**STAT5120, Allen Baumgarten, Building the Regression Model, Pt. I**

1. In each case below, indicate whether a more flexible model (like a high-degree polynomial) would tend to perform better or worse on a test set than a more inflexible one (like a low-degree polynomial). Explain your reasoning.

(a) The number of observations n is large and the number of predictor variables is small. A flexible, higher-order polynomial model (HOPM) would perform better. One reason is that in having a smaller number of predictors we could try some higher-order terms without taxing our computing resources (if you happen to own IBM’s Watson supercomputer, this would be a non-issue). Flexible models are advantageous in that they allow us to explore other possibilities for fitting the data. With smaller numbers of predictors, we can do this.

(b) There is a very large number of predictor variables, but a small number of observations n. In this case, we might opt for an Inflexible, low-degree model (ILDM) since having a larger number of predictors, and then multiplying them by adding many high-order terms, would be cumbersome at best. Better to select, or at least start with, a simpler modeling scenario.

(c) The apparent dependence of the response on the predictor variables is highly non-linear. This is (even for me) a “No-brainer:” if we detect a curvilinear trend in the response vs. predictors, we would need a HOPM with its higher-order terms to fine-tune and fit the curved response in the data better than a ILDM would.

(d) The variance of the error terms is very high. If the variance of the errors is very high, perhaps even higher than the odds of the Dolphins NOT making the Play Offs, then we would likely need a HOPM to fit this data. Polynomials could fit to the response better and reduce variance in the error terms.[[1]](#footnote-1)

2. Answer the following.

(a) True or false, and explain. If a regression model has high bias, it is unlikely that collecting more data to train/build the model will increase its performance on a validation or test set (with respect to, say, SSE; MSE, or R2). TRUE. Bias happens when a model is “under-specified” due to the omission of important predictor variables. When a model lacks one or more important predictor variables, and n is already large enough, adding more observations will probably not do much more than reduce the standard errors of the terms.

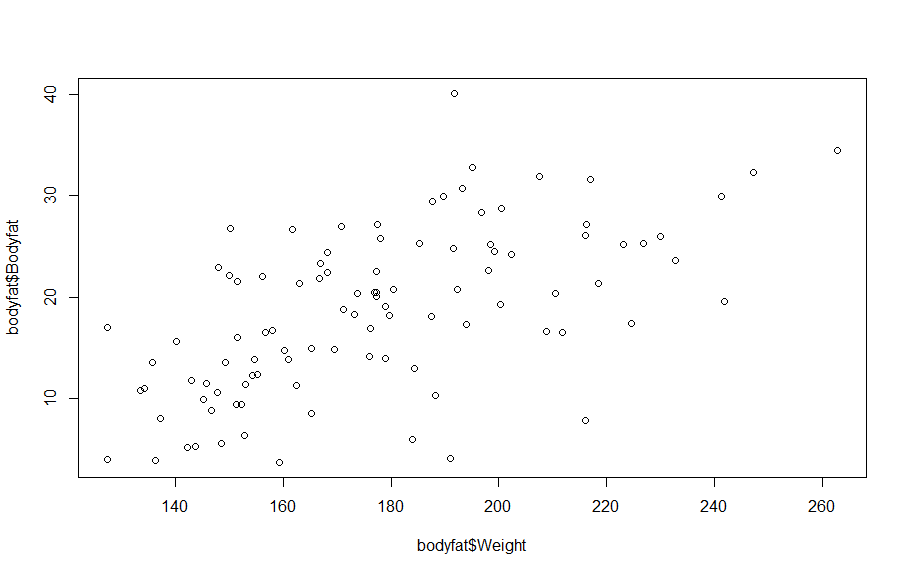
Commenting on bias due to variable selection, Rawlings et al state, “Assume that the correct model involves *t* independent variables but that a subset of *p* variables (chosen randomly or on the basis of external information) is used in the regression equation. Let Xp and βp denote submatrices of X and β that relate to the *p* selected variables. β̂p denotes the least squares estimate of βp obtained from the p-variate subset model. Similarly, Ŷpi, Ŷpredpi, and MS(Resp) denote the estimated mean for the *i*th observation, the prediction for the *i*th observation, and the means squared residual, respectively, obtained from the p-variate subset model. Hocking (1976) summarizes the following properties: 1. MS(Resp) is a *positively* biased estimate of σ2 unless the true regression coefficients for all deleted variables are zero. 2. β̂p is a biased estimate of βp and Ŷpi is a biased estimate of ε(Yi) unless the true regression coefficient for each deleted variable is zero or, in the case of β̂p, each deleted variable is orthogonal to the *p* retained variables. 3. β̂p, Ŷpi, and Ŷpredpi are generally *less* variable than the corresponding statistics obtained from the t-variate model. 4. There are conditions under which the mean squared errors (variance plus squared bias) of β̂p, Ŷpi, and Ŷpredpi are smaller than the variances of the estimates obtained under the t-variate model.”[[2]](#footnote-2)

