LINEAR MODELS PROJECT REPORT

Group 20

Gregory van Kruijsdijk-R0622004

Sanne Kerckhoffs-R0667062

Adhithya Unni Narayanan-R0776057

Cristián Andrés Frigolett Catalán-R0826230

Natalia Paz Bravo Gómez-R0837039*

Ine Aertgeerts-R0589446**

*Natalia Paz Bravo Gómez: Did not contribute to the project.

**Ine Aertgeerts: No participation, due to unenrollment of the course.

Description:

The analysis is conducted on the data collected from patients that would undergo a prostatectomy (removal of a part of the prostate gland) in order to investigate the association between Prostate Specific Antigen(psa) and a set of prognostic variables which are given below. The report contains details of the analysis such as the exploratory analysis of the dataset, model building, model comparison and interpretation of the chosen model.

The model is built via a training dataset and validated via a validation dataset, in order to split the data into these two datasets a set.seed(0142520) statement is used. The data contains six variables, see Table 1 for a description of each variable. The dependent variable of the model is Prostate Specific Antigen (psa), it will be predicted by a combination of the five prognostic variables (i.e. age, volume, caps, gleason and invasion).

Table 1. Overview of variables and descriptions in dataset.

Variable	Description
Psa	Serum Prostate Specific Antigen (ng/ml)
Age	Age of the male patient
Volume	Volume prostate cancer volume (cc)
Caps	The amount of penetration of the cancer cell through the prostate's capsule (cm)
Gleason	Gleason score, which indicates the severity of the disease (1, 2, or 3; larger = more severe)
Invasion	Indicator whether the cancer has spread into the seminal vesicles (1 = yes, 0 = no)

Exploratory Analysis:

To begin with, for each variable a histogram and a boxplot is plotted (Figure 1). It can be seen in the boxplots that there are no outliers in the data. The variable psa is skewed to the right, with a modus at 5. The variable age seems uniformly distributed. The variable volume is normally

distributed, with a slight skew to the left. The variable caps is skewed to the right, with a modus at 1. The categorical variable gleason has three levels (1,2 and 3). Gleason level 2 has the highest frequency, followed by level 3 and then level 1. The boxplot shows that gleason level 1 has the smallest variance and mean on the variable psa. On the other hand gleason level 3 has the biggest variance and gleason level 2 has the highest mean. The categorical variable invasion has two levels (0 and 1). Invasion score 1 has the highest frequency. The boxplot shows that invasion level 1 has the highest variance and mean on psa compared to invasion level 0.

