Homework 6

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have practice 1) multiple linear regression; 2) quality of fit measures (R^2, RSE); 3) collinearity; 4) variable selection; 4) R coding.

Problem #1

This question involves the use of multiple linear regression on the Auto data set of ISLR library

Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables (except name) as the predictors. Use the summary() function to print the results.

```
lm.obj <- lm(mpg ~ . - name, data = ISLR::Auto)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = ISLR::Auto)
##
## 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
```

1) Formulate the H0 and Ha hypotheses (using parameter notation) for testing whether the overall model is significant. Which part of summary() output corresponds to this test? Is the model significant?
$H_0 : \beta_1 = ... = \beta_7 = 0$

$H_a : \{at\ least\ one\ \beta_j \neq 0,\ j = 1, ..., 7\}$
The f-statistic section at the bottom corresponds to model significance. The model is statistically significant because the p-value (≈ 0) is less than most common significance levels ($\alpha = 0.001, 0.01, 0.05$).

2) Which predictors appear to have a statistically significant relationship to the response? Just list them.
weight, year, and origin appear to have a statistically significant relationship to mpg at the $alpha = 0.001$ level. In addition to the previous predictors, displacement appears to additionally have a statistically significant relationship to mpg at the $\alpha = 0.01$ level.

3) Interpret the effect of car's weight on its miles per gallon.
There is statistically significant linear relationship between weight (a predictor) and mpg (the response value) when holding all other predictors constant because the p-value (≈ 0) is less than all common significance levels ($\alpha = 0.001, 0.01, 0.05$). For 1 pound increase in weight, holding all other predictors constant, mpg will decrease by 0.006474 miles per gallon, on average

4) For the effect from part 3, proceed to report and interpret the 95% confidence interval.

```
confint(lm.obj)["weight",]
```

```
##        2.5 %       97.5 %
## -0.007756074 -0.005192013
```

We are 95% confident that for 1 pound increase in weight, holding all other predictors constant, the average mpg will decrease by between 0.005192 and 0.0077561 miles per gallon.

5) Report and interpret both quality-of-fit metrics.

```
summary(lm.obj)$r.squared # R-Squared
summary(lm.obj)$sigma # RSE
```

- R^2: Our linear regression model explains 82.1478076 percent of variety in mpg.
- RSE: Our predicted mpg misses the true population mpg by 3.3276824 miles per gallon, on average

Problem #2

This problem will deal with cystfibr data example of ISwR package. In particular, we will be building a model to predict pemax (patient's maximum respiratory pressure) based on other physical characteristics.

1) Proceed to fit the following multiple linear regression model: $pemax \sim\ .$

```
lm(pemax ~ ., data = ISwR::cystfibr) %>%
  lm.summary()
```

a. Comment on the 1) overall model significance; 2) significance of any individual predictors. Why do you think this is happening (name the main issue)?
The overall linear model is significant at the $\alpha = 0.05$ level, indicating at least one important variable, despite no individual predictor being significant. This paradox is due to the issue of collinearity - some predictors are strongly correlated with each other and are thus fighting for effect on the model.

b. Proceed to address the issue observed in part (a) via studying a correlation matrix of predictors, modifying the model accordingly. Fit the modified model, comment on its 1) overall model significance; 2) significance of any individual predictors.

Table 1: pemax ∼ .