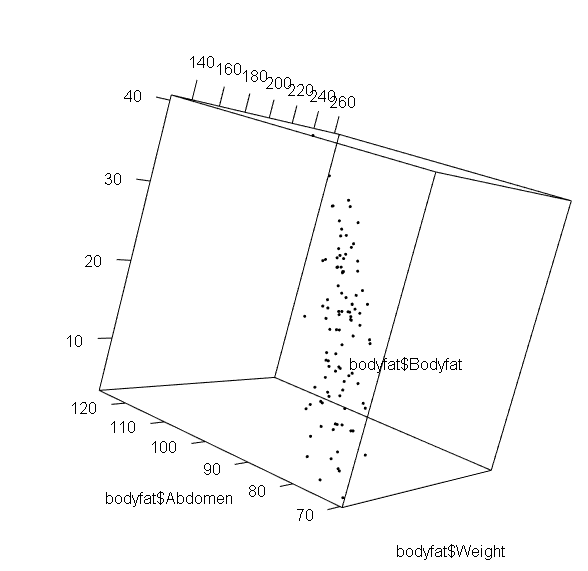
(b) What do you think will happen to the variance of an over-fitted regression model as the size of the training set increases? Variance will be very high. This is because with an over-fitted model, the already-existing data points are highly anticipated by the model coefficients. However, when new data points which do NOT have the same values as the originals are entered into the model, the model “misses” these new points and variance occurs in the (new) errors that result.

(c) Suppose you build polynomial regressions in one variable (y^ = 0 + p􀀀1k=1 kxk), and you want to choose the polynomial degree by evaluating the performance of your models (using say, MSE) on a validation or test set. Once you choose the final model, would you expect the test MSE to be higher or lower than the training MSE? Why? The MSE measures the mean squared errors of the fitted response vs. the actual responses. When we train and then test a model, we used a majority of our data, say, 75% of the observations to train the model, that is, estimate our coefficients, and then use the remaining 25% to test how those estimates work. If we have estimated well and iterated some ideas for the best model fit, we would expect our MSE to be lower than the training on the training set. If MSE was higher we would simple stick with that fitted result rather than pick a final model with a higher MSE.

(d) Suppose when you add flexibility to your model by adding higher order terms or more predictor variables that you begin to see the test or validation error (MSE or SSE evaluated on the test or validation set) begin to increase away from the training error. What kind of a problem are your models experiencing? One possibility would be that we have a collinearity problem developing.

3. Simpson's Paradox.[[3]](#footnote-3) Open the *Bodyfat* data in R and refer to the R examples at the beginning of this set of notes. Plot Bodyfat vs Weight. The plot seems to indicate body fat percentage tends to increase as weight increases. Now make a 3D plot of Bodyfat vs. Weight and Abdomen.





Describe how Simpson's Paradox is apparent in this context. Recall how the sign of the Weight coefficient changed when we included the Abdomen variable in the regression model. Simpson’s Paradox is the phenomenon of seeing a trend in a data relationship seemingly reverse trend as other variates are considered. In the Encyclopedia Britannica we read, “Simpson’s paradox, also called Yule-Simpson effect, in statistics, an effect that occurs when the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables. Simpson’s paradox is important for three critical reasons. First, people often expect statistical relationships to be immutable. They often are not. The relationship between two variables might increase, decrease, or even change direction depending on the set of variables being controlled. Second, Simpson’s paradox is not simply an obscure phenomenon of interest only to a small group of statisticians. Simpson’s paradox is actually one of a large class of association paradoxes. Third, Simpson’s paradox reminds researchers that causal inferences, particularly in nonexperimental studies, can be hazardous. Uncontrolled and even unobserved variables that would eliminate or reverse the association observed between two variables might exist.”[[4]](#footnote-4)

We see in this example set how the data seems to curve back around after we add the *abdomen* variable. Initially, as *weight* increases so, too does *bodyfat*. But almost paradoxically after the *abdomen* variable is introduced, *bodyfat* seems to curve back around.[[5]](#footnote-5)

4. Open the *prostate* data from the faraway package. Model l*psa* as the response and all other variables as predictors. Implement the methods of best subsets as well as forward and backward stepwise selection to determine “best" models, comparing them based on the performance measures Cp, BIC, etc.

Rawlings et al remark that, “Alternative variable selection methods have been developed that identify good (although not necessarily the best) subset models, with considerably less computing than is required for all possible regressions. These methods are referred to as stepwise regression methods. The subset models are identified sequentially by adding or deleting, depending on the method, the one variable that has the greatest impact on the residual sum of squares. These stepwise methods are not guaranteed to find the ‘best’ subset for each subset size, and the results produced by different methods may not agree with each other.”[[6]](#footnote-6)

“Backward” elimination starts with all variables and systematically eliminates variables and testing the fit of the model with each iteration. “Forward” stepwise is the same process but in the opposite direction with one variable being added to the sequentially.[[7]](#footnote-7)

Backward elimination was done and resulting diagnostics examined, beginning with a full model:

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi + prostate$lcp +

prostate$gleason + prostate$pgg45)

Residuals:

Min 1Q Median 3Q Max

-1.7331 -0.3713 -0.0170 0.4141 1.6381

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.669337 1.296387 0.516 0.60693

prostate$lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*

prostate$lweight 0.454467 0.170012 2.673 0.00896 \*\*

prostate$age -0.019637 0.011173 -1.758 0.08229 .

prostate$lbph 0.107054 0.058449 1.832 0.07040 .

