Predicting Cancer Recurrence Times From Cell Images

Jeffrey Liang

November 20, 2020

Introduction

This data set was obtained from the publicly available UCI Machine Learning Repository (Dua & Graff, 2019). Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. William H. Wolberg (University of Wisconsin, Clinical Sciences Center) since 1984, and the data set includes only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Cases are divided into two groups: one where recurrence of cancer occurred, and one where cancer did not recur. Given our question of interest, we will exclude cases where cancer did not recur. This reduced our data set from 198 to 47 independent cases. Although this data set was referenced by many papers, after extensive searching we could not find or determine the units by which the data were presented. Therefore, our analysis offers values in terms of the original units used in the data set (which remain unknown, except for time until recurrence).

We would also like to note that we are excluding the attribute “ID Number”, provided for each case in the data set, because we believe that it is not a significant predictor of the time til recurrence. Also, there are numerous cell nuclei present in each FNA image. The dataset provided the means, standard errors, and “worst” or maximum (mean of the three largest values) values for each variable, calculated from all nuclei in the image. Our analysis will use the mean values for each variable.

Questions of Interest

We consider the following research questions:

1. Assuming that breast cancer recurs in a given patient, we want to know if we are able to predict, with confidence, the time til recurrence (in months) considering the following variables computed from FNA images of malignant breast masses:

a. radius (mean of distances from center to points on the perimeter)
b. texture (standard deviation of gray-scale values)
c. perimeter, or circumference
d. area
e. smoothness (local variation in radius lengths)
f. compactness (perimeter2 / area - 1.0)
g. concavity (severity of concave portions of the contour)
h. concave points (number of concave portions of the contour)
i. symmetry
j. fractal dimension (“coastline approximation” - 1.0)

2. Once we establish our model, can we determine which of the variables included are significant predictors of the response? And if a predictor is significant, then what is the nature of the relationship between it and the

time until recurrence (eg. positively or negatively related)?

Regression Method

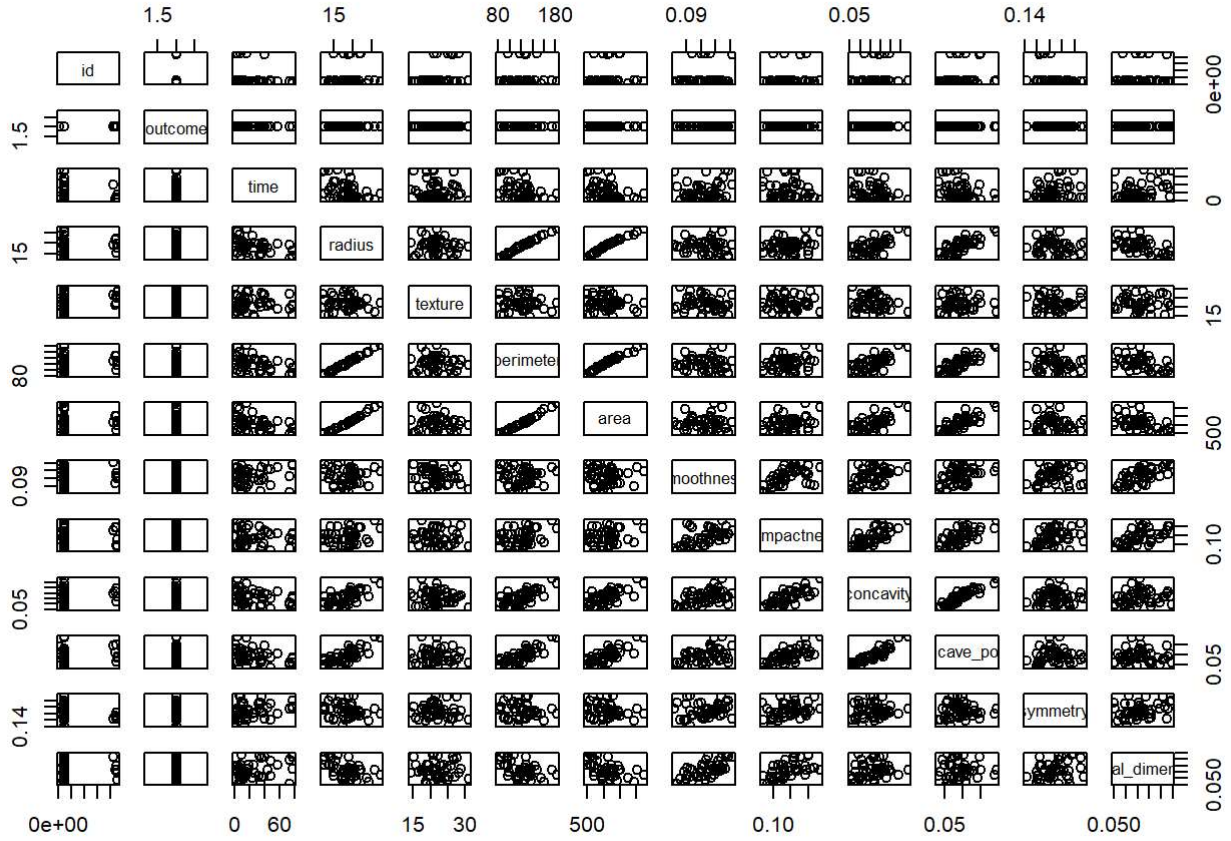
We begin by building a regression model based on the variables listed above. To do so, we utilize the method of stepwise multiple regression, using Akaike's Information Criterion (AIC) is a guideline for selecting a model. We will also perform "best subsets regression" with respect to adjusted R-squared and compare the model with the one obtained with AIC. Due to the nature of this data set (image analysis of cell masses), we expect lower R-squared values. Thus it makes sense for the model with the highest adjusted R-squared to be used. Hamilton et al. (2015) state that "when comparing radiographic parameters or associating surgical technical factors, values of R2 are reported in the 0.2 to 0.4 range." Then, the model will be checked for the four LINE conditions (specified later), and appropriate transformations will be applied. Afterward, we can observe the reported coefficient of determination of the model and see if the predictors play a significant role in determining the time until recurrence.

