STAT 8210 – Applied Regression Analysis

Homework 4

March 16, 2020

Connor Armstrong

***1. PROBLEM 7.18.*** *An article in the Journal of Pharmaceutical Sciences presents data on the observed mole fraction solubility of a solute at a constant temperature, along with x1 = dispersion partial solubility, x2 = dipolar partial solubility, and x3 = hydrogen bonding Hansen partial solubility. The response y is the negative logarithm of the mole fraction solubility.*

***7.18.a*** *Fit a complete quadratic model to the data. 7.18 Part a: Hints: Note that a complete quadratic model means that in addition to the intercept, each predictor is included with a linear effect, each predictor is included with a quadratic effect, and each interaction between the linear effects is included. Inclusion of quadratic effects in SAS and R is explicitly explained in the Chapter 7 notes; and remember that interactions in Proc Reg require creating the interaction terms in a data step and interactions in R use a colon (:) between the variable names. Expectations for writeup: You should copy and paste the computer output that would be used to construct the estimated regression equation, but you also need to write out the estimated regression equation.*

R Code:

# Question 7.18 Create quadratic variables and interaction terms

attach(q1)

names(q1)

x11 <- x1\*x1

x22 <- x2\*x2

x33 <- x3\*x3

x12 <- x1\*x2

x13 <- x1\*x3

x23 <- x2\*x3

head(q1)

# Question 7.18: Build model and check coefficients

q1.model <-lm(y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23)

summary(q1.model)

R Output:

Call:

lm(formula = y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 +

x23)

Residuals:

Min 1Q Median 3Q Max

-0.063213 -0.037282 -0.001113 0.016738 0.122539

Coefficients:

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

(Intercept) -1.769364 1.286976 -1.375 0.1881

x1 0.420799 0.294173 1.430 0.1718

x2 0.222453 0.130742 1.701 0.1082

x3 -0.127995 0.070245 -1.822 0.0872 .

x11 -0.019325 0.016797 -1.150 0.2668

x22 -0.007449 0.012048 -0.618 0.5451

x33 0.000824 0.001441 0.572 0.5754

x12 -0.019876 0.012037 -1.651 0.1182

x13 0.009151 0.007621 1.201 0.2473

x23 0.002576 0.007039 0.366 0.7192

---

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

Residual standard error: 0.06092 on 16 degrees of freedom

Multiple R-squared: 0.9169, Adjusted R-squared: 0.8702

F-statistic: 19.63 on 9 and 16 DF, p-value: 5.051e-07

The estimated regression equation is:

***7.18.b*** *Test for significance of regression, and construct t statistics for each model parameter. Interpret these results. Part b: Expectations for writeup: Provide the table with the F test statistic and p-value for significance of regression, restate the test statistic and p-value, and then state your conclusion in context. Then refer to the part a output which has the t-tests, and state for each of the estimated coefficients (except the intercept) whether or not it is statistically significant and then this conclusion in the context of explaining the negative logarithm of the mole fraction solubility.*

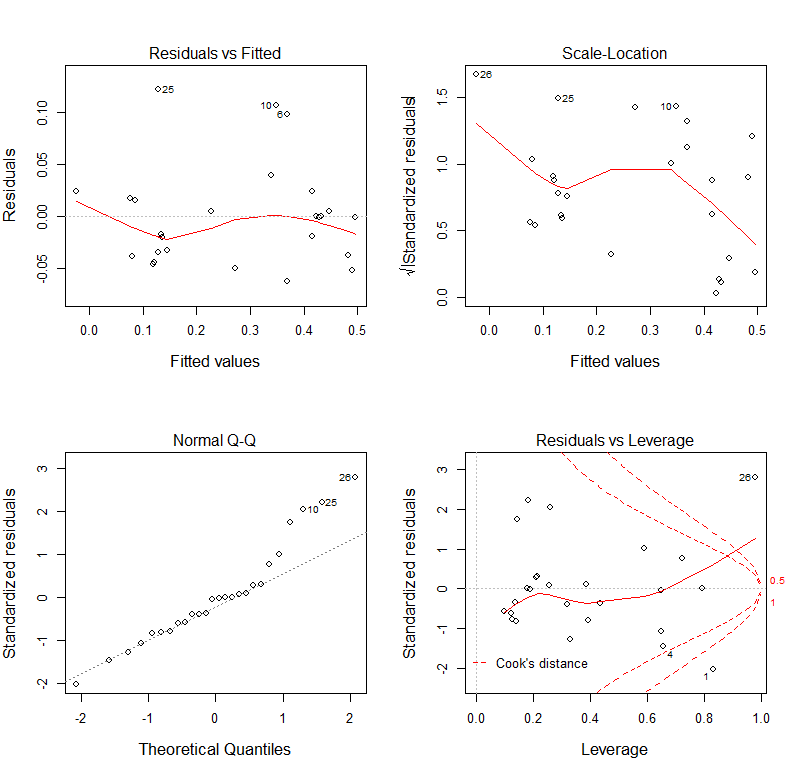
Given a p-value of 5.051e-07 for the test of significance of regression in the R output above, there is sufficient evidence to reject the null hypothesis that all coefficients equal zero. In other words, some of the variation in y can be explained by at least one of the independent variables (x’s).

