Homework 2

Allan Kimaina

February 22, 2018

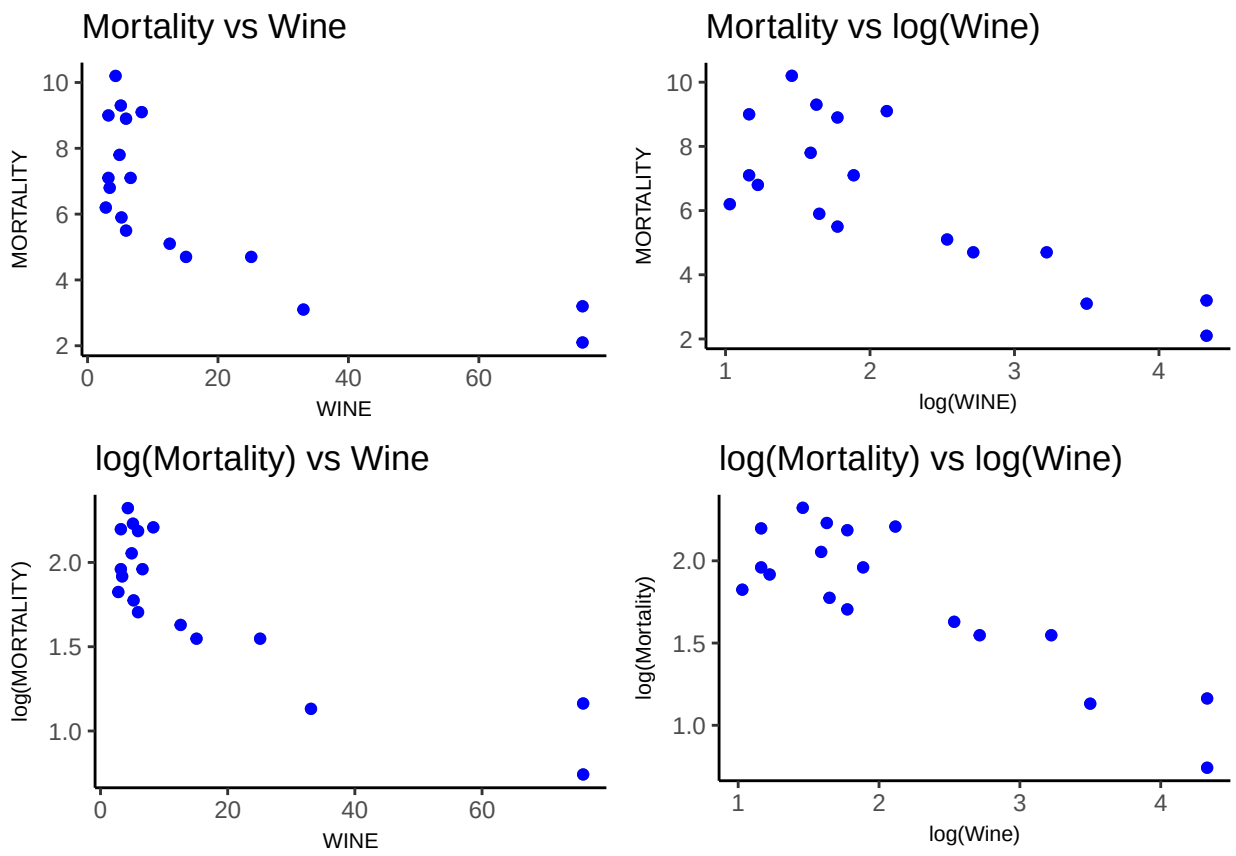
Question 1

a. Using the wine data from the previous homework, create two new variables by logarithmically transforming both wine and mortality

Table 1:

	COUNTRY	WINE	MORTALITY	logMORTALITY	logWINE
1	Norway	2.800	6.200	1.825	1.030
2	Scotland	3.200	9	2.197	1.163
3	England	3.200	7.100	1.960	1.163
4	Ireland	3.400	6.800	1.917	1.224
5	Finland	4.300	10.200	2.322	1.459
6	Canada	4.900	7.800	2.054	1.589
7	UnitedStates	5.100	9.300	2.230	1.629
8	Netherlands	5.200	5.900	1.775	1.649
9	NewZealand	5.900	8.900	2.186	1.775
10	Denmark	5.900	5.500	1.705	1.775
11	Sweden	6.600	7.100	1.960	1.887
12	Australia	8.300	9.100	2.208	2.116
13	Belgium	12.600	5.100	1.629	2.534
14	Germany	15.100	4.700	1.548	2.715
15	Austria	25.100	4.700	1.548	3.223
16	Switzerland	33.100	3.100	1.131	3.500
17	Italy	75.900	3.200	1.163	4.329
18	France	75.900	2.100	0.742	4.329

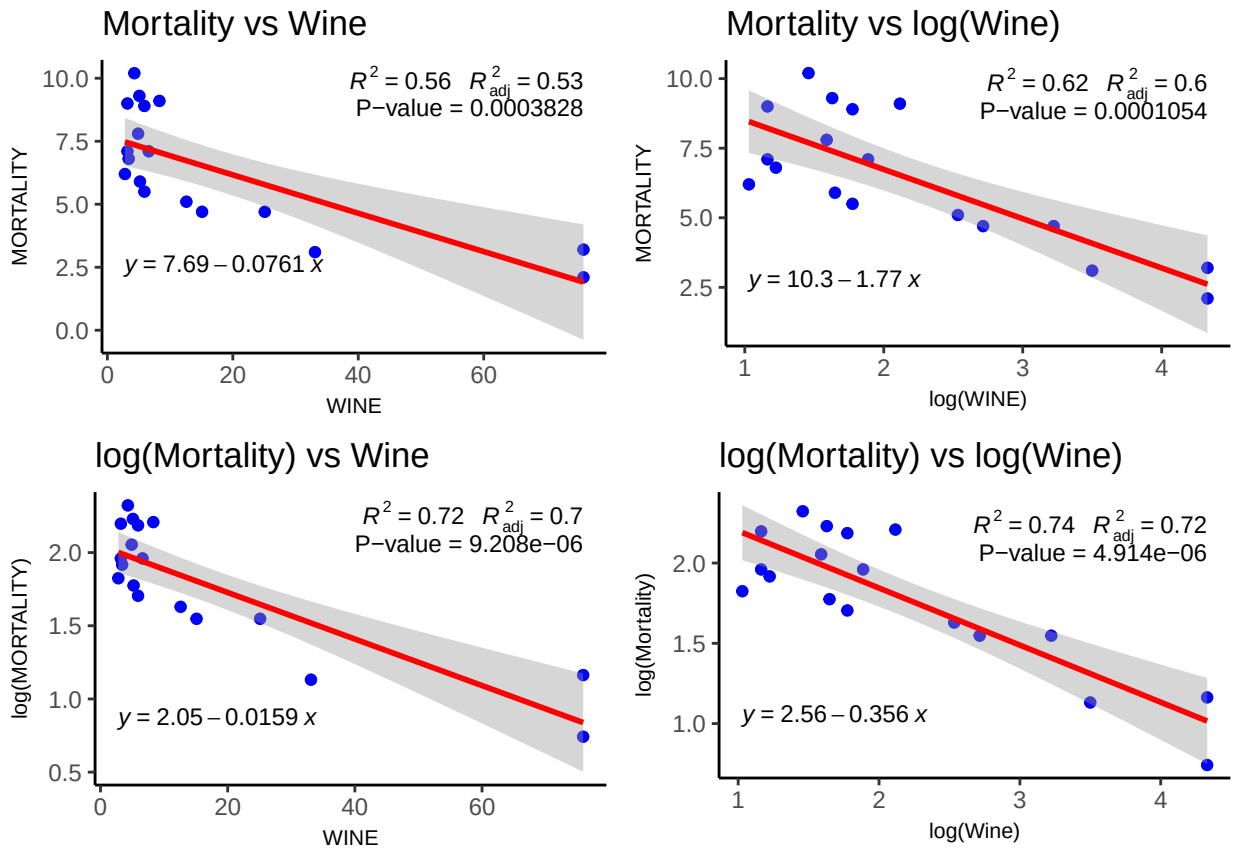
b. Plot mortality vs. wine; log mortality vs. wine, mortality vs. log wine and log mortality vs. log wine. Which one looks more linear?



- From the above scatter plot, the logarithmic transformation of both dependent variable (log of Mortality) and independent variable (log of Wine) results in the most linear relationship compared to the other plots.
- Transforming both predictor and response variable we get the best linear relationship

c. Fit four linear regression models corresponding to the four plots and report the regression equation and R-squared. Which model do you think is best?

- **Mortality vs Wine:** $HeartAttackMortality = 7.69 - 0.0761 * WineConsumption + \epsilon$
- **Mortality vs log(Wine):** $HeartAttackMortality = 10.280 - 1.771 * log(WineConsumption) + \epsilon$
- **log(Mortality) vs Wine:** $log(HeartAttackMortality) = 2.0453 - 0.0159 * WineConsumption + \epsilon$
- **log(Mortality) vs log(Wine):** $log(HeartAttackMortality) = 2.5556 - 0.3556 * log(WineConsumption) + \epsilon$



Graphically we notice that the regression line for the log-transformed model (ln(y) ~ ln(x)) covers most data-points as well as having a more precise and uniform CI band compared with the other 3 models.

Table 2: Comparison of the 4 Models

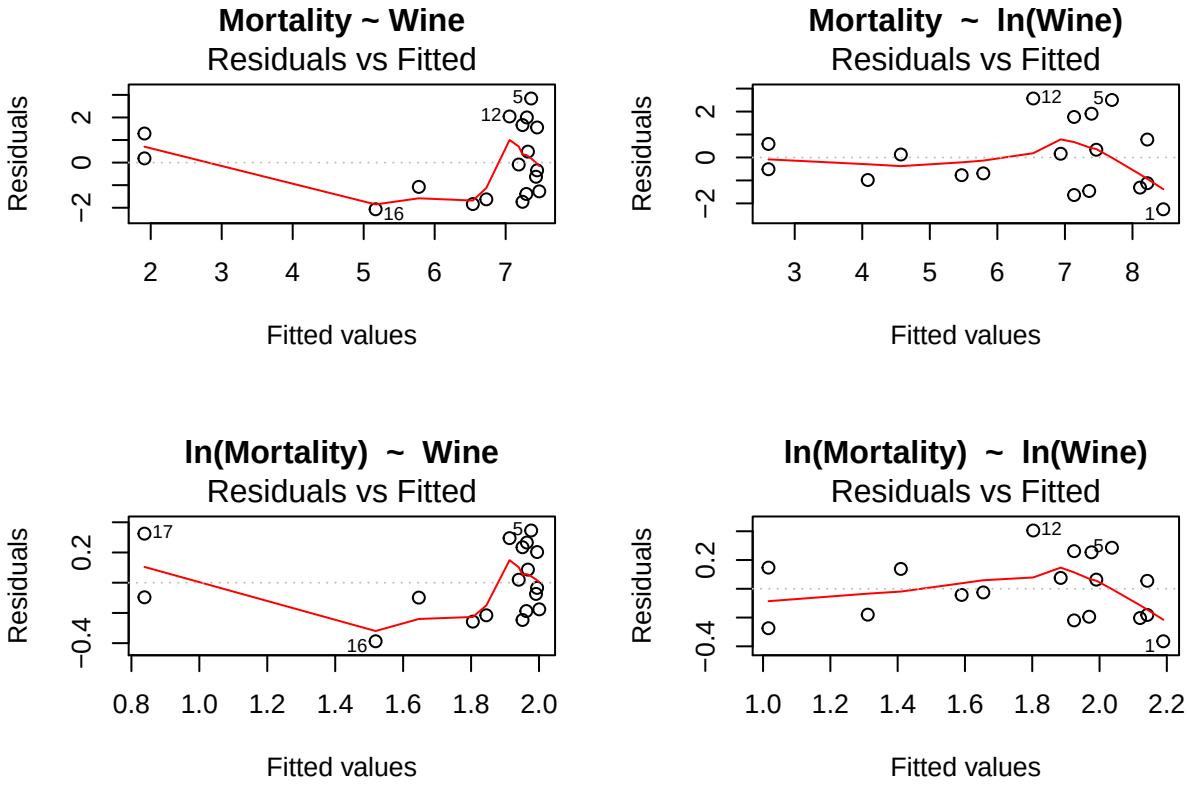
	-			
	MORTALITY		logMORTALITY	
	y x	y ln(x)	ln(y) x	ln(y) vs ln(x)
	(1)	(2)	(3)	(4)
WINE	−0.076*** (0.017)		−0.016*** (0.002)	
logWINE		−1.771*** (0.347)		−0.356*** (0.053)
Constant	7.687*** (0.473)	10.280*** (0.832)	2.045*** (0.069)	2.556*** (0.127)
Observations	18	18	18	18
R^2	0.556	0.620	0.718	0.738
Adjusted R^2	0.528	0.596	0.700	0.722
Residual Std. Error (df = 16)	1.619	1.498	0.238	0.229
F Statistic (df = 1; 16)	20.026***	26.090***	40.639***	45.170***
Note:				*p<0.1; **p<0.05; ***p<0.01
				t-values have been hidden!

Looking at y~x and y~ln(x) models, we have really low explanatory power (R-squared). In y~x model, only 56% of the variability in heart attack mortality is explained by wine consumption. While in y~ln(x) model only 64% of the variability in the response is explained by the predictor.

Comparing the explanatory power between the ln(y)~x model and ln(y)~ln(x) model, we get a higher explanatory power in ln(y)~ln(x) model. Over 73.6% of the variability in heart attack mortality is explained by wine consumption in ln(y)~ln(x) model while only 73.6% of the variability in heart attack mortality is explained by wine consumption in ln(y)~x model

For these reasons, **the best fit model** is ln(y)~ln(x) model because it has the highest explanatory model compared to the other models. Furthermore, it has a p-value that is very significant (approximately 4.914e-06)

Model Diagnostics



Looking at the residual plots, ln(Mortality) ~ ln(Wine) model exhibits acceptable randomness and unpredictability. From the plot, the residuals bounce randomly around the 0 line and roughly forms a horizontal band around the 0 line without any noticeable pattern. This suggests that the variances of the error terms are approximately equal.

The other models seem not to follow the rule of residual heteroskedasticity which is a crucial component of any regression model. Residuals in the other models deviate from linearity and equality of standard deviations assumptions. Hence this gives us more reason to select the model with logarithmic transformation of both predictor and response as the best fit model

Conclusion

In summary, the model with the natural log of heartAttackMortality as the response and the natural log of wineConsumption as the predictor exhibits a relationship with the highest explanatory power. This relationship appears to be linear and at the same time with error terms that appear independent and normally

distributed with somewhat equal variances. Therefore we have sufficient evidence that this model [ln(y)~ln(x)] is the best.

d. Interpret each regression model by describing the change in mortality for a given change in the predictor. Use an increase of 10 units of wine consumption for linear wine and a doubling of wine consumption for logarithmic wine

Mortality vs Wine

- An increase of wine consumption by 10 Liters per person will decrease Ischemic heart attack mortality by 0.761 per 1000 person

Mortality vs log(Wine)

- Doubling wine consumption will decrease Ischemic heart attack mortality rate by 1.227 per 1000 person

log(Mortality) vs Wine

medianProportion = exp(-0.0159*10) = 0.8529964* *%change = (exp(-0.0159*10)-1)*100 = -14.7%

- An increase of wine consumption by 10 Liters per person will decrease Ischemic heart attack mortality by a median proportion of 0.8529964
- An increase of wine consumption by 10 Liters per person will decrease Ischemic heart attack mortality by 14.7%

log(Mortality) vs log(Wine)

%change = 1-exp(-0.3556*log(2)))*100 = -21.84555

- Doubling wine consumption will decrease Ischemic heart attack mortality by 21.85%

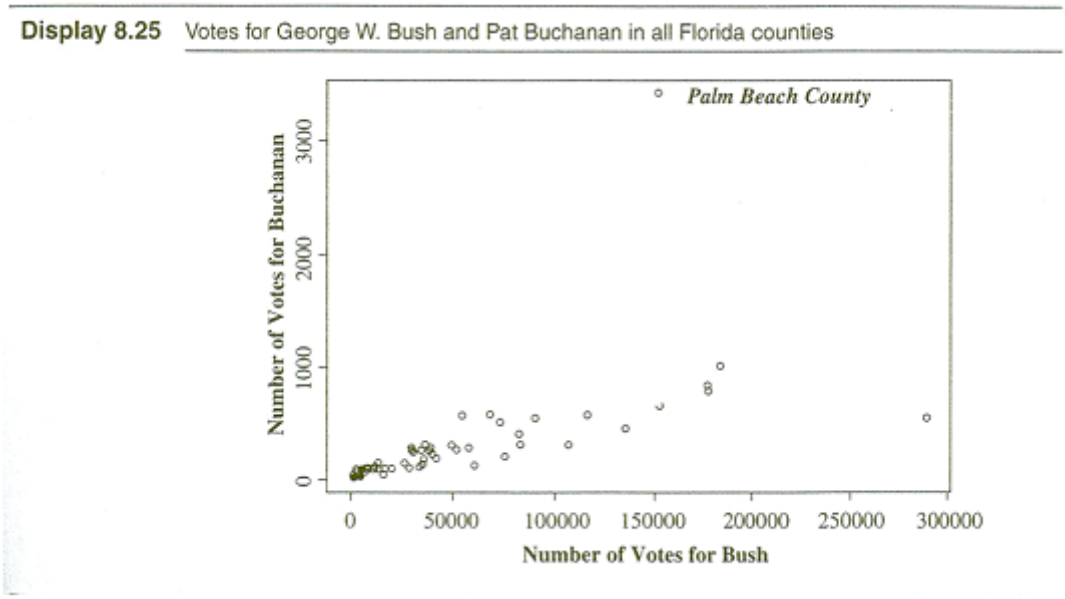


Figure 1: Caption for the picture.

Question 2: The Dramatic U.S. Presidential Election of 2000

The U.S. presidential election of November 7, 2000 was one of the closest in history. As returns were counted on election night it became clear that the outcome in the state of Florida would determine the next president. At one point in the evening, television networks projected that the state was carried by the Democratic nominee, Al Gore, but a retraction of the projection followed a few hours later. Then, early in the morning of November 8, the networks projected that the Republican nominee, George W. Bush, had carried Florida and won the presidency. Gore called Bush to concede. While on route to his concession speech, though, the Florida count changed rapidly in his favor. The networks once again reversed their projection, and Gore called Bush to retract his concession. When the roughly 6 million Florida votes had been counted

a. The data in File Bush.xls contain the numbers of votes for Buchanan and Bush in all 67 counties in Florida. What evidence is there in the scatterplot of Display 8.25 that Buchanan received more votes than expected in Palm Beach County? Analyze the data without Palm Beach County results to obtain an equation for predicting Buchanan votes from Bush votes. Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result-assuming the relationship is the same in this county as in the others. If it is assumed that Buchanan's actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval? (Consider transformation.)

From the scatter plot (Display 8.25) we can vividly see that Palm Beach County is one of the outlier because it has been separated away from the trend of other observation. In fact, if we regress Buchanan on Bush using the full data set with this potential outlier, the studentized residual and cooks distance flags this datapoint because it influences the regression model to such an extent that the estimated regression equation is pulled towards this datapoint.

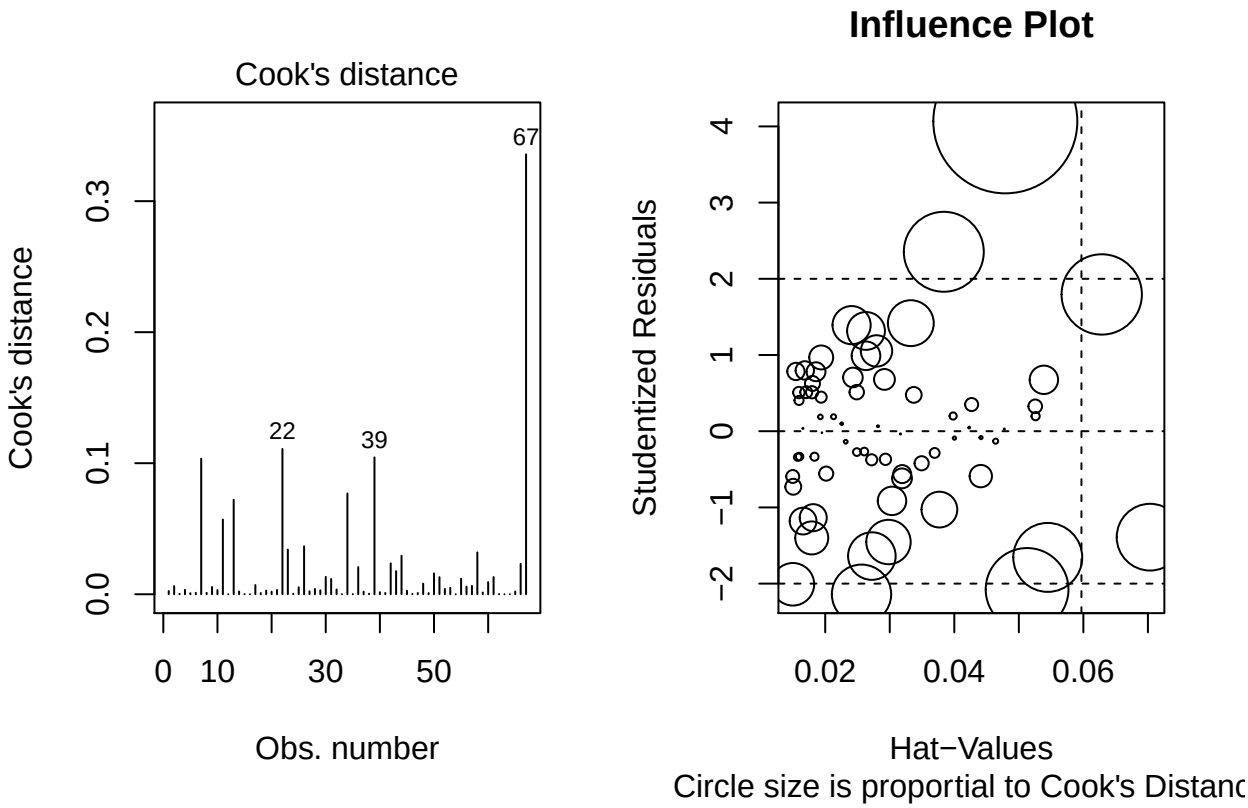
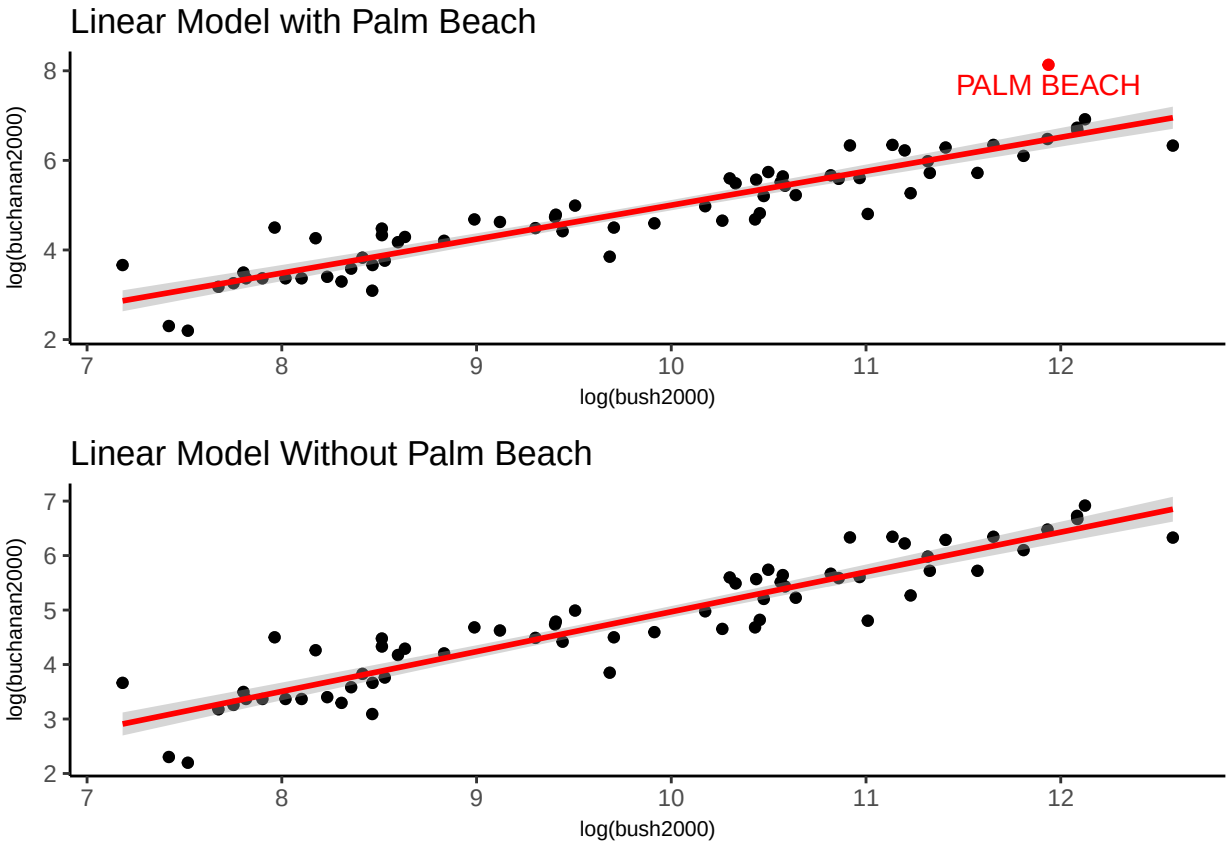


Table 3: Models With and without outlier

	-	
	buchanan2000	
	Without Palm Beach	With Palm Beach
	(1)	(2)
Observations	67	66
R^2	0.851	0.866
Adjusted R^2	0.848	0.864
Residual Std. Error	0.467 (df = 65)	0.420 (df = 64)
F Statistic	370.615*** (df = 1; 65)	413.022*** (df = 1; 64)
Note:		*p<0.1; **p<0.05; ***p<0.01
		Coefficients have been removed!