After establishing our model, we will perform hypothesis tests, testing the reduced models against the full models using ANOVA tables. The second research question can also be answered through linear regression. We could further employ a hypothesis tests for each predictor in the model, to see if a single predictor has a significant relationship with the time until recurrence. Then, for the variables that were significant, we could estimate confidence intervals for the slope coefficients to determine the magnitude of the relationship.

Regression Model

Correlation and Multicollinearity

We begin with plotting each potential predictor against the response variable using the pairs() functions in R. This gives a scatterplot matrix allowing us to visually analyze the correlations between the potential predictors, as well as the response. The results are as follows:



From the scatterplot matrix, we observe several relationships between potential predictors:

- Strong positive linear relationships between: radius and perimeter, radius and area, and perimeter and area. These relationships match our intuition as these 3 predictors are directly related to each other geometrically in a cell nucleus. There is also a strong positive correlation between concavity and number of concave points (concave_points), which again matches our intuition.
- Weak positive linear relationships between time and the predictors smoothness and symmetry.
- Non-linear relationships between time and the predictors radius, perimeter, area, concavity, and concave points.

We found that the correlation coefficients between the predictors demonstrating strong positive linear relationships in the scatterplot matrix were extremely high. Furthermore, the VIF (variance inflation factors) for radius, perimeter, and area were extremely high (1369.502, 1411.260, 76.844 respectively), as well as concavity and concave points (9.774, 23.218 respectively). Finally, fractal dimension was shown to have a VIF of 12.613.

Because radius, perimeter and area were so highly correlated, perimeter and area were removed as potential predictors for the linear regression model. It was reasoned that including radius as a predictor should provide sufficient information regarding perimeter and area due to their geometric relationship. Concave points were also removed as potential predictors as it was assumed to be highly correlated with concavity. Fractal dimension was also removed, due to its VIF being greater than 10. We get the following result:

```
vif(fit)
```

```
##      radius     texture  smoothness compactness   concavity    symmetry
##    3.465357    1.068495    2.195060    3.725060    7.451421    1.239343
```

```
fit = lm(time ~ radius + texture + smoothness + compactness + concavity + symmetry)
```

Building the Model

To determine which variables are best suited for predicting the Y values (time), we will perform a stepwise regression using the step() function in R. The resulting best fitting model is given as follows, using the AIC as a criterion:

$$(Time)_i = \beta_0 + \beta_1(Radius)_i + \beta_2(Smoothness)_i + \beta_3(Compactness)_i + \beta_4(Symmetry)_i + \epsilon_i$$