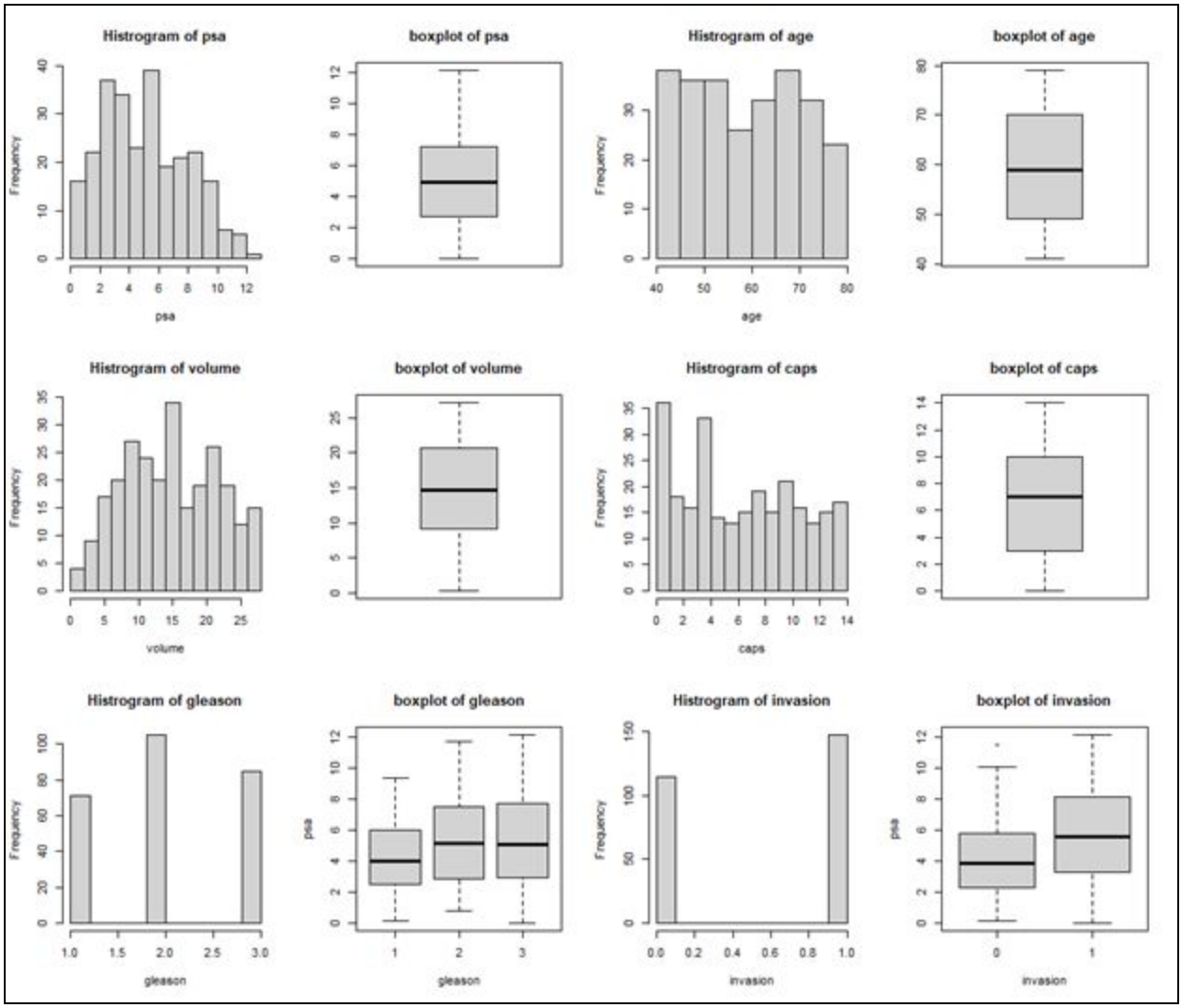

Figure 1. Histogram and boxplot for every variable in the full dataset.

The correlation matrix of all variables in the data can be found in table 2. All variables, except age, have a moderate to high correlation with psa. All other correlations are low. It is surprising that there is no correlation found between psa and age, because literature claims that there often is a positive correlation between psa and age (Littlejohns, Travis, Key, & Allen, 2016). Also

surprising is the high negative correlation between psa and caps, literature often finds a positive correlation between psa and caps (Horninger, Reissigl, Rogatsch, Volgger, Studen, Klocker, & Bartsch, 2000).

Table 2. Correlation between all the variables.

	psa	age	invasion	caps	volume	gleason
psa	1.00000000					
age	-0.04525817	1.00000000				
invasion	0.25539342	-0.06107316	1.000000000			
caps	-0.48459172	-0.02484152	0.098027507	1.00000000		
volume	0.65655935	-0.09027599	-0.001872019	0.06583616	1.000000000	
gleason	0.15075188	-0.04424759	-0.048928778	0.02803899	-0.101133929	1.00000000

Base Model:

For exploratory purposes, we will start by constructing a baseline model, which is a model with no higher order terms, and is constructed using ordinary least squares regression.

Before conducting the regression analysis, we scaled the continuous variables. This was done since the explanatory variables had different scales. We then build our linear regression model. The model here will be:

$$psa = \beta_0 + \beta_1 * age + \beta_2 * invasion + \beta_3 * caps + \beta_4 * volume + \beta_5 * gleason.$$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.183633   0.153793  -7.696 3.79e-12 ***
age               0.004289   0.002322   1.847   0.0672 .
volume            0.731569   0.027283  26.814  < 2e-16 ***
caps             -0.603016   0.027524 -21.909  < 2e-16 ***
factor(gleason)2  0.793934   0.068435  11.601  < 2e-16 ***
factor(gleason)3  0.668494   0.069087   9.676  < 2e-16 ***
factor(invasion)1 0.713408   0.054468  13.098  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3031 on 124 degrees of freedom
Multiple R-squared:  0.9124,   Adjusted R-squared:  0.9081
F-statistic: 215.1 on 6 and 124 DF,  p-value: < 2.2e-16
```

Output 1. Fit summary of the ols model.

We chose to omit the age variable from the future models. This is motivated by two reasons. First of all, as highlighted by above output, age adds only a small contribution to the model. Secondly, in recent research it seemed to be the consensus that age had no correlation with

PSA in cancer patients (Kirollos, 1997; Littlejohns, Travis, Key, & Allen, 2016). In light of keeping this model as parsimonious as possible, we will agree with these experts and remove age from the model.

Otherwise, the model fits the data well as can be noted that 91.24% of the variation in the data is explained by the model (indicated by coefficient of determination R^2).

We first started by checking whether the Gauss Markov conditions hold. For this, we considered the diagnostic plot (Figure 2). Normality might be an issue, as highlighted by the Q-Q plot. Moreover, the residuals did not pass the Shapiro-Wilk test (p-value= 0.0086). However, some outliers were detected, and more worryingly there seemed to be a strong presence of heteroscedasticity, as seen by the cone shape in the lower left plot.

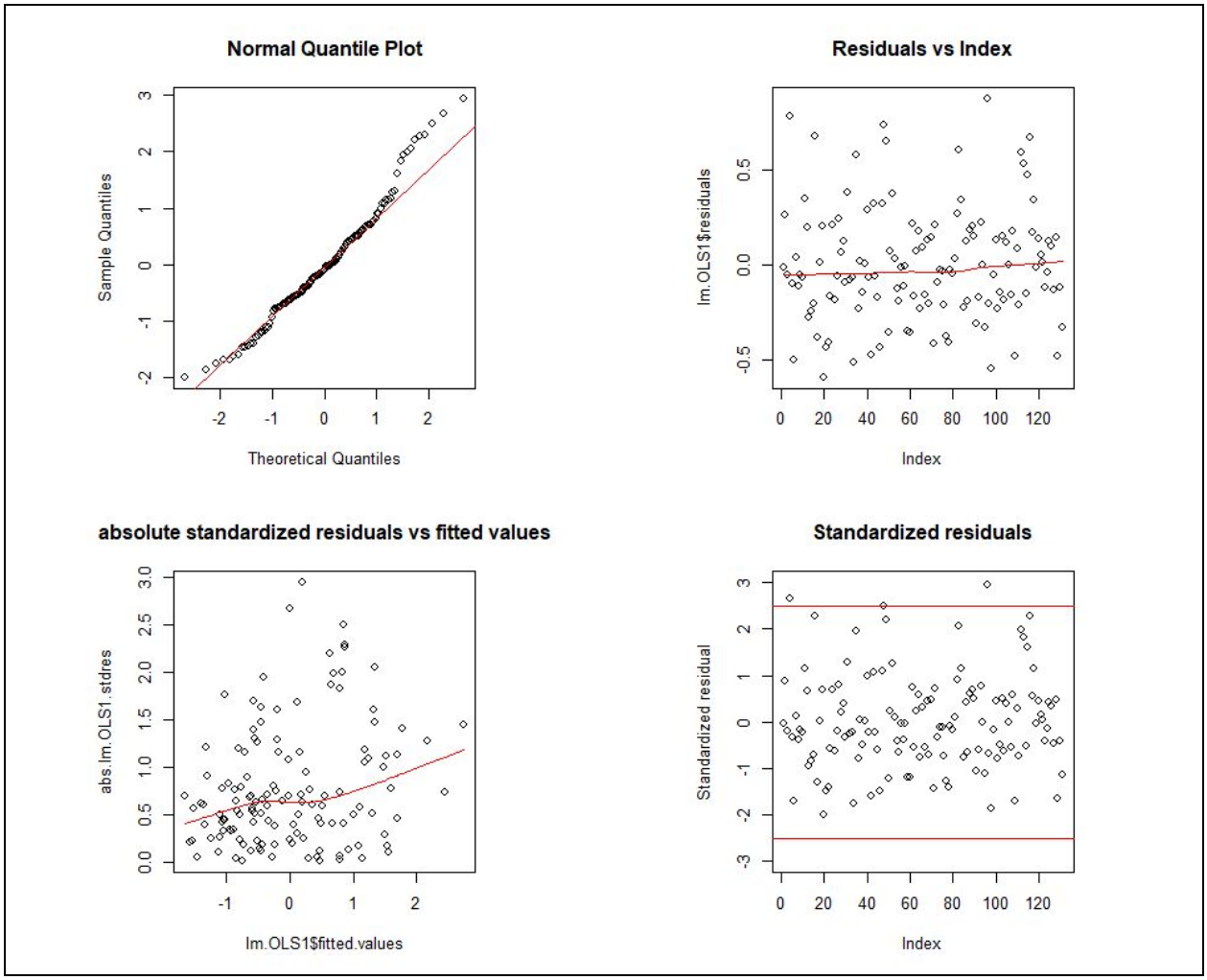

Figure 2. Diagnostic plots of the baseline model.

We also checked if any of the variables have a non-linear effect on psa using partial residual plots and it was seen that there were no such non-linear effects.

Furthermore, we also studied whether there was multicollinearity among the variables using simple correlations and Variance Inflation Factor (VIF) values.

```
                 age     invasion       caps      volume     gleason
age       1.000000000
invasion  0.004187901  1.000000000
caps      0.025897509  0.151370999 1.00000000
volume   -0.146877853  0.041566749 0.03771935  1.00000000
gleason  -0.047029971 -0.011173113 0.12275584 -0.14628434  1.00000000
```

Output 2. Simple correlations between the variables.

Output 2 shows the simple correlation between the predictor variables and it can be found that there are no strong correlations within the predictor variables. This provides some evidence that there is no multicollinearity present in our data set.

```
> vif
                     GVIF Df GVIF^(1/(2*Df))
age              1.029072  1        1.014432
volume           1.053093  1        1.026203
caps             1.071789  1        1.035272
factor(gleason)  1.087006  2        1.021076
factor(invasion) 1.043547  1        1.021542
> mean(vif)
[1] 1.093535
```

Output 3. VIF values of the baseline model.

From Output 3, the VIF values also do not show any signs of multicollinearity as the individual values are all close to 1 and so is their mean.

To test how well the model performs on the validation data, we calculated the MSEP and compared it with the MSE.

MSEP	MSE	MSEP-MSE
0.0635	0.0869	-0.0233

We see that the Mean Square Error (MSE) and the Predicted Mean Square Error (MSEP) are relatively close together, indicating a decent fit. However, there is certainly room for

improvement. We also check the Predicted Residual Error Sum of Squares (PRESS) criterion as a summary measure of the model fit. The PRESS value for the baseline model is 12.80117.

Using this model as our baseline, we now consider the following questions.

1. What problems are present, and should be accommodated by the models?

2. Should any interaction terms be added? If yes, which one? Is the interaction term logical?

3. What are possible solutions to the problems presented?

We first consider the first question. The main problem we should handle is the presence of heteroscedasticity. This will be the main motivation for the choice of models presented later. Furthermore, we would like to improve the generalization of the model, using the difference between MSE and MSEP as a metric. Finally, we should always keep an eye on keeping the model as parsimonious as possible, since potential clients from clinical backgrounds might be more interested in getting an interpretable model compared to a slightly better complex model.

For the second question, we use stepwise AIC to conduct variable selection. Based on the output, we decided that the interaction term between caps and gleason is the most relevant. Notice that other interaction terms are still included in the final model, but due to the small influence that their omission has on the AIC, we choose to omit them either way for the sake of keeping the model simple. We will consider a model with the gleason and caps interaction term included. Notice that this interaction term is also interpretable. We also conducted a forward selection using the F-statistic, and obtained the same final model.