Furthermore, if we remove this potential outlier from the dataset and regress Buchanan on Bush, our regression model changes substantially in the sense that residual standard error is inflated from 413 to 371. This inflation highly increases the width of confidence and prediction intervals



Using the regression model without palm beach to generate a 95% prediction interval for the number of Buchanan votes in the Palm Beach, we see that the expected votes for Buchanan given the corresponding votes for Bush (152,846) is 592.377 which highly deviates from the observed votes of 3407

Table 4: Prediction for Buchanana

	fit	lwr	upr
1	592.377	250.800	1,399.164

In fact, the observed votes for Buchanan (3407) is way far from our 95 % prediction interval of [250.8, 1399.164]. Therefore we have sufficient evidence that the Buchanan's actual count contains a number of votes intended for Gore because of the ballot paper candidate arrangement error. Using our prediction model, the extra votes for Buchanan (3407-592) was meant for Gore hence decreasing Buchanan votes by 592 and increasing Gores votes from 268945 to 268353

b. Analyze the data in Ex1222 and write a statistical summary predicting the number of Buchanan, votes in Palm Beach Country that were not intended for him. It would be appropriate to describe any unverifiable assumptions used in applying the prediction equation for this purpose. (Suggestion: Find a model for predicting Buchanan's 2000 vote from other variables, excluding Palm Beach County, which is listed last in the data set. Consider a transformation of all counts.)

We used Stepwise AIC for predictor variable selection and after evaluating the relative importance measure of each variable, we came up with the best model that is in line with the objective of this analysis.

Variable Selection

The full model containing all potential predictor variables were passed into the step function. The StepAIC function iteratively searched best predictors by dropping or adding one X predictor variable at a time. In each iteration, AIC of the models were calculated and the model that had the lowest AIC was retained for the next iteration.

Full Model: buchanan2000~bush2000+gore2000+nader2000+browne2000+total2000+clinton96+dole96+perot96+buchanan96p+reform.reg+total.reg

After the step iterations were completed the full model was reduced to a model with 7 predictors:

StepWise AIC Model: : buchanan2000 ~ nader2000 + browne2000 + total2000 + clinton96 + perot96 + buchanan96p + total.reg

The rule of thumb for a good predictive accuracy has always been to look at no more variables than 10% of the total number of observation. This model is in check with this rule, therefore, there is no need to remove any more predictors. In fact looking at these variables individually with relation to 2000 and 1996 elections we have much more relevancy in context compared to the other predictors

Table 5: Best Model

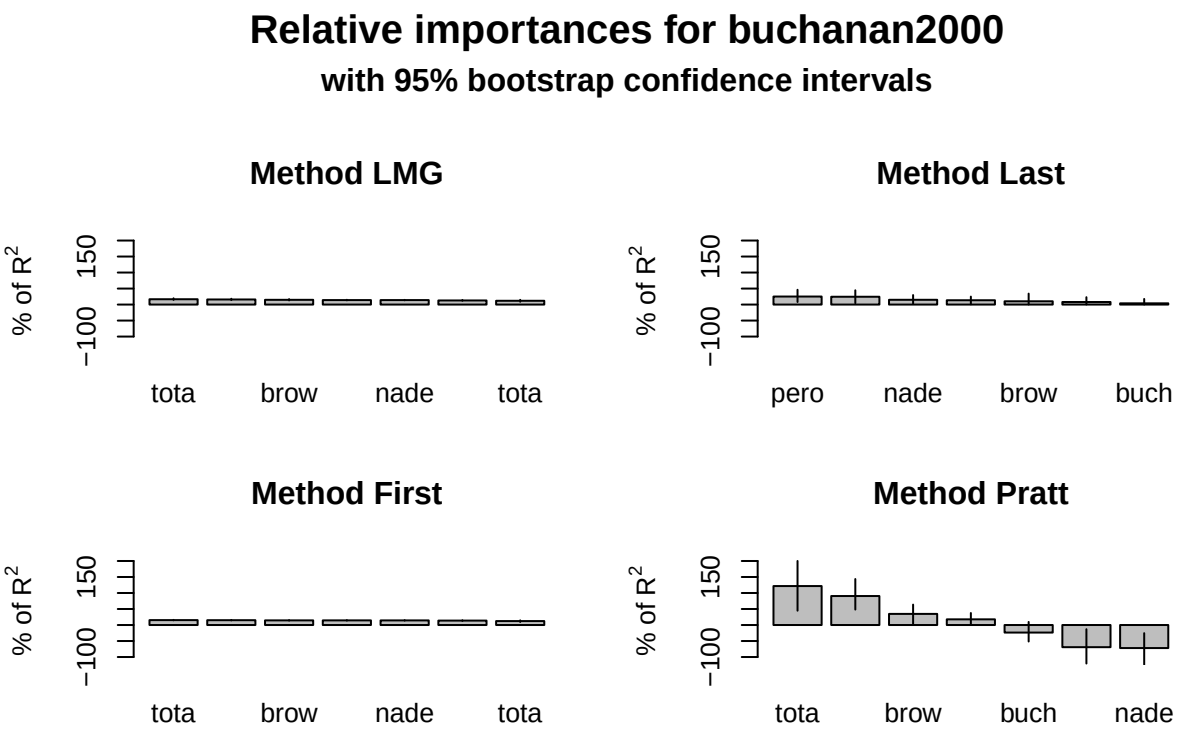
	-
	buchanan2000
nader2000	-0.491^{***} (-0.826, -0.155)
browne2000	0.254^{**} (0.044, 0.464)
total2000	0.899^{***} (0.417, 1.380)
clinton96	-0.522^{***} (-0.898, -0.146)
perot96	0.768^{***} (0.362, 1.175)
buchanan96p	-0.146 (-0.343, 0.051)
total.reg	0.149^{**} (0.009, 0.289)
Constant	-4.629^{***} (-6.665, -2.593)
Observations	66
R^2	0.898
Adjusted R^2	0.886
Residual Std. Error	0.384 (df = 58)
F Statistic	73.312^{***} (df = 7; 58)
Note:	*p<0.1; **p<0.05; ***p<0.01

This model has very high explanatory power, with R^2 of 0.898 and Adjusted R^2 of 0.886. Over 89.8% of the variability in Buchanan votes for 2000 is explained by selected predictors. Furthermore, most of the

predictors have really low p-values except for browne2000

Relative Importance Measure

After building the model using stepwise variable selection, We went ahead and evaluated the relative importance of each predictor in the model.



Using Lindeman, Merenda, and Gold method, all the predictors selected from the stepWise regression had significant contribution percentage into the model, therefore, we did not have any reason to remove any of the variables identified by stepwiseAIC.

Prediction

Table 6: Best Prediction for Buchanana

	fit	lwr	upr
67	431.903	164.101	1,136.736

Assumptions

- Due to the fact we were modeling for predictive regression, we didn't assess multicollinearity
- Relative importance is not as effective when regressors are correlated. We did not evaluate collinearity

- We did not have a variable for total1996 and observations for other candidates in 1996. For us to standardize vote counts for the year 1996 and 2000, we needed these variables. Therefore we assumed that the data was standardized in the sense that the total number of votes for 1996 and 2000 were the same.

We used the above model to predict Buchanan's 2000 vote in Palm Beach County. Among the candidate predictors include in this model were 2000 votes for Browne and Nader, 1996 votes for Dole, Clinton, Perot, and Buchanan. With a 95% prediction interval of about [164.101, 1136.736], we got an expected number of votes for Buchanan in Palm Beach County as 431.903

The Observed number of votes for Buchanan was 3407 which was way higher than expected number of votes of 432. Furthermore, this observation is way far from our 95 % prediction interval of about [164,1137]. This can only mean that Buchanan's actual count contains a number of votes intended for Gore because of the ballot paper candidate arrangement error. Using our prediction model, part of the extra votes for Buchanan (3407 - 432) was meant for Gore. From the physical evidence presented and the statistical inference made, we have sufficient evidence that the expected number of votes for Gore was supposed to be approximately 271917 (268945+2972) instead of 268945.

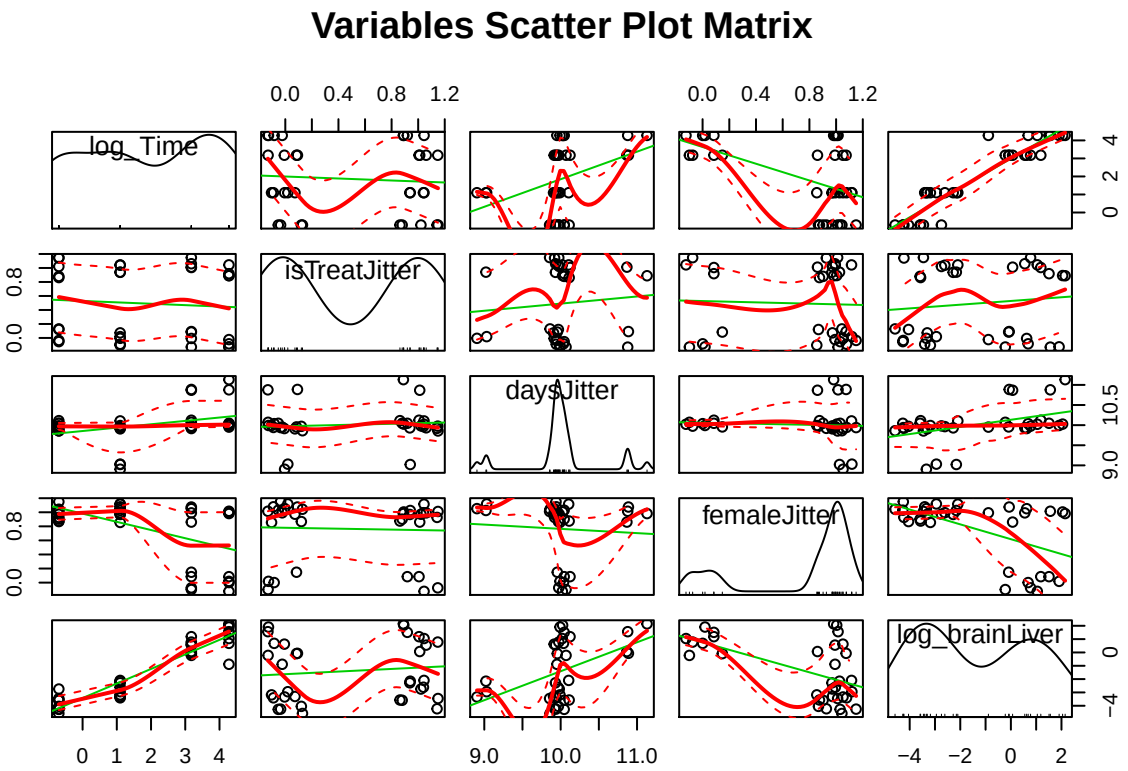
Question 3

a. Compute "Jittered" versions of-treatment, days after inoculation, and an indicator variable for females by adding small random numbers to each (uniform random numbers between -.15 and .15 work well). Or you could use the jitter function.

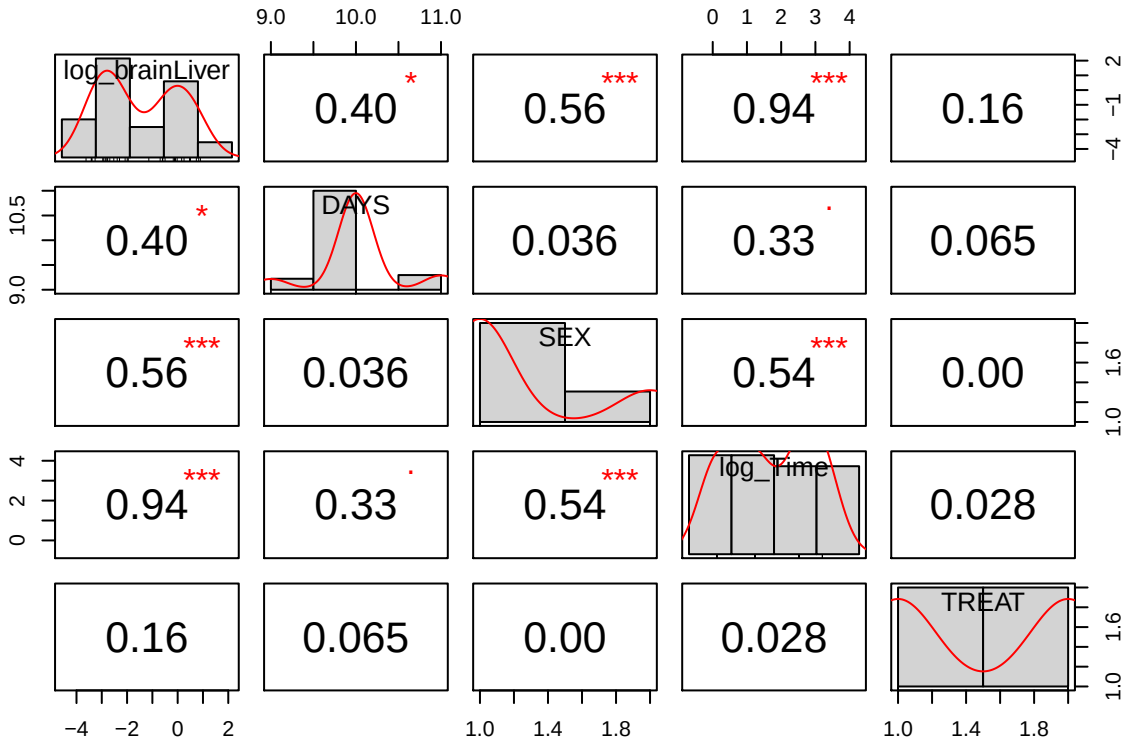
Table 7:

	BRAIN	brainLiver	isTreatJitter	daysJitter	femaleJitter	log_brainLiver	log_Time
1	41,081	0.028	0.863	10.111	0.875	-3.568	-0.693
2	44,286	0.028	0.880	10.046	0.903	-3.588	-0.693
3	102,926	0.064	1.021	9.982	0.972	-2.745	-0.693
4	25,927	0.015	1.142	9.920	0.927	-4.227	-0.693
5	42,643	0.032	1.149	9.855	1.011	-3.456	-0.693
6	31,342	0.018	0.121	9.970	0.868	-4.045	-0.693
7	22,815	0.014	-0.053	10.046	1.030	-4.271	-0.693
8	16,629	0.010	0.129	9.861	1.043	-4.578	-0.693
9	22,315	0.014	-0.035	9.939	1.150	-4.252	-0.693
10	77,961	0.074	1.040	9.963	0.994	-2.610	1.099
11	73,178	0.102	0.937	9.932	0.969	-2.280	1.099
12	76,167	0.123	1.044	10.105	1.115	-2.097	1.099
13	123,730	0.116	0.940	9.021	1.026	-2.156	1.099
14	25,569	0.035	-0.005	8.904	1.052	-3.340	1.099
15	33,803	0.033	-0.089	9.942	1.078	-3.406	1.099
16	24,512	0.037	0.073	9.907	0.973	-3.305	1.099
17	50,545	0.053	0.019	9.031	1.118	-2.945	1.099
18	50,690	0.042	-0.103	9.990	0.855	-3.181	1.099
19	84,616	1.733	1.060	9.934	0.991	0.550	3.178
20	55,153	3.266	1.006	10.123	0.085	1.184	3.178
21	48,829	2.180	1.147	9.975	-0.074	0.779	3.178
22	89,454	1.071	1.012	10.867	0.861	0.069	3.178
23	37,928	1.866	-0.007	9.915	1.051	0.624	3.178
24	12,816	0.802	-0.129	10.070	-0.099	-0.221	3.178
25	23,734	0.917	0.084	9.956	0.148	-0.087	3.178
26	31,097	0.936	0.091	10.887	1.075	-0.066	3.178
27	35,395	8.545	0.889	11.131	0.982	2.145	4.277
28	18,270	7.728	0.891	10.025	1.005	2.045	4.277
29	5,625	2.842	1.044	10.022	-0.124	1.045	4.277
30	7,497	4.519	0.918	10.046	0.085	1.508	4.277
31	6,250	6.735	-0.127	9.991	0.018	1.907	4.277
32	11,519	4.754	-0.131	10.878	1.015	1.559	4.277
33	3,184	1.980	-0.084	10.037	0.001	0.683	4.277
34	1,334	0.411	-0.023	9.951	0.992	-0.888	4.277

b. Obtain a matrix of scatter plots for the following variables: log sacrifice time, treatment (jittered), days after inoculation (jittered), sex (jittered), and the log of the brain tumor-to-liver antibody ratio. Use the function pairs in the graphics package or scatterplotMatrix in the car package.



c. Obtain a matrix of the correlation coefficients among the same five variables (not jittered).



d. On the basis of this, what can be said about the relationship between the covariates (sex and days after inoculation), the response, and the design variables (treatment and sacrifice time.

We have a very strong and very significant relationship between log response variable (ratio between antibodies in brain and liver) and the log of sacrifice time (r=.54). However, the relationship between the log of response and treatment is weak and insignificant. On the other hand, we have a moderately strong and significant relationship between log response variable and days after inoculation (r=.4). There is a strong and significant relationship between log response and sex (r=.56). We do not have a correlation between the covariates (sex and days after inoculation), their relationship is weak and insignificant. Also, the relationship between days after inoculation and log of sacrifice time is weak (r=.33) with a marginal p-value. We also have a weak insignificant signal between Days after inoculation and treatment. On the other hand, the relationship between sex and log of sacrifice time was moderately strong and very significant (r=.54). However, we do not have any signal between sex and treatment (r=0). Finally, there is no correlation between the design variable (treatment and log of sacrifice time)

In summary, we don't have collinearity within the design variables (treatment and log of sacrifice time) as well as within the covariates (sex and days after inoculation). However, we do have collinearity between sex and log of sacrifice time. The relationship between the log response variable and all the predictors are mostly strong and significant except for the treatment predictor variable.

e. Fit the regression of the log response (brain tumor-to-liver antibody ratio) on an indicator variable for treatment and on sacrifice time treated as a factor with four levels (include three indicator variables, for sacrifice time == 3, 24, and 72 hours). Use the model to find the estimated mean of the log response at each of the eight treatment combinations (all combinations of the two infusions and the four sacrifice times).

Table 8:

	-
	log_brainLiver
TIMEFactor3	1.134^{***} (0.640, 1.628)
TIMEFactor24	4.257^{***} (3.749, 4.765)
TIMEFactor72	5.154^{***} (4.646, 5.662)
TREATNS	-0.797^{***} (-1.156, -0.437)
Constant	-3.505^{***} (-3.888, -3.122)
Observations	34
R^2	0.951
Adjusted R^2	0.944
Residual Std. Error	0.533 (df = 29)
F Statistic	139.646^{***} (df = 4; 29)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9:

	TIMEFactor	TREAT	fit	lwr	upr
1	0.5	BD	-3.505	-4.666	-2.344
2	3	BD	-2.371	-3.538	-1.203
3	24	BD	0.752	-0.419	1.923
4	72	BD	1.649	0.478	2.820
5	0.5	NS	-4.302	-5.469	-3.134
6	3	NS	-3.168	-4.328	-2.007
7	24	NS	-0.044	-1.215	1.126
8	72	NS	0.852	-0.319	2.023

f. Let X represent log of sacrifice time. Fit the regression of the log response on an indicator variable for treatment, X, X2, and X3. Use the estimated model to find the estimated mean of the log response at each of the eight treatment combinations.

Table 10:

	-
	log_brainLiver
TREATNS	-0.846^{***} (-1.270, -0.422)
log_Time	1.098^{***} (0.987, 1.209)
Constant	-3.009^{***} (-3.370, -2.649)
Observations	34
R^2	0.926
Adjusted R^2	0.921
Residual Std. Error	0.631 (df = 31)
F Statistic	194.187^{***} (df = 2; 31)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 11:

	log_Time	TREAT	fit	lwr	upr
1	-0.693	BD	-3.770	-5.125	-2.416
2	1.099	BD	-1.803	-3.129	-0.477
3	3.178	BD	0.480	-0.853	1.813
4	4.277	BD	1.686	0.332	3.040
5	-0.693	NS	-4.616	-5.973	-3.259
6	1.099	NS	-2.649	-3.976	-1.322
7	3.178	NS	-0.366	-1.698	0.965
8	4.277	NS	0.840	-0.511	2.191

g. Why are the answers to parts (5) and (6) the same?