prostate$svi 0.766157 0.244309 3.136 0.00233 \*\*

prostate$lcp -0.105474 0.091013 -1.159 0.24964

prostate$gleason 0.045142 0.157465 0.287 0.77503

prostate$pgg45 0.004525 0.004421 1.024 0.30886

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7084 on 88 degrees of freedom

Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234

F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi + prostate$lcp +

prostate$gleason)

Residuals:

Min 1Q Median 3Q Max

-1.78803 -0.36933 0.00302 0.43436 1.62160

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.02416 1.13313 0.021 0.98304

prostate$lcavol 0.57471 0.08712 6.597 2.92e-09 \*\*\*

prostate$lweight 0.45260 0.17005 2.662 0.00923 \*\*

prostate$age -0.01812 0.01108 -1.636 0.10542

prostate$lbph 0.10886 0.05844 1.863 0.06579 .

prostate$svi 0.79783 0.24241 3.291 0.00143 \*\*

prostate$lcp -0.07488 0.08599 -0.871 0.38619

prostate$gleason 0.14591 0.12292 1.187 0.23837

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7086 on 89 degrees of freedom

Multiple R-squared: 0.6506, Adjusted R-squared: 0.6232

F-statistic: 23.68 on 7 and 89 DF, p-value: < 2.2e-16

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi + prostate$lcp)

Residuals:

Min 1Q Median 3Q Max

-1.82853 -0.40741 0.01695 0.47159 1.59040

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.93487 0.83577 1.119 0.26630

prostate$lcavol 0.58765 0.08663 6.783 1.2e-09 \*\*\*

prostate$lweight 0.41808 0.16792 2.490 0.01462 \*

prostate$age -0.01511 0.01081 -1.398 0.16546

prostate$lbph 0.11381 0.05842 1.948 0.05452 .

prostate$svi 0.78256 0.24261 3.226 0.00175 \*\*

prostate$lcp -0.04118 0.08135 -0.506 0.61392

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7102 on 90 degrees of freedom

Multiple R-squared: 0.6451, Adjusted R-squared: 0.6215

F-statistic: 27.27 on 6 and 90 DF, p-value: < 2.2e-16

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi)

Residuals:

Min 1Q Median 3Q Max

-1.83505 -0.39396 0.00414 0.46336 1.57888

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.95100 0.83175 1.143 0.255882

prostate$lcavol 0.56561 0.07459 7.583 2.77e-11 \*\*\*

prostate$lweight 0.42369 0.16687 2.539 0.012814 \*

prostate$age -0.01489 0.01075 -1.385 0.169528

prostate$lbph 0.11184 0.05805 1.927 0.057160 .

prostate$svi 0.72095 0.20902 3.449 0.000854 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7073 on 91 degrees of freedom

Multiple R-squared: 0.6441, Adjusted R-squared: 0.6245

F-statistic: 32.94 on 5 and 91 DF, p-value: < 2.2e-16

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph)

Residuals:

Min 1Q Median 3Q Max

-1.4885 -0.4241 -0.0001 0.4031 1.8073

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.73074 0.87703 0.833 0.4069

prostate$lcavol 0.69854 0.06754 10.343 <2e-16 \*\*\*

prostate$lweight 0.45770 0.17617 2.598 0.0109 \*

prostate$age -0.01371 0.01137 -1.206 0.2309

prostate$lbph 0.08404 0.06080 1.382 0.1702

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.748 on 92 degrees of freedom

Multiple R-squared: 0.5976, Adjusted R-squared: 0.5801

F-statistic: 34.15 on 4 and 92 DF, p-value: < 2.2e-16

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age)

Residuals:

Min 1Q Median 3Q Max

-1.60894 -0.44897 -0.02805 0.45602 1.91756

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.146917 0.772374 0.190 0.84956

prostate$lcavol 0.687819 0.067418 10.202 < 2e-16 \*\*\*

prostate$lweight 0.549941 0.163838 3.357 0.00114 \*\*

prostate$age -0.009486 0.011003 -0.862 0.39082

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7517 on 93 degrees of freedom

Multiple R-squared: 0.5892, Adjusted R-squared: 0.576

F-statistic: 44.47 on 3 and 93 DF, p-value: < 2.2e-16

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight)

Residuals:

Min 1Q Median 3Q Max

-1.61965 -0.50778 -0.02095 0.52291 1.89885

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.30262 0.56904 -0.532 0.59612

prostate$lcavol 0.67753 0.06626 10.225 < 2e-16 \*\*\*

prostate$lweight 0.51095 0.15726 3.249 0.00161 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7506 on 94 degrees of freedom

Multiple R-squared: 0.5859, Adjusted R-squared: 0.5771

F-statistic: 66.51 on 2 and 94 DF, p-value: < 2.2e-16

> summary(regmod\_1)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol)

Residuals:

Min 1Q Median 3Q Max

-1.67625 -0.41648 0.09859 0.50709 1.89673

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.50730 0.12194 12.36 <2e-16 \*\*\*

prostate$lcavol 0.71932 0.06819 10.55 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7875 on 95 degrees of freedom

Multiple R-squared: 0.5394, Adjusted R-squared: 0.5346

F-statistic: 111.3 on 1 and 95 DF, p-value: < 2.2e-16

Examining these models, we can ascertain that by including and excluding individual variables, the model using cavol, a variable commonly seen to have a significant p-value in all iterations, has the lowest R2adj of all of them. This model might be the best choice.

