6020Midterm

Johan Boer