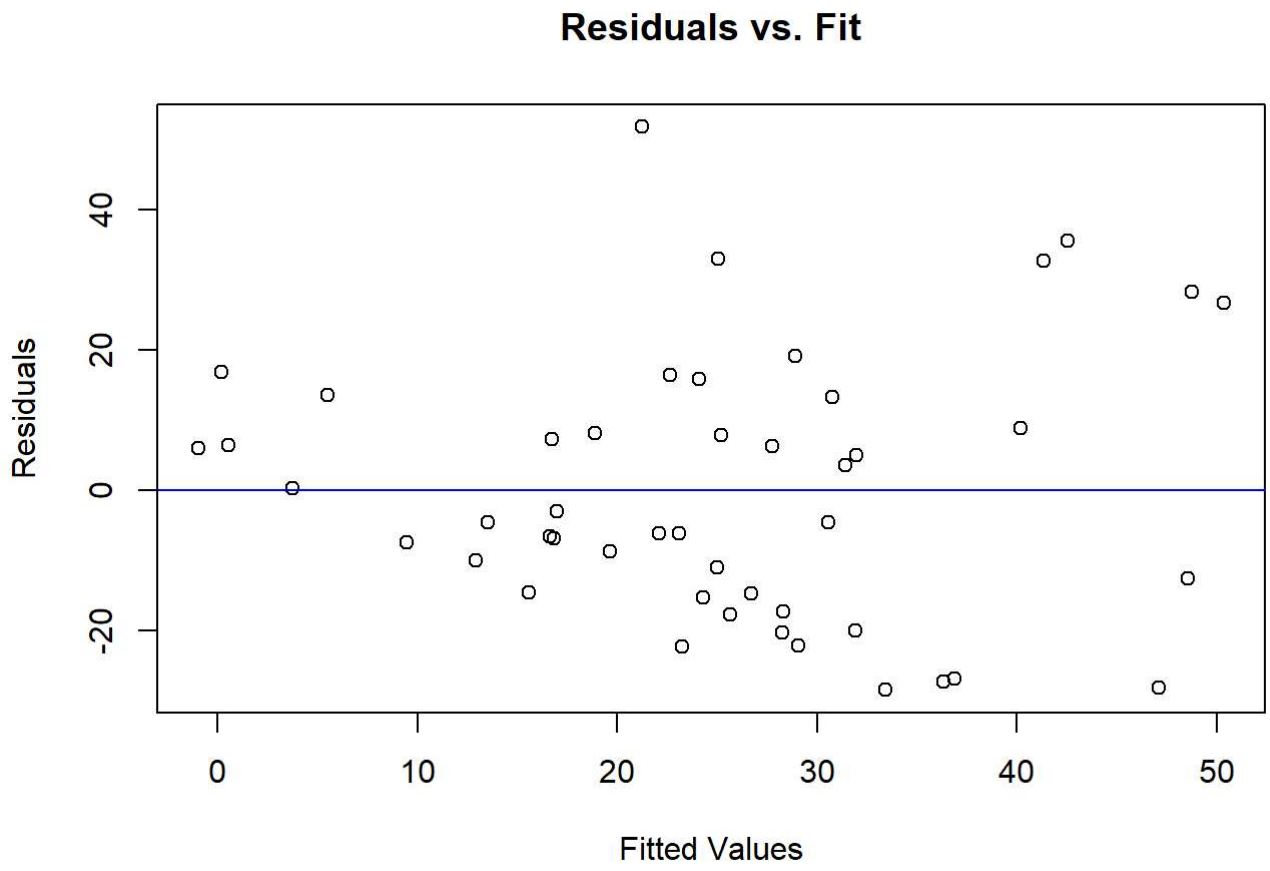
Further information on the model can be found in the appendix.

As mentioned above, we also performed best subsets regression using the “leaps” package in R and found that the same exact model yielded the highest adjusted R-squared value. Therefore, we continued with this model for the rest of our analysis.