5. Open the *divusa* data from the faraway package. Model *divorce* as the response and all other variables as predictors. Implement the methods of best subsets as well as forward and backward stepwise selection to determine “best" models, comparing them based on the performance measures

Cp, BIC, etc.

Opened and loaded the divusa dataset and used the glm() function to generate the AIC statistics besides the R2adj and SSE:

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab + divusa$marriage + divusa$birth + divusa$military)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.9087 -0.9212 -0.0935 0.7447 3.4689

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 380.14761 99.20371 3.832 0.000274 \*\*\*

divusa$year -0.20312 0.05333 -3.809 0.000297 \*\*\*

divusa$unemployed -0.04933 0.05378 -0.917 0.362171

divusa$femlab 0.80793 0.11487 7.033 1.09e-09 \*\*\*

divusa$marriage 0.14977 0.02382 6.287 2.42e-08 \*\*\*

divusa$birth -0.11695 0.01470 -7.957 2.19e-11 \*\*\*

divusa$military -0.04276 0.01372 -3.117 0.002652 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.288535)

Null deviance: 2442.5 on 76 degrees of freedom

Residual deviance: 160.2 on 70 degrees of freedom

AIC: 290.93

Number of Fisher Scoring iterations: 2

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab + divusa$marriage + divusa$birth)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.3483 -0.9429 0.0430 0.8912 3.8625

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 285.53474 100.07330 2.853 0.00567 \*\*

divusa$year -0.15213 0.05378 -2.829 0.00607 \*\*

divusa$unemployed -0.02846 0.05654 -0.503 0.61626

divusa$femlab 0.69620 0.11564 6.021 6.92e-08 \*\*\*

divusa$marriage 0.12832 0.02416 5.310 1.20e-06 \*\*\*

divusa$birth -0.11711 0.01557 -7.520 1.29e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.56942)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 182.43 on 71 degrees of freedom

AIC: 298.93

Number of Fisher Scoring iterations: 2

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab + divusa$marriage)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.2553 -1.5614 -0.6785 1.2788 4.9885

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 454.99121 129.77695 3.506 0.000787 \*\*\*

divusa$year -0.25045 0.06944 -3.607 0.000568 \*\*\*

divusa$unemployed 0.18883 0.06468 2.919 0.004681 \*\*

divusa$femlab 1.01139 0.14345 7.051 8.91e-10 \*\*\*

divusa$marriage 0.11355 0.03206 3.542 0.000700 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.551885)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 327.74 on 72 degrees of freedom

AIC: 342.04

Number of Fisher Scoring iterations: 2

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.2112 -1.6888 -0.0618 1.0481 8.6792

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 324.83352 133.95070 2.425 0.0178 \*

divusa$year -0.17447 0.07107 -2.455 0.0165 \*

divusa$unemployed 0.04437 0.05404 0.821 0.4143

divusa$femlab 0.77071 0.13596 5.669 2.7e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.27198)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 384.85 on 73 degrees of freedom

AIC: 352.41

Number of Fisher Scoring iterations: 2

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.7892 -1.8315 0.1415 1.5944 7.2843

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.224e+02 2.838e+01 -14.886 <2e-16 \*\*\*

divusa$year 2.225e-01 1.443e-02 15.416 <2e-16 \*\*\*

divusa$unemployed -5.241e-03 6.356e-02 -0.082 0.935

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.489987)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 554.26 on 74 degrees of freedom

AIC: 378.5

Number of Fisher Scoring iterations: 2

Call:

glm(formula = divusa$divorce ~ divusa$year)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.7828 -1.8092 0.1592 1.6292 7.3048

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -422.97530 27.29465 -15.50 <2e-16 \*\*\*

divusa$year 0.22280 0.01394 15.98 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.3908)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 554.31 on 75 degrees of freedom

AIC: 376.51

Number of Fisher Scoring iterations: 2

Keeping year and femlab results in a comparably high AIC score along with an R2adj = .8367. This iteration avoids including the insignificant variables while maintaining a nice AIC score and R2adj

> summary(regmod\_2b <-glm(divusa$divorce ~ divusa$year+divusa$femlab))

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$femlab)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.2545 -1.6205 -0.0812 0.9239 8.4775

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 312.1191 132.7594 2.351 0.0214 \*

divusa$year -0.1675 0.0704 -2.379 0.0200 \*

divusa$femlab 0.7526 0.1339 5.622 3.16e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.248768)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 388.41 on 74 degrees of freedom

AIC: 351.12

Number of Fisher Scoring iterations: 2

6. Load and read the documentation for the *Boston* data set from the MASS package. We want to build a model to predict the per capita crime rate. Implement the methods of (1) best subsets, (2) forward and (3) backward stepwise selection, (4) LOOCV, and (5) k-fold cross validation to build models (use k = 10). Compare and discus the models.