```
Step:  AIC=-391.96
psa ~ factor(invasion) + caps + volume + factor(gleason)
+ volume:factor(gleason) +
    caps:factor(gleason) + factor(invasion):volume

                          Df Sum of Sq    RSS     AIC
<none>                                  5.0429 -391.96
+ factor(invasion):caps    1    0.0326 5.0103 -390.79
- factor(invasion):volume  1    0.1328 5.1757 -390.63
- volume:factor(gleason)   2    0.2308 5.2737 -390.23
+ I(caps^2)               1    0.0093 5.0336 -390.19
+ I(volume^2)             1    0.0080 5.0349 -390.16
- caps:factor(gleason)     2    3.6610 8.7039 -326.10
```

Output 4. Final result of step AIC.

Finally, we propose three models to accommodate these problems. Our first model will be a model based on reweighted least squares, in order to try to deal with the heteroscedasticity. Next, we will consider another reweighted least squares model, but this time include the interaction term mentioned above. Finally, we will consider a robust regression method.

As a final remark, we recall that there were outliers present, as indicated by the diagnostic plot (Figure 2) above. We chose to delete observations from the training set when the absolute standardized residual was above 2.5. For the first two models, we will use this updated data set.

Model 1:

Considering the heteroscedasticity detected in our baseline model, we implemented a weighted least squares model. This approach is more suitable to the studied data as it does not assume an equal variance, weighting the different variance across the sample.

$$psa = \beta_{0(WLS)} + \beta_{1(WLS)} * invasion + \beta_{2(WLS)} * caps + \beta_{3(WLS)} * volume + \beta_{4(WLS)} * gleason.$$

The output of the model is the following:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.96577    0.05209  -18.54   <2e-16 ***
volume             0.71989    0.02657   27.09   <2e-16 ***
caps              -0.62295    0.02576  -24.18   <2e-16 ***
factor(gleason)2   0.88236    0.06274   14.06   <2e-16 ***
factor(gleason)3   0.63614    0.06257   10.17   <2e-16 ***
factor(invasion)1  0.73790    0.04865   15.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.29 on 120 degrees of freedom
Multiple R-squared:  0.9139,   Adjusted R-squared:  0.9103
F-statistic: 254.9 on 5 and 120 DF,  p-value: < 2.2e-16
```

Output 5. Fit summary of the weighted least square model without interaction.

All predictors are statistically significant even considering a Bonferroni correction. Furthermore, we see that there is an improvement in the fit, with an adjusted R squared of approximately 0.92.

Next, we check the model assumptions. We do this once again based on the diagnostic plots (Figure 3). The normal Q-Q plot reveals that the normality assumption seems to be acceptable. The null hypothesis of the Shapiro-Wilk test cannot be rejected (p-value= 0.33). The top right plot indicates that there is error independence, and there is no sign of a non-linear effect. The bottom left plot shows a good improvement towards homoscedasticity, though there is still room for further improvement. Notice that the fitted lowess line seems to decrease towards higher fitted values. This trend can be due to only having few extreme value observations. Finally, the bottom right plot displays no noticeable outliers.