Model Validity

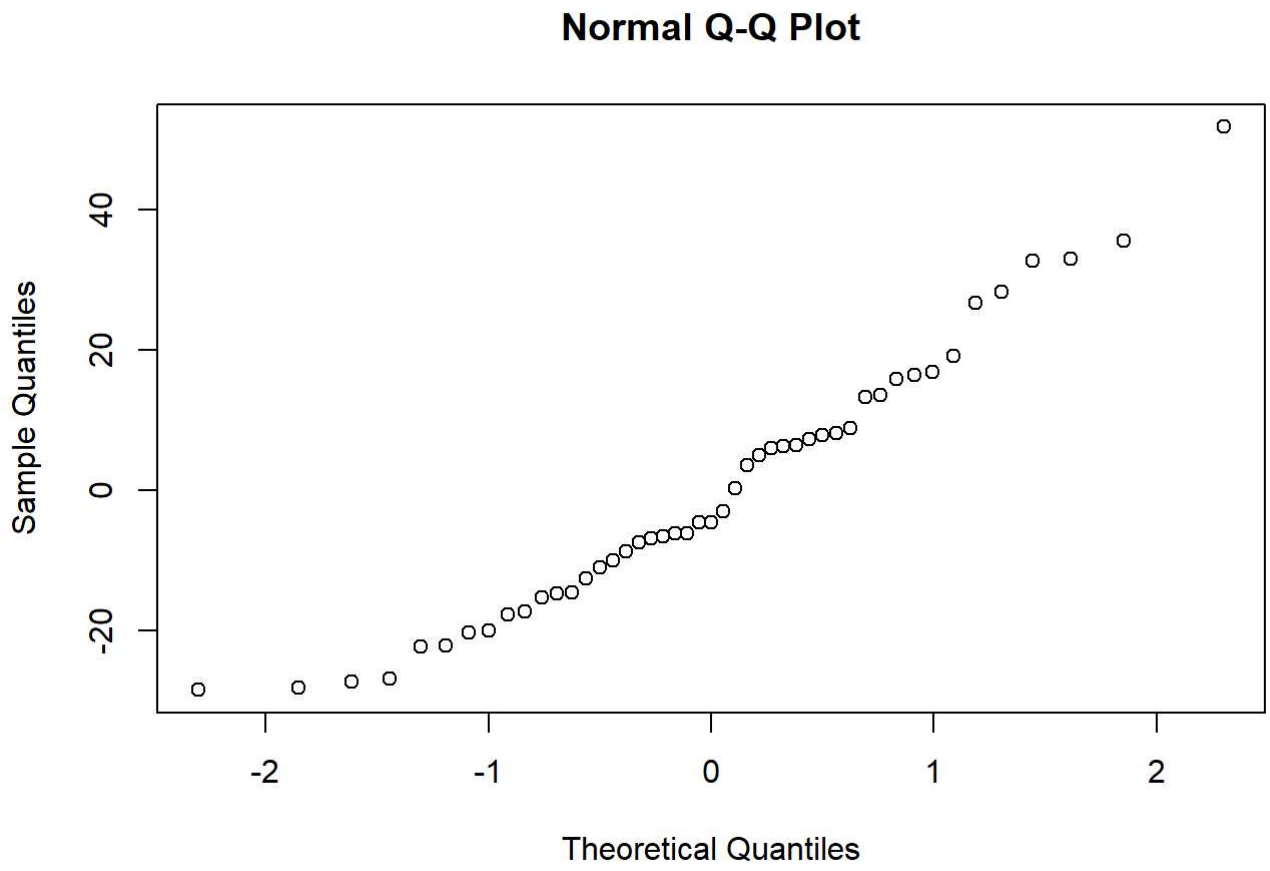
After obtaining the regression model, the four “LINE” conditions were checked according to the following:

1. “L” - Linearity of the mean response function, by analyzing a “residuals vs. fitted values” plot, and visually checking for a vertical mean of 0 as the plot moves left to right.