Loaded the Boston dataset and ran iterations of models to find the best fit. Begain with a full model to find statistically significant variables:

Residuals:

Min 1Q Median 3Q Max

-10.129 -2.031 -0.438 1.000 74.887

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.845372 6.645791 3.287 0.00108 \*\*

Boston$zn 0.048614 0.018631 2.609 0.00935 \*\*

Boston$indus -0.053696 0.083334 -0.644 0.51965

Boston$chas -0.742473 1.182266 -0.628 0.53029

Boston$nox -10.599854 5.282238 -2.007 0.04533 \*

Boston$rm 0.137741 0.588242 0.234 0.81496

Boston$age 0.010751 0.017065 0.630 0.52897

Boston$dis -1.040772 0.280481 -3.711 0.00023 \*\*\*

Boston$rad 0.609621 0.087264 6.986 9.21e-12 \*\*\*

Boston$tax -0.004453 0.005149 -0.865 0.38756

Boston$ptratio -0.301584 0.185884 -1.622 0.10535

Boston$black -0.008034 0.003668 -2.190 0.02897 \*

Boston$medv -0.243237 0.054450 -4.467 9.84e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.451 on 493 degrees of freedom

Multiple R-squared: 0.4509, Adjusted R-squared: 0.4376

F-statistic: 33.74 on 12 and 493 DF, p-value: < 2.2e-16

Six variables were seen to have statistical significance in relation to the predictor. A reduced model was run to see how these variables alone would fair, using a backward elimination approach:

Call:

lm(formula = Boston$crim ~ Boston$zn + Boston$nox + Boston$dis +

Boston$rad + Boston$black + Boston$medv)

Residuals:

Min 1Q Median 3Q Max

-10.240 -1.915 -0.376 0.852 75.438

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 14.642639 3.709443 3.947 9.04e-05 \*\*\*

Boston$zn 0.053963 0.017305 3.118 0.001923 \*\*

Boston$nox -9.238768 4.477580 -2.063 0.039597 \*

Boston$dis -0.992811 0.255075 -3.892 0.000113 \*\*\*

Boston$rad 0.499838 0.044036 11.351 < 2e-16 \*\*\*

Boston$black -0.008711 0.003612 -2.412 0.016237 \*

Boston$medv -0.195990 0.037685 -5.201 2.90e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.452 on 499 degrees of freedom

Multiple R-squared: 0.444, Adjusted R-squared: 0.4373

F-statistic: 66.42 on 6 and 499 DF, p-value: < 2.2e-16

It was found that these six variables didn’t result in a stronger model than even the full one but did nevertheless provide at least some explanatory power.

APPENDIX: R SCRIPTS IMPLEMENTED

**Question 3:**

> library(readxl)

> bodyfat <- read\_excel("c:/Users/baumgaral/Data/bodyfat.xlsx")

> head(bodyfat)

# A tibble: 6 x 10

Bodyfat Age Weight Height Neck Chest Abdomen Ankle Biceps Wrist

<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 32.3 41. 247. 73.5 42.1 117. 116. 26.3 37.3 19.7

2 22.5 31. 177. 71.5 36.2 101. 92.4 24.6 30.1 18.2

3 22.0 42. 156. 69.0 35.5 97.8 86.0 24.0 31.2 17.4

4 12.3 23. 154. 67.8 36.2 93.1 85.2 21.9 32.0 17.1

5 20.5 46. 177. 70.0 37.2 99.7 95.6 22.5 29.1 17.7

> plot(bodyfat$Bodyfat ~ bodyfat$Weight)

> library(scatterplot3d)

> library(readxl)

> plot3d(bodyfat$Weight, bodyfat$Abdomen, bodyfat$Bodyfat)

**Question 4:**

> library(faraway)

> head(prostate)

lcavol lweight age lbph svi lcp gleason pgg45 lpsa

1 -0.5798185 2.7695 50 -1.386294 0 -1.38629 6 0 -0.43078

2 -0.9942523 3.3196 58 -1.386294 0 -1.38629 6 0 -0.16252

3 -0.5108256 2.6912 74 -1.386294 0 -1.38629 7 20 -0.16252

4 -1.2039728 3.2828 58 -1.386294 0 -1.38629 6 0 -0.16252

5 0.7514161 3.4324 62 -1.386294 0 -1.38629 6 0 0.37156

6 -1.0498221 3.2288 50 -1.386294 0 -1.38629 6 0 0.76547

> regmod1\_full <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight+prostate$age+prostate$lbph+prostate$svi+prostate$lcp+prostate$gleason+prostate$pgg45)

> regmod1\_6 <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight+prostate$age+prostate$lbph+prostate$svi+prostate$lcp+prostate$gleason)

> regmod1\_5 <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight+prostate$age+prostate$lbph+prostate$svi+prostate$lcp)

> regmod1\_4 <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight+prostate$age+prostate$lbph+prostate$svi)

> regmod1\_3 <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight+prostate$age+prostate$lbph)

> regmod\_7 <- regmod1\_6

> regmod\_6 <- regmod1\_5

> regmod\_5 <- regmod1\_4

> regmod\_4 <- regmod1\_3

> regmod\_3 <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight+prostate$age)

> regmod\_2 <- lm(prostate$lpsa ~ prostate$lcavol+prostate$lweight)

> regmod\_1 <- lm(prostate$lpsa ~ prostate$lcavol)