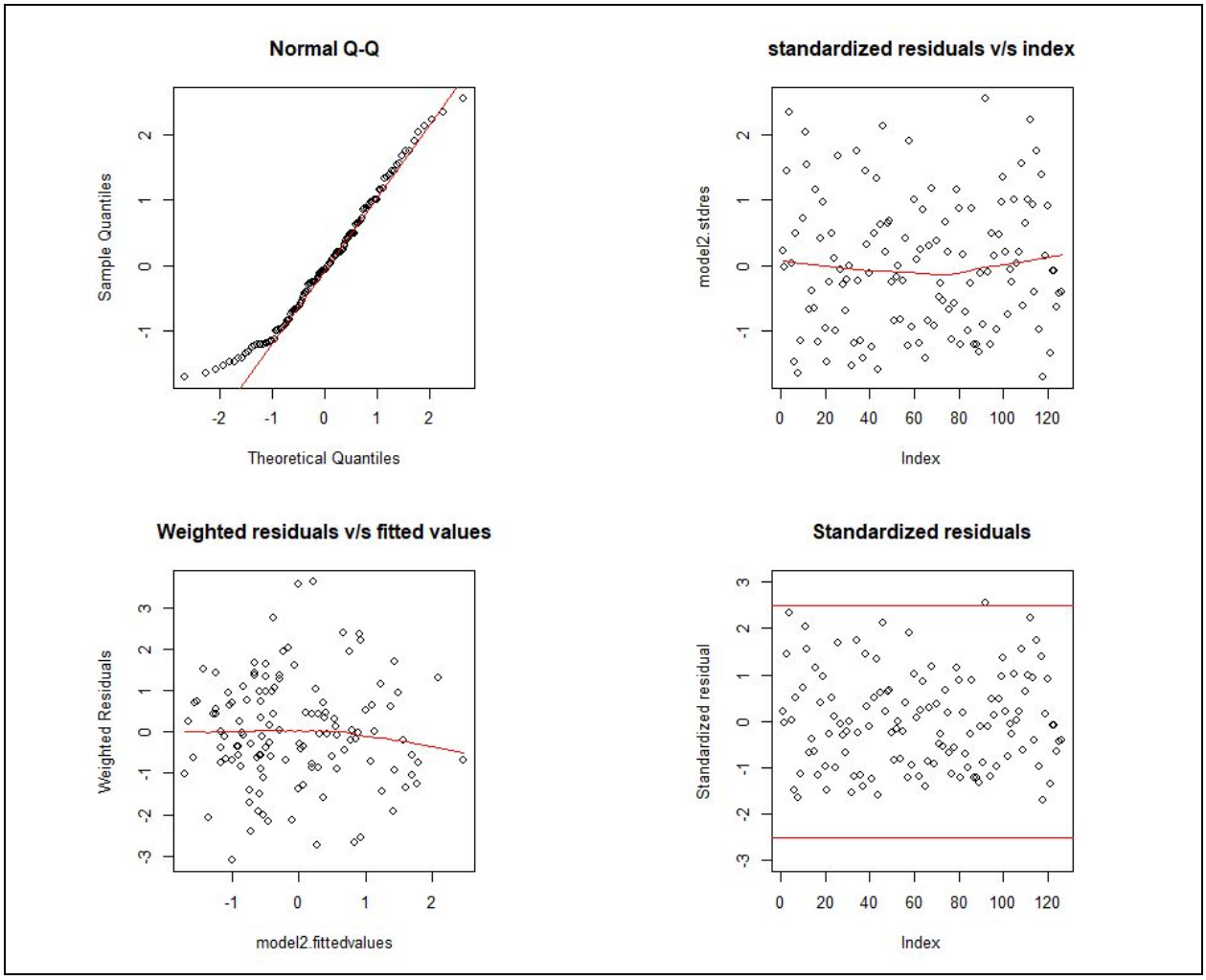

Figure 3. Diagnostic plots of model 1.

We now validate the model, using the validation set.

MSEP	MSE	MSEP-MSE

0.0681	0.0828	-0.0147

In order to validate our model, we compared the performance of the model over a training sample to the performance over a second validation sample. The mean square error is approximately 0.07 and the predicted mean square error is approximately 0.08. There seems to be only a slight difference between the MSE and the MSEP, indicating that there is no evidence of overfitting. Thus the model is suitable for predicting new data. The PRESS value for this model is 10.72894.

Model 2:

As previously mentioned this second model weighted least squares model includes the interaction between caps and gleason, motivated by the variable selection performed in the baseline model.

$$psa=\beta_{0(WLS)} + \beta_{1(WLS)}*invasion + \beta_{2(WLS)}*caps + \beta_{3(WLS)}*volume + \beta_{4(WLS)}*gleason + \beta_{5(WLS)}*caps:gleason.$$