Table 12:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	8.233				
2	31	12.330	-2	-4.098	7.217	0.003

Since the origin model is already linear, further log transformation wouldn't change any of the predictions by much. Transformation doesn't have much effect if the data points don't vary too much. In fact, plotting out a scatter plot for data of the 2 models indicates that the range of data points don't vary and spread a lot as we would have expected in a classical scenario where variable transformation was necessary.

h. Fit the regression of the log response (brain tumor-to-liver antibody ratio) on all covariates, the treatment indicator, and sacrifice time, treated as a factor with four levels (include three indicator variables, for sacrifice time == 3, 24, and 72 hours).

Table 13:

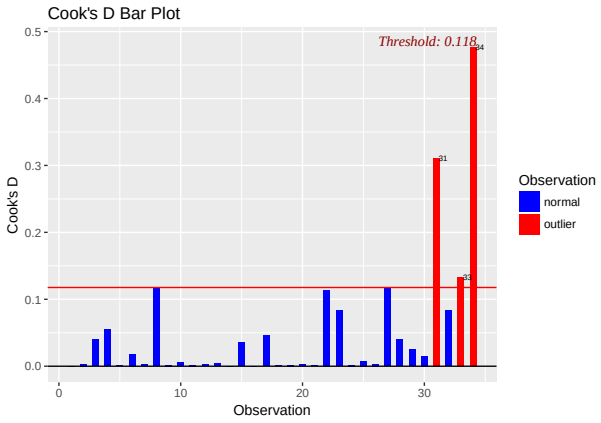
	-
	log_brainLiver
TREATNS	-0.831^{***} (-1.218, -0.444)
TIMEFactor3	1.089^{***} (0.513, 1.666)
TIMEFactor24	4.114^{***} (3.453, 4.775)
TIMEFactor72	5.137^{***} (4.468, 5.805)
SEXM	-0.036 (-0.737, 0.666)
WEIGHT	0.002 (-0.008, 0.011)
DAYS	0.019 (-0.533, 0.572)
LOSS	-0.048^{*} (-0.102, 0.006)
TUMOR	0.001 (-0.001, 0.004)
Constant	-4.064 (-10.227, 2.100)
Observations	34
R^2	0.957
Adjusted R^2	0.941
Residual Std. Error	0.547 (df = 24)
F Statistic	59.258^{***} (df = 9; 24)
Note:	*p<0.1; **p<0.05; ***p<0.01

i. Obtain a set of case influence statistics, including a measure of influence, the leverage, and the studentized residua1.

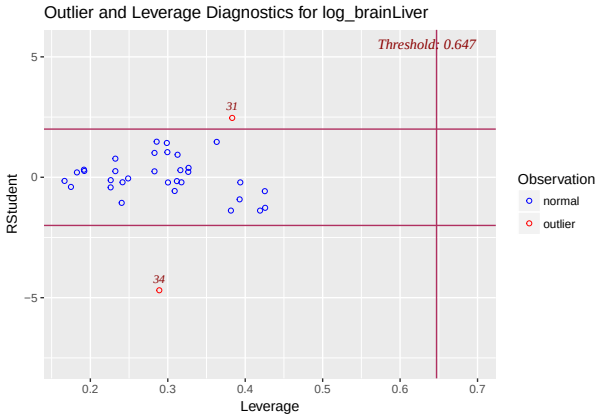
Table 14:

	dffit	cov.r	cook.d	hat	studResidual
1	-0.06773	1.8184	0.00048	0.16667	-0.15144
2	-0.18453	1.73035	0.00353	0.17494	-0.40073
3	0.63507	1.38136	0.04029	0.28285	1.01123
4	-0.73812	1.7591	0.05484	0.39278	-0.91775
5	0.09536	1.83973	0.00095	0.18266	0.20172
6	0.42539	1.54232	0.0184	0.23231	0.7733
7	0.15475	2.07858	0.00249	0.28284	0.24642
8	-1.09187	1.35591	0.11627	0.42574	-1.26809
9	0.12766	1.83909	0.0017	0.19218	0.26173
10	-0.2269	1.83261	0.00533	0.22619	-0.41968
11	0.13953	1.93875	0.00203	0.23232	0.25363
12	0.15328	1.81387	0.00244	0.19171	0.31473
13	0.20082	2.15548	0.00419	0.31631	0.29524
14	-0.02816	2.03529	8e-05	0.2488	-0.04892
15	-0.59814	1.24736	0.03558	0.24047	-1.06302
16	-0.06492	1.96582	0.00044	0.2263	-0.12003
17	0.68159	1.37696	0.04629	0.29948	1.04244
18	-0.11695	1.98067	0.00142	0.24156	-0.20724
19	-0.14133	2.14341	0.00208	0.30024	-0.21576
20	0.15602	2.22333	0.00253	0.32647	0.22409
21	-0.10701	2.19946	0.00119	0.31175	-0.159
22	-1.0865	1.1128	0.11372	0.38152	-1.38334
23	0.93414	0.86531	0.08316	0.28562	1.47736
24	-0.14313	2.20051	0.00213	0.31765	-0.20978
25	0.2714	2.12889	0.00764	0.32683	0.38951
26	-0.17095	2.47526	0.00304	0.39365	-0.21216
27	1.10933	0.98032	0.1174	0.36318	1.46895
28	0.63064	1.53313	0.03998	0.31262	0.93513
29	-0.49652	2.30647	0.02536	0.42531	-0.57717
30	-0.38116	1.9246	0.01495	0.30889	-0.57014
31	1.94181	0.23855	0.3113	0.38319	2.46364
32	0.92963	0.93731	0.08287	0.29875	1.42427
33	-1.17553	1.18431	0.13311	0.41916	-1.3838
34	-2.9917	0.00261	0.4772	0.28905	-4.69189

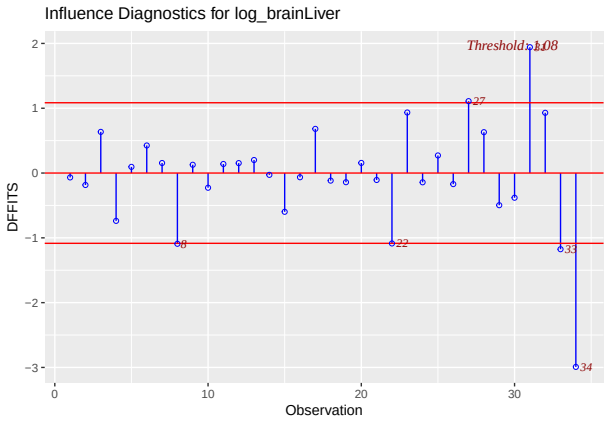
Cook's Distance



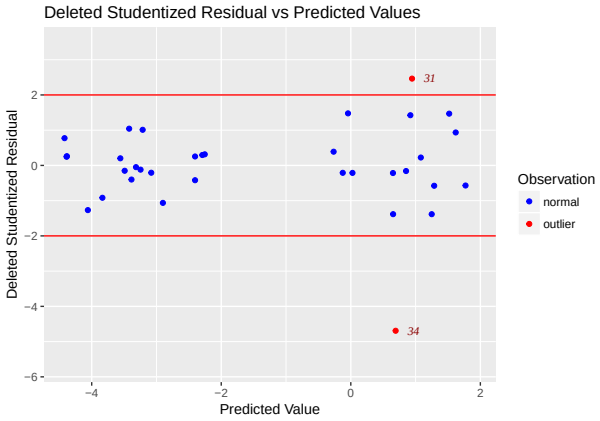
Leverage Plot



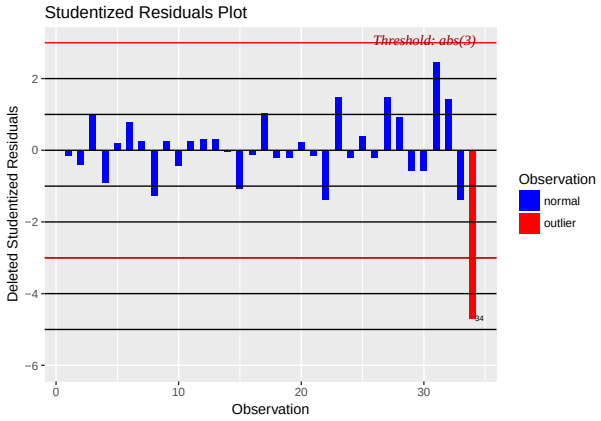
DFFITS Plot - difference in fits



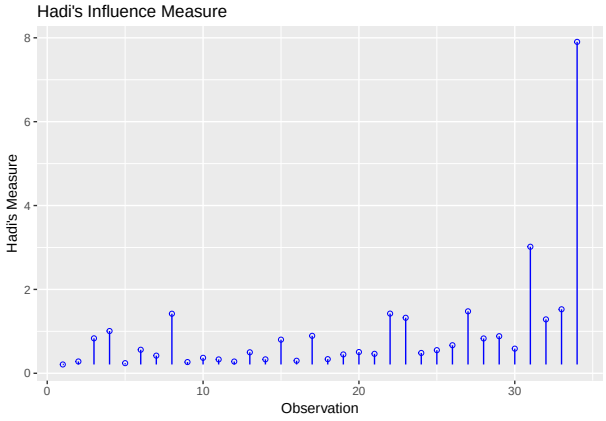
Deleted Stud.Residual vs Fitted Values



Studentized Residual Plot



Hadi Plot



*** Discuss whether any influential observations or outliers occur with respect to this fit.**

- The cook's distance plot revealed that we have 3 regression outliers which strongly influences fitted values of our model. These outliers include observation 31, 33, and 34.
- DFFIT diagnostics plot indicated that we have 6 influential data points. Observation 31, 33 and 34 had the strongest influence while observation 27, 22 and 8 had marginal (acceptable) influence on our model
- From the Studentized residual plot, we found out that 1 observation which had studentized residual greater than 3 (absolute cutoff) indicating that observation 34 is an outlier that has a large effect on the overall residual
- Leverage plot indicated that we did not have any leverage observations since we did not have any observation with an extreme value on a predictor variable
- Deleted Studentized Residual vs Fitted Value plot indicated that we had 2 observations which pulled the estimated regression line towards themselves. This implies that if observation 31 and 34 are deleted from the model then the overall observed response would be much closer to the predicted response.

In summary, two observations had the highest influence on our regression model implying that if included in the model, they would substantially change the estimate of our coefficients. Observation 31 had a cook distance of 0.3113, hat value of 0.38319, dffit of 1.94181 and studentized residual of 2.46364. Observation 34 had a cook distance of 0.4772, hat value of 0.28905, dffit of -2.9917 and studentized residual of -4.69189

Question 4

Question 4

$(X_i, Y_{i1}, \ldots, Y_{in_i})$, $i = 1, \ldots, r$

Assuming that $X > 0$ and

$\text{Var}(Y|X=t) = X^q \sigma^2$ and q is known,

Let

$$X^{(i)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_i & x_i & \cdots & x_i \end{bmatrix}^T_{2 \times n_i}$$

$$X = \left(X^{(1)T}, X^{(2)T}, \ldots, X^{(r)T}\right)^T$$

$$Y_i = \left(Y_{i1}, Y_{i2}, \ldots Y_{in_i}\right)^T$$

$$Y = \left(Y_1^T, Y_2^T, \ldots, Y_r^T\right)^T$$

We know that $\varepsilon \sim N(0, \sigma^2)_n$ therefore

$$\varepsilon_i = [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i}]^T$$

$$\varepsilon = [\varepsilon_1^T, \varepsilon_2^T, \ldots, \varepsilon_r^T]^T$$

Since this is a fixed-x regression setting

$Y = X\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 W)_n$, $n = \sum_{i=1}^{r} n_i$

Also we know

$\text{Var}(Y|x=t) = X^q \sigma^2$ then this can be expressed as

$\text{Var}(Y|x=t) = \text{Var}(X\beta + \epsilon | x=t) = \text{Var}(\varepsilon | x=t)$

therefore :-

$$\begin{bmatrix} x_1^q & & & \\ & x_1^q & & \\ & & \ddots & \\ & & & x_r^q \end{bmatrix}^{-\frac{1}{2}}$$

$\therefore W = \text{diag}\{w_1, w_2 \ldots w_r\}$, where

$w_i = \text{diag}\{x_i^q, x_1^q, \ldots x_r^q\}$

Figure 2:

Since $X_i > 0$ then there is nonsingular matrix K, such that

$$W = K K^T$$

$\Rightarrow$ we can calculate $K = K^T = W^{\frac{1}{2}} = \text{diag}\{w_1^{\frac{1}{2}}, w_2^{\frac{1}{2}}, \ldots, w_r^{\frac{1}{2}}\}$

where $w_i^{\frac{1}{2}} = \text{diag}\{x_i^{q/2}, x_i^{q/2}, \ldots, x_i^{q/2}\}_{n \times r_i}$

and $K^{-1} = (K^T)^{-1} = W^{-\frac{1}{2}} = \text{diag}\{w_1^{-\frac{1}{2}}, w_2^{-\frac{1}{2}}, \ldots w_r^{-\frac{1}{2}}\}$

where $w_i^{-\frac{1}{2}} = \text{diag}\{x_i^{-\frac{q}{2}}, x_i^{\frac{q}{2}}, \ldots x_i^{-\frac{q}{2}}\}_{n \times r_i}$

$$K^{-1}Y = K^{-1} X\beta + K^{-1}\epsilon \quad \text{---- (1)}$$

Let $M^* = K^{-1}Y$, $P = K^{-1}X$, $\eta = K^{-1}\epsilon$ then (1) can be evaluated as

$$M^* = P\beta + \eta \quad \text{---- (2)}$$

We know that

$$\beta = (\beta_0, \beta_1) \text{ then}$$

$$M^* = \sigma^2 (K^{-1}K)(K^{-1}K)^T = \sigma^2 I_n,$$

Hence the Least Square estimator of β $(\hat{\beta})$:-

$$\hat{\beta} = (P^T P)^{-1} P^T M^* = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$$

Figure 3:

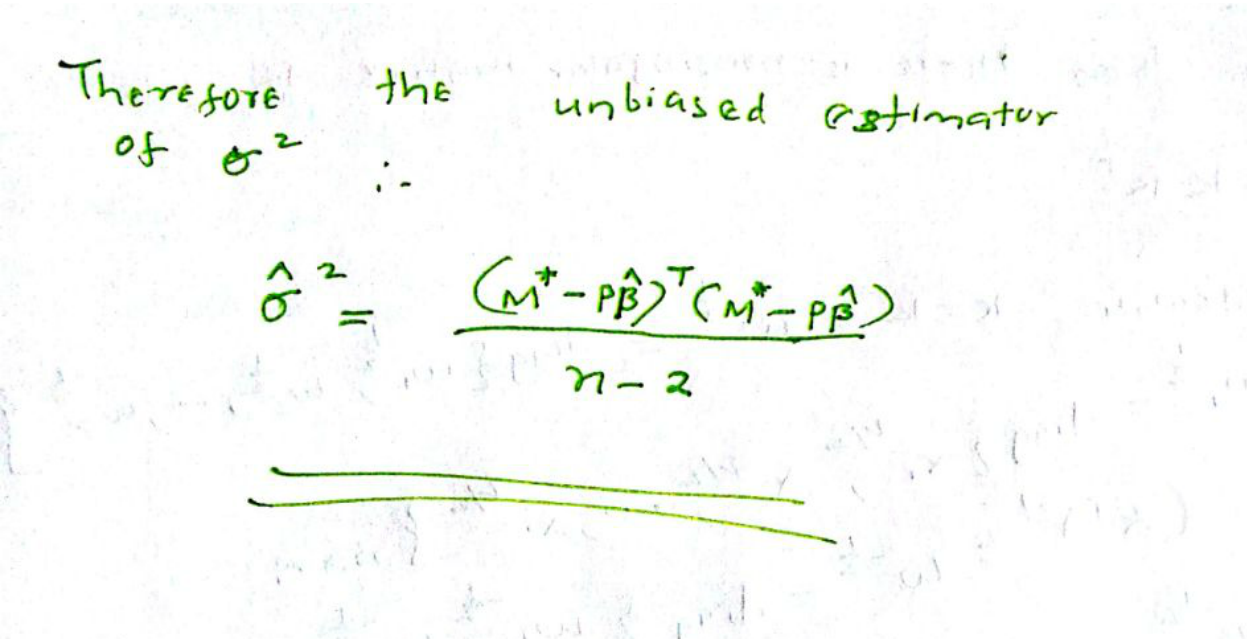

Therefore the unbiased estimator of σ^2 :.
$\hat{\sigma}^2 = \frac{(M^* - P\hat{\beta})^T (M^* - P\hat{\beta})}{n-2}$

Figure 4:

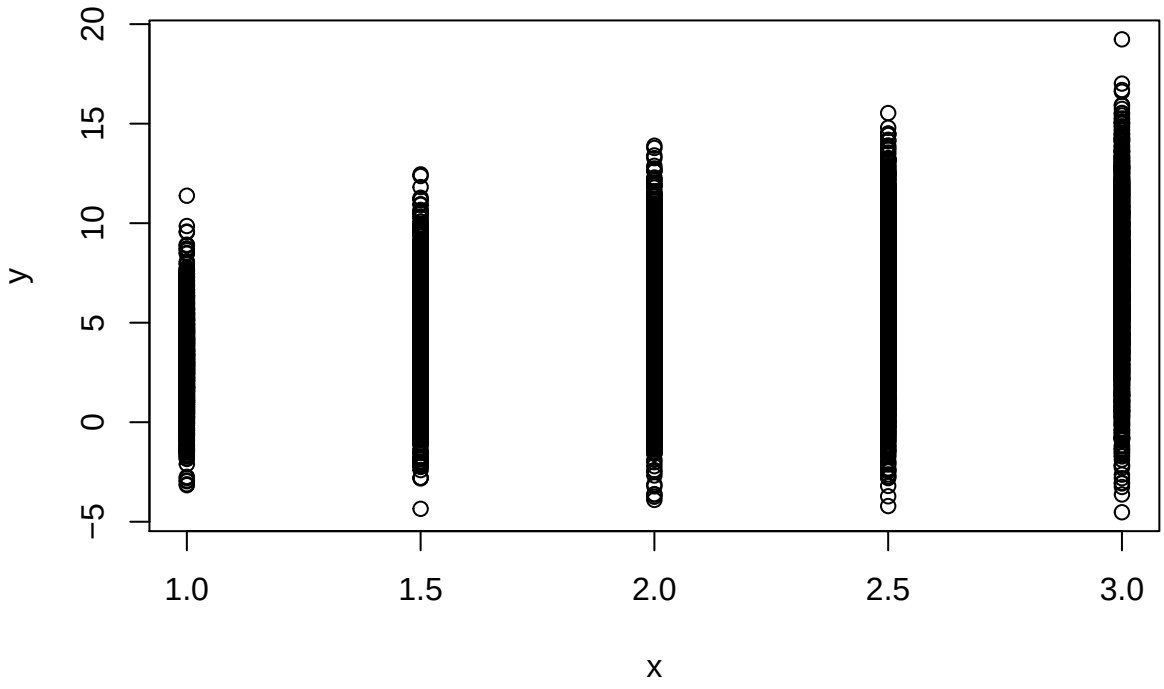
Question 5

LMQ Function

```
library(stats4)
library(dplyr)

lmq<-function(x,y){

  set.seed(120) #reproducability

  beta0=1 # init starting point
  beta1=1 # init starting point
  sigma=1 # init starting point
  q=1 # init starting point

  method = "L-BFGS-B" # or BFGS

  # par(mfrow = c(1,1))
  # plot(x,y)

  logLikelihood <- function(beta0,beta1,sigma,q){
    beta= t(c(beta0,beta1)) # transpose and convert to matrix
    resid = y - beta0 - beta1*x
    R = dnorm(resid, mean = 0, sd=sqrt(x^(q)*sigma^2), log=TRUE)
    -sum(R)
  }

  mle(logLikelihood, list(beta0=beta0, beta1=beta1, sigma=sigma, q=q), method =method)
}
```

Test Run 1: Simulate data with beta0=1.4, beta1=1.8 , sigma=2, and q=1

```
set.seed(120) #reproducability
# test run1
x = rep(seq(1,3,by=0.5),2000)
y = 1.4 + 1.8*x + rnorm(length(x),0,2*x^(1/2))
plot(x,y)
```

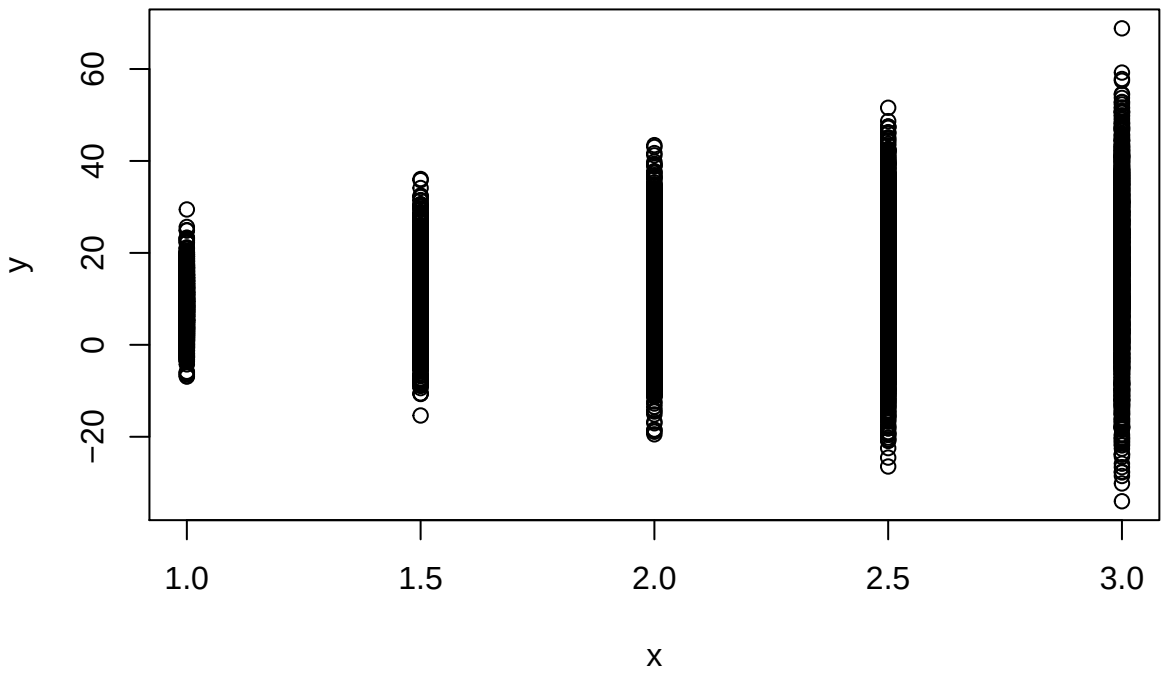



```
lmq(x,y)
```

```
##
## Call:
## mle(minuslogl = logLikelihood, start = list(beta0 = beta0, beta1 = beta1,
##     sigma = sigma, q = q), method = method)
##
## Coefficients:
##     beta0     beta1     sigma         q
## 1.2661222 1.8813507 1.9973277 0.9892181
```

