which is also highly significant because the p-value is $< 2.2 \cdot 10^{-16}$. The intercept, as with all models above, is highly significant with a p-value $< 2.2 \cdot 10^{-16}$ and a value of -70.012. The adjusted R^2 is 0.337, which means that 33.7% of the variance in the data can be explained by the determined coefficient. This is lower than for all other coefficients, which implies that the resulting model is not as good as the other models.

d) Perform a multiple linear regression with *mpg* as the response and all other variables except *name* as the predictors. Compare the full model to those generated in (c) in terms of their model fit.

Residuals	Min	-9.5903							
	1Q	-2.1565							
	Median	-0.1169							
	3Q	1.8690							
	Max	13.0604							
Coefficients:		Intercept	cylinders	displacement	horsepower	weight	acceleration	year	origin
	Estimate	-17.218435	-0.493376	0.019896	-0.016951	-0.006474	0.080576	0.750773	1.426141
	Std. Error	4.644294	0.323282	0.007515	0.013787	0.000652	0.098845	0.050973	0.278136
	t-value	-3.707	-1.526	2.647	-1.230	-9.929	0.815	14.729	5.127
	Pr(>|t|)	0.00024***	0.12780	0.00844***	0.21963	$<2 * 10^{-16}$***	0.41548	$<2 * 10^{-16}$***	$4.67 * 10^{-7}$
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		$2.2 * 10^{-16}$							

Table 5: Summary of the multiple regression model (Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1)

What can you observe in the different models concerning the significance of the relationship between response and individual predictors?

Taken by itself, each simple linear regression model examined above for had p-values which implied highly significant relationships between the predictors in question and the response. However, when taken all features into account simultaneously, this changes. There are still variables with highly significant coefficients (i.e. the intercept, as well as the coefficients for "weight", "year" and "origin"), but for the other coefficients, the p-values have increased, causing the coefficient to have less of a statistical significance (for "displacement"), or no statistical significance (for "cylinders" and "acceleration"). This implies that the features with non-significant coefficients are less suitable to predict "mpg" than the features with highly significant coefficients.

The p-value $= 2.2 * 10^{-16}$, this means that the probability of the null hypothesis being true is close to 0. $\longrightarrow$ there is a relationship between the features (predictors) and mpg.
For comparing the individual features, take the coefficient p-values (Pr(>|t|) into account. For p= 0.05 as the threshold for significance, all variable except cylinders, horsepower and acceleration have a statistically significant relationship with the response mpg (all values with the *'s in the table 5).

Does the sign of the coefficient tell you about the relationship between the predictor and the response?

The regression coefficients are displayed in table 5 in the column "Estimate". Those coefficients represent the difference in the predicted value of the response variable for each one-unit change in the predictor variable, assuming all other predictor variables are held constant.

As one can see in table 5, the regression coefficient of the variable "year" is 0.75077, which means that an increase of 1 unit in "year" is associated with an increase of 0.75077 in the response ("mpg").
In contrast, the feature "weight" has a negative coefficient of -0.006474, which means that the effect of an increase of 1 unit in "weight" is associated with the decrease of -0.006474 in the response ("mpg").
Thus, the sign of the coefficient determines the direction of the change in the response (increase or decrease).

e) **Investigate potential problems with the fitted model from the previous subtask:**

- Does the residual plot suggest any non-linearity in the data?

  In figure 3a the residual plot shows that there is a U-shape pattern in the residuals. This U-shape might indicate that the data is non-linear. In addition to this, the variance of the residuals is not constant. With increasing input values, the variance of the residuals increases. The residuals are not homoscedastic. This implies that the data is not adequately represented using a linear model.

- Are there any outliers?

  Data will be an outlier if the standardized residual is outside the range [-3,3], so three time the standard deviation in each direction [2]. The figure 3b the scale-location plot displays in our example values in a range of [0,2]. This means that because $\sqrt{3} \approx 1.73$, there are a few outliers with values above that threshold.

- Are there any high-leverage points?