> summary(regmod1\_full)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi + prostate$lcp +

prostate$gleason + prostate$pgg45)

Residuals:

Min 1Q Median 3Q Max

-1.7331 -0.3713 -0.0170 0.4141 1.6381

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.669337 1.296387 0.516 0.60693

prostate$lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*

prostate$lweight 0.454467 0.170012 2.673 0.00896 \*\*

prostate$age -0.019637 0.011173 -1.758 0.08229 .

prostate$lbph 0.107054 0.058449 1.832 0.07040 .

prostate$svi 0.766157 0.244309 3.136 0.00233 \*\*

prostate$lcp -0.105474 0.091013 -1.159 0.24964

prostate$gleason 0.045142 0.157465 0.287 0.77503

prostate$pgg45 0.004525 0.004421 1.024 0.30886

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7084 on 88 degrees of freedom

Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234

F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

> summary(regmod\_7)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi + prostate$lcp +

prostate$gleason)

Residuals:

Min 1Q Median 3Q Max

-1.78803 -0.36933 0.00302 0.43436 1.62160

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.02416 1.13313 0.021 0.98304

prostate$lcavol 0.57471 0.08712 6.597 2.92e-09 \*\*\*

prostate$lweight 0.45260 0.17005 2.662 0.00923 \*\*

prostate$age -0.01812 0.01108 -1.636 0.10542

prostate$lbph 0.10886 0.05844 1.863 0.06579 .

prostate$svi 0.79783 0.24241 3.291 0.00143 \*\*

prostate$lcp -0.07488 0.08599 -0.871 0.38619

prostate$gleason 0.14591 0.12292 1.187 0.23837

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7086 on 89 degrees of freedom

Multiple R-squared: 0.6506, Adjusted R-squared: 0.6232

F-statistic: 23.68 on 7 and 89 DF, p-value: < 2.2e-16

> summary(regmod\_6)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi + prostate$lcp)

Residuals:

Min 1Q Median 3Q Max

-1.82853 -0.40741 0.01695 0.47159 1.59040

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.93487 0.83577 1.119 0.26630

prostate$lcavol 0.58765 0.08663 6.783 1.2e-09 \*\*\*

prostate$lweight 0.41808 0.16792 2.490 0.01462 \*

prostate$age -0.01511 0.01081 -1.398 0.16546

prostate$lbph 0.11381 0.05842 1.948 0.05452 .

prostate$svi 0.78256 0.24261 3.226 0.00175 \*\*

prostate$lcp -0.04118 0.08135 -0.506 0.61392

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7102 on 90 degrees of freedom

Multiple R-squared: 0.6451, Adjusted R-squared: 0.6215

F-statistic: 27.27 on 6 and 90 DF, p-value: < 2.2e-16

> summary(regmod\_5)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph + prostate$svi)

Residuals:

Min 1Q Median 3Q Max

-1.83505 -0.39396 0.00414 0.46336 1.57888

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.95100 0.83175 1.143 0.255882

prostate$lcavol 0.56561 0.07459 7.583 2.77e-11 \*\*\*

prostate$lweight 0.42369 0.16687 2.539 0.012814 \*

prostate$age -0.01489 0.01075 -1.385 0.169528

prostate$lbph 0.11184 0.05805 1.927 0.057160 .

prostate$svi 0.72095 0.20902 3.449 0.000854 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7073 on 91 degrees of freedom

Multiple R-squared: 0.6441, Adjusted R-squared: 0.6245

F-statistic: 32.94 on 5 and 91 DF, p-value: < 2.2e-16

> summary(regmod\_4)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age + prostate$lbph)

Residuals:

Min 1Q Median 3Q Max

-1.4885 -0.4241 -0.0001 0.4031 1.8073

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.73074 0.87703 0.833 0.4069

prostate$lcavol 0.69854 0.06754 10.343 <2e-16 \*\*\*

prostate$lweight 0.45770 0.17617 2.598 0.0109 \*

prostate$age -0.01371 0.01137 -1.206 0.2309

prostate$lbph 0.08404 0.06080 1.382 0.1702

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.748 on 92 degrees of freedom

Multiple R-squared: 0.5976, Adjusted R-squared: 0.5801

F-statistic: 34.15 on 4 and 92 DF, p-value: < 2.2e-16

> summary(regmod\_3)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight +

prostate$age)

Residuals:

Min 1Q Median 3Q Max

-1.60894 -0.44897 -0.02805 0.45602 1.91756

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.146917 0.772374 0.190 0.84956

prostate$lcavol 0.687819 0.067418 10.202 < 2e-16 \*\*\*

prostate$lweight 0.549941 0.163838 3.357 0.00114 \*\*

prostate$age -0.009486 0.011003 -0.862 0.39082

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7517 on 93 degrees of freedom

Multiple R-squared: 0.5892, Adjusted R-squared: 0.576

F-statistic: 44.47 on 3 and 93 DF, p-value: < 2.2e-16

> summary(regmod\_2)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol + prostate$lweight)

Residuals:

Min 1Q Median 3Q Max

-1.61965 -0.50778 -0.02095 0.52291 1.89885

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.30262 0.56904 -0.532 0.59612

prostate$lcavol 0.67753 0.06626 10.225 < 2e-16 \*\*\*

prostate$lweight 0.51095 0.15726 3.249 0.00161 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7506 on 94 degrees of freedom