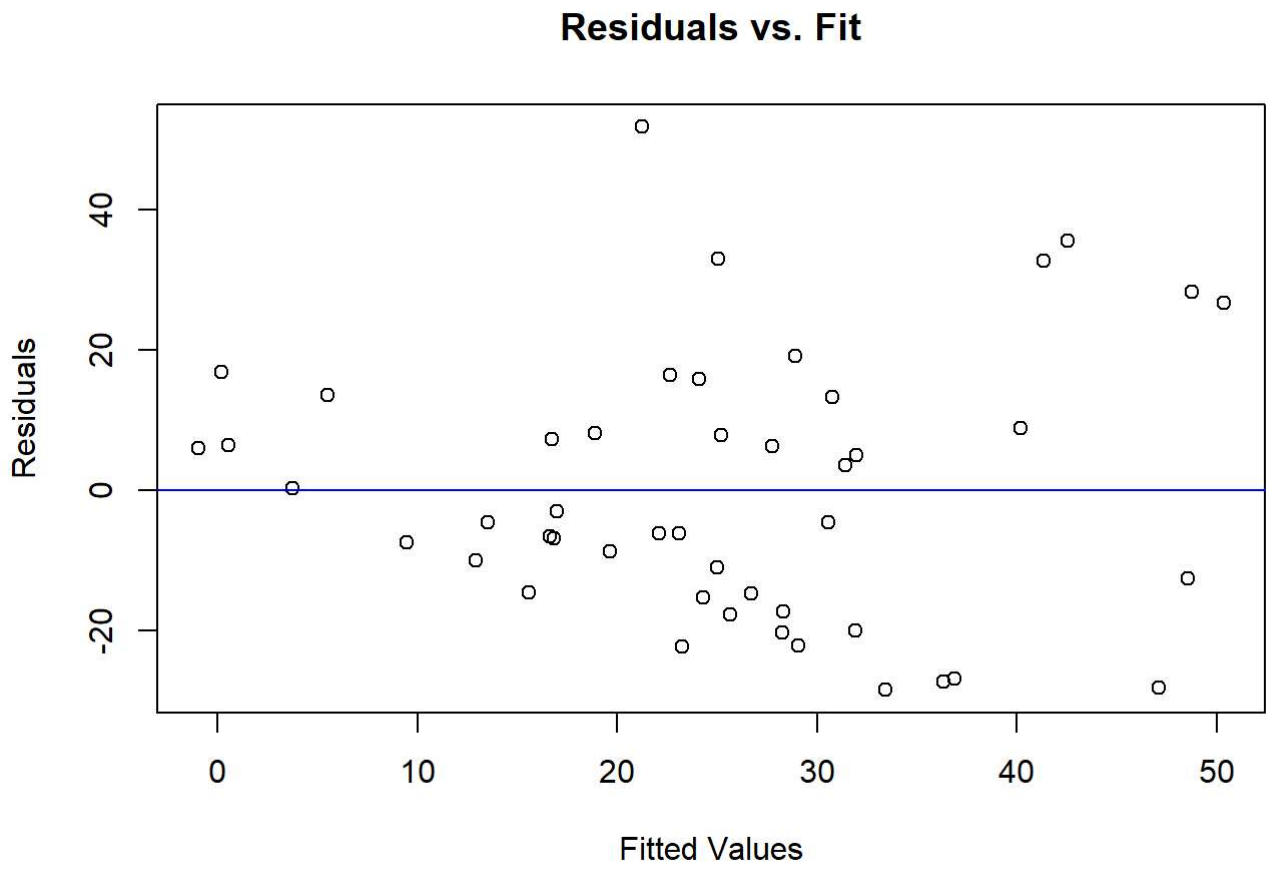
With this plot we can (subjectively) judge that the residuals “bounce randomly” around the 0 line, indicating that our expectation function is linear.

2. “I” - Independence of the errors by analyzing the “residual vs order” plot, and looking for a random pattern that implies independence. However, since it is not known the order in which the data was collected, we must simply assume that the errors are independent of each other, and free of serial correlation.
3. “N” - Normally distributed errors using the normal Q-Q plot of the residuals, and checking that the resulting plot is approximately linear.



With this plot, we are able to see that the relationship is relatively linear, which implies that condition that the error terms are normally distributed is met.

4. “E” - Equal variances throughout the errors of the predictors, and checking that the residuals roughly form around the “horizontal band” around the 0 line.



Using the same residual vs fitted values plot, we can see that the residuals are around the 0 line, which indicates that the variance of the error terms are approximately equal.

Since our model has satisfied the four LINE conditions, there is no requirement for transformations on the data or model.

Results and Interpretation

We can now move on to answering the updated questions of interest:

1. Can the time until recurrence (in months) be predicted by the following variables: radius, smoothness, compactness, and symmetry?

Observing the coefficient of determination, we can see that the adjusted R-squared value is 0.244, which means that 24.4% of the variation of the time until recurrence is explained by considering the radius, smoothness, compactness, and symmetry of the images of the breast masses. Furthermore, the associated F-statistic of 4.712 with p-value of 0.003121 does not offer a strong significant relationship between the response time and the predictors radius, smoothness, compactness, and symmetry. Using the residual standard error, we can see that the actual time until recurrence can deviate from the regression line by 19.75 months. Taking a look at the coefficient standard errors, the smallest error is given by the radius with 0.985, whereas the largest error is given by the smoothness with 356.635. The coefficient standard can tell us the average amount the coefficient estimates vary from the actual, so we can see that the radius characteristic is our strongest variable in this case. Additionally, since the coefficient t-values show us how many standard deviations our coefficient estimates are away from 0, the values of each predictor offer potential in declaring a relationship between time and the predictors radius, smoothness, compactness, and symmetry.