Test Run 1: Simulate data with beta0=6, beta1=3 , sigma=5, and q=2

```
set.seed(120) #reproducability

# test run2
x = rep(seq(1,3,by=0.5),2000)
y = 6 + 3*x + rnorm(length(x),0,5*x^(2/2))
plot(x,y)
```




```
lmq(x,y)
```

```
##
## Call:
## mle(minuslogl = logLikelihood, start = list(beta0 = beta0, beta1 = beta1,
##     sigma = sigma, q = q), method = method)
##
## Coefficients:
##    beta0    beta1    sigma        q
## 5.553277 3.292531 4.992762 1.989496
```

Question 6

Question 6

We know that

$$Y = (Y^{(1)}, Y^{(2)})^T \quad \text{where } Y \text{ is a vector of response variable}$$

1st regression equation

$$\mu_i(t) = E(Y^{(i)} \mid X = t), \quad i = 1, 2$$

2nd regression equation

$$\mu_i(t) = \beta_{0i} + \beta_{1i} t, \quad i = 1, 2$$

Then the ratio of $\mu_2(t) / \mu_1(t)$

$$\frac{\mu_2(t)}{\mu_1(t)} = \frac{\hat{\beta}_{02} + \hat{\beta}_{12} t}{\hat{\beta}_{01} + \hat{\beta}_{11} t}$$

Assuming $\varepsilon^{(i)} \sim N(0, \sigma^2 I_n)$, then

$$Y^{(i)} = X\beta^{(i)} + \varepsilon^{(i)}$$

where $\beta^{(i)} = (\beta_{0i}, \beta_{1i})^T$

$$\varepsilon^{(i)} = (\varepsilon_1^{(i)}, \varepsilon_2^{(i)}, \ldots, \varepsilon_n^{(i)})^T$$

$$Y^{(i)} = (Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)})^T$$

$$X = \begin{bmatrix} 1 & 1, \ldots 1 \\ x_1, & x_2, \ldots x_n \end{bmatrix}^T$$

Figure 5:

$$\hat{\beta}^{(i)} = (X^T X)^{-1} X^T Y^{(i)}$$

$$\text{Var}(\hat{\beta}^{(i)}) = (X^T X)^{-1} X^T \cdot \text{var}(Y^{(i)}) \times (X^T X)^{-1}$$

Therefore :-

$$\text{Var}(\hat{\beta}^{(i)}) = \sigma^2 (X^T X)^{-1}$$

$$(X^T X) = \begin{bmatrix} n & \Sigma x_j \\ \Sigma x_j & \Sigma x_j^2 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{n \Sigma x_j^2 - (\Sigma x_j)^2} \begin{bmatrix} \Sigma x_j^2 & -\Sigma x_j \\ -\Sigma x_j & n \end{bmatrix}$$

Therefore

$$\text{Var}(\hat{\beta}^{(i)}) = \frac{\sigma^2}{n \Sigma (x_j - \bar{x})^2} \begin{bmatrix} \Sigma x_j^2 & -\Sigma x_j \\ -\Sigma x_j & n \end{bmatrix}$$

Assuming $\text{cov}(\varepsilon_j^{(1)}, \varepsilon_j^{(2)}) = (c\sigma^2)$ where c is a constant then

$$\Rightarrow \text{cov}(\varepsilon^{(1)}, \varepsilon^{(2)}) = c\sigma^2 \cdot 1_n \cdot 1_n^T \text{ where}$$

$$1_n = (1, 1, \cdots, 1)^T$$

$$\Rightarrow \hat{\beta}^{(1)} = (X^T X)^{-1} X^T Y^{(i)} = (X^T X)^{-1} X^T X \beta^{(1)} + (X^T X)^{-1} X^T \varepsilon^{(1)}$$

$$\Rightarrow \text{Cov}(\hat{\beta}^{(1)}, \hat{\beta}^{(2)}) = (X^T X)^{-1} X^T \text{cov}(\varepsilon^{(1)}, \varepsilon^{(2)}) \times (X^T X)^{-1}$$

$$= c\sigma^2 (X^T X)^{-1} (X^T 1_n)(X^T 1_n)^T \cdot (X^T X)^{-1}$$

Figure 6:

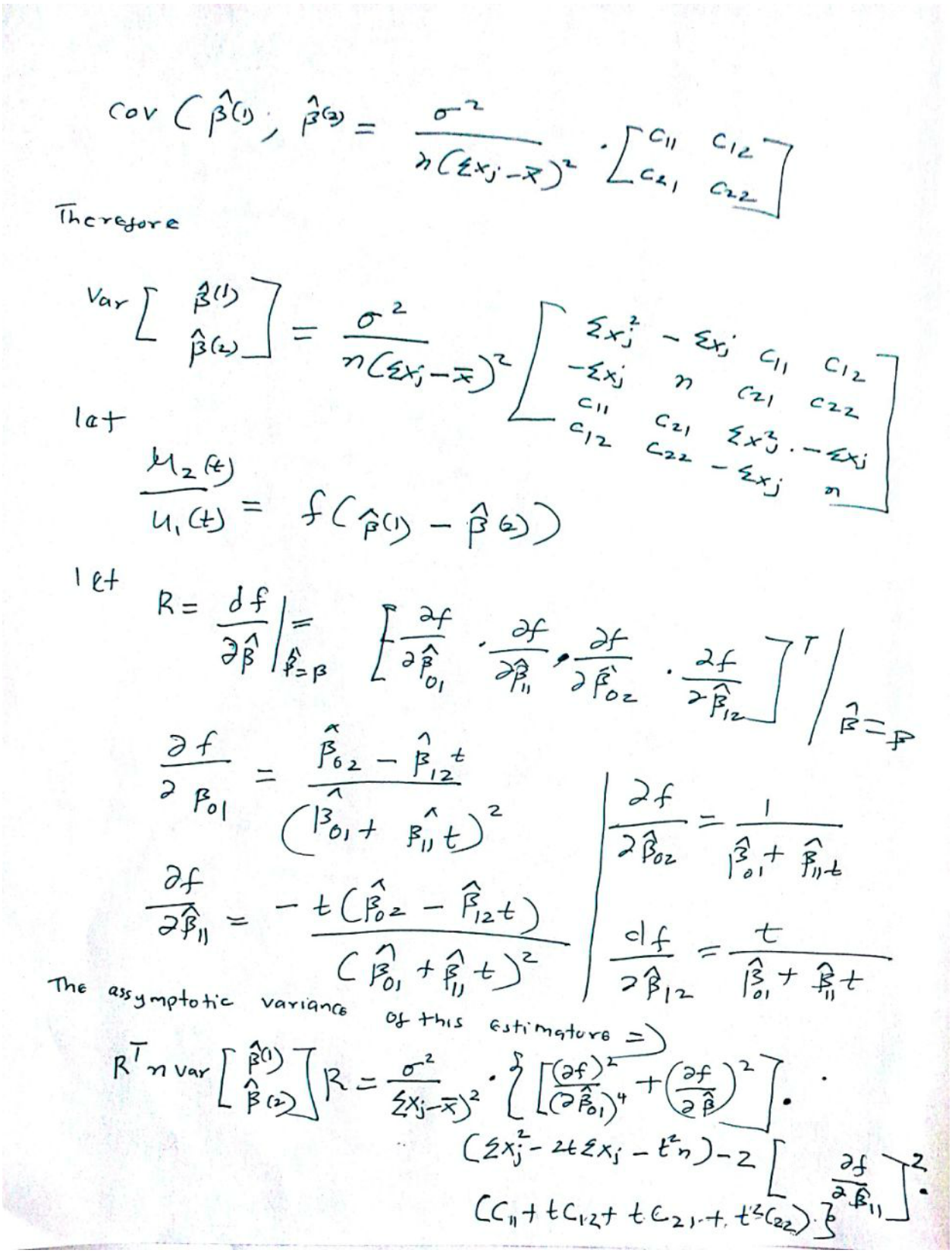

$\mathrm{cov}(\hat{\beta}(1), \hat{\beta}(2)) = \frac{\sigma^2}{n(\Sigma x_j - \bar{x})^2} \cdot \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$
Therefore
$\mathrm{Var}\begin{bmatrix} \hat{\beta}(1) \\ \hat{\beta}(2) \end{bmatrix} = \frac{\sigma^2}{n(\Sigma x_j - \bar{x})^2} \begin{bmatrix} \Sigma x_j^2 - \Sigma x_j & c_{11} & c_{12} \\ -\Sigma x_j & n & c_{21} & c_{22} \\ c_{11} & c_{21} & \Sigma x_j^3 \cdot -\Sigma x_j \\ c_{12} & c_{22} & -\Sigma x_j & n \end{bmatrix}$
let
$\frac{M_2(t)}{M_1(t)} = f(\hat{\beta}(1) - \hat{\beta}(2))$
let
$R = \frac{df}{\partial \hat{\beta}}\Big|_{\hat{\beta}=\beta} \left[\frac{\partial f}{\partial \hat{\beta}_{01}} \cdot \frac{\partial f}{\partial \hat{\beta}_{11}}, \frac{\partial f}{\partial \hat{\beta}_{02}} \cdot \frac{\partial f}{\partial \hat{\beta}_{12}}\right]^T \Big|_{\hat{\beta}=\beta}$
$\frac{\partial f}{\partial \beta_{01}} = \frac{\hat{\beta}_{02} - \hat{\beta}_{12} t}{(\hat{\beta}_{01} + \hat{\beta}_{11} t)^2}$
$\frac{\partial f}{\partial \hat{\beta}_{11}} = \frac{-t(\hat{\beta}_{02} - \hat{\beta}_{12} t)}{(\hat{\beta}_{01} + \hat{\beta}_{11} t)^2}$
$\frac{\partial f}{\partial \hat{\beta}_{02}} = \frac{1}{\hat{\beta}_{01} + \hat{\beta}_{11} t}$
$\frac{df}{\partial \hat{\beta}_{12}} = \frac{t}{\hat{\beta}_{01} + \hat{\beta}_{11} t}$
The assymptotic variance of this estimatore =)
$R^T n \,\mathrm{Var}\begin{bmatrix} \hat{\beta}(1) \\ \hat{\beta}(2) \end{bmatrix} R = \frac{\sigma^2}{\Sigma x_j - \bar{x})^2} \cdot \left\{ \left[\frac{(\partial f)^2}{(\partial \hat{\beta}_{01})^4} + \left(\frac{\partial f}{\partial \hat{\beta}}\right)^2\right] \cdot (\Sigma x_j^2 - 2t\Sigma x_j - t^2 n) - 2\left[\frac{\partial f}{\partial \hat{\beta}_{11}}\right]^2 \cdot (C_{11} + tC_{12} + tC_{21} + t^2(22)) \right\}$

Figure 7:

Source Code

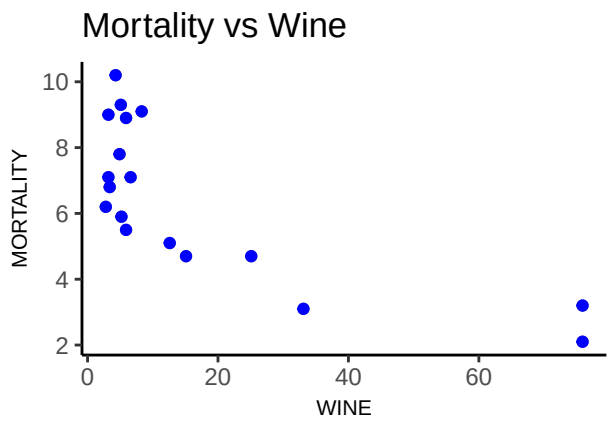
```
#' ---
#' title: "Homework 2"
#' author: "Allan Kimaina"
#' date: "February 22, 2018"
#' header-includes:
#' - \usepackage{pdflscape}
#' - \newcommand{\blandscape}{\begin{landscape}}
#' - \newcommand{\elandscape}{\end{landscape}}
#' output:
#'   pdf_document: default
#'   html_document: default
#' ---
#'
#'
#knitr::opts_chunk$set(echo = F)

# load package
library(dplyr)
library(car)
library(sjPlot)
library(sjmisc)
library(sjlabelled)
library(ggpubr)
library(ggpmisc)
library("gridExtra")
library(stargazer)
library(e1071)
library(jtools)
library(effects)
library(multcompView)
library(ggplot2)
library(ggrepel)
library(MASS)
library(broom)
library(ggcorrplot)
library(leaps)
library(relaimpo)
library(olsrr)


# load all data
wine = read.csv("data/Wine.csv")
bush = read.csv("data/Bush.csv")
bloodBrain = read.csv("data/BloodBrain.csv")

# clean bush
colnames(bush)[1] <- "county" # name the 1st column using standard
col2cvt <- 2:ncol(bush) # get all numeric columns
bush[,col2cvt] <- lapply(bush[,col2cvt],function(x){as.numeric(gsub(",", "", x))}) # convert to numeric

# clean wine
```

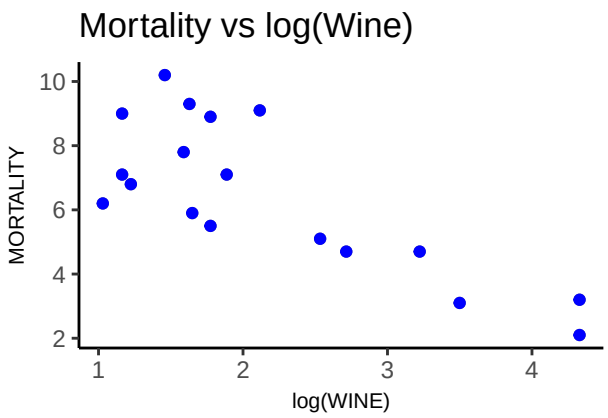
```r
colnames(wine)[1] <- "COUNTRY"

# clean wine
colnames(bloodBrain)[1] <- "BRAIN"

#'
#' \onecolumn
#'
#' # Question 1
#'
#' ## a. Using the wine data from the previous homework, create two new variables by logarithmically tr
#'
#'

wine$logMORTALITY <- log(wine$MORTALITY)
wine$logWINE <- log(wine$WINE)

# print
stargazer(wine,
          header=F,
          type = "latex",
          no.space = T,
          summary = F,
          single.row = T
          )
```

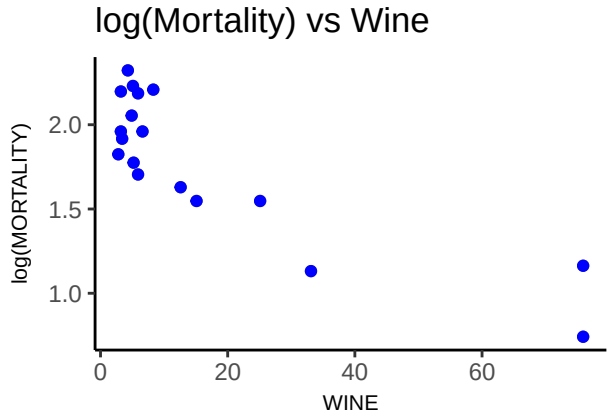
```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & COUNTRY & WINE & MORTALITY & logMORTALITY & logWINE \\
## \hline \\[-1.8ex]
## 1 & Norway & $2.800$ & $6.200$ & $1.825$ & $1.030$ \\
## 2 & Scotland & $3.200$ & $9$ & $2.197$ & $1.163$ \\
## 3 & England & $3.200$ & $7.100$ & $1.960$ & $1.163$ \\
## 4 & Ireland & $3.400$ & $6.800$ & $1.917$ & $1.224$ \\
## 5 & Finland & $4.300$ & $10.200$ & $2.322$ & $1.459$ \\
## 6 & Canada & $4.900$ & $7.800$ & $2.054$ & $1.589$ \\
## 7 & UnitedStates & $5.100$ & $9.300$ & $2.230$ & $1.629$ \\
## 8 & Netherlands & $5.200$ & $5.900$ & $1.775$ & $1.649$ \\
## 9 & NewZealand & $5.900$ & $8.900$ & $2.186$ & $1.775$ \\
## 10 & Denmark & $5.900$ & $5.500$ & $1.705$ & $1.775$ \\
## 11 & Sweden & $6.600$ & $7.100$ & $1.960$ & $1.887$ \\
## 12 & Australia & $8.300$ & $9.100$ & $2.208$ & $2.116$ \\
## 13 & Belgium & $12.600$ & $5.100$ & $1.629$ & $2.534$ \\
## 14 & Germany & $15.100$ & $4.700$ & $1.548$ & $2.715$ \\
## 15 & Austria & $25.100$ & $4.700$ & $1.548$ & $3.223$ \\
## 16 & Switzerland & $33.100$ & $3.100$ & $1.131$ & $3.500$ \\
## 17 & Italy & $75.900$ & $3.200$ & $1.163$ & $4.329$ \\
## 18 & France & $75.900$ & $2.100$ & $0.742$ & $4.329$ \\
## \hline \\[-1.8ex]
```

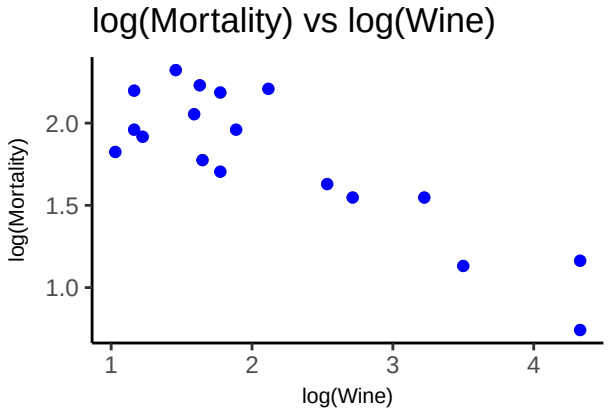
```
## \end{tabular}
## \end{table}
#'
#'
#' ## b. Plot mortality vs. wine; log mortality vs. wine, mortality vs. log wine and log mortality vs.
#'
#'

wine.model.plot = ggplot(wine, aes(y=MORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    labs(x='WINE', y='MORTALITY',
         title = "Mortality vs Wine")+
   # scale_x_log10() +
   # scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



wine.model.log.x.plot=ggplot(wine, aes(y=MORTALITY, x=logWINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    labs(x='log(WINE)', y='MORTALITY',
         title = "Mortality vs log(Wine)")+

    #scale_x_log10() +
    #scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



wine.model.log.y.plot=ggplot(wine, aes(y=logMORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    labs(x='WINE', y='log(MORTALITY)',
         title = "log(Mortality) vs Wine")+

   # scale_x_log10() +
   # scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
```

```
      axis.title.y = element_text( size=8)
    )




wine.model.log.xy.plot =ggplot(wine, aes(y=logMORTALITY, x=logWINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    labs(x='log(Wine)', y='log(Mortality)',
         title = "log(Mortality) vs log(Wine)")+

   #scale_x_log10() +
    #scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



  grid.arrange(wine.model.plot,
               wine.model.log.x.plot,
               wine.model.log.y.plot,
               wine.model.log.xy.plot,
               ncol =2.,
               nrow = 2.)
```







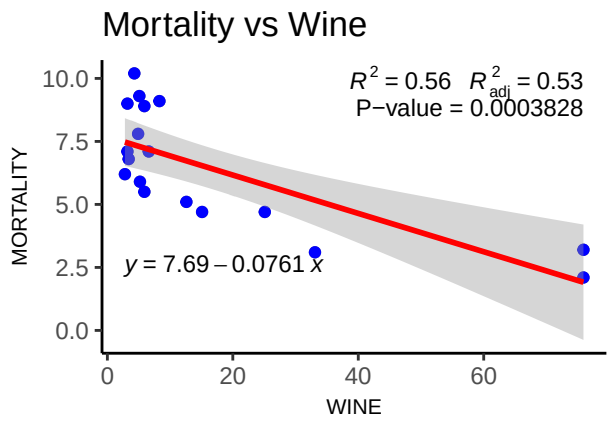


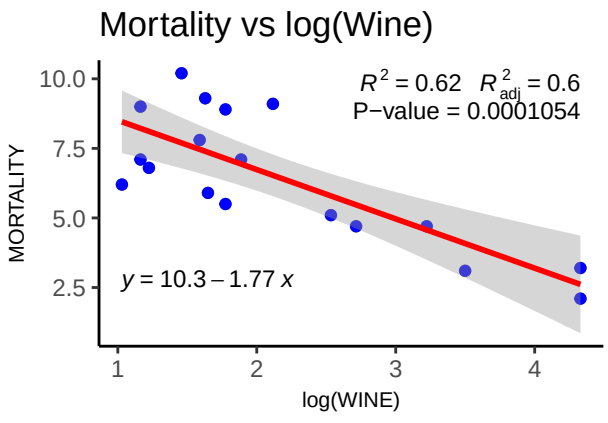

```
#'
#'
#' - From the above scatter plot, the logarithmic transformation of both dependent variable (log of Mor
#' - Transforming both predictor and response variable we get the best linear relationship
#'
#' ## c. Fit four linear regression models corresponding to the four plots and report the regression eq
#'
#'
#' * **Mortality vs Wine:** \(HeartAttackMortality =  7.69- 0.0761*WineConsumption + \epsilon\)
#' * **Mortality vs log(Wine):**  \(HeartAttackMortality =  10.280 - 1.771 *log(WineConsumption) + \eps
#' * **log(Mortality) vs Wine:**  \(log(HeartAttackMortality) =  2.0453 - 0.0159 *WineConsumption + \ep
#' * **log(Mortality) vs log(Wine):** \(log(HeartAttackMortality) =  2.5556 - 0.3556*log(WineConsumptio
#'
#'
#'

wine.model.plot.1=ggplot(wine, aes(y=MORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='WINE', y='MORTALITY',
         title = "Mortality vs Wine")+
    stat_poly_eq(aes(label = paste( ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                 label.x.npc = "right", label.y.npc = 0.87, geom = "text",
                 formula =  y ~x, parse = TRUE, size = 3)+
   stat_poly_eq(aes(label = paste(..eq.label.., sep = "~~~")),
```

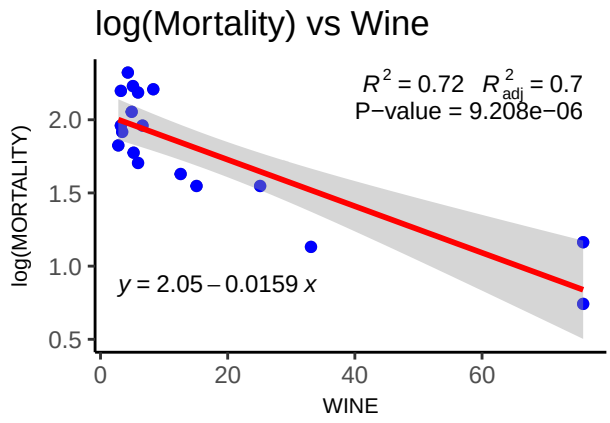
```
                label.x.npc = "left", label.y.npc = 0.001, geom = "text",
                formula =  y ~x, parse = TRUE, size = 3)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y~x),
                    geom = "text", size = 3,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
   # scale_x_log10() +
   # scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



wine.model.plot.2=ggplot(wine, aes(y=MORTALITY, x=logWINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='log(WINE)', y='MORTALITY',
         title = "Mortality vs log(Wine)")+
    stat_poly_eq(aes(label = paste( ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                label.x.npc = "right", label.y.npc = 0.87, geom = "text",
                formula =  y ~x, parse = TRUE, size = 3)+
   stat_poly_eq(aes(label = paste(..eq.label.., sep = "~~~")),
                label.x.npc = "left", label.y.npc = 0.03, geom = "text",
                formula =  y ~x, parse = TRUE, size = 3)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y~x),
                    geom = "text", size = 3,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
    #scale_x_log10() +
    #scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



wine.model.plot.3=ggplot(wine, aes(y=logMORTALITY, x=WINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='WINE', y='log(MORTALITY)',
```