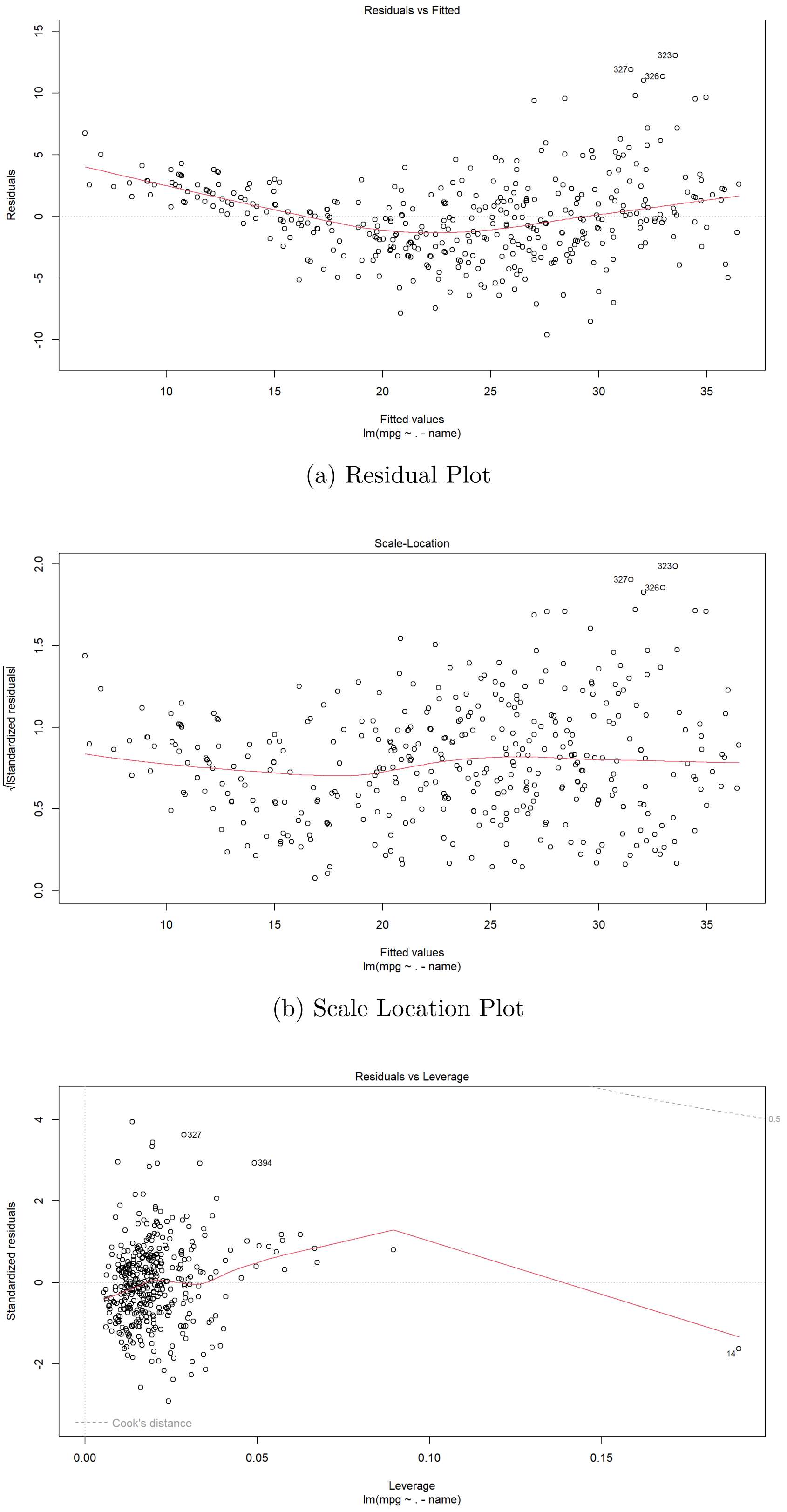
  Taking the formula from the slides for lecture 2: *Linear Regression*, a point with high leverage is defined as a "data point i with "extreme" x_i". High leverage is indicated if the following equation is satisfied:

  $$h_i >> (p+1)/n$$

  Inserting $p = 7$ and $n = 392$ into the formula:

  $$h_i >> 8/392 \approx 0.0204$$

  As h_i has to be a lot larger than the term on the right side, we interpret this as "twice as large" [4]. Since there are no points with a leverage this high, there seem to be no points with high leverage (see Figure 3c).