	Coefficients						*Overall Model*	
	estimate	2.5 %	97.5 %	std.error	statistic	p.value		value
(Intercept)	176.058	-305.417	657.534	225.891	0.779	0.44787	RSS	9731.25
age	-2.542	-12.777	7.693	4.802	-0.529	0.60428	RSE	25.471
sex	-3.737	-36.689	29.215	15.460	-0.242	0.81228	R^2	0.637
height	-0.446	-2.372	1.479	0.903	-0.494	0.62846	p.value	3.195e-02
weight	2.993	-1.287	7.273	2.008	1.490	0.15683		
bmp	-1.745	-4.207	0.717	1.155	-1.510	0.1517		
fev1	1.081	-1.223	3.385	1.081	1.000	0.33328		
rv	0.197	-0.221	0.615	0.196	1.004	0.33136		
frc	-0.308	-1.358	0.741	0.492	-0.626	0.54047		
tlc	0.189	-0.877	1.254	0.500	0.377	0.71116		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4: pemax ∼ . - age - height - rv

	Coefficients						*Overall Model*	
	estimate	2.5 %	97.5 %	std.error	statistic	p.value		value
(Intercept)	57.500	-81.404	196.404	66.116	0.870	0.39592	RSS	10542.993
sex	4.831	-21.549	31.211	12.556	0.385	0.70494	RSE	24.202
weight	1.804	0.857	2.752	0.451	4.000	0.00084 ***	R^2	0.607
bmp	-1.639	-2.997	-0.280	0.647	-2.534	0.02076 *	p.value	5.145e-03
fev1	1.697	-0.021	3.414	0.818	2.075	0.05257 .		
frc	0.161	-0.308	0.630	0.223	0.722	0.47946		
tlc	0.215	-0.665	1.095	0.419	0.514	0.61375		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
cor.summary(ISwR::cystfibr)
```

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
age	NA	0	0.93	0.91	0	0	0.00	0.00	0	0
height	0.93	0	NA	0.92	0	0	0.00	0.00	0	0
weight	0.91	0	0.92	NA	0	0	0.00	0.00	0	0
rv	0.00	0	0.00	0.00	0	0	NA	0.91	0	0
frc	0.00	0	0.00	0.00	0	0	0.91	NA	0	0

```
cor(ISwR::cystfibr)["pemax", c(1, 3, 4, 7, 8)] %>% data.frame(pemax = .) %>% t() %>%
  kableExtra::kable(booktabs = TRUE)
```

	age	height	weight	rv	frc
pemax	0.6134741	0.5992195	0.635222	-0.3155501	-0.4172078

age, weight, and height are are strongly correlated with each other. Similarly, rv and frc are strongly correlated with each other. One one out of each strongly correlated predictor group needs to be retained; I chose the one with the strongest correlation to pemax, the response variable.

```
lm(pemax ~ . - age - height - rv, data = ISwR::cystfibr) %>%
  lm.summary()
```


The overall linear model is significant at the $\alpha = 0.01$ level; bmp is significant at the $\alpha = 0.05$ level; weight is significant at the $\alpha = 0.001$ level.

Table 5: pemax ∼ . - weight - frc - height

	Coefficients						*Overall Model*	
	estimate	2.5 %	97.5 %	std.error	statistic	p.value		value
(Intercept)	-83.525	-257.499	90.448	82.808	-1.009	0.3265	RSS	11516.239
age	5.038	2.316	7.760	1.296	3.889	0.00108 **	RSE	25.294
sex	4.991	-22.138	32.120	12.913	0.386	0.70367	R^2	0.571
bmp	-0.403	-1.588	0.782	0.564	-0.715	0.48397	p.value	1.028e-02
fev1	1.931	0.293	3.570	0.780	2.476	0.02345 *		
rv	0.114	-0.098	0.325	0.101	1.127	0.2745		
tlc	0.465	-0.385	1.315	0.405	1.149	0.26543		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
car::vif(lm(pemax ~ . - weight, data = ISwR::cystfibr)) # frc highest
```

```
##       age       sex    height       bmp      fev1        rv       frc       tlc
##  8.097571  2.029182  7.595539  2.730462  4.205260 10.332505 15.814231  2.177076
```

```
car::vif(lm(pemax ~ . - weight - frc, data = ISwR::cystfibr)) # height highest
```

```
##      age      sex   height      bmp     fev1       rv      tlc
## 7.341695 1.606561 7.595520 1.794168 2.870202 2.836471 1.768577
```

```
car::vif(lm(pemax ~ . - weight - frc - height, data = ISwR::cystfibr))
```

```
##      age      sex      bmp     fev1       rv      tlc
## 1.611582 1.605444 1.718765 2.861478 2.814628 1.768466
```

```
lm(pemax ~ . - weight - frc - height, data = ISwR::cystfibr) %>%
  lm.summary()
```