The t-statistics for each model parameter are expressed in the R output above and are summarized in the table below.

|  |  |  |
| --- | --- | --- |
| **Term** | **T-Value** | **P-Value** |
| Constant | -1.37 | 0.188 |
| x1 | 1.43 | 0.172 |
| x2 | 1.7 | 0.108 |
| x3 | -1.82 | 0.087 |
| x1\*x1 | -1.15 | 0.267 |
| x2\*x2 | -0.62 | 0.545 |
| x3\*x3 | 0.57 | 0.575 |
| x1\*x2 | -1.65 | 0.118 |
| x1\*x3 | 1.2 | 0.247 |
| x2\*x3 | 0.37 | 0.719 |

None of the p-values are below the standard significance level of α=0.05. This is a partial or marginal test, given that for the multiple regression model each regressor depends on the others. These values explain the contribution of each regressor given the others in the model. A partial F-test can be implemented to determine which variables do not contribute meaningfully to the model. There is likely some multicollinearity present in the model which should be investigated.

***7.18.c*** *Plot residuals and comment on the model adequacy. Part c: Expectations for writeup: Include the residual plots and your comments on them. Pay attention to model assumptions and influential points. Make sure to conclude with whether or not you think the model is reasonable, and if not give recommendations for fixing it.*



The standard diagnostic plots for R are shown in the figures above. There are multiple causes for concern in these plots. There is a clear departure from normality in the quantiles plot, with clear outliers (Observations 26, 25, and 10) and potential leverage points (Observations 1 and 26). The data appears to have reasonably homogenous spread around the predicted value of the dependent variable y therefore it is unlikely that there is a major problem with the homogeneity of variance assumption.

There are fundamental assumption violations with this model. Multicollinearity is a likely cause for the assumption violations. VIF values will help to determine the extent to which certain variables are correlated while a partial F-test can provide insight into whether the removing variables of most concern improve the model.

***7.18.d*** *Use the extra sum of squares method (partial F test) to test the contribution of all second order terms to the model. Part d: Hints: Note the interaction effects and quadratic effects are both considered second-order terms. Expectations for writeup: Include the test statistic, p-value, and the conclusion in context.*

The following contains both R-code and output for the 6 alternative models with partial F-tests comparing the full quadratic model to the alternative models having removed one quadratic term from each. A summary of the conclusions of the 6 partial F-tests are displayed in Table 7.18.1.

> # Question 7.18 Partial F-tests for each quadratic term

> q1.model11 <-lm(y ~ x1 + x2 + x3 + x22 + x33 + x12 + x13 + x23)

> summary(q1.model11)

Call:

lm(formula = y ~ x1 + x2 + x3 + x22 + x33 + x12 + x13 + x23)

Residuals:

Min 1Q Median 3Q Max

-0.071697 -0.036541 -0.009632 0.014398 0.121903

Coefficients:

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

(Intercept) -0.320589 0.268134 -1.196 0.2483

x1 0.084149 0.030529 2.756 0.0135 \*

x2 0.239429 0.131137 1.826 0.0855 .

x3 -0.141600 0.069899 -2.026 0.0588 .

x22 -0.009727 0.011997 -0.811 0.4287

x33 0.001050 0.001441 0.728 0.4764

x12 -0.020495 0.012139 -1.688 0.1096

x13 0.010157 0.007643 1.329 0.2014

x23 0.002981 0.007097 0.420 0.6797

---

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

Residual standard error: 0.0615 on 17 degrees of freedom

Multiple R-squared: 0.9101, Adjusted R-squared: 0.8678

F-statistic: 21.51 on 8 and 17 DF, p-value: 1.906e-07

> anova(q1.model11, q1.model)#Partial F-test for removing x11

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x22 + x33 + x12 + x13 + x23

Model 2: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 0.064299

2 16 0.059386 1 0.0049128 1.3236 0.2668

Conclusion:

> q1.model22 <-lm(y ~ x1 + x2 + x3 + x11 + x33 + x12 + x13 + x23)

> summary(q1.model22)

Call:

lm(formula = y ~ x1 + x2 + x3 + x11 + x33 + x12 + x13 + x23)

Residuals:

Min 1Q Median 3Q Max

-0.057549 -0.034182 -0.007483 0.014451 0.123269

Coefficients:

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

(Intercept) -1.898916 1.246517 -1.523 0.146

x1 0.451132 0.284733 1.584 0.132

x2 0.170678 0.098560 1.732 0.101

x3 -0.107557 0.060843 -1.768 0.095 .

x11 -0.021031 0.016265 -1.293 0.213

x33 0.001129 0.001329 0.849 0.408

x12 -0.017311 0.011093 -1.561 0.137

x13 0.008037 0.007269 1.106 0.284

x23 -0.001035 0.003856 -0.268 0.792

---

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

Residual standard error: 0.05981 on 17 degrees of freedom

Multiple R-squared: 0.915, Adjusted R-squared: 0.8749

F-statistic: 22.86 on 8 and 17 DF, p-value: 1.202e-07

> anova(q1.model22, q1.model)#Partial F-test for removing x22

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x11 + x33 + x12 + x13 + x23

Model 2: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 0.060805

2 16 0.059386 1 0.0014187 0.3822 0.5451

>

> q1.model33 <-lm(y ~ x1 + x2 + x3 + x11 + x22 + x12 + x13 + x23)

> summary(q1.model33)

Call:

lm(formula = y ~ x1 + x2 + x3 + x11 + x22 + x12 + x13 + x23)

Residuals:

Min 1Q Median 3Q Max

-0.069253 -0.033820 -0.002736 0.022454 0.124276

Coefficients:

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

(Intercept) -1.836728 1.255945 -1.462 0.162

x1 0.440165 0.286373 1.537 0.143

x2 0.214776 0.127450 1.685 0.110

x3 -0.131142 0.068629 -1.911 0.073 .

x11 -0.020631 0.016308 -1.265 0.223

x22 -0.009806 0.011094 -0.884 0.389

x12 -0.018506 0.011560 -1.601 0.128

x13 0.009269 0.007466 1.241 0.231

x23 0.005734 0.004277 1.341 0.198

---

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

Residual standard error: 0.05971 on 17 degrees of freedom

Multiple R-squared: 0.9153, Adjusted R-squared: 0.8754

F-statistic: 22.95 on 8 and 17 DF, p-value: 1.169e-07

> anova(q1.model33, q1.model)#Partial F-test for removing x33

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x11 + x22 + x12 + x13 + x23

Model 2: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 0.060600

2 16 0.059386 1 0.0012134 0.3269 0.5754

>

> q1.model12 <-lm(y ~ x1 + x2 + x3 + x11 + x22 + x33 + x13 + x23)

> summary(q1.model12)

Call:

lm(formula = y ~ x1 + x2 + x3 + x11 + x22 + x33 + x13 + x23)

Residuals:

Min 1Q Median 3Q Max

-0.082374 -0.034857 -0.007385 0.019015 0.112075

Coefficients:

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

(Intercept) -1.6783921 1.3495092 -1.244 0.230

x1 0.4217270 0.3087488 1.366 0.190

x2 0.0207144 0.0488525 0.424 0.677

x3 -0.0683897 0.0632464 -1.081 0.295

x11 -0.0205637 0.0176115 -1.168 0.259

x22 -0.0005924 0.0118700 -0.050 0.961

x33 0.0003500 0.0014822 0.236 0.816

x13 0.0032951 0.0070800 0.465 0.648

x23 0.0015932 0.0073616 0.216 0.831

Residual standard error: 0.06394 on 17 degrees of freedom

Multiple R-squared: 0.9028, Adjusted R-squared: 0.8571

F-statistic: 19.74 on 8 and 17 DF, p-value: 3.617e-07

> anova(q1.model12, q1.model)#Partial F-test for removing x12

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x13 + x23

Model 2: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 0.069506

2 16 0.059386 1 0.01012 2.7265 0.1182

>

> q1.model13 <-lm(y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x23)

> summary(q1.model13)

Call:

lm(formula = y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x23)

Residuals:

Min 1Q Median 3Q Max

-0.04982 -0.03593 -0.01917 0.01162 0.12075

Coefficients:

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

(Intercept) -2.0844543 1.2762097 -1.633 0.1208

x1 0.4779355 0.2940473 1.625 0.1225

x2 0.1520929 0.1183856 1.285 0.2161

x3 -0.0468420 0.0193991 -2.415 0.0273 \*

x11 -0.0216365 0.0169016 -1.280 0.2177

x22 -0.0040254 0.0118568 -0.340 0.7384

x33 0.0008705 0.0014592 0.597 0.5587

x12 -0.0131498 0.0107921 -1.218 0.2397

x23 0.0006722 0.0069469 0.097 0.9240

---

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

Residual standard error: 0.06171 on 17 degrees of freedom

Multiple R-squared: 0.9095, Adjusted R-squared: 0.8669

F-statistic: 21.35 on 8 and 17 DF, p-value: 2.016e-07

> anova(q1.model13, q1.model)#Partial F-test for removing x13

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x23

Model 2: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 0.064738

2 16 0.059386 1 0.0053517 1.4419 0.2473

>

> q1.model23 <-lm(y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13)

> summary(q1.model23)

Call:

lm(formula = y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13)

Residuals:

Min 1Q Median 3Q Max

-0.058820 -0.034237 -0.005368 0.013013 0.122059

Coefficients:

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

(Intercept) -1.8130356 1.2483637 -1.452 0.1646

x1 0.4285361 0.2858403 1.499 0.1522

x2 0.2067881 0.1203493 1.718 0.1039

x3 -0.1174682 0.0624320 -1.882 0.0771 .

x11 -0.0196317 0.0163429 -1.201 0.2461

x22 -0.0037898 0.0065498 -0.579 0.5704

x33 0.0012378 0.0008704 1.422 0.1731

x12 -0.0195038 0.0116848 -1.669 0.1134

x13 0.0085232 0.0072338 1.178 0.2549

---

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

Residual standard error: 0.05935 on 17 degrees of freedom

Multiple R-squared: 0.9163, Adjusted R-squared: 0.8768

F-statistic: 23.25 on 8 and 17 DF, p-value: 1.06e-07

> anova(q1.model23, q1.model)#Partial F-test for removing x23

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13

Model 2: y ~ x1 + x2 + x3 + x11 + x22 + x33 + x12 + x13 + x23

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 0.059883

2 16 0.059386 1 0.00049712 0.1339 0.7192

Table 7.18.1: Summary of Partial F-tests

|  |  |  |
| --- | --- | --- |
| Removed Term | F-Statistic | P-value |
| X11 | 1.3236 | 0.2668 |
| X22 | 0.3822 | 0.5451 |
| X33 | 0.3269 | 0.5754 |
| X12 | 2.7265 | 0.1182 |
| X13 | 1.4419 | 0.2473 |
| X23 | 0.1339 | 0.7192 |

All p-values are not significant at the standard 0.05 significance level. With that said, there is insufficient evidence to reject the null hypothesis for each F-test that each respective quadratic term does not contribute meaningfully to the regression model on its own.

***2. PROBLEM 7.19.*** *Consider the quadratic regression model from problem 7.18. Find the variance inflation factors and comment on multicollinearity in this model. Expectations for writeup: include the output with the VIFs of each included parameter estimate. State for which variables (if any) multicollinearity seems to be a problem and how you made that determination. What are your recommendations for correcting any multicollinearity issues?*



The standard criteria for multicollinearity is a presence of a VIF which is greater than 5. Every variable in this regression model has a VIF greater than 5, therefore there is a strong presence of multicollinearity in this model. Some of the variables should be removed from the model to improve the reliability and performance of the model. The variable with the highest degree of multicollinearity in this model is x3. Referencing the conclusions of the F-test in problem 7.18, the best course of action would likely be to sequentially remove the quadratic terms with the highest p-values and reinspect the VIF’s and other model parameters until the multicollinearity issue is remedied.

***3. PROBLEM 8.6.*** *Consider the National Football League data in Table B.1. Build a linear regression model relating the number of games won to the yards gained rushing by opponents x8, the percentage of rushing plays x7, and a modification of the turnover differential x5. Specifically, let the turnover differential be an indicator variable whose value is determined by whether the actual turnover differential is positive, negative, or zero. What conclusions can you draw about the effect of turnovers on the number of games won? General hints: Remember that indicator variables in R are easily created using as.numeric() with a logical condition as the argument to the function. Also, remember that for testing equality, you use two equal signs right next to each other.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Y** | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **X8** | **X9** |
| Games won | Rushing Yards | Passing Yards | Punting Average | Field Goal % | Turnover Differential | Penalty Yards | % Rushing | Opp Rush yds | Opp Pass yds |