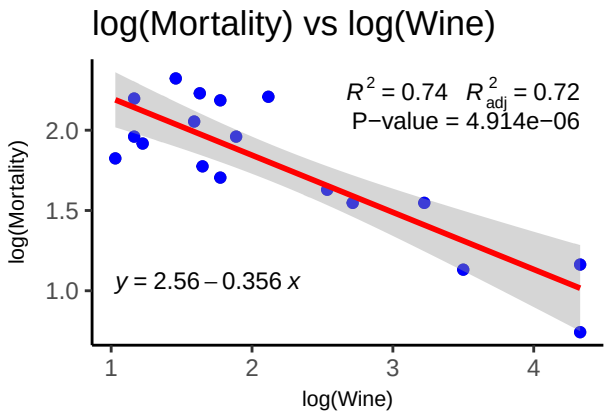
```
        title = "log(Mortality) vs Wine")+
    stat_poly_eq(aes(label = paste( ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                 label.x.npc = "right", label.y.npc = 0.87, geom = "text",
                 formula =  y ~x, parse = TRUE, size = 3)+
   stat_poly_eq(aes(label = paste(..eq.label.., sep = "~~~")),
                 label.x.npc = "left", label.y.npc = 0.03, geom = "text",
                 formula =  y ~x, parse = TRUE, size = 3)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y~x),
                    geom = "text", size = 3,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
   # scale_x_log10() +
    #scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )



wine.model.plot.4=ggplot(wine, aes(y=logMORTALITY, x=logWINE)) +
    geom_point(alpha = .5) +
    geom_point(color = "blue") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='log(Wine)', y='log(Mortality)',
         title = "log(Mortality) vs log(Wine)")+
    stat_poly_eq(aes(label = paste( ..rr.label.., ..adj.rr.label.., sep = "~~~")),
                 label.x.npc = "right", label.y.npc = 0.87, geom = "text",
                 formula =  y ~x, parse = TRUE, size = 3)+
  stat_poly_eq(aes(label = paste(..eq.label.., sep = "~~~")),
                 label.x.npc = "left", label.y.npc = 0.15, geom = "text",
                 formula =  y ~x, parse = TRUE, size = 3)+
    stat_fit_glance(method = "lm",
                    method.args = list(formula = y~x),
                    geom = "text", size = 3,
                    label.y.npc = 0.85, label.x.npc = "right",
                    aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
   #scale_x_log10() +
    #scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )

  grid.arrange(wine.model.plot.1,
               wine.model.plot.2,
               wine.model.plot.3,
               wine.model.plot.4,
```

```
              ncol =2.,
              nrow = 2.)
```









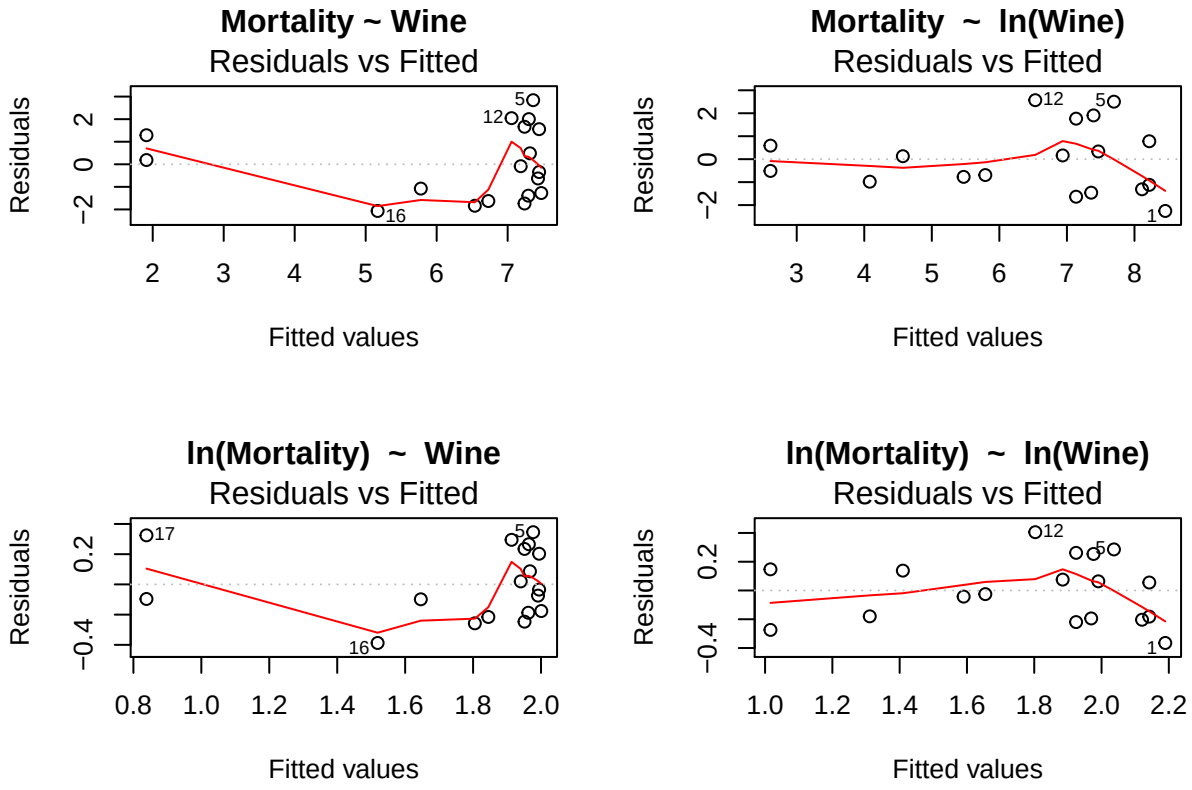

```
#'
#'
#' Graphically we notice that the regression line for the log-transformed model (ln(y) ~ ln(x)) covers
#'
#'
wine.model <- lm(MORTALITY~WINE, data=wine )
wine.model.log.x <- lm( MORTALITY~logWINE, data=wine )
wine.model.log.y <- lm( logMORTALITY~WINE, data=wine)
wine.model.log.xy <- lm( logMORTALITY~logWINE, data=wine )


stargazer(wine.model,wine.model.log.x,wine.model.log.y,wine.model.log.xy,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          ci = F,
         #keep = c("\\bprecip\\b"),
          title = "Comparison of the 4 Models",
          column.labels = c( "y~x", "y~ln(x)","ln(y)~x", "ln(y) vs ln(x)"),
          notes = "t-values have been hidden!",
            dep.var.caption  = "-" ,  # Bold
          single.row = T
```

```
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Comparison of the 4 Models}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{4}{c}{-} \\
## \cline{2-5}
## \\[-1.8ex] & \multicolumn{2}{c}{MORTALITY} & \multicolumn{2}{c}{logMORTALITY} \\
##  & y~x & y~ln(x) & ln(y)~x & ln(y) vs ln(x) \\
## \\[-1.8ex] & (1) & (2) & (3) & (4)\\
## \hline \\[-1.8ex]
##  WINE & $-$0.076$^{***}$ (0.017) &  & $-$0.016$^{***}$ (0.002) &  \\
##   logWINE &  & $-$1.771$^{***}$ (0.347) &  & $-$0.356$^{***}$ (0.053) \\
##   Constant & 7.687$^{***}$ (0.473) & 10.280$^{***}$ (0.832) & 2.045$^{***}$ (0.069) & 2.556$^{***}$
##  \hline \\[-1.8ex]
## Observations & 18 & 18 & 18 & 18 \\
## R$^{2}$ & 0.556 & 0.620 & 0.718 & 0.738 \\
## Adjusted R$^{2}$ & 0.528 & 0.596 & 0.700 & 0.722 \\
## Residual Std. Error (df = 16) & 1.619 & 1.498 & 0.238 & 0.229 \\
## F Statistic (df = 1; 16) & 20.026$^{***}$ & 26.090$^{***}$ & 40.639$^{***}$ & 45.170$^{***}$ \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{4}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
##  & \multicolumn{4}{r}{t-values have been hidden!} \\
## \end{tabular}
## \end{table}
```

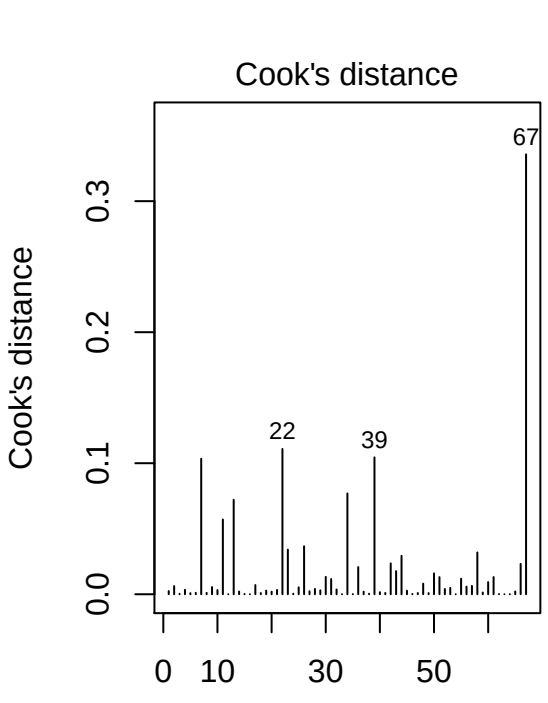
```
#'
#'
#' Looking at y~x and y~ln(x) models, we have really low explanatory power (R-squared). In  y~x model,
#'
#' Comparing the explanatory power between the  ln(y)~x model and ln(y)~ln(x) model, we get a higher ex
#'
#' For these reasons,  **the best fit model** is ln(y)~ln(x) model because it has the highest explanato
#'
#' ### Model Diagnostics
#'
#'
par( mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
# model 1
plot(wine.model, main = "Mortality ~ Wine",  which=c(1))


# model 2
plot(wine.model.log.x, main = "Mortality  ~  ln(Wine)",  which=c(1))

# model 3
plot(wine.model.log.y, main = "ln(Mortality)  ~  Wine",  which=c(1))

# model 4
```

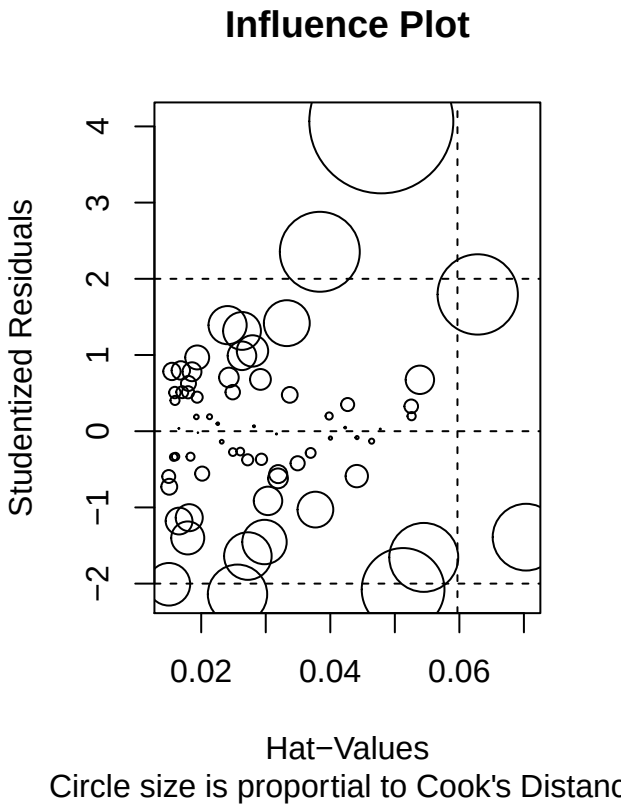
```r
plot(wine.model.log.xy, main = "ln(Mortality)  ~  ln(Wine)",  which=c(1))
```




```r
#'
#'
#' Looking at the residual plots,  ln(Mortality)  ~  ln(Wine) model exhibits acceptable randomness and
#'
#' The other models seem not to follow the rule of residual heteroskedasticity which is a crucial compo
#'
#' ### Conclusion
#'
#' In summary, the model with the natural log of heartAttackMortality as the response and the natural l
#'
#'
#' ## d. Interpret each regression model by describing the change in mortality for a given change in th
#'
#'
#' ### Mortality vs Wine
#'
#' - An increase of wine consumption by 10 Liters per person will decrease Ischemic heart attack mortal
#'
#' ### Mortality vs log(Wine)
#'
#' - Doubling wine consumption will decrease Ischemic heart attack mortality rate by 1.227 per 1000 per
#'
#' ### log(Mortality) vs Wine
#'
```

```
#' medianProportion = exp(-0.0159*10) =  0.8529964
#' %change = (exp(-0.0159*10)-1)*100 = -14.7%
#'
#' - An increase of wine consumption by 10 Liters per person will decrease Ischemic heart attack mortal
#' - An increase of wine consumption by 10 Liters per person will decrease Ischemic heart attack mortal
#'
#' ### log(Mortality) vs log(Wine)
#'
#' %change = 1-exp(-0.3556*log(2))*100 = -21.84555
#'
#' - Doubling wine consumption will decrease Ischemic heart attack mortality by 21.85%
#'
#' \onecolumn
#'
#' # Question 2: The Dramatic U.S. Presidential Election of 2000
#'
#' The U.S. presidential election of November 7, 2000 was one of the closest in history. As returns wer
#'
#'
#'
#'
#' ## a. The data in File Bush.xls contain the numbers of votes for Buchanan and Bush in all 67 countie
#'
#' ![Caption for the picture.](scatter.PNG)
#'
#' From the scatter plot  (Display 8.25) we can vividly see that Palm Beach County is one of the outlie
#'
#'
bush <- bush %>%
  mutate(total96=clinton96+dole96+perot96+buchanan96p)

# transformation
bush.log <- bush
bush.log$reform.reg = bush.log$reform.reg+0.00000001
for(i in 2:ncol(bush.log)){
  bush.log[,i] <- log(bush.log[,i])
}

bush.withPalmBeach.model <- lm(buchanan2000~bush2000, data=bush.log)

par( mfrow = c(1, 2))

# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(bush)-length(bush.withPalmBeach.model$coefficients)-2))
plot(bush.withPalmBeach.model, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(bush.withPalmBeach.model, id.method="identify", main="Influence Plot", sub="Circle size is
```





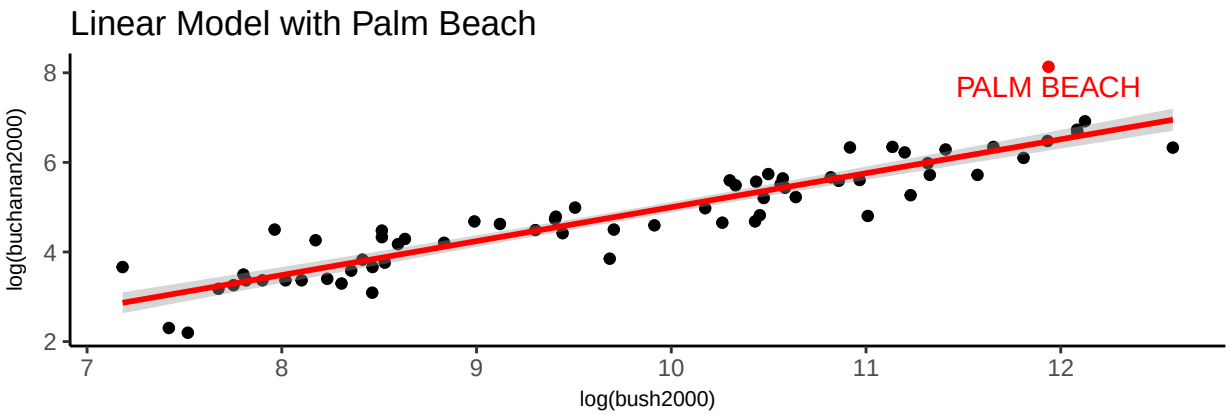

```
#'
#'
#'
#'
#'


# What evidence is there in the scatterplot of Display 8.25 that Buchanan received more votes than expe
# - Palm Beach County is an outlier

bushPalmBeach <- bush.log %>%
                     filter(county=='PALM BEACH')


bush.withoutPalmBeach <- bush.log %>%
                           filter(county!='PALM BEACH')


bush.withoutPalmBeach.model <- lm(buchanan2000~bush2000, data=bush.withoutPalmBeach)


stargazer(bush.withPalmBeach.model, bush.withoutPalmBeach.model,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
```