(a) Residual Plot

(b) Scale Location Plot

(c) Residuals vs Leverage

Figure 3: MLR-Model investigation

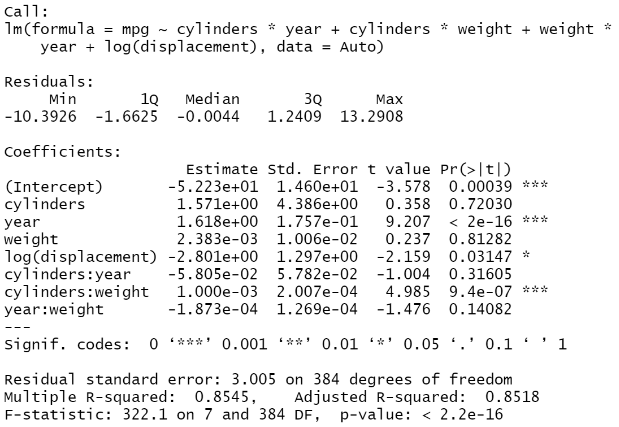
f) Generate three linear models that are all based on all pairwise interaction terms (X_1X_2) for *cylinders, weight,* and *year*, and differ by a non-linear transformations for the displacement variable. Use $log(X)$, $\sqrt{X}$, X^2, one for each of the three models. Perform inference on the three models and comment on your findings.

When we compare the performance of the models set up in Task d) with the models set up here, we can clearly see an improved fit to the data, as the previous adjusted R^2 was 0.82 and the new models each have an adjusted R^2 several points higher, at around 0.85. Interestingly, the R^2 is very similar across the models, with the three values ranging from 0.85 to 0.8518 (see Figure 4). Correspondingly, the Residual Standard Error also takes very similar values, ranging from 3.005 to 3.023.

Looking at the interaction terms, it becomes clear that only one of them, namely "cylinders" and "weight", are statistically significant for inferring "mpg", as it is highly significant in every model. The interaction terms of "cylinders" and "year" as well as "year" and "weight" are not statistically significant.

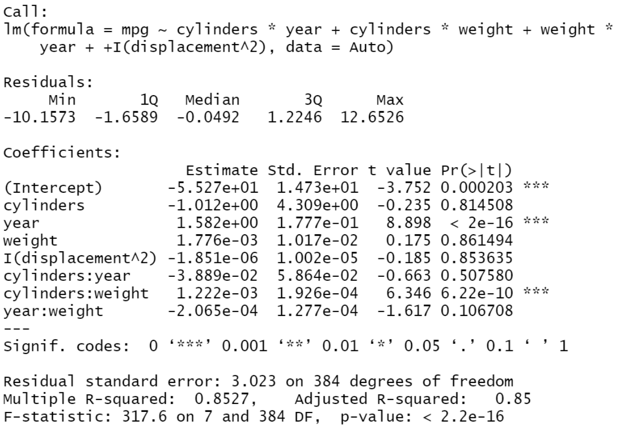
Apart from the interaction terms, the nature of the nonlinear term of displacement seemed to have little impact on the goodness of the model fit, as this is the only difference between the three models. The U-shape that was initially visible in the residuals is now flattened (see Figure 5), suggesting that the model is better fit to represent the data, but the "Residuals vs. Fitted" plots in Figure 5 are exceedingly similar across all models. This also goes for the other plots for each model.

To summarize, introducing pairwise interaction terms and a nonlinear transformation term improved the model fit, but the nature of the interaction term does not seem to influence the fit. In addition to this, of all interaction terms, only "cylinders"/"weight" seems to be statistically significant for inferring "mpg".