2. Which predictors (out of radius, smoothness, compactness, and symmetry) can be deemed significant? And if a predictor is significant, then what is the nature of the relationship between it and the time until recurrence (eg. positively or negatively related)?

We were able to answer this question by hypothesis testing reduced models against full models for each variable, and obtaining ANOVA tables for each. For each variable, we performed a hypothesis test with the reduced model (not containing the predictor in question) against the full model (containing radius, smoothness, compactness and symmetry). For a 95% level of confidence, we failed to reject the null hypothesis for smoothness, compactness, symmetry (F-test p-values: 0.064, 0.110, and 0.118 respectively). On the other hand, we were able to reject the null hypothesis for radius (F-test p-values: 0.031). This suggested the radius is a significant predictor for time until recurrence in our model. To further investigate the relationship between radius with the response, we estimated a confidence interval for its slope coefficient. The result was, with 95% confidence, the slope coefficient was between -4.182 and -0.206.

Conclusion

In conclusion, our regression analysis has shown that the time until the recurrence of breast cancer (in months) is slightly influenced by the radius, smoothness, compactness, and symmetry of the cell nuclei in the breast mass images. While we lacked conclusive evidence of a strong relationship between our response and predictor variables, it is apparent that there is still a relationship between the variables that we are able to use to predict the time until recurrence. From our hypothesis tests described above, we concluded that with our regression model, radius (the mean radius of all cell nuclei present in the image) is a significant predictor of time until recurrence. Furthermore, the slope coefficient was estimated to be a negative value of relatively small magnitude; this indicates that for every unit increase in radius, the time until recurrence decreased by a small amount. Since the observations of our data were taken during follow ups, it is possible that the time since recurrence was not precise in when cancer cells began to recur. Also, it is possible that for non-recurrence patients, recurrence took place after the follow up and the study was not able to accurately collect data. Thus, the addition of another follow up study may be beneficial in order to improve our regression analysis model. We also acknowledge that leaving out non-recurrent cases in our analysis may have been detrimental, as patients may still experience recurrence after the time the data was collected. Also, the event of patients not experiencing recurrence events may depend on our chosen predictors as well. Therefore, in possible future studies we would like to include “recurrence or non-recurrence (at time of study)” as a possible categorical predictor.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml (http://archive.ics.uci.edu/ml)]. Irvine, CA: University of California, School of Information and Computer Science.

Hamilton, D. F., Ghert, M., & Simpson, A. H. R. W. (2015). Interpreting regression models in clinical outcome studies. Bone & Joint Research, 4(9), 152-153. doi: 10.1302/2046-3758.49.2000571

Appendix

Stepwise Regression

Start: AIC=288.69 time ~ radius + texture + smoothness + compactness + concavity + symmetry

```
         Df Sum of Sq   RSS    AIC
```

- concavity 1 84.04 16312 286.93
- texture 1 97.10 16326 286.97
- compactness 1 264.03 16492 287.44
- radius 1 407.65 16636 287.85 16228 288.69
- symmetry 1 874.10 17103 289.15
- smoothness 1 1222.70 17451 290.10

Step: AIC=286.93 time ~ radius + texture + smoothness + compactness + symmetry

```
          Df Sum of Sq   RSS    AIC
```

- texture 1 75.02 16387 285.14 16312 286.93
- symmetry 1 931.44 17244 287.54
- compactness 1 973.28 17286 287.65
- smoothness 1 1241.51 17554 288.38
- radius 1 1995.42 18308 290.35

Step: AIC=285.14 time ~ radius + smoothness + compactness + symmetry

```
          Df Sum of Sq   RSS    AIC
```

16387 285.14 - symmetry 1 992.12 17380 285.91 - compactness 1 1041.28 17429 286.04 - smoothness 1 1412.69 17800 287.03 - radius 1 1934.94 18322 288.39

Call: lm(formula = time ~ radius + smoothness + compactness + symmetry)