Multiple R-squared: 0.5859, Adjusted R-squared: 0.5771

F-statistic: 66.51 on 2 and 94 DF, p-value: < 2.2e-16

> summary(regmod\_1)

Call:

lm(formula = prostate$lpsa ~ prostate$lcavol)

Residuals:

Min 1Q Median 3Q Max

-1.67625 -0.41648 0.09859 0.50709 1.89673

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.50730 0.12194 12.36 <2e-16 \*\*\*

prostate$lcavol 0.71932 0.06819 10.55 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7875 on 95 degrees of freedom

Multiple R-squared: 0.5394, Adjusted R-squared: 0.5346

F-statistic: 111.3 on 1 and 95 DF, p-value: < 2.2e-16

**Question 5:**

> summary(regmod\_full)

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab + divusa$marriage + divusa$birth + divusa$military)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.9087 -0.9212 -0.0935 0.7447 3.4689

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 380.14761 99.20371 3.832 0.000274 \*\*\*

divusa$year -0.20312 0.05333 -3.809 0.000297 \*\*\*

divusa$unemployed -0.04933 0.05378 -0.917 0.362171

divusa$femlab 0.80793 0.11487 7.033 1.09e-09 \*\*\*

divusa$marriage 0.14977 0.02382 6.287 2.42e-08 \*\*\*

divusa$birth -0.11695 0.01470 -7.957 2.19e-11 \*\*\*

divusa$military -0.04276 0.01372 -3.117 0.002652 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.288535)

Null deviance: 2442.5 on 76 degrees of freedom

Residual deviance: 160.2 on 70 degrees of freedom

AIC: 290.93

Number of Fisher Scoring iterations: 2

> summary(regmod\_5)

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab + divusa$marriage + divusa$birth)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.3483 -0.9429 0.0430 0.8912 3.8625

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 285.53474 100.07330 2.853 0.00567 \*\*

divusa$year -0.15213 0.05378 -2.829 0.00607 \*\*

divusa$unemployed -0.02846 0.05654 -0.503 0.61626

divusa$femlab 0.69620 0.11564 6.021 6.92e-08 \*\*\*

divusa$marriage 0.12832 0.02416 5.310 1.20e-06 \*\*\*

divusa$birth -0.11711 0.01557 -7.520 1.29e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.56942)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 182.43 on 71 degrees of freedom

AIC: 298.93

Number of Fisher Scoring iterations: 2

> summary(regmod\_4)

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab + divusa$marriage)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.2553 -1.5614 -0.6785 1.2788 4.9885

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 454.99121 129.77695 3.506 0.000787 \*\*\*

divusa$year -0.25045 0.06944 -3.607 0.000568 \*\*\*

divusa$unemployed 0.18883 0.06468 2.919 0.004681 \*\*

divusa$femlab 1.01139 0.14345 7.051 8.91e-10 \*\*\*

divusa$marriage 0.11355 0.03206 3.542 0.000700 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.551885)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 327.74 on 72 degrees of freedom

AIC: 342.04

Number of Fisher Scoring iterations: 2

> summary(regmod\_3)

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed + divusa$femlab)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.2112 -1.6888 -0.0618 1.0481 8.6792

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 324.83352 133.95070 2.425 0.0178 \*

divusa$year -0.17447 0.07107 -2.455 0.0165 \*

divusa$unemployed 0.04437 0.05404 0.821 0.4143

divusa$femlab 0.77071 0.13596 5.669 2.7e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.27198)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 384.85 on 73 degrees of freedom

AIC: 352.41

Number of Fisher Scoring iterations: 2

> summary(regmod\_2)

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$unemployed)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.7892 -1.8315 0.1415 1.5944 7.2843

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.224e+02 2.838e+01 -14.886 <2e-16 \*\*\*

divusa$year 2.225e-01 1.443e-02 15.416 <2e-16 \*\*\*

divusa$unemployed -5.241e-03 6.356e-02 -0.082 0.935

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.489987)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 554.26 on 74 degrees of freedom

AIC: 378.5

Number of Fisher Scoring iterations: 2

> summary(regmod\_1)

Call:

glm(formula = divusa$divorce ~ divusa$year)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.7828 -1.8092 0.1592 1.6292 7.3048

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -422.97530 27.29465 -15.50 <2e-16 \*\*\*

divusa$year 0.22280 0.01394 15.98 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.3908)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 554.31 on 75 degrees of freedom

AIC: 376.51

Number of Fisher Scoring iterations: 2

> summary(regmod\_2b <-glm(divusa$divorce ~ divusa$year+divusa$femlab))

Call:

glm(formula = divusa$divorce ~ divusa$year + divusa$femlab)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.2545 -1.6205 -0.0812 0.9239 8.4775

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 312.1191 132.7594 2.351 0.0214 \*

divusa$year -0.1675 0.0704 -2.379 0.0200 \*

divusa$femlab 0.7526 0.1339 5.622 3.16e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.248768)