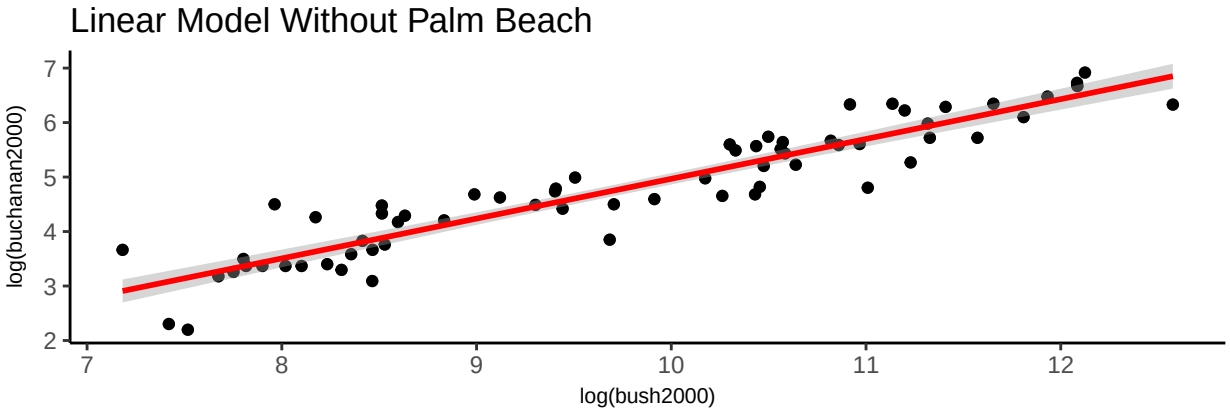
```
          ci = TRUE,
         keep = c("\\bprecip\\b"),
          title = "Models With and without outlier",
          column.labels = c( "Without Palm Beach", "With Palm Beach"),
          notes = "Coefficients have been removed!",
            dep.var.caption  = "-" ,  # Bold
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Models With and without outlier}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{2}{c}{-} \\
## \cline{2-3}
## \\[-1.8ex] & \multicolumn{2}{c}{buchanan2000} \\
##  & Without Palm Beach & With Palm Beach \\
## \\[-1.8ex] & (1) & (2)\\
## \hline \\[-1.8ex]
## \hline \\[-1.8ex]
## Observations & 67 & 66 \\
## R$^{2}$ & 0.851 & 0.866 \\
## Adjusted R$^{2}$ & 0.848 & 0.864 \\
## Residual Std. Error & 0.467 (df = 65) & 0.420 (df = 64) \\
## F Statistic & 370.615$^{***}$ (df = 1; 65) & 413.022$^{***}$ (df = 1; 64) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{2}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
##  & \multicolumn{2}{r}{Coefficients have been removed!} \\
## \end{tabular}
## \end{table}
```

```
#'
#'
#' Furthermore, if we remove this potential outlier from the dataset and regress Buchanan on Bush, our
#'
#'

bush.withoutPalmBeach.plot = ggplot(bush.withoutPalmBeach,
                                    aes(y=buchanan2000, x=bush2000
                                        )) +
    geom_point(alpha = .5) +
     geom_point(color = "black") +
    stat_smooth(method = "lm",  color="red")+
    labs(x='log(bush2000)', y='log(buchanan2000)',
         title = "Linear Model Without Palm Beach")+
    # stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., ..adj.rr.label.., sep = "~~~")),
    #              label.x.npc = "right", label.y.npc = 0.17, geom = "text",
    #              formula =  y ~x, parse = TRUE, size = 3)+
    # stat_fit_glance(method = "lm",
```

```
    #                 method.args = list(formula = y~x),
    #                 geom = "text", size = 3,
    #                 label.y.npc = 0.15, label.x.npc = "right",
    #                 aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
   # scale_x_log10() +
   # scale_y_log10()+
    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )


  bush.withPalmBeach.plot = ggplot(bush.log, aes(y=buchanan2000, x=bush2000, color=ifelse(((abs(buchanan
    geom_point(alpha = .5) +
    geom_point() +
    scale_color_manual(guide=FALSE, values=c("red", "black")) + #turn off the legend, define the colors
    geom_text_repel(data = subset(bush.log, buchanan2000 > 8), aes(label = county))+
    stat_smooth(method = "lm",  color="red")+
    labs(x='log(bush2000)', y='log(buchanan2000)',
         title = "Linear Model with Palm Beach")+
    # stat_poly_eq(aes(label = paste(..eq.label.., ..rr.label.., ..adj.rr.label.., sep = "~~~")),
    #              label.x.npc = "right", label.y.npc = 0.40, geom = "text",
    #              formula =  y ~x, parse = TRUE, size = 3)+
    # stat_fit_glance(method = "lm",
    #                 method.args = list(formula = y~x),
    #                 geom = "text", size = 3,
    #                 label.y.npc = 0.25, label.x.npc = "right",
    #                 aes(label = paste("P-value = ", signif(..p.value.., digits = 4), sep = "")))+
   # scale_x_log10() +
   # scale_y_log10()+

    theme_classic()+ guides(fill=FALSE)+
    theme(
      axis.title = element_text( size=8),
      axis.title.x = element_text( size=8),
      axis.title.y = element_text( size=8)
    )
# Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result-assu


 grid.arrange(bush.withPalmBeach.plot,bush.withoutPalmBeach.plot , ncol = 1,
              nrow = 2)
```






```
#'
#'
#'
#'
#' Using the regression model without palm beach to generate a 95% prediction interval for the number o
#'
#'

bushPalmBeach.predict <- predict(bush.withoutPalmBeach.model,bushPalmBeach, interval = "prediction")

stargazer(exp(bushPalmBeach.predict), title="Prediction for Buchanana",   type = "latex",   header=FALS
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Prediction for Buchanana}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & fit & lwr & upr \\
## \hline \\[-1.8ex]
## 1 & $592.377$ & $250.800$ & $1,399.164$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
#'
#'
#'
#' In fact, the observed votes for Buchanan (3407) is way far from our 95 % prediction interval of [250
#'
#' ## b. Analyze the data in Ex1222 and write a statistical summary predicting the number of Buchanan,
#'
#' We used Stepwise AIC for predictor variable selection and after evaluating the relative importance m
#'
#' ### Variable Selection
#'
#' The full model containing all potential predictor variables were passed into the step function. The
#'
#' **Full Model: ** buchanan2000~bush2000+gore2000+nader2000+browne2000+total2000+clinton96+ dole96+per
#'
#' After the step iterations were completed the full model was reduced to a model with
#' 7 predictors:
#'
#' **StepWise AIC Model: **: buchanan2000 ~ nader2000 + browne2000 + total2000 + clinton96 + perot96 +
#'
#' The rule of thumb for a good predictive accuracy has always been to look at no more variables than 1
#'
#'


bush.withoutPalmBeach.allSubset.model <- lm(buchanan2000~bush2000+gore2000+nader2000+browne2000+total200
                                    dole96+perot96+buchanan96p+reform.reg+total.reg, data=bush.withoutPa

#plot(bush.withoutPalmBeach.stepAIC.model)
#summary(bush.withoutPalmBeach.allSubset.model)
#names(bush.withoutPalmBeach)

bush.withoutPalmBeach.stepAIC  <- stepAIC(bush.withoutPalmBeach.allSubset.model, direction="both")
```

```
## Start:  AIC=-114.03
## buchanan2000 ~ bush2000 + gore2000 + nader2000 + browne2000 +
##     total2000 + clinton96 + dole96 + perot96 + buchanan96p +
##     reform.reg + total.reg
##
##               Df Sum of Sq    RSS     AIC
## - dole96       1   0.02534 8.1775 -115.83
## - bush2000     1   0.02877 8.1809 -115.80
## - gore2000     1   0.06020 8.2124 -115.55
## - total2000    1   0.11357 8.2657 -115.12
## - reform.reg   1   0.18596 8.3381 -114.54
## - clinton96    1   0.23665 8.3888 -114.14
## <none>                     8.1522 -114.03
## - nader2000    1   0.45550 8.6077 -112.44
## - total.reg    1   0.45825 8.6104 -112.42
## - buchanan96p  1   0.60967 8.7618 -111.27
## - browne2000   1   0.84899 9.0012 -109.49
## - perot96      1   1.28578 9.4379 -106.36
##
## Step:  AIC=-115.83
```

```
## buchanan2000 ~ bush2000 + gore2000 + nader2000 + browne2000 +
##     total2000 + clinton96 + perot96 + buchanan96p + reform.reg +
##     total.reg
##
##               Df Sum of Sq    RSS     AIC
## - bush2000     1   0.01987 8.1974 -117.67
## - gore2000     1   0.09613 8.2736 -117.05
## - total2000    1   0.12732 8.3048 -116.81
## - reform.reg   1   0.17616 8.3537 -116.42
## <none>                     8.1775 -115.83
## - clinton96    1   0.25498 8.4325 -115.80
## - total.reg    1   0.43430 8.6118 -114.41
## - nader2000    1   0.47771 8.6552 -114.08
## + dole96       1   0.02534 8.1522 -114.03
## - buchanan96p  1   0.61230 8.7898 -113.06
## - browne2000   1   0.82880 9.0063 -111.45
## - perot96      1   1.28337 9.4609 -108.20
##
## Step:  AIC=-117.67
## buchanan2000 ~ gore2000 + nader2000 + browne2000 + total2000 +
##     clinton96 + perot96 + buchanan96p + reform.reg + total.reg
##
##               Df Sum of Sq    RSS     AIC
## - reform.reg   1   0.17807 8.3754 -118.25
## - gore2000     1   0.19713 8.3945 -118.10
## - clinton96    1   0.24179 8.4392 -117.75
## <none>                     8.1974 -117.67
## - total.reg    1   0.42328 8.6207 -116.34
## - nader2000    1   0.45785 8.6552 -116.08
## + bush2000     1   0.01987 8.1775 -115.83
## + dole96       1   0.01644 8.1809 -115.80
## - buchanan96p  1   0.63120 8.8286 -114.77
## - browne2000   1   0.89544 9.0928 -112.82
## - perot96      1   1.35981 9.5572 -109.54
## - total2000    1   1.78421 9.9816 -106.67
##
## Step:  AIC=-118.25
## buchanan2000 ~ gore2000 + nader2000 + browne2000 + total2000 +
##     clinton96 + perot96 + buchanan96p + total.reg
##
##               Df Sum of Sq     RSS     AIC
## - gore2000     1   0.16100  8.5364 -118.99
## - clinton96    1   0.25686  8.6323 -118.25
## <none>                      8.3754 -118.25
## + reform.reg   1   0.17807  8.1974 -117.67
## - buchanan96p  1   0.46193  8.8374 -116.70
## + bush2000     1   0.02177  8.3537 -116.42
## + dole96       1   0.00865  8.3668 -116.31
## - nader2000    1   0.51957  8.8950 -116.28
## - total.reg    1   0.58084  8.9563 -115.82
## - browne2000   1   0.80962  9.1851 -114.16
## - perot96      1   1.27102  9.6465 -110.92
## - total2000    1   1.71053 10.0860 -107.98
##
```

```
## Step:  AIC=-118.99
## buchanan2000 ~ nader2000 + browne2000 + total2000 + clinton96 +
##     perot96 + buchanan96p + total.reg
##
##               Df Sum of Sq     RSS     AIC
## <none>                      8.5364 -118.99
## - buchanan96p  1   0.31244  8.8489 -118.62
## + gore2000     1   0.16100  8.3754 -118.25
## + reform.reg   1   0.14194  8.3945 -118.10
## + dole96       1   0.13722  8.3992 -118.06
## + bush2000     1   0.09403  8.4424 -117.72
## - total.reg    1   0.64363  9.1801 -116.19
## - browne2000   1   0.82930  9.3657 -114.87
## - clinton96    1   1.08738  9.6238 -113.08
## - nader2000    1   1.20657  9.7430 -112.27
## - total2000    1   1.96845 10.5049 -107.30
## - perot96      1   2.02449 10.5609 -106.94
```

```
bush.withoutPalmBeach.stepAIC.anova <- bush.withoutPalmBeach.stepAIC$anova



# final model results
bush.WPB.stepAIC.best.model <-lm(buchanan2000 ~ nader2000 + browne2000 + total2000 + clinton96 +
    perot96 + buchanan96p + total.reg, data=bush.withoutPalmBeach)


#'
#'
#'
stargazer(bush.WPB.stepAIC.best.model,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          ci = TRUE,
        # keep = c("\\bprecip\\b"),
          title = "Best Model",

         # notes = "Coefficients have been removed!",
            dep.var.caption  = "-" ,  # Bold
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Best Model}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{-} \\
## \cline{2-2}
## \\[-1.8ex] & buchanan2000 \\
```

```
## \hline \\[-1.8ex]
##  nader2000 & $-$0.491$^{***}$ ($-$0.826, $-$0.155) \\
##   browne2000 & 0.254$^{**}$ (0.044, 0.464) \\
##   total2000 & 0.899$^{***}$ (0.417, 1.380) \\
##   clinton96 & $-$0.522$^{***}$ ($-$0.898, $-$0.146) \\
##   perot96 & 0.768$^{***}$ (0.362, 1.175) \\
##   buchanan96p & $-$0.146 ($-$0.343, 0.051) \\
##   total.reg & 0.149$^{**}$ (0.009, 0.289) \\
##   Constant & $-$4.629$^{***}$ ($-$6.665, $-$2.593) \\
##  \hline \\[-1.8ex]
## Observations & 66 \\
## R$^{2}$ & 0.898 \\
## Adjusted R$^{2}$ & 0.886 \\
## Residual Std. Error & 0.384 (df = 58) \\
## F Statistic & 73.312$^{***}$ (df = 7; 58) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#'
#'
#' This model has very high explanatory power, with R^2 of 0.898 and Adjusted R^2 of 0.886. Over 89.8% 
#'
#' ### Relative Importance Measure
#'
#' After building the model using stepwise variable selection, We went ahead and evaluated the relative
#'
#'
#bush.WPB.relimp<- calc.relimp(bush.WPB.stepAIC.best.model,type=c("lmg"), rela=TRUE)

# Bootstrap Measures of Relative Importance (1000 samples)
#bush.WPB.stepAIC.best.boot <- boot.relimp(bush.WPB.stepAIC.best.model, b = 1000, type = c("lmg",
#  "last", "first", "pratt"), rank = TRUE,
 # diff = TRUE, rela = TRUE)
#booteval.relimp(bush.WPB.stepAIC.best.boot) # print result
#plot(booteval.relimp(bush.WPB.stepAIC.best.boot,sort=TRUE)) # plot result

#'
#'
#' Using Lindeman, Merenda, and Gold method, all the predictors selected from the stepWise regression h
#'
#'
#'
#' ### Prediction
#'
#'
#'

bush.WPB.stepAIC.predicted <-predict(bush.WPB.stepAIC.best.model,
                                     bush.log[67,],
                                     se.fit=T, interval="pred")
```

```r
stargazer(exp(bush.WPB.stepAIC.predicted$fit), title="Best Prediction for Buchanana",   type = "latex",
```

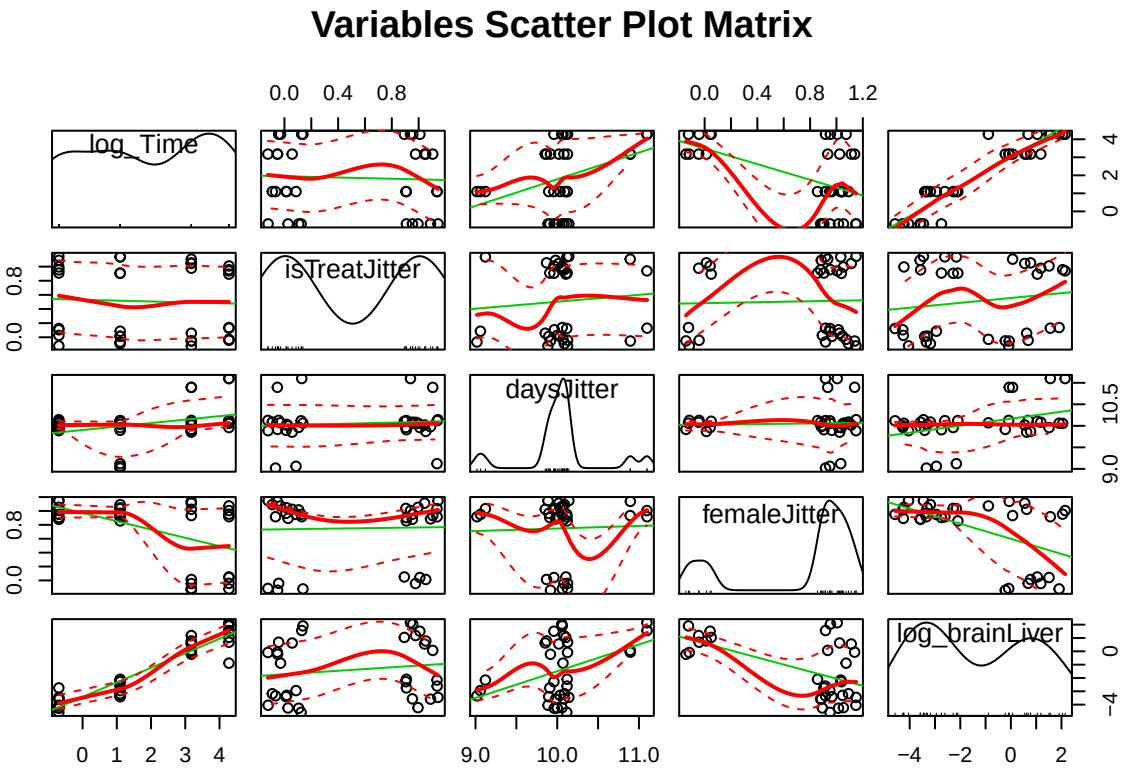
```
##
## \begin{table}[!htbp] \centering
##   \caption{Best Prediction for Buchanana}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & fit & lwr & upr \\
## \hline \\[-1.8ex]
## 67 & $431.903$ & $164.101$ & $1,136.736$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```r
#'
#'
#' ### Assumptions
#'
#' - Due to the fact we were modeling for predictive regression, we didn't assess multicollinearity
#' - Relative importance is not as effective when regressors are correlated. We did not evaluate collin
#' - We did not have a variable for total1996 and observations for other candidates in 1996. For us to
#'
#'
#' We used the above model to predict Buchanan's 2000 vote in Palm Beach County. Among the candidate pr
#'
#' The Observed number of votes for Buchanan was 3407 which was way higher than expected number of vote
#'
#'
#'
#'
#'
#'
#' \onecolumn
#'
#' # Question 3
#'
#' ## a. Compute "Jittered" versions of-treatment, days after inoculation, and an indicator variable fo
#'
#'
bloodBrain$brainLiver= bloodBrain$BRAIN/bloodBrain$LIVER
bloodBrain=bloodBrain%>%
  mutate(

    isTreatJitter=jitter(ifelse(TREAT=='BD',1,0),amount = 0.15),
    daysJitter=jitter(DAYS,amount = 0.15),
    femaleJitter=jitter(ifelse(SEX=='F',1,0),amount = 0.15)
    )

bloodBrain$log_brainLiver= log(bloodBrain$brainLiver)
bloodBrain$log_Time= log(bloodBrain$TIME)
```

```
stargazer(bloodBrain[c(1,10:ncol(bloodBrain))],
          header=F,
          type = "latex",
          no.space = T,
          summary = F,
          single.row = T
          )
```

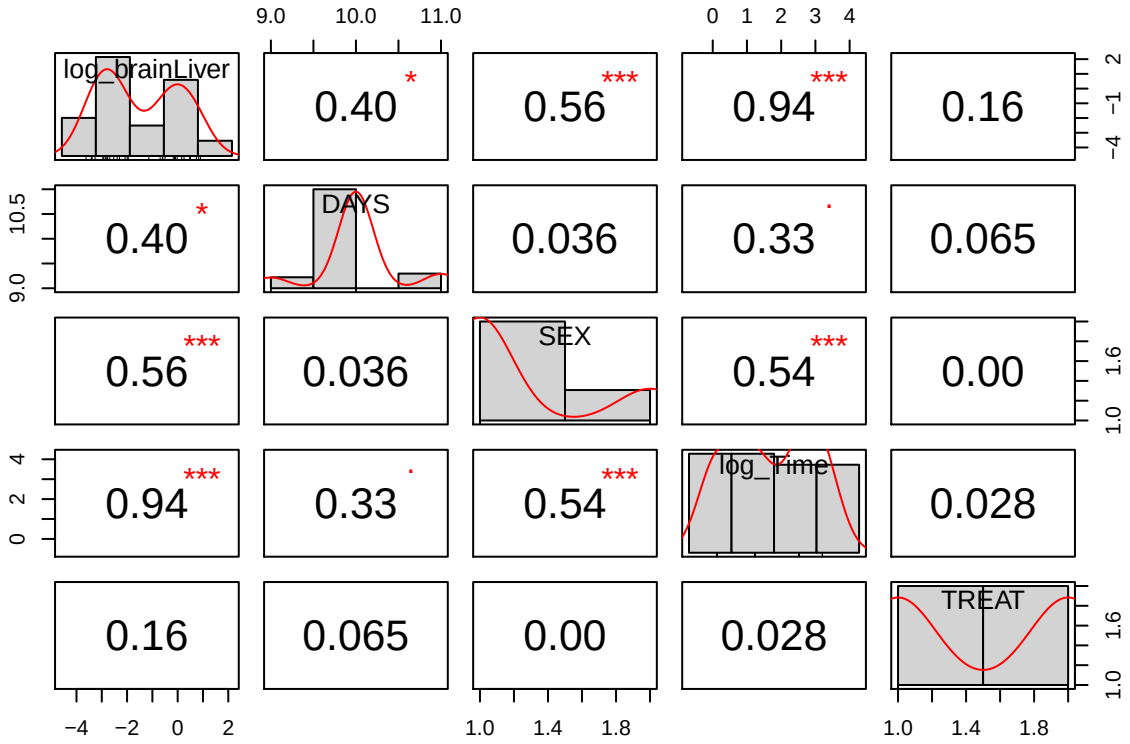
```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & BRAIN & brainLiver & isTreatJitter & daysJitter & femaleJitter & log\_brainLiver & log\_Time \\
## \hline \\[-1.8ex]
## 1 & $41,081$ & $0.028$ & $0.953$ & $10.084$ & $0.880$ & $-$3.568$ & $-$0.693$ \\
## 2 & $44,286$ & $0.028$ & $1.089$ & $9.980$ & $0.885$ & $-$3.588$ & $-$0.693$ \\
## 3 & $102,926$ & $0.064$ & $0.909$ & $9.909$ & $0.905$ & $-$2.745$ & $-$0.693$ \\
## 4 & $25,927$ & $0.015$ & $0.995$ & $10.060$ & $1.053$ & $-$4.227$ & $-$0.693$ \\
## 5 & $42,643$ & $0.032$ & $1.144$ & $10.142$ & $1.149$ & $-$3.456$ & $-$0.693$ \\
## 6 & $31,342$ & $0.018$ & $-$0.122$ & $9.887$ & $1.148$ & $-$4.045$ & $-$0.693$ \\
## 7 & $22,815$ & $0.014$ & $0.106$ & $9.964$ & $1.015$ & $-$4.271$ & $-$0.693$ \\
## 8 & $16,629$ & $0.010$ & $0.124$ & $10.117$ & $0.951$ & $-$4.578$ & $-$0.693$ \\
## 9 & $22,315$ & $0.014$ & $0.027$ & $10.000$ & $0.893$ & $-$4.252$ & $-$0.693$ \\
## 10 & $77,961$ & $0.074$ & $1.138$ & $10.121$ & $0.916$ & $-$2.610$ & $1.099$ \\
## 11 & $73,178$ & $0.102$ & $0.905$ & $9.960$ & $1.037$ & $-$2.280$ & $1.099$ \\
## 12 & $76,167$ & $0.123$ & $0.908$ & $10.122$ & $0.857$ & $-$2.097$ & $1.099$ \\
## 13 & $123,730$ & $0.116$ & $1.137$ & $9.122$ & $1.033$ & $-$2.156$ & $1.099$ \\
## 14 & $25,569$ & $0.035$ & $-$0.067$ & $9.018$ & $0.917$ & $-$3.340$ & $1.099$ \\
## 15 & $33,803$ & $0.033$ & $0.012$ & $9.896$ & $0.965$ & $-$3.406$ & $1.099$ \\
## 16 & $24,512$ & $0.037$ & $0.017$ & $10.034$ & $1.046$ & $-$3.305$ & $1.099$ \\
## 17 & $50,545$ & $0.053$ & $0.086$ & $9.060$ & $0.949$ & $-$2.945$ & $1.099$ \\
## 18 & $50,690$ & $0.042$ & $-$0.091$ & $10.089$ & $1.092$ & $-$3.181$ & $1.099$ \\
## 19 & $84,616$ & $1.733$ & $1.044$ & $10.019$ & $1.112$ & $0.550$ & $3.178$ \\
## 20 & $55,153$ & $3.266$ & $0.977$ & $10.082$ & $-$0.045$ & $1.184$ & $3.178$ \\
## 21 & $48,829$ & $2.180$ & $1.053$ & $9.870$ & $0.014$ & $0.779$ & $3.178$ \\
## 22 & $89,454$ & $1.071$ & $1.105$ & $10.893$ & $0.934$ & $0.069$ & $3.178$ \\
## 23 & $37,928$ & $1.866$ & $0.057$ & $9.853$ & $1.024$ & $0.624$ & $3.178$ \\
## 24 & $12,816$ & $0.802$ & $-$0.051$ & $9.917$ & $-$0.141$ & $-$0.221$ & $3.178$ \\
## 25 & $23,734$ & $0.917$ & $-$0.123$ & $10.126$ & $-$0.116$ & $-$0.087$ & $3.178$ \\
## 26 & $31,097$ & $0.936$ & $-$0.053$ & $10.895$ & $1.138$ & $-$0.066$ & $3.178$ \\
## 27 & $35,395$ & $8.545$ & $0.939$ & $11.098$ & $1.006$ & $2.145$ & $4.277$ \\
## 28 & $18,270$ & $7.728$ & $0.957$ & $10.056$ & $0.954$ & $2.045$ & $4.277$ \\
## 29 & $5,625$ & $2.842$ & $0.893$ & $10.106$ & $0.050$ & $1.045$ & $4.277$ \\
## 30 & $7,497$ & $4.519$ & $1.009$ & $9.960$ & $0.042$ & $1.508$ & $4.277$ \\
## 31 & $6,250$ & $6.735$ & $0.138$ & $10.062$ & $-$0.123$ & $1.907$ & $4.277$ \\
## 32 & $11,519$ & $4.754$ & $0.129$ & $11.099$ & $0.917$ & $1.559$ & $4.277$ \\
## 33 & $3,184$ & $1.980$ & $-$0.043$ & $10.127$ & $-$0.042$ & $0.683$ & $4.277$ \\
## 34 & $1,334$ & $0.411$ & $-$0.029$ & $10.071$ & $1.073$ & $-$0.888$ & $4.277$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
#'
#'
#' ## b. Obtain a matrix of scatter plots for the following variables: log sacrifice time, treatment (j
#'
#'
#pairs(bloodBrain[11:15],pch=21)
scatterplotMatrix(~log_Time+isTreatJitter+daysJitter+femaleJitter+log_brainLiver, data=bloodBrain,
    main="Variables Scatter Plot Matrix")
```




```
#'
#'
#' ## c. Obtain a matrix of the correlation coefficients among the same five variables (not jittered).
#'
#'
#library(ggcorrplot)

#data(mtcars)
#corr <- cor(bloodBrain[c(10,4,5,6,3)])
# Second Correlogram Example
# Correlations with significance levels
#library(Hmisc)
#mtcars is a data frame
#rcorr(as.matrix(bloodBrain[c(10,4,5,6,3)]))
  hist.panel = function (x, ...=NULL ) {

        par(new = TRUE)
```

```
        hist(x,

             col = "light gray",

             probability = TRUE,

             axes = FALSE,

             main = "",

             breaks = "FD")

        lines(density(x, na.rm=TRUE),

              col = "red",

              lwd = 1)

        #lines(f, col="blue", lwd=1, lty=1) how to add gaussian normal overlay?

        rug(x)

  }

corr.panel <- function(x, y, digits=2, prefix="", cex.cor)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

    test <- cor.test(x,y)
    # borrowed from printCoefmat
    Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                  cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                  symbols = c("***", "**", "*", ".", " "))

    text(0.5, 0.5, txt, cex = 2)
    text(.8, .8, Signif, cex=1.5, col=2)
}



pairs(bloodBrain[c(14,5,6,15,4)], lower.panel=corr.panel, upper.panel=corr.panel,  diag.panel=hist.panel
```




```
#'
#'
#'
#' ## d.    On the basis of this, what can be said about the relationship between the covariates (sex a
#'
#'
#' We have a very strong and very significant relationship between log response variable (ratio between
#'
#' In summary, we don't have collinearity within the design variables (treatment and log of sacrifice t
#'
#' \onecolumn
#'
#' ## e.    Fit the regression of the log response (brain tumor-to-liver antibody ratio) on an indicato
#'
#'

bloodBrain$TIMEFactor<-as.factor(bloodBrain$TIME)
bloodBrain.model <- lm(log_brainLiver~TIMEFactor+TREAT,data=bloodBrain)

stargazer(bloodBrain.model,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          ci = TRUE,
        # keep = c("\\bprecip\\b"),
```

```
          title = "",

         # notes = "Coefficients have been removed!",
           dep.var.caption  = "-" ,  # Bold
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{-} \\
## \cline{2-2}
## \\[-1.8ex] & log\_brainLiver \\
## \hline \\[-1.8ex]
##  TIMEFactor3 & 1.134$^{***}$ (0.640, 1.628) \\
##   TIMEFactor24 & 4.257$^{***}$ (3.749, 4.765) \\
##   TIMEFactor72 & 5.154$^{***}$ (4.646, 5.662) \\
##   TREATNS & $-$0.797$^{***}$ ($-$1.156, $-$0.437) \\
##   Constant & $-$3.505$^{***}$ ($-$3.888, $-$3.122) \\
##  \hline \\[-1.8ex]
## Observations & 34 \\
## R$^{2}$ & 0.951 \\
## Adjusted R$^{2}$ & 0.944 \\
## Residual Std. Error & 0.533 (df = 29) \\
## F Statistic & 139.646$^{***}$ (df = 4; 29) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
bloodBrain.treatment <- data.frame(
  TIMEFactor=as.factor(rep(c(.5,3,24,72),2)),
   TREAT=as.factor(c(rep("BD",4),rep("NS",4)))
)

bloodBrain.model.predicted <- cbind(bloodBrain.treatment,
data.frame(predict(bloodBrain.model,bloodBrain.treatment, interval = "prediction")))

stargazer(bloodBrain.model.predicted,
          header=F,
          type = "latex",
          no.space = T,
          summary = F,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
```

```
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & TIMEFactor & TREAT & fit & lwr & upr \\
## \hline \\[-1.8ex]
## 1 & 0.5 & BD & $$-$3.505$ & $$-$4.666$ & $$-$2.344$ \\
## 2 & 3 & BD & $$-$2.371$ & $$-$3.538$ & $$-$1.203$ \\
## 3 & 24 & BD & $0.752$ & $$-$0.419$ & $1.923$ \\
## 4 & 72 & BD & $1.649$ & $0.478$ & $2.820$ \\
## 5 & 0.5 & NS & $$-$4.302$ & $$-$5.469$ & $$-$3.134$ \\
## 6 & 3 & NS & $$-$3.168$ & $$-$4.328$ & $$-$2.007$ \\
## 7 & 24 & NS & $$-$0.044$ & $$-$1.215$ & $1.126$ \\
## 8 & 72 & NS & $0.852$ & $$-$0.319$ & $2.023$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
#'
#' \onecolumn
#'
#' ## f.    Let X represent log of sacrifice time. Fit the regression of the log response on an indicat
#'
#'
bloodBrain.model.2 <- lm(log_brainLiver~TREAT+log_Time+log_Time^2+log_Time^3,data=bloodBrain)

stargazer(bloodBrain.model.2,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          ci = TRUE,
        # keep = c("\\bprecip\\b"),
          title = "",

         # notes = "Coefficients have been removed!",
            dep.var.caption  = "-" ,  # Bold
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{-} \\
## \cline{2-2}
## \\[-1.8ex] & log\_brainLiver \\
## \hline \\[-1.8ex]
##  TREATNS & $-$0.846$^{***}$ ($-$1.270, $-$0.422) \\
##   log\_Time & 1.098$^{***}$ (0.987, 1.209) \\
##   Constant & $-$3.009$^{***}$ ($-$3.370, $-$2.649) \\
```