The indicator variables for determining whether the turnover differential is positive, negative, or zero are coded as follows:

|  |  |  |
| --- | --- | --- |
| **X51** | **X52** | **Definition** |
| 0 | 0 | X5<0 |
| 0 | 1 | X5=0 |
| 1 | 0 | X5>0 |
| 1 | 1 | Undefined, Impossible |

x51 <- as.numeric(x5>0) # 0 if less than or equal to 0, 1 if more than 0

x52 <- as.numeric(x5==0) # 0 if not equal to 0, 1 if 0

Two models were computed, one without interactions and one with interactions. A partial F-test having a p-value of 0.39 indicates that the model with interactions is not significantly better than the model without interactions, and will therefore not be used for this exercise.

> x51 <- as.numeric(x5>0) # 0 if less than or equal to 0, 1 if more than 0

> x52 <- as.numeric(x5==0) # 0 if not equal to 0, 1 if 0

> View(q2)

> x52

[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0

> x51

[1] 1 1 1 0 1 1 1 0 0 0 1 1 0 1 1 0 0 1 1 1 0 0 0 1 0 0 0 0

> x517 <- x51\*x7

> x527 <- x52\*x7

> x518 <- x51\*x8

> x528 <- x51\*x8

> x78 <- x7\*x8

>

> q2.model1 <- lm(y ~ x51 + x52 + x7 + x8)#linear model with no interactions

> q2.model2 <- lm(y ~ x51 + x52 + x7 + x8 + x517 + x527 + x518 + x528 + x78)#linear model with interactions

> summary(q2.model1)

Call:

lm(formula = y ~ x51 + x52 + x7 + x8)

Residuals:

Min 1Q Median 3Q Max

-3.2352 -1.6697 -0.0034 1.3071 5.6650

Coefficients:

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

(Intercept) 19.813559 9.688921 2.045 0.05247 .

x51 1.872527 0.962129 1.946 0.06394 .

x52 -0.460504 2.466185 -0.187 0.85351

x7 -0.006825 0.118841 -0.057 0.95470

x8 -0.006337 0.001719 -3.685 0.00122 \*\*

---

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

Residual standard error: 2.337 on 23 degrees of freedom

Multiple R-squared: 0.6158, Adjusted R-squared: 0.549

F-statistic: 9.216 on 4 and 23 DF, p-value: 0.0001349

> summary(q2.model2)

Call:

lm(formula = y ~ x51 + x52 + x7 + x8 + x517 + x527 + x518 + x528 +

x78)

Residuals:

Min 1Q Median 3Q Max

-3.3520 -1.3889 -0.4104 1.5108 4.9027

Coefficients: (2 not defined because of singularities)

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

(Intercept) 1.272e+01 3.496e+01 0.364 0.720

x51 -1.238e+01 2.009e+01 -0.616 0.545

x52 -8.446e-01 2.528e+00 -0.334 0.742

x7 2.285e-01 5.946e-01 0.384 0.705

x8 -2.172e-03 1.381e-02 -0.157 0.877

x517 4.499e-02 2.596e-01 0.173 0.864

x527 NA NA NA NA

x518 5.481e-03 3.480e-03 1.575 0.131

x528 NA NA NA NA

x78 -1.243e-04 2.433e-04 -0.511 0.615

Residual standard error: 2.329 on 20 degrees of freedom

Multiple R-squared: 0.6683, Adjusted R-squared: 0.5522

F-statistic: 5.757 on 7 and 20 DF, p-value: 0.000935

> anova(q2.model1, q2.model2)

Analysis of Variance Table

Model 1: y ~ x51 + x52 + x7 + x8

Model 2: y ~ x51 + x52 + x7 + x8 + x517 + x527 + x518 + x528 + x78

Res.Df RSS Df Sum of Sq F Pr(>F)

1 23 125.62

2 20 108.45 3 17.17 1.0554 0.39

*What conclusions can you draw about the effect of turnovers on the number of games won?*

The effect of turnovers, when coded to whether the turnover differential was positive, negative, or 0, has a marginally significant effect on number of games won for this dataset. The coded variable x51, indicating those observations of x5 which are positive, has a p-value of 0.06394. The coded variable x51, indicating those observations of x5 which are zero, has a very non-significant p-value of 0.85351, and does therefore not contribute meaningfully to the model.

The most significant factor is x8, which has a p-value of 0.00122, and contributes most meaningfully to the model predicting the number of games won from the selected variables.