The overall linear model is significant at the $\alpha = 0.05$ level; fev1 is significant at the $\alpha = 0.05$ level; age is significant at the $\alpha = 0.01$ level.

d. In you own words, why does collinearity prevent us from accurately estimating effects of collinear predictors on the response variable?
Collinearity prevents us from accurately estimating effects of collinear predictors on the response variable because it is tough to change one predictor while holding a collinear predictor constant since they are correlated. Even if there are observations which meet said criteria, the sample is small and therefore contains more uncertainty.

2) Proceed to fit the following two models:

- Full model: $pemax \sim \ sex + weight + height + rv + frc$
- Reduced model: $pemax \sim \ sex + height + rv$

Table 6: Standard Error Comparison

	height	rv
Full Model	0.677	0.165
Reduced Model	0.320	0.082

```
models <- list(lm(pemax ~ sex + weight + height + rv + frc, data = ISwR::cystfibr), lm(pemax ~ sex + hei
names(models) <- c("Full Model", "Reduced Model")
modelsummary::modelsummary(models, output = "latex", estimate = "std.error", statistic = NULL, coef_map
```

a. Comment on what happens to standard errors for $\hat{\beta_{height}}$ and $\hat{\beta_{frc}}$ coefficients when going from the full to reduced model. Why does this happen?
The standard errors shrank when going from the full model to the reduced model. Collinear predictors fight for effect on the response and therefore have more uncertainty which is reflected in increased standard error of respective coefficient estimates (i.e. standard errors get bloated). By removing a collinear predictor, the standard error of the remaining predictor is improved (i.e. shrunk).

b. Proceed to use VIF criteria in order to get from full model down to the reduced model. Which variable is dropped first? Second? Why?

```
car::vif(lm(pemax ~ sex + weight + height + rv + frc, data = ISwR::cystfibr))
```

```
##      sex   weight   height       rv      frc
## 1.113463 7.734391 7.597122 7.193580 7.144279
```

Weight is dropped first because it has the highest vif value.

```
car::vif(lm(pemax ~ sex + height + rv + frc, data = ISwR::cystfibr))
```

```
##      sex   height       rv      frc
## 1.113227 1.646786 6.268064 6.682708
```

Frc should be dropped second because it has the highest vif value after removing weight.

```
car::vif(lm(pemax ~ sex + height + rv, data = ISwR::cystfibr))
```

```
##      sex   height       rv
## 1.079776 1.480463 1.553273
```

No more predictors should be dropped because all the VIF values are below 5.

Problem #3

For Auto data set from ISLR library. Proceed to conduct variable selection via backward AIC approach:

```
lm.obj <- lm(mpg ~ . - name, data = ISLR::Auto)
step(lm.obj)
```

```
## Start:  AIC=950.5
## mpg ~ (cylinders + displacement + horsepower + weight + acceleration +
##     year + origin + name) - name
##
##                Df Sum of Sq    RSS     AIC
## - acceleration  1      7.36 4259.6  949.18
## - horsepower    1     16.74 4269.0  950.04
## <none>                      4252.2  950.50
## - cylinders     1     25.79 4278.0  950.87
## - displacement  1     77.61 4329.8  955.59
## - origin        1    291.13 4543.3  974.46
## - weight        1   1091.63 5343.8 1038.08
## - year          1   2402.25 6654.5 1124.06
##
## Step:  AIC=949.18
## mpg ~ cylinders + displacement + horsepower + weight + year +
##     origin
##
##                Df Sum of Sq    RSS     AIC
## <none>                      4259.6  949.18
## - cylinders     1     27.27 4286.8  949.68
## - horsepower    1     53.80 4313.4  952.10
## - displacement  1     73.57 4333.1  953.89
## - origin        1    292.02 4551.6  973.17
## - weight        1   1310.43 5570.0 1052.32
## - year          1   2396.17 6655.7 1122.13
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year + origin, data = ISLR::Auto)
##
## Coefficients:
##  (Intercept)     cylinders  displacement    horsepower        weight
##   -15.563492     -0.506685      0.019269     -0.023895     -0.006218
##         year        origin
##     0.747516      1.428242
```