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.98489    0.02206 -44.653  < 2e-16 ***
volume                0.74411    0.01267  58.720  < 2e-16 ***
caps                 -0.86722    0.01987 -43.650  < 2e-16 ***
factor(gleason)2      0.78261    0.03139  24.930  < 2e-16 ***
factor(gleason)3      0.70699    0.02969  23.812  < 2e-16 ***
factor(invasion)1     0.71409    0.02087  34.216  < 2e-16 ***
caps:factor(gleason)2 0.48651    0.02979  16.331  < 2e-16 ***
caps:factor(gleason)3 0.16539    0.02427   6.816 4.01e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.225 on 120 degrees of freedom
Multiple R-squared:  0.9841,    Adjusted R-squared:  0.9832
F-statistic:  1062 on 7 and 120 DF,  p-value: < 2.2e-16
```

Output 4. Fit summary of WLS model with one interaction term.

Output 4 shows a highly significant effect of the interaction terms between caps and gleason on psa, the effect of the other predictors is highly significant as well. It can be noted from this output an improvement regarding the model‘s fit, the adjusted R-squared for this model is approximately 0.98.

If we analyse the plots displayed in Figure 4, we notice that the Normal Q-Q plot displays normality for most part except deviations in the tails. The null hypothesis of the Shapiro-Wilk test cannot be rejected (p-value= 0.052). In the bottom left plot we can notice a decrease in

heteroscedasticity. The top right plot shows independency of the error terms and no signs of a non-linear effect. From the bottom right plot we can see that there are no outlier residual observations.

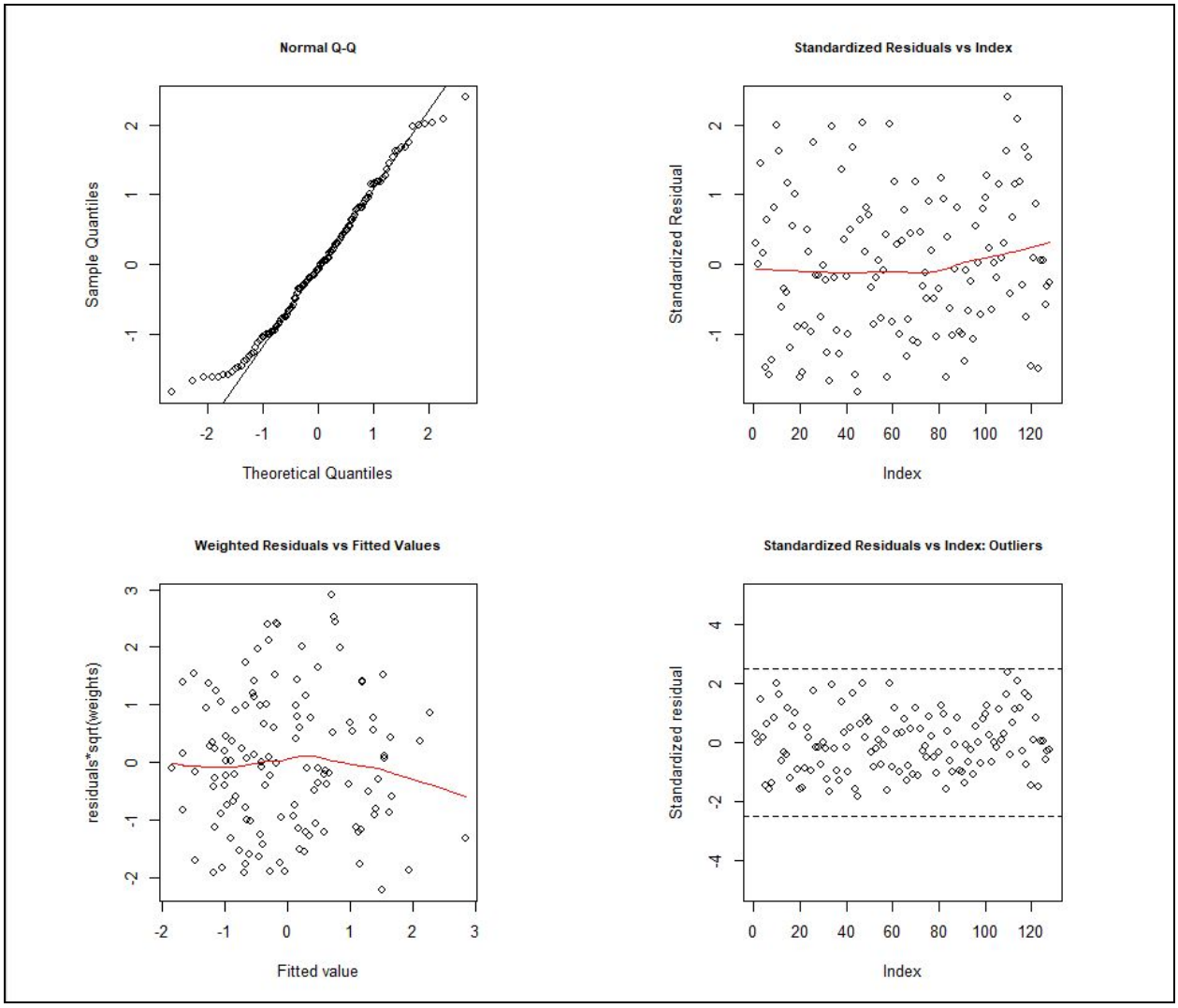

Figure 4. Diagnostic plots of model 2.

For this second model the MSE is rather low (i.e. 0.043) and the difference between MSE and the MSEP is small (i.e. -0.008), suggesting a good performance on new data and no overfitting. The PRESS value for this second model is 5.886098.

MSEP	MSE	MSEP-MSE
0.035	0.043	-0.008

Model 3:

For Model 3, we use robust regression which is another method to deal with the heteroscedasticity in the data as it does not assume that the variance is independent of the predictor variables. Robust regression function R analyses the data in SM method using the function lmrob. The analysis is performed on non-scaled data with no outliers removed. The third model can be given as:

$$psa=\beta_0 + \beta_1*invasion + \beta_2*caps + \beta_3*volume + \beta_4*gleason.$$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.91973    0.05420 -16.970  < 2e-16 ***
volume             0.71148    0.03012  23.623  < 2e-16 ***
caps              -0.59905    0.03974 -15.075  < 2e-16 ***
factor(gleason)2   0.77972    0.07683  10.149  < 2e-16 ***
factor(gleason)3   0.64446    0.07313   8.813 8.69e-15 ***
factor(invasion)1  0.67844    0.05568  12.186  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.2598
Multiple R-squared:   0.92,    Adjusted R-squared:  0.9168
Convergence in 14 IRWLS iterations
```

Output 6. Fit summary of lmrob function.

Output 6 shows that all predictors are statistically significant. It has a high adjusted R^2 value of 0.92. The robust residual error value is quite small which indicates a good fit.

Further the diagnostic plots (Figure 5) are considered to see if the model satisfies the model assumptions. The Normal Q-Q plot shows normality between quantiles -1 and 1, but there is deviation in the right tail of the plot. The Shapiro-Wilk test also doesn't agree with the normality assumption in the model as it has a p-value of 0.004. The bottom left plot shows that heteroscedasticity seems to be present in the model. The top right plot shows no signs of dependency between error terms. Finally in the bottom left plot it is possible to observe two outlier observations.