Null deviance: 2442.53 on 76 degrees of freedom

Residual deviance: 388.41 on 74 degrees of freedom

AIC: 351.12

Number of Fisher Scoring iterations: 2

> summary(regmod\_2b <-lm(divusa$divorce ~ divusa$year+divusa$femlab))

Call:

lm(formula = divusa$divorce ~ divusa$year + divusa$femlab)

Residuals:

Min 1Q Median 3Q Max

-3.2545 -1.6205 -0.0812 0.9239 8.4775

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 312.1191 132.7594 2.351 0.0214 \*

divusa$year -0.1675 0.0704 -2.379 0.0200 \*

divusa$femlab 0.7526 0.1339 5.622 3.16e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.291 on 74 degrees of freedom

Multiple R-squared: 0.841, Adjusted R-squared: 0.8367

F-statistic: 195.7 on 2 and 74 DF, p-value: < 2.2e-16

**Question 6:**

> library(MASS)

> head(Boston)

crim zn indus chas nox rm age dis rad tax ptratio black lstat medv

1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98 24.0

2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14 21.6

3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03 34.7

4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94 33.4

5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33 36.2

6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21 28.7

> regmod\_full <- lm(Boston$crim ~ Boston$zn + Boston$indus + Boston$chas + Boston$nox + Boston$rm + Boston$age + Boston$dis + Boston$dis + Boston$rad + Boston$tax + Boston$ptratio + Boston$black + Boston$black + Boston$medv)

> summary(regmod\_full)

Call:

lm(formula = Boston$crim ~ Boston$zn + Boston$indus + Boston$chas +

Boston$nox + Boston$rm + Boston$age + Boston$dis + Boston$dis +

Boston$rad + Boston$tax + Boston$ptratio + Boston$black +

Boston$black + Boston$medv)

Residuals:

Min 1Q Median 3Q Max

-10.129 -2.031 -0.438 1.000 74.887

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.845372 6.645791 3.287 0.00108 \*\*

Boston$zn 0.048614 0.018631 2.609 0.00935 \*\*

Boston$indus -0.053696 0.083334 -0.644 0.51965

Boston$chas -0.742473 1.182266 -0.628 0.53029

Boston$nox -10.599854 5.282238 -2.007 0.04533 \*

Boston$rm 0.137741 0.588242 0.234 0.81496

Boston$age 0.010751 0.017065 0.630 0.52897

Boston$dis -1.040772 0.280481 -3.711 0.00023 \*\*\*

Boston$rad 0.609621 0.087264 6.986 9.21e-12 \*\*\*

Boston$tax -0.004453 0.005149 -0.865 0.38756

Boston$ptratio -0.301584 0.185884 -1.622 0.10535

Boston$black -0.008034 0.003668 -2.190 0.02897 \*

Boston$medv -0.243237 0.054450 -4.467 9.84e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.451 on 493 degrees of freedom

Multiple R-squared: 0.4509, Adjusted R-squared: 0.4376

F-statistic: 33.74 on 12 and 493 DF, p-value: < 2.2e-16

> regmod\_reduced <- lm(Boston$crim ~ Boston$zn + Boston$nox + Boston$dis + Boston$rad + Boston$black + Boston$medv)

> summary(regmod\_reduced)

Call:

lm(formula = Boston$crim ~ Boston$zn + Boston$nox + Boston$dis +

Boston$rad + Boston$black + Boston$medv)

Residuals:

Min 1Q Median 3Q Max

-10.240 -1.915 -0.376 0.852 75.438

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 14.642639 3.709443 3.947 9.04e-05 \*\*\*

Boston$zn 0.053963 0.017305 3.118 0.001923 \*\*

Boston$nox -9.238768 4.477580 -2.063 0.039597 \*

Boston$dis -0.992811 0.255075 -3.892 0.000113 \*\*\*

Boston$rad 0.499838 0.044036 11.351 < 2e-16 \*\*\*

Boston$black -0.008711 0.003612 -2.412 0.016237 \*

Boston$medv -0.195990 0.037685 -5.201 2.90e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.452 on 499 degrees of freedom

Multiple R-squared: 0.444, Adjusted R-squared: 0.4373

F-statistic: 66.42 on 6 and 499 DF, p-value: < 2.2e-16

1. This wouldn’t help the Dolphins’ post-season prospects but that’s another discussion for another day. [↑](#footnote-ref-1)
2. Rawlings, John O., Sastry G. Pantula, and David A. Dickey, *Applied Regression Analysis: A Research Tool*, 2nd ed., (Springer-Verlag: New York, 1998), 208-09. [↑](#footnote-ref-2)
3. This is not to be confused with “Sampson’s Paradox” which goes something like, “What in the world did I ever see in that gal?” (sorry) [↑](#footnote-ref-3)
4. “Simpson’s Paradox,” *Encyclopedia Britannica*, accessed on 4/7/18 at: https://www.britannica.com/topic/Simpsons-paradox [↑](#footnote-ref-4)
5. This same paradox can be observed in the seemingly backward trajectory of the Miami Heat’s prospects for playoff standing after they introduced a great but aging Dwayne Wade back onto the roster. Great player but aging as all greats do. [↑](#footnote-ref-5)
6. Rawlings, 213. [↑](#footnote-ref-6)
7. Ibid. [↑](#footnote-ref-7)