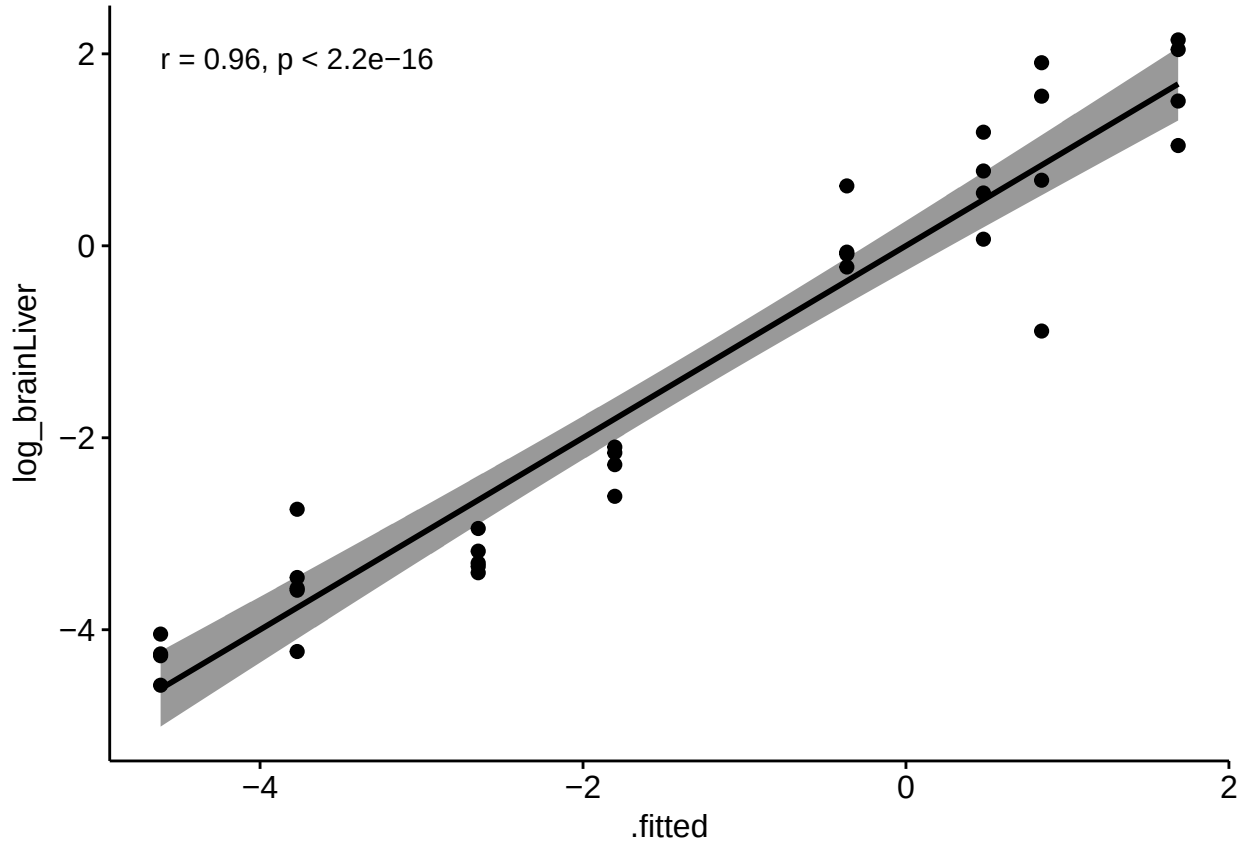
```
##  \hline \\[-1.8ex]
## Observations & 34 \\
## R$^{2}$ & 0.926 \\
## Adjusted R$^{2}$ & 0.921 \\
## Residual Std. Error & 0.631 (df = 31) \\
## F Statistic & 194.187$^{***}$ (df = 2; 31) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
bloodBrain.treatment.2 <- data.frame(
  log_Time=log(rep(c(.5,3,24,72),2)),
   TREAT=as.factor(c(rep("BD",4),rep("NS",4)))
)

bloodBrain.model.2.predicted <- cbind(bloodBrain.treatment.2,
data.frame(predict(bloodBrain.model.2,bloodBrain.treatment.2, interval = "prediction")))



stargazer(bloodBrain.model.2.predicted,
          header=F,
          type = "latex",
          no.space = T,
          summary = F,
          single.row = T
          )
```

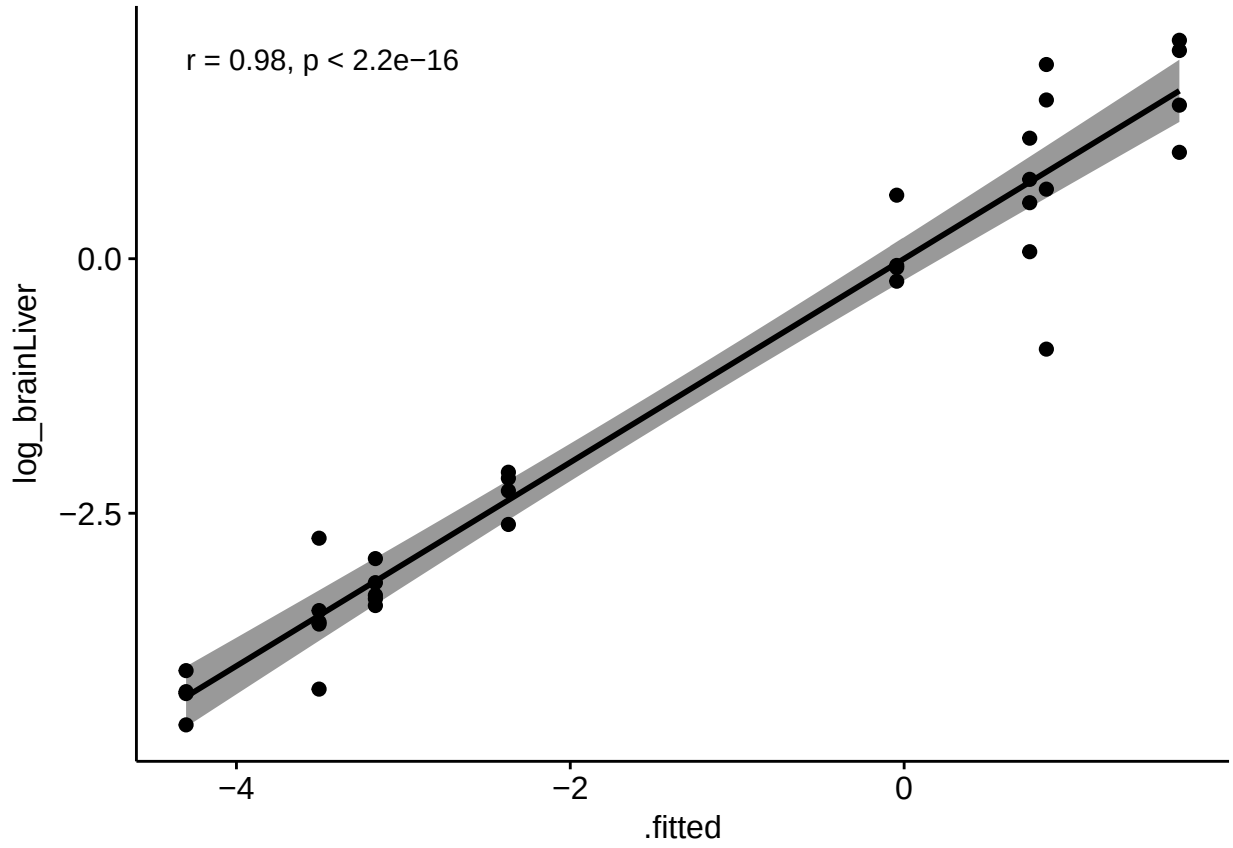
```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & log\_Time & TREAT & fit & lwr & upr \\
## \hline \\[-1.8ex]
## 1 & $$-$0.693$ & BD & $$-$3.770$ & $$-$5.125$ & $$-$2.416$ \\
## 2 & $1.099$ & BD & $$-$1.803$ & $$-$3.129$ & $$-$0.477$ \\
## 3 & $3.178$ & BD & $0.480$ & $$-$0.853$ & $1.813$ \\
## 4 & $4.277$ & BD & $1.686$ & $0.332$ & $3.040$ \\
## 5 & $$-$0.693$ & NS & $$-$4.616$ & $$-$5.973$ & $$-$3.259$ \\
## 6 & $1.099$ & NS & $$-$2.649$ & $$-$3.976$ & $$-$1.322$ \\
## 7 & $3.178$ & NS & $$-$0.366$ & $$-$1.698$ & $0.965$ \\
## 8 & $4.277$ & NS & $0.840$ & $$-$0.511$ & $2.191$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
#'
#'
#' ## g.    Why are the answers to parts (5) and (6) the same?
#'
```

```
#'
augment(bloodBrain.model.2, bloodBrain) %>%
  ggscatter(
  x = ".fitted",
  y = "log_brainLiver",
  conf.int = T,
  cor.coef = T,
  cor.method = "pearson",
  add = "reg.line"

  )
```




```
augment(bloodBrain.model, bloodBrain) %>%
  ggscatter(
  x = ".fitted",
  y = "log_brainLiver",
  conf.int = T,
  cor.coef = T,
  cor.method = "pearson",
  add = "reg.line"

  )
```




```
#'
#'
#'
#'
bloodBrain.model.anova <- anova(bloodBrain.model, bloodBrain.model.2)
stargazer(bloodBrain.model.anova,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & Res.Df & RSS & Df & Sum of Sq & F & Pr(\textgreater F) \\
## \hline \\[-1.8ex]
## 1 & $29$ & $8.233$ & $$ & $$ & $$ & $$ \\
## 2 & $31$ & $12.330$ & $$-$2$ & $$-$4.098$ & $7.217$ & $0.003$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```r
# there is no significant difference between the 4 models beccause all the p-values are extremly high
#'
#'
#' Since the origin model is already linear, further log transformation wouldn't change any of the pred
#'
#' ## h.    Fit the regression of the log response (brain tumor-to-liver antibody ratio) on all covaria
#'
#'
bloodBrain.model.3 = lm(log_brainLiver~ TREAT+ TIMEFactor+ SEX+ WEIGHT+ DAYS+ LOSS + TUMOR, data=bloodBr
bloodBrain.model.3.df=augment(bloodBrain.model.3,bloodBrain)
#model4.df
stargazer(bloodBrain.model.3,
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          ci = TRUE,
        # keep = c("\\bprecip\\b"),
          title = "",

         # notes = "Coefficients have been removed!",
            dep.var.caption  = "-" ,  # Bold
          single.row = T

          )
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{-} \\
## \cline{2-2}
## \\[-1.8ex] & log\_brainLiver \\
## \hline \\[-1.8ex]
##  TREATNS & $-$0.831$^{***}$ ($-$1.218, $-$0.444) \\
##   TIMEFactor3 & 1.089$^{***}$ (0.513, 1.666) \\
##   TIMEFactor24 & 4.114$^{***}$ (3.453, 4.775) \\
##   TIMEFactor72 & 5.137$^{***}$ (4.468, 5.805) \\
##   SEXM & $-$0.036 ($-$0.737, 0.666) \\
##   WEIGHT & 0.002 ($-$0.008, 0.011) \\
##   DAYS & 0.019 ($-$0.533, 0.572) \\
##   LOSS & $-$0.048$^{*}$ ($-$0.102, 0.006) \\
##   TUMOR & 0.001 ($-$0.001, 0.004) \\
##   Constant & $-$4.064 ($-$10.227, 2.100) \\
##  \hline \\[-1.8ex]
## Observations & 34 \\
## R$^{2}$ & 0.957 \\
## Adjusted R$^{2}$ & 0.941 \\
## Residual Std. Error & 0.547 (df = 24) \\
## F Statistic & 59.258$^{***}$ (df = 9; 24) \\
## \hline
```

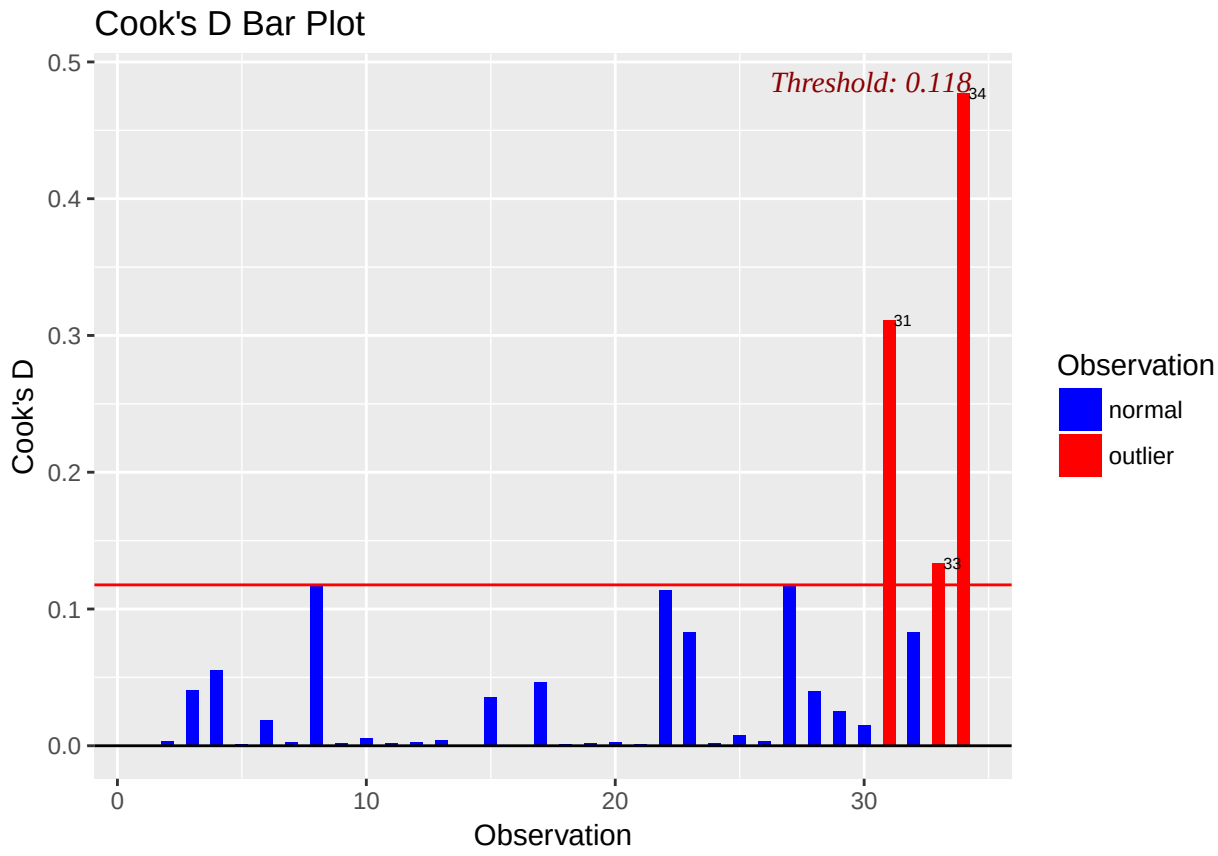
```
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```
#'
#'
#' ## i.    Obtain a set of case influence statistics, including a measure of influence, the leverage,
#'
#'
# influencial points
bloodBrain.model.3.influence <- influence.measures(bloodBrain.model.3)
bloodBrain.model.3.influence.df <- data.frame(bloodBrain.model.3.influence[[1]])

# studentized
bloodBrain.model.3.studres <- studres(bloodBrain.model.3)

bloodBrain.model.3.influence.df$studResidual<-bloodBrain.model.3.studres

 as_tibble(bloodBrain.model.3.influence.df)%>%
  mutate_if(is.numeric, funs(round(., 5))) %>%
     dplyr::select(dffit,cov.r,cook.d,hat,studResidual)%>%
        stargazer(
          header=F,
          type = "latex",
          summary = F,
          no.space = T,
          single.row = T
          )
```