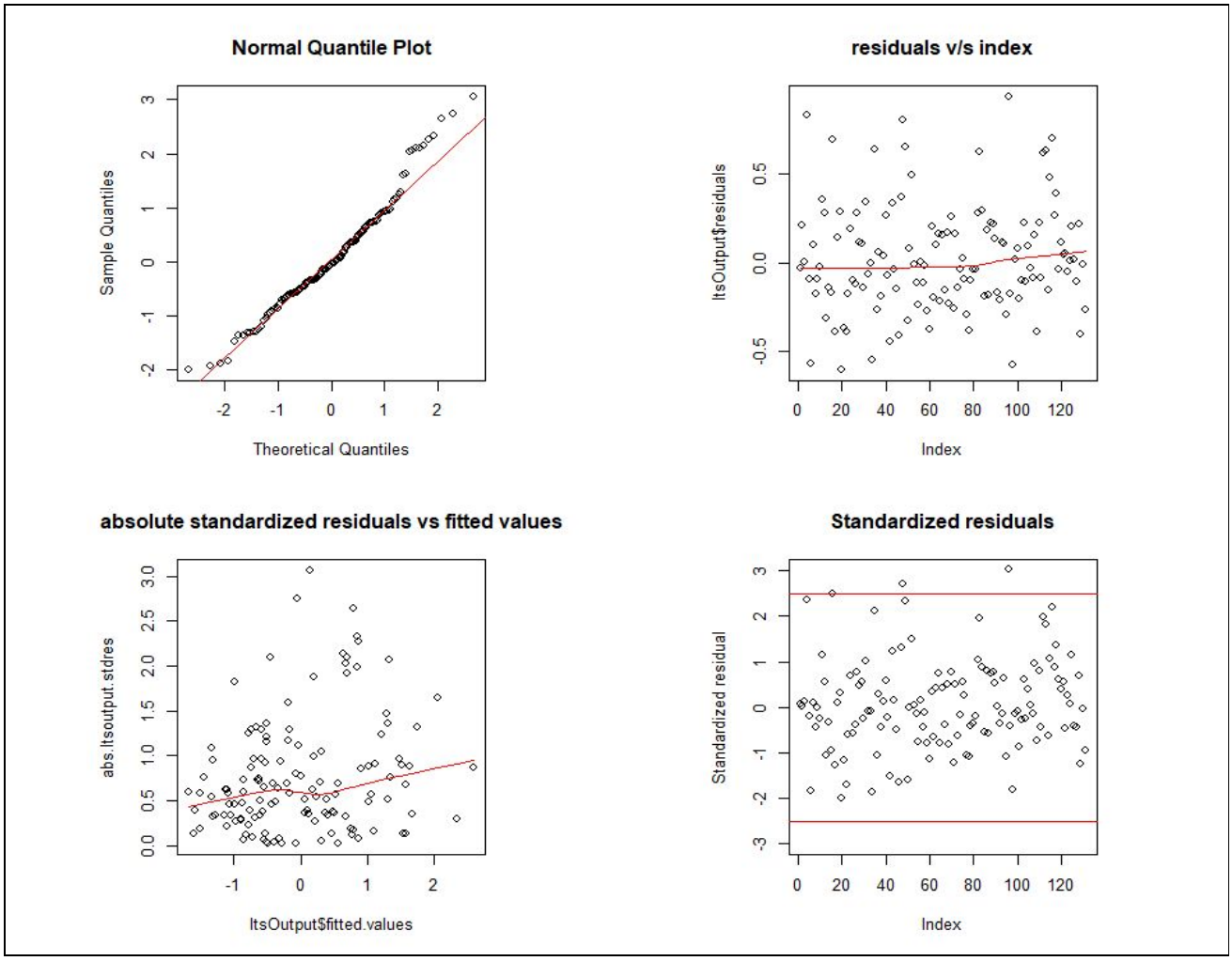

Figure 5. Diagnostic plots of model 3.

We now validate the model, using the validation dataset.

MSEP	MSE	MSEP-MSE
0.0652	0.0903	-0.0251

The table shows the MSEP, MSE and their difference. The difference between MSE and the MSEP is small, suggesting a good performance on new data and no overfitting. The PRESS value for Model 3 is 12.7462.

Model Comparison:

Table 3. Comparison of models.

Model	Adjusted R^2	MSE	MSEP	PRESS
Baseline Model	0.908	0.0869	0.0635	12.801
Model 1	0.9103	0.0828	0.0681	10.782
Model 2	0.9832	0.043	0.035	5.8806
Model 3	0.9168	0.0903	0.0652	12.7462

As indicated by Table 3, we notice that the second model performs the best with respect to all the criteria. Therefore, we choose model two as our final model. We found that it generalizes remarkably well to the test data, and that its fit values all indicate that the model is well supported by the data. We also observed that this model satisfied the Gauss Markov conditions to an acceptable degree: normality of the residuals was attained, and the problem of heteroscedasticity was minimized.

Given that the data is from a medical context, the fact that the model generalized well to the test data is favourable, indicating that it might be useful for diagnostic purposes.

Interpretation of the chosen model:

Since the variables were standardized all the estimated effects are in terms of standard deviation, i.e. the effect coefficients correspond to the effect of the regression variable over psa when the independent variable increases in one standard deviation.