Coefficients: (Intercept) radius smoothness compactness symmetry
-27.695 -2.194 678.603 -161.344 245.747

Best Subsets Regression (using Adjusted R2)

Subset selection object 6 Variables (and intercept) Forced in Forced out radius FALSE FALSE texture FALSE FALSE smoothness FALSE FALSE compactness FALSE FALSE concavity FALSE FALSE symmetry FALSE FALSE 1 subsets of each size up to 6 Selection Algorithm: exhaustive radius texture smoothness compactness concavity symmetry 1 (1) “*” " " " " " " " " " " "
2 (1) " " " " “*” " " “*” " "
3 (1) " " " " “*” " " “*” “*”
4 (1) “*” " " “*” “*” " " “*”
5 (1) “*” “*” “*” “*” " " “*”
6 (1) “*” “*” “*” “*” “*” “*”

summary.subsets$adjr2 [1] 0.2005028 0.2139779 0.2388103 0.2440232 0.2291301 0.2139290

Correlation Matrix

```
             time radius texture perimeter   area smoothness compactness concavity
```

time 1.000 -0.467 -0.094 -0.466 -0.438 0.221 -0.142 -0.284 radius -0.467 1.000 -0.044 0.997 0.992 -0.106 0.275 0.650 texture -0.094 -0.044 1.000 -0.047 -0.064 -0.167 -0.054 -0.146 perimeter -0.466 0.997 -0.047 1.000 0.990 -0.065 0.345 0.696 area -0.438 0.992 -0.064 0.990 1.000 -0.083 0.285 0.660 smoothness 0.221 -0.106 -0.167 -0.065 -0.083 1.000 0.533 0.483 compactness -0.142 0.275 -0.054 0.345 0.285 0.533 1.000 0.774 concavity

-0.284 0.650 -0.146 0.696 0.660 0.483 0.774 1.000 concave_points -0.352 0.789 -0.081 0.823 0.794 0.429 0.685 0.922 symmetry 0.221 -0.134 -0.079 -0.115 -0.147 0.163 0.337 0.105 fractal_dimension 0.309 -0.528 -0.021 -0.471 -0.497 0.706 0.567 0.162

```
              concave_points symmetry fractal_dimension
```

time -0.352 0.221 0.309 radius 0.789 -0.134 -0.528 texture -0.081 -0.079 -0.021 perimeter 0.823 -0.115 -0.471 area 0.794 -0.147 -0.497 smoothness 0.429 0.163 0.706 compactness 0.685 0.337 0.567 concavity 0.922 0.105 0.162 concave_points 1.000 0.022 0.012 symmetry 0.022 1.000 0.270 fractal_dimension 0.012 0.270 1.000

ANOVA Tables for Significant Predictors

Analysis of Variance Table

Model 1: time ~ radius + smoothness + compactness + symmetry Model 2: time ~ smoothness + compactness + symmetry Res.Df RSS Df Sum of Sq F Pr(>F)
1 42 16387
2 43 18322 -1 -1934.9 4.9591 0.03137 * — Signif. codes: 0 ‘***’ 0.001 ”** 0.01 ”* 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table

Model 1: time ~ radius + smoothness + compactness + symmetry Model 2: time ~ radius + compactness + symmetry Res.Df RSS Df Sum of Sq F Pr(>F)
1 42 16387
2 43 17800 -1 -1412.7 3.6206 0.06394 . — Signif. codes: 0 ‘***’ 0.001 ”** 0.01 ”* 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table

Model 1: time ~ radius + smoothness + compactness + symmetry Model 2: time ~ radius + smoothness + symmetry Res.Df RSS Df Sum of Sq F Pr(>F) 1 42 16387
2 43 17429 -1 -1041.3 2.6687 0.1098

Analysis of Variance Table

Model 1: time ~ radius + smoothness + compactness + symmetry Model 2: time ~ radius + smoothness + compactness Res.Df RSS Df Sum of Sq F Pr(>F) 1 42 16387
2 43 17380 -1 -992.12 2.5427 0.1183

Confidence Intervals for Slope Coefficients confint(lm(time ~ radius + smoothness + compactness + symmetry), level = 0.95) 2.5 % 97.5 % (Intercept) -129.162141 73.7718717 radius -4.181454 -0.2057056 smoothness -41.115090 1398.3204814 compactness -360.657852 37.9705708 symmetry -65.263969 556.7574336