1) Which R function allows us to do that? Which variable(s) ended up being dropped from the model? step with the default parameter values allows for the execution of variable selection via backward AIC approach. Acceleration was the only variable which was dropped.

2) Explain what is meant by "Df", "Sum of Sq", "RSS" and "AIC" in the tables outputted by step() function.

- df: by dropping a predictor (a parameter to be estimated), we gain a "degree of freedom"
- Sum of S: Increase in RSS by going from the full model to the reduced model
- RSS: RSS for the reduced model

3) Explain why the algorithm stopped (!) on that particular subset of variables. Hint: What does "" represent? Why is it of interest?)
$< none >$ corresponds to the current full-model since no variables are added or dropped. The resulting table from performing one step is sorted in ascending order according to the AIC value for the model listed on the left. The most optimized model is therefore always the first one listed. The algorithm stopped on that particular subset of variables because $< none >$ (the full-model) was on top (i.e. the full-model had the lowest AIC).

Problem #4 (BONUS)

This question involves the use of linear regression on the Advertising data set. Proceed to calculate RSS and R^2 for the following three models:

- $Sales \sim\ TV$
- $Sales \sim\ TV + radio$
- $Sales \sim\ TV + radio + newspaper$

```
df <- readr::read_csv(sprintf("https://docs.google.com/uc?id=%s&export=download", "1UJIu7Ku3rRWTnFJpK4u

models <- list(lm(sales ~ TV, data = df),
     lm(sales ~ TV + radio, data = df),
     lm(sales ~ TV + radio + newspaper, data = df))
names(models) <- sapply(models, function(x) as.character(x$call[2]) )

sapply(models, function(x) with(summary(x), c(RSS = sum(residuals^2), "$R^2$" = r.squared))) %>% as.data
```

Table 7: Model Comparison

	RSS	R^2
sales ~ TV	2102.531	0.612
sales ~ TV + radio	556.914	0.897
sales ~ TV + radio + newspaper	556.825	0.897

1) Did the RSS decrease (or at least didn't increase) every time you added an extra variable? Why do you think that is?
Dropping a variable and its coefficient β_j always increases the RSS. RSS metrics represents the *amount of variability in Y left* ***UNEXPLAINED*** *by the model*. More predictors means there are more parameters to base the prediction off of, and therefore less uncertainty. For predictors which have no effect on the response value ($\beta_j = 0$), removal would not change the RSS as the predictor explains no variability in the first place.

2) Did the R^2 increase (or at least didn't decrease) every time you added an extra variable? Why do you think that is? Hint: use the definition of R^2 + part (a).
R^2 represents the *% of variability in Y that's* ***EXPLAINED*** *by the regression model*. It is calculated by $R^2 = \frac{TSS-RSS}{TSS}$. TSS is always the same regardless of which predictors are included/excluded from a model for a particular response variable because it ignores the predictors and only looking at the response values and the mean of that response variable; TSS it is static. $R^2 = \frac{TSS-RSS}{TSS} = \frac{constant-RSS}{constant}$. Because TSS can be considered a constant, R^2 will increase if RSS decreases and visa-versa. As explained earlier, RSS will decrease when variables are added. Therefore, R^2 will increase (or at least not decrease) when variables are added.

higher end value