From Output 4 we can observe that the prognostic variables volume, caps, gleason and the interaction between caps and gleason are significantly associated with psa. It can be noted that caps have a negative effect in psa, but when gleason has a value of 2 or 3 this negative effect is attenuated. Furthermore, the volume of the tumor is observed to have a positive effect on the amount of psa detected. We can notice as well that the gleason score 2 and 3 seem to have a similar positive effect on psa, compared to the reference level, which is consistent with the findings in our descriptive analyses (see boxplot of gleason in Figure 1). In a similar way a positive effect of the invasion factor on psa is observed.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.98489    0.02206 -44.653  < 2e-16 ***
volume                0.74411    0.01267  58.720  < 2e-16 ***
caps                 -0.86722    0.01987 -43.650  < 2e-16 ***
factor(gleason)2      0.78261    0.03139  24.930  < 2e-16 ***
factor(gleason)3      0.70699    0.02969  23.812  < 2e-16 ***
factor(invasion)1     0.71409    0.02087  34.216  < 2e-16 ***
caps:factor(gleason)2 0.48651    0.02979  16.331  < 2e-16 ***
caps:factor(gleason)3 0.16539    0.02427   6.816 4.01e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.225 on 120 degrees of freedom
Multiple R-squared:  0.9841,    Adjusted R-squared:  0.9832
F-statistic:  1062 on 7 and 120 DF,  p-value: < 2.2e-16
```

Repeated (**Output 4.** Fit summary of WLS model with one interaction term.)

Final Remarks:

In the process of attaining the three models previously presented, additional models were tested. These intermediate models are worth mentioning given that they help to better understand how the final three models were achieved.

A Box Cox transformed model was tested to comply better with the Gauss Markov conditions. One of the downsides was that the suggested transformation was very close to the identity, and it has the additional issue of interpretability. Given that the dependent variable was transformed, it also alters the relation of psa with all of the independent variables.

A least square regression model with one and two interaction terms was also tested, to assess if heteroscedasticity was due to an unaccounted interaction term. This model was unsuccessful in reducing heteroscedasticity.

A weighted least square model with two interaction terms was also tested. The fit was marginally better when compared to including a single interaction term, and due to the added complexity of having two interaction terms we decided to continue with only one interaction term. Also, the condition for normality of the model with two interactions was not satisfactory.

Further research questions:

We remark that the data has been gathered from men who were in an advanced stage of prostate cancer. A possible question is how well the model generalizes to earlier stages of cancer. If it is found that this model does not generalize well, possible research questions are the construction of a better model for this subgroup of patients, and also investigate why the model does not generalize well. This could lead to newer insights regarding the values of prognostic parameters between advanced and early stages of prostate cancer.

Another proposed question is based on the evolution of the psa levels after the prostatectomy. This is often discussed in the literature, and it was noticed that this evolution is not uniform (Naselli, Introini, Andreatta, Spina, Truini, & Puppo, 2009). The psa levels of some patients dropped to the base level after the surgery, whilst other patients had barely any change in their psa levels. A possible question is whether this evolution can be predicted using the initial value of psa post-prostatectomy and the values of the other prognostic variables. This could give insight on the efficacy of the surgeries for patients with differing scores on these prognostic variables.

When using a psa-test one must be cautious that screening for prostate cancer via a psa-test leads to only a small reduction in disease-specific mortality over 10 years but does not affect overall mortality (Ilic, Djulbegovic, Jung, Hwang, Zhou, et al, 2018). Another severe limitation of the psa-test is the high incidence of false positive and false negative results (Roland, Neal, & Buckley, 2018). Because of the high false positive results the psa-test should always be accompanied by a digital rectal examination and a normal psa result should not be regarded as reassuring if there are other symptoms suggesting urogenital cancer or metastatic disease (Roland, Neal, & Buckley, 2018). Before considering a psa-test it is important to beware if the potential short and long term harms of screening, including complications from biopsies and

subsequent treatment, as well as the risk of overdiagnosis and overtreatment (Ilic, Djulbegovic, Jung, Hwang, Zhou, et al, 2018).

The associations between prostate cancer and risk factors found in observational studies could be affected by detection bias (i.e. factors that are associated with men choosing to undergo psa testing will, in turn, be associated with increased prostate cancer incidence; Littlejohns, Travis, Key, & Allen, 2016)

Appendix A: References:

Horninger, W., Reissigl, A., Rogatsch, H., Volgger, H., Studen, M., Klocker, H., & Bartsch, G. (2000). Prostate cancer screening in the tyrol, austria: Experience and results. *European Journal of Cancer, 36,* 1322-1335. https://doi.org/10.1016/S0959-8049(00)00113-1

Ilic, D., Djulbegovic, M., Jung, J.H., Hwang, E.C., Zhou, Q., et al. (2018). Prostate cancer screening with prostate-specific antigen (PSA) test: A systematic review and meta-analysis. *BMJ, 362,* K3519. https://doi.org/10.1136/bmj.k3519

Littlejohns, T.J., Travis, R.C., Key, T.J., & Allen, N.E. (2016). Lifestyle factors and prostate-specific antigen (PSA) testing in UK Biobank: Implications for epidemiological research. *Cancer Epidemiology, 45,* 40-46. https://doi.org/10.1016/j.canep.2016.09.010

Kirollos, M.M, (1997). Statistical Review and Analysis of the Relationship Between Serum Prostate Specific Antigen and Age. *Journal of Urology, 158,* 143-145. https://doi.org/10.1097/00005392-199707000-00044

Naselli, A., Introini, C., Andreatta, R., Spina, B., Truini, M., & Puppo, P.. (2009). Prognostic factors of persistently detectable PSA after radical prostatectomy. *International Journal of Urology, 16,* 82-86. https://doi.org/10.1111/j.1442-2042.2008.02198.x

Roland, M., Neal, D., & Buckley, R. (2018). What should doctors say to men asking for a PSA test? *BMJ (Online), 362,* K3702. https://doi.org/10.1136/bmj.k3702