Assignment 7

Jenni Simon
09-116-005

Exercise 1

a) I generated the following three models with rising complexity:

1. $mpg \sim weight$
2. $mpg \sim poly(weight, 2)$
3. $mpg \sim poly(weight + cylinders + horsepower + year + displacement, 3)$

b) Performing 10-fold cross-validation I obtained the following cross-validation estimate of prediction error:

Tab. 1: Cross-validation estimate of prediction error.

Model	**Error-estimate**
1	18.9
2	17.6
3	17.1

c) To compare the three models we can (for example) perform pairwise t-tests between all the models. Results of the t-test are shown in Table 2. We can observe that model 3 performs significantly better than the other two models (small p-values).

Tab. 2: Results of the t-tests comparing the three linear models.

Models	t	p-value	95% confidence interval	Mean difference
1 vs 2	4	0.004	[0.529, 2.122]	1.33
1 vs 3	4	0.002	[0.824, 2.678]	1.75
2 vs 3	3	0.02	[0.0886, 0.7627]	0.426

Exercise 2 To estimate the error-rate of the model, I held out 20% of the data in a test set and trained the model on the remaining 80%. I trained two models, one containing all the available features (including all the SE and Max values...) and another using only the original values (i.e. "Radius", "Texture", "Perimeter", "Area", "Smooth", "Compact", "Concavity", "Concave", "Symmetry", "Fractal").

Listing 1: Summary of the logistic regression model.

```
Call:
glm(formula = Diagnostic ~ ., family = binomial, data = train,
    control = glm.control(maxit = 100))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9385  -0.1236  -0.0274   0.0023   2.9526

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.0560    15.5291    0.00   0.9971
Radius         0.4154     4.0561    0.10   0.9184
Texture        0.3895     0.0784    4.97  6.8e-07 ***
Perimeter     -0.5706     0.5470   -1.04   0.2969
Area           0.0476     0.0213    2.24   0.0253 *
Smooth        51.9680    33.5203    1.55   0.1211
Compact       22.1566    22.1140    1.00   0.3164
Concavity     -0.6457     9.4057   -0.07   0.9453
Concave      116.2626    35.7437    3.25   0.0011 **
Symmetry      19.6644    11.8512    1.66   0.0971 .
Fractal     -122.3228    98.0166   -1.25   0.2120
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 608.16  on 454  degrees of freedom
Residual deviance: 108.19  on 444  degrees of freedom
AIC: 130.2

Number of Fisher Scoring iterations: 9
```

The results of the smaller model are shown in Listing 1. We see that in this model the three predictors "Texture", "Concave" and "Area" are most significant. The positive coefficient for these three predictors suggests that with high values for "Texture", "Concave" or "Area" and all other values being equal, a diagnosis of "Malignant" is more likely.

Results for the big model are shown in Listing 2. We can observe that in this case the model was able to perfectly or quasi-perfectly separate the two classes on the training set (given by the very small residual deviance).

The error-rate estimated on the test-set was **8.77%** for the smaller model and **5.26%** for the bigger model.

Listing 2: Summary of the logistic regression model.

```
Call:
glm(formula = Diagnostic ~ ., family = binomial, data = train,
    control = glm.control(maxit = 100))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-8.21e-06  -2.10e-08  -2.10e-08   2.10e-08   1.11e-05

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.38e+02   3.12e+07       0        1
Radius       -1.14e+03   9.73e+06       0        1
Texture       2.48e+01   4.81e+05       0        1
Perimeter     1.91e+01   5.03e+05       0        1
Area          9.39e+00   8.34e+04       0        1
Smooth        2.50e+03   6.78e+07       0        1
Compact      -8.10e+03   8.25e+07       0        1
Concavity     4.80e+02   5.15e+07       0        1
Concave       1.55e+04   5.07e+07       0        1
Symmetry     -3.78e+03   1.65e+07       0        1
Fractal       8.55e+03   2.34e+08       0        1
RadiusSE      6.81e+02   2.08e+07       0        1
TextureSE    -6.51e+01   1.27e+06       0        1
PerimeterSE  -2.72e+02   2.54e+06       0        1
AreaSE        2.48e+01   2.00e+05       0        1
SmoothSE     -2.16e+04   9.71e+07       0        1
CompactSE     2.49e+04   2.54e+08       0        1
ConcavitySE  -1.56e+04   1.16e+08       0        1
ConcaveSE     7.16e+04   3.16e+08       0        1
SymmetrySE   -2.49e+04   2.56e+08       0        1
FractalSE    -1.81e+05   6.14e+08       0        1
RadiusMax     2.47e+02   6.70e+06       0        1
TextureMax    1.39e+01   2.78e+05       0        1
PerimeterMax  2.19e+01   2.20e+05       0        1
AreaMax      -2.06e+00   5.68e+04       0        1
SmoothMax     9.32e+02   3.94e+07       0        1
CompactMax   -3.25e+03   4.60e+07       0        1
ConcavityMax  2.47e+03   3.13e+07       0        1
ConcaveMax   -1.20e+03   2.20e+07       0        1
SymmetryMax   4.38e+03   3.09e+07       0        1
FractalMax    1.88e+04   5.75e+07       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.0341e+02  on 454  degrees of freedom
Residual deviance: 8.5555e-10  on 424  degrees of freedom
AIC: 62

Number of Fisher Scoring iterations: 34
```