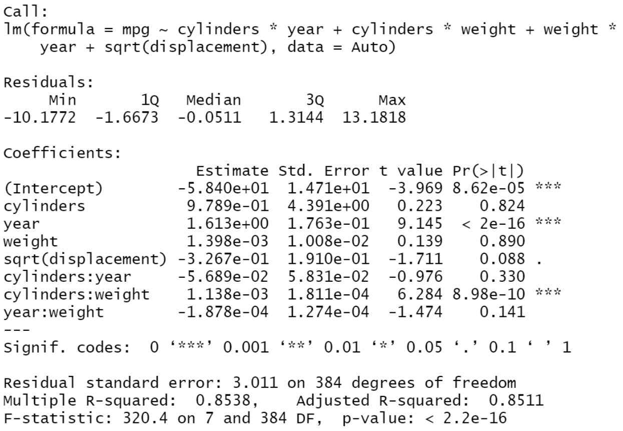
```
Call:
lm(formula = mpg ~ cylinders * year + cylinders * weight + weight *
    year + log(displacement), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3926  -1.6625  -0.0044   1.2409  13.2908

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.223e+01  1.460e+01  -3.578  0.00039 ***
cylinders          1.571e+00  4.386e+00   0.358  0.72030
year               1.618e+00  1.757e-01   9.207  < 2e-16 ***
weight             2.383e-03  1.006e-02   0.237  0.81282
log(displacement) -2.801e+00  1.297e+00  -2.159  0.03147 *
cylinders:year    -5.805e-02  5.782e-02  -1.004  0.31605
cylinders:weight   1.000e-03  2.007e-04   4.985  9.4e-07 ***
year:weight       -1.873e-04  1.269e-04  -1.476  0.14082
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.005 on 384 degrees of freedom
Multiple R-squared:  0.8545,    Adjusted R-squared:  0.8518
F-statistic: 322.1 on 7 and 384 DF,  p-value: < 2.2e-16
```

(a) Inference of model 1 (pairwise interaction of cylinders, weight and year) with nonlinear transformation log(displacement)



```
Call:
lm(formula = mpg ~ cylinders * year + cylinders * weight + weight *
    year + +I(displacement^2), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1573  -1.6589  -0.0492   1.2246  12.6526

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.527e+01  1.473e+01  -3.752 0.000203 ***
cylinders         -1.012e+00  4.309e+00  -0.235 0.814508
year               1.582e+00  1.777e-01   8.898  < 2e-16 ***
weight             1.776e-03  1.017e-02   0.175 0.861494
I(displacement^2) -1.851e-06  1.002e-05  -0.185 0.853635
cylinders:year    -3.889e-02  5.864e-02  -0.663 0.507580
cylinders:weight   1.222e-03  1.926e-04   6.346 6.22e-10 ***
year:weight       -2.065e-04  1.277e-04  -1.617 0.106708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.023 on 384 degrees of freedom
Multiple R-squared:  0.8527,    Adjusted R-squared:   0.85
F-statistic: 317.6 on 7 and 384 DF,  p-value: < 2.2e-16
```

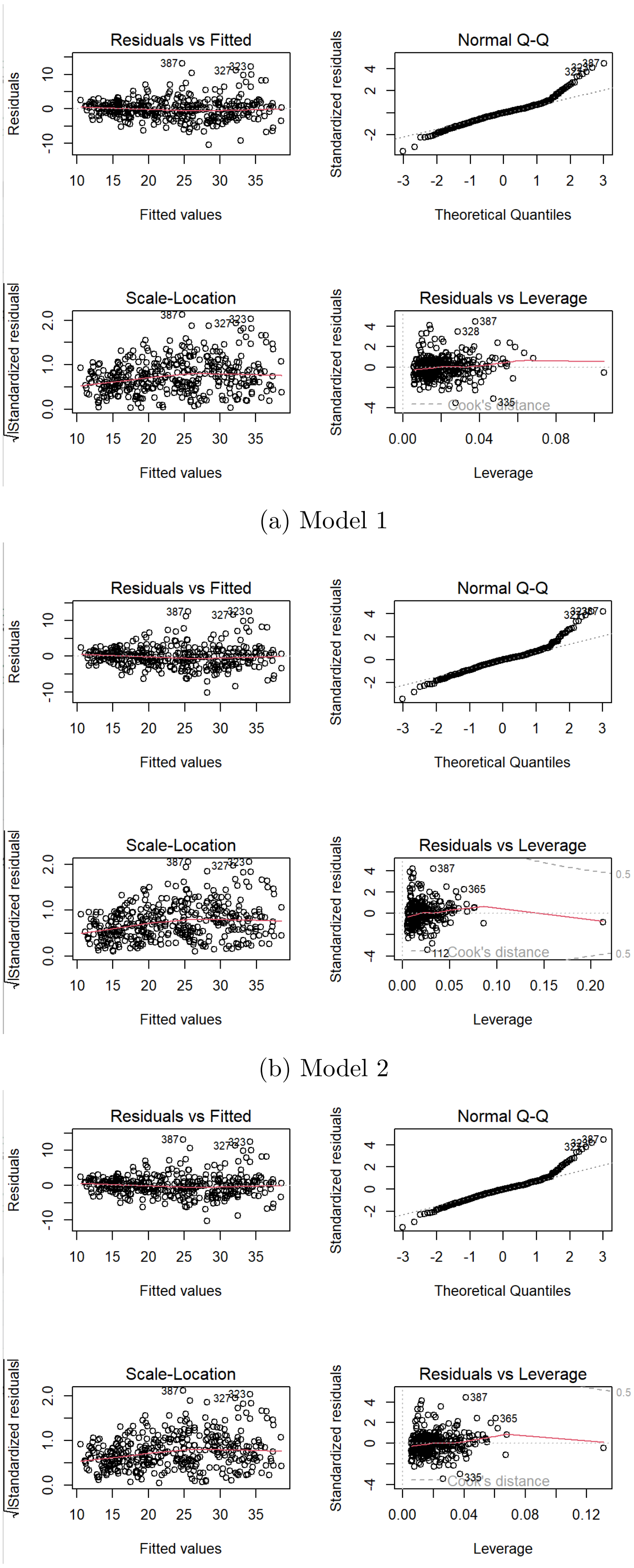
(b) Inference of model 2 (pairwise interaction of cylinders, weight and weight) with nonlinear transformation I(displacement^2)



```
Call:
lm(formula = mpg ~ cylinders * year + cylinders * weight + weight *
    year + sqrt(displacement), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1772  -1.6673  -0.0511   1.3144  13.1818

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.840e+01  1.471e+01  -3.969 8.62e-05 ***
cylinders           9.789e-01  4.391e+00   0.223    0.824
year                1.613e+00  1.763e-01   9.145  < 2e-16 ***
weight              1.398e-03  1.008e-02   0.139    0.890
sqrt(displacement) -3.267e-01  1.910e-01  -1.711    0.088 .
cylinders:year     -5.689e-02  5.831e-02  -0.976    0.330
cylinders:weight    1.138e-03  1.811e-04   6.284 8.98e-10 ***
year:weight        -1.878e-04  1.274e-04  -1.474    0.141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.011 on 384 degrees of freedom
Multiple R-squared:  0.8538,    Adjusted R-squared:  0.8511
F-statistic: 320.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

(c) Inference of model 3 (pairwise interaction of cylinders, weight and year) with nonlinear transformation sqrt(displacement)

Figure 4: Models with pairwise interaction terms with nonlinear transformation



(a) Model 1

(b) Model 2

(c) Model 3

Figure 5: Residual Plots of the three different models

References

[1] Greg Shakhnarovich. https://web.archive.org/web/20140821063842/http://ttic.uchicago.edu/~gregory/courses/wis-ml2012/lectures/biasVarDecom.pdf, 2011.

[2] The Pennsylvania State University. 9.3 - identifying outliers (unusual y values). https://online.stat.psu.edu/stat462/node/172/#:~:text=The%20good%20thing%20about%20standardized,some%20to%20be%20an%20outlier., 2018.

[3] Wikipedia. Bias-variance tradeoff. https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff, 2022.

[4] Wikipedia. Leverage (statistics). https://en.wikipedia.org/wiki/Leverage_(statistics), 2022.