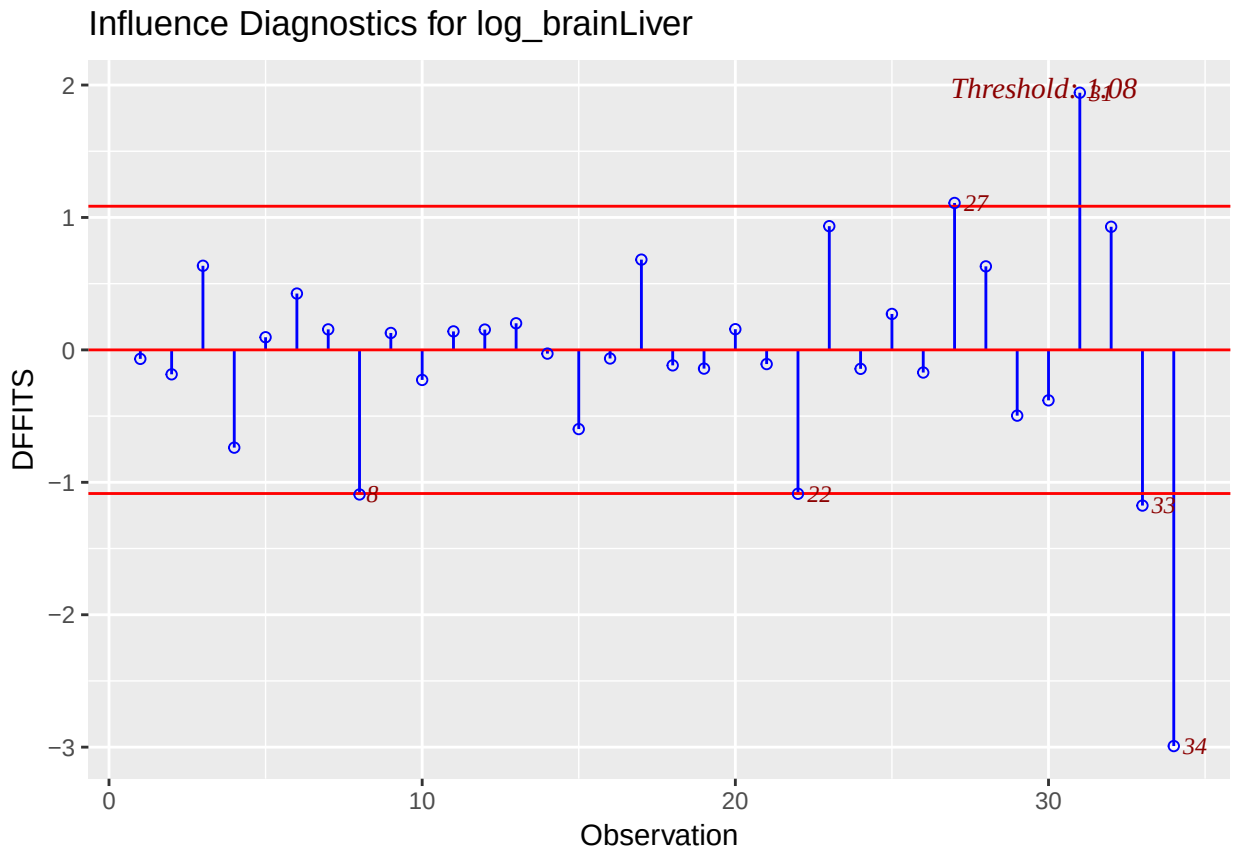
```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & dffit & cov.r & cook.d & hat & studResidual \\
## \hline \\[-1.8ex]
## 1 & -0.06773 & 1.8184 & 0.00048 & 0.16667 & -0.15144 \\
## 2 & -0.18453 & 1.73035 & 0.00353 & 0.17494 & -0.40073 \\
## 3 & 0.63507 & 1.38136 & 0.04029 & 0.28285 & 1.01123 \\
## 4 & -0.73812 & 1.7591 & 0.05484 & 0.39278 & -0.91775 \\
## 5 & 0.09536 & 1.83973 & 0.00095 & 0.18266 & 0.20172 \\
## 6 & 0.42539 & 1.54232 & 0.0184 & 0.23231 & 0.7733 \\
## 7 & 0.15475 & 2.07858 & 0.00249 & 0.28284 & 0.24642 \\
## 8 & -1.09187 & 1.35591 & 0.11627 & 0.42574 & -1.26809 \\
## 9 & 0.12766 & 1.83909 & 0.0017 & 0.19218 & 0.26173 \\
## 10 & -0.2269 & 1.83261 & 0.00533 & 0.22619 & -0.41968 \\
## 11 & 0.13953 & 1.93875 & 0.00203 & 0.23232 & 0.25363 \\
## 12 & 0.15328 & 1.81387 & 0.00244 & 0.19171 & 0.31473 \\
## 13 & 0.20082 & 2.15548 & 0.00419 & 0.31631 & 0.29524 \\
## 14 & -0.02816 & 2.03529 & 8e-05 & 0.2488 & -0.04892 \\
## 15 & -0.59814 & 1.24736 & 0.03558 & 0.24047 & -1.06302 \\
## 16 & -0.06492 & 1.96582 & 0.00044 & 0.2263 & -0.12003 \\
```

```
## 17 & 0.68159 & 1.37696 & 0.04629 & 0.29948 & 1.04244 \\
## 18 & -0.11695 & 1.98067 & 0.00142 & 0.24156 & -0.20724 \\
## 19 & -0.14133 & 2.14341 & 0.00208 & 0.30024 & -0.21576 \\
## 20 & 0.15602 & 2.22333 & 0.00253 & 0.32647 & 0.22409 \\
## 21 & -0.10701 & 2.19946 & 0.00119 & 0.31175 & -0.159 \\
## 22 & -1.0865 & 1.1128 & 0.11372 & 0.38152 & -1.38334 \\
## 23 & 0.93414 & 0.86531 & 0.08316 & 0.28562 & 1.47736 \\
## 24 & -0.14313 & 2.20051 & 0.00213 & 0.31765 & -0.20978 \\
## 25 & 0.2714 & 2.12889 & 0.00764 & 0.32683 & 0.38951 \\
## 26 & -0.17095 & 2.47526 & 0.00304 & 0.39365 & -0.21216 \\
## 27 & 1.10933 & 0.98032 & 0.1174 & 0.36318 & 1.46895 \\
## 28 & 0.63064 & 1.53313 & 0.03998 & 0.31262 & 0.93513 \\
## 29 & -0.49652 & 2.30647 & 0.02536 & 0.42531 & -0.57717 \\
## 30 & -0.38116 & 1.9246 & 0.01495 & 0.30889 & -0.57014 \\
## 31 & 1.94181 & 0.23855 & 0.3113 & 0.38319 & 2.46364 \\
## 32 & 0.92963 & 0.93731 & 0.08287 & 0.29875 & 1.42427 \\
## 33 & -1.17553 & 1.18431 & 0.13311 & 0.41916 & -1.3838 \\
## 34 & -2.9917 & 0.00261 & 0.4772 & 0.28905 & -4.69189 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

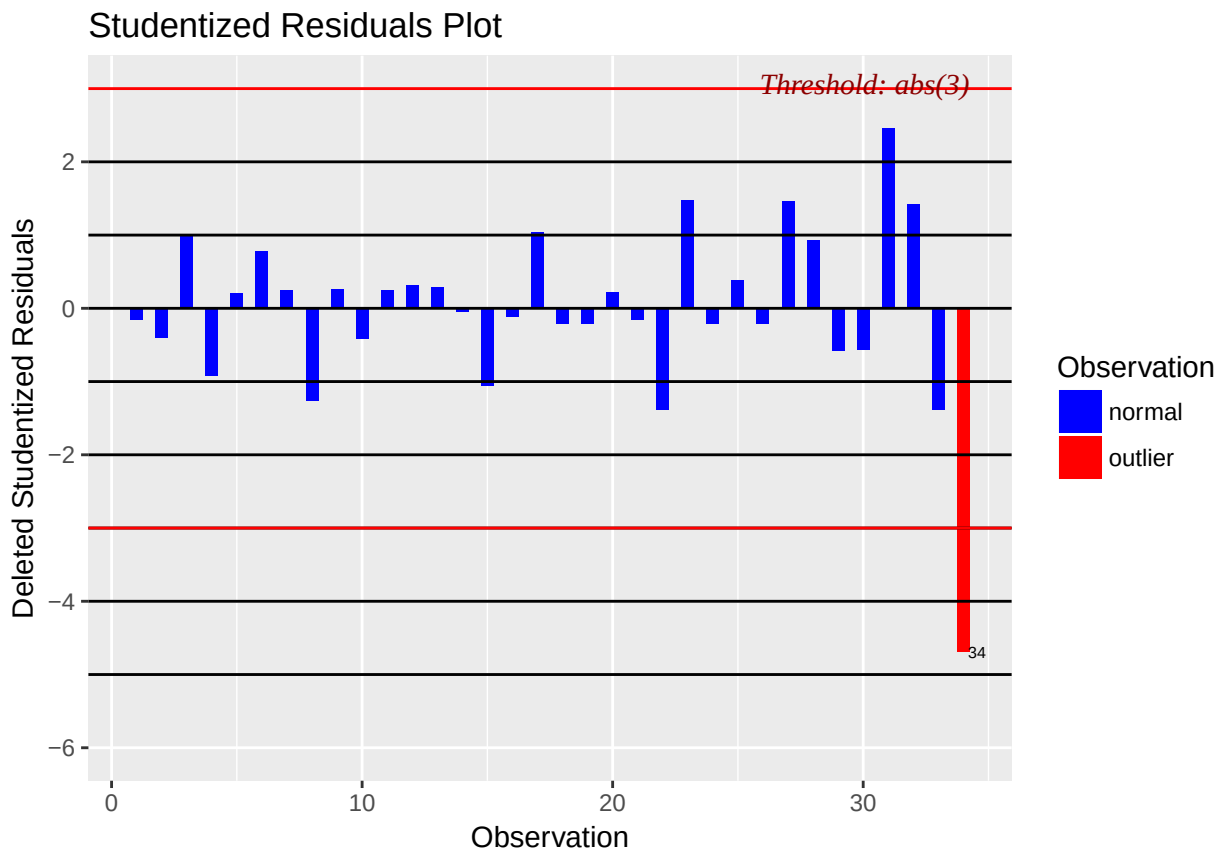
```
#'
#'
#' \twocolumn
#'
#' ### Cook's Distance
#'
#'

ols_cooksd_barplot(bloodBrain.model.3)
```



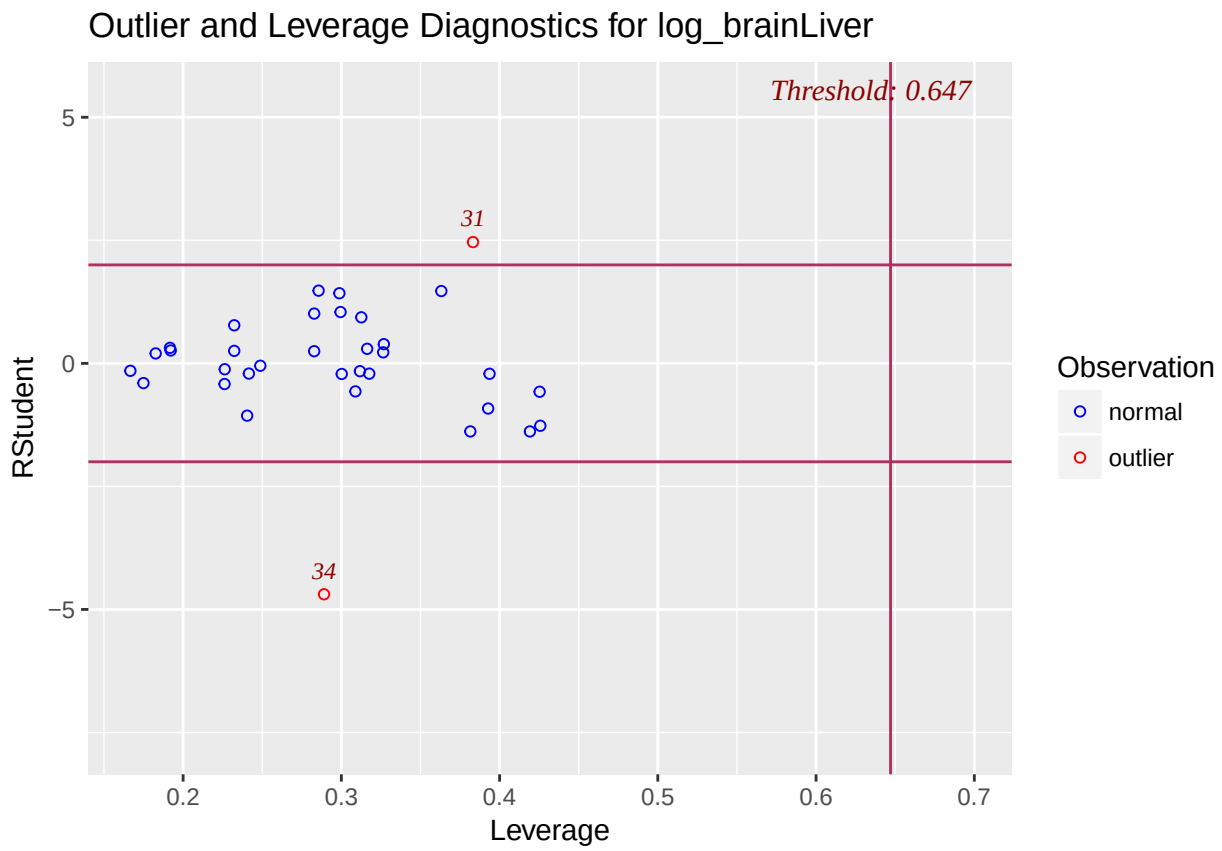

```
#'
#'
#'
#' ### DFFITS Plot  - difference in fits
#'
#'


ols_dffits_plot(bloodBrain.model.3)
```



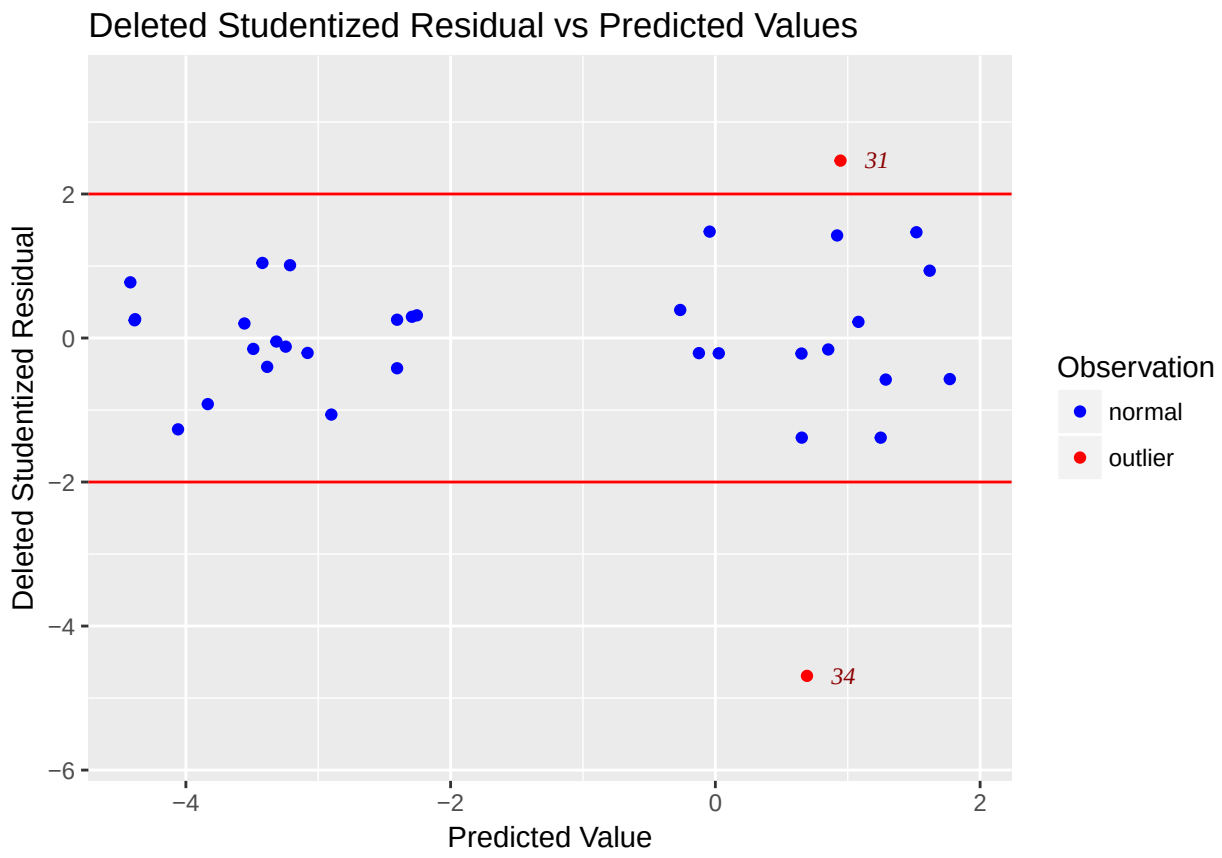

```
#'
#'
#'
#'
#' ### Studentized Residual Plot
#'
#'

ols_srsd_plot(bloodBrain.model.3)
```



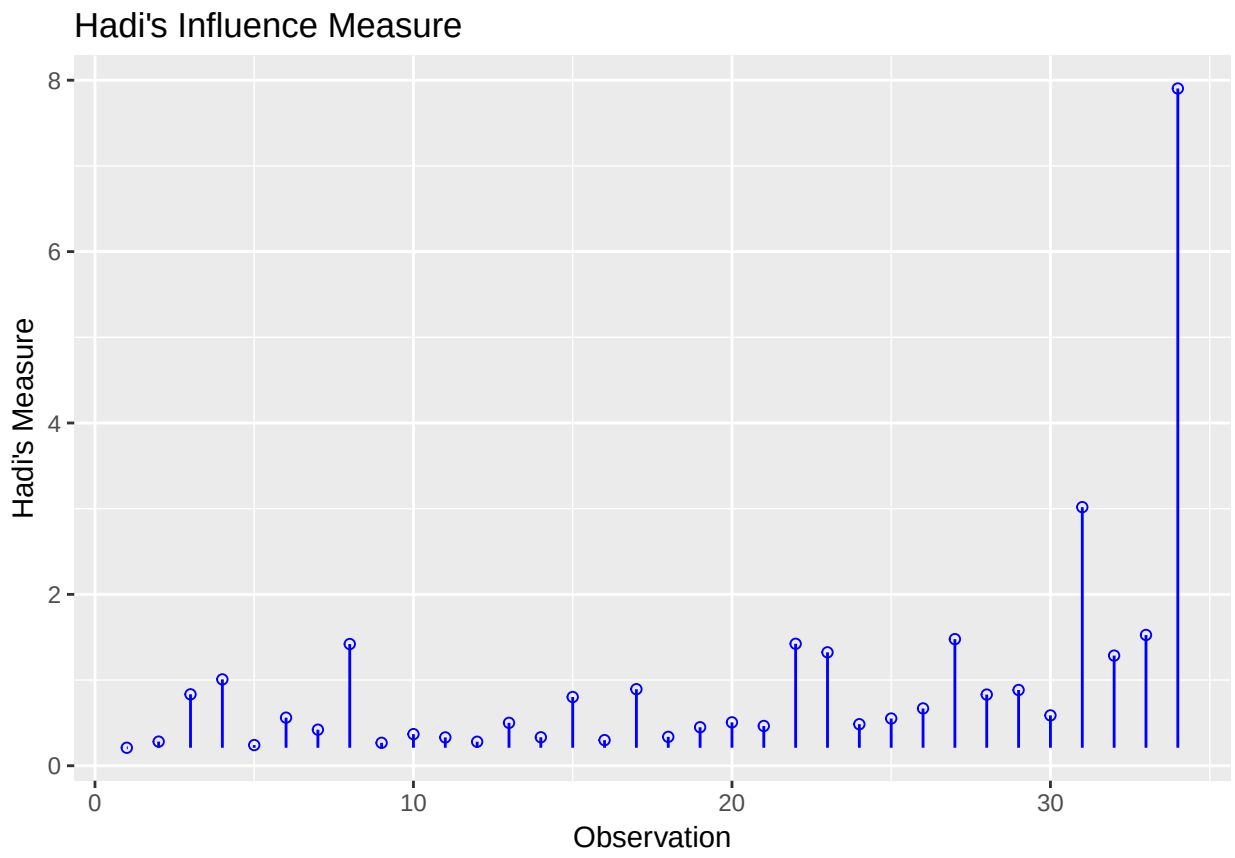

```
#'
#'
#'
#' ### Leverage Plot
#'
#'

ols_rsdlev_plot(bloodBrain.model.3)
```




```
#'
#'
#'
#' ### Deleted Stud.Residual vs Fitted Values
#'
#'

ols_dsrvsp_plot(bloodBrain.model.3)
```




```
#'
#'
#'
#'
#' ### Hadi Plot
#'
#'

ols_hadi_plot(bloodBrain.model.3)
```



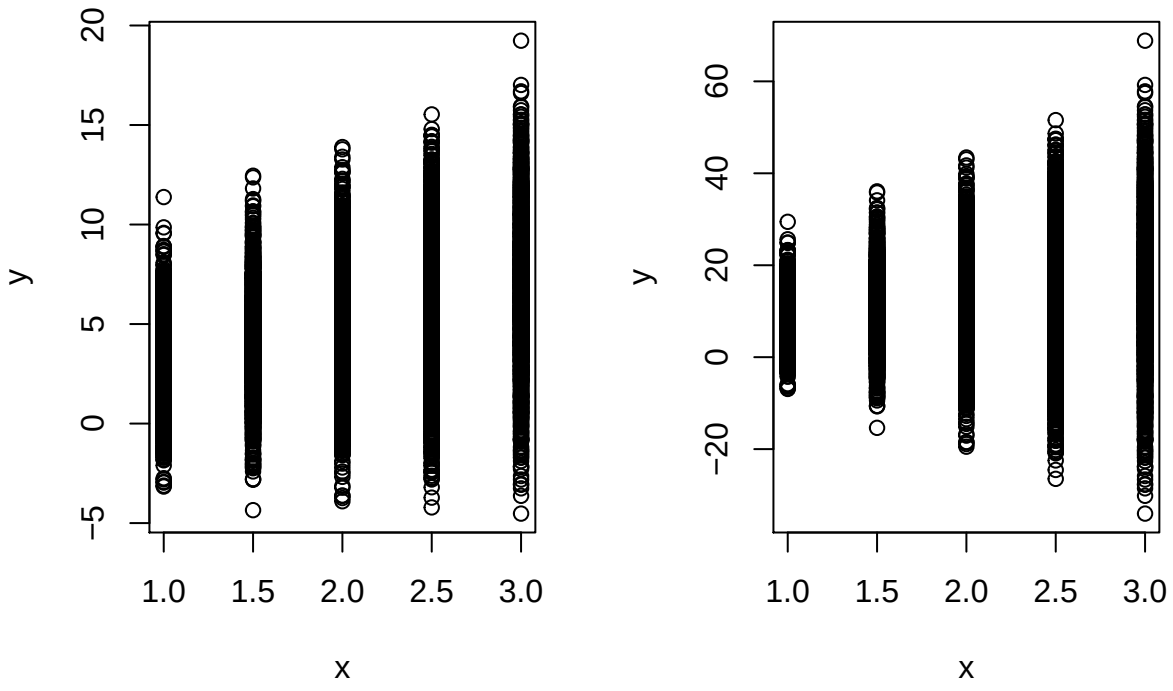

```
#'
#'
#'
#' \onecolumn
#'
#' ## * Discuss whether any influential observations or outliers occur with respect to this fit.
#'
#' - The cook's distance plot revealed that we have 3 regression outliers which strongly influences fit
#'
#' - DFFIT diagnostics plot indicated that we have 6 influential data points. Observation 31, 33 and 34
#'
#' - From the Studentized residual plot, we found out that 1 observation which had studentized residual
#'
#' - Leverage plot indicated that we did not have any leverage observations since we did not have any o
#'
#' - Deleted Studentized Residual vs Fitted Value plot indicated that we had 2 observations which pulle
#'
#'
#' In summary, two observations had the highest influence on our regression model implying that if incl
#'
#' \onecolumn
#'
#' # Question 4
#'
#' ![](pic/1.png)
#'
```

```
#' ![](pic/2.png)
#'
#'
#' ![](pic/3.png)
#'
#' \onecolumn
#'
#' # Question 5
#'
#' ## LMQ Function
#'
#'

library(stats4)
library(dplyr)

lmq<-function(x,y){

  set.seed(120) #reproducability

  beta0=1 # init starting point
  beta1=1 # init starting point
  sigma=1 # init starting point
  q=1 # init starting point

  method = "L-BFGS-B" # or BFGS

  # par(mfrow = c(1,1))
  # plot(x,y)

  logLikelihood <- function(beta0,beta1,sigma,q){
    beta= t(c(beta0,beta1)) # transpose and convert to matrix
    resid = y - beta0 - beta1*x
    R = dnorm(resid, mean = 0, sd=sqrt(x^(q)*sigma^2), log=TRUE)
    -sum(R)
  }

  mle(logLikelihood, list(beta0=beta0, beta1=beta1, sigma=sigma, q=q), method =method)
}
#'
#'
#'
#' ## Test Run 1: Simulate data with beta0=1.4, beta1=1.8 , sigma=2, and q=1
#'
#'

set.seed(120) #reproducability
# test run1
x = rep(seq(1,3,by=0.5),2000)
y = 1.4 + 1.8*x + rnorm(length(x),0,2*x^(1/2))
plot(x,y)
lmq(x,y)
```

```
##
```

```
## Call:
## mle(minuslogl = logLikelihood, start = list(beta0 = beta0, beta1 = beta1,
##     sigma = sigma, q = q), method = method)
##
## Coefficients:
##     beta0     beta1     sigma         q
## 1.2661222 1.8813507 1.9973277 0.9892181
```

```
#'
#'
#' ## Test Run 1: Simulate data with beta0=6, beta1=3 , sigma=5, and q=2
#'
#'

set.seed(120) #reproducability

# test run2
x = rep(seq(1,3,by=0.5),2000)
y = 6 + 3*x + rnorm(length(x),0,5*x^(2/2))
plot(x,y)
```




```
lmq(x,y)
```

```
##
## Call:
## mle(minuslogl = logLikelihood, start = list(beta0 = beta0, beta1 = beta1,
##     sigma = sigma, q = q), method = method)
```

```
##
## Coefficients:
##    beta0    beta1    sigma        q
## 5.553277 3.292531 4.992762 1.989496
```

```
#'
#'
#' \onecolumn
#'
#' # Question 6
#'
#'
#' ![](pic/4.png)
#'
#'
#'
#' ![](pic/5.png)
#'
#' ![](pic/6.png)
#'
#' \onecolumn
#'
#' # Source Code
#'
#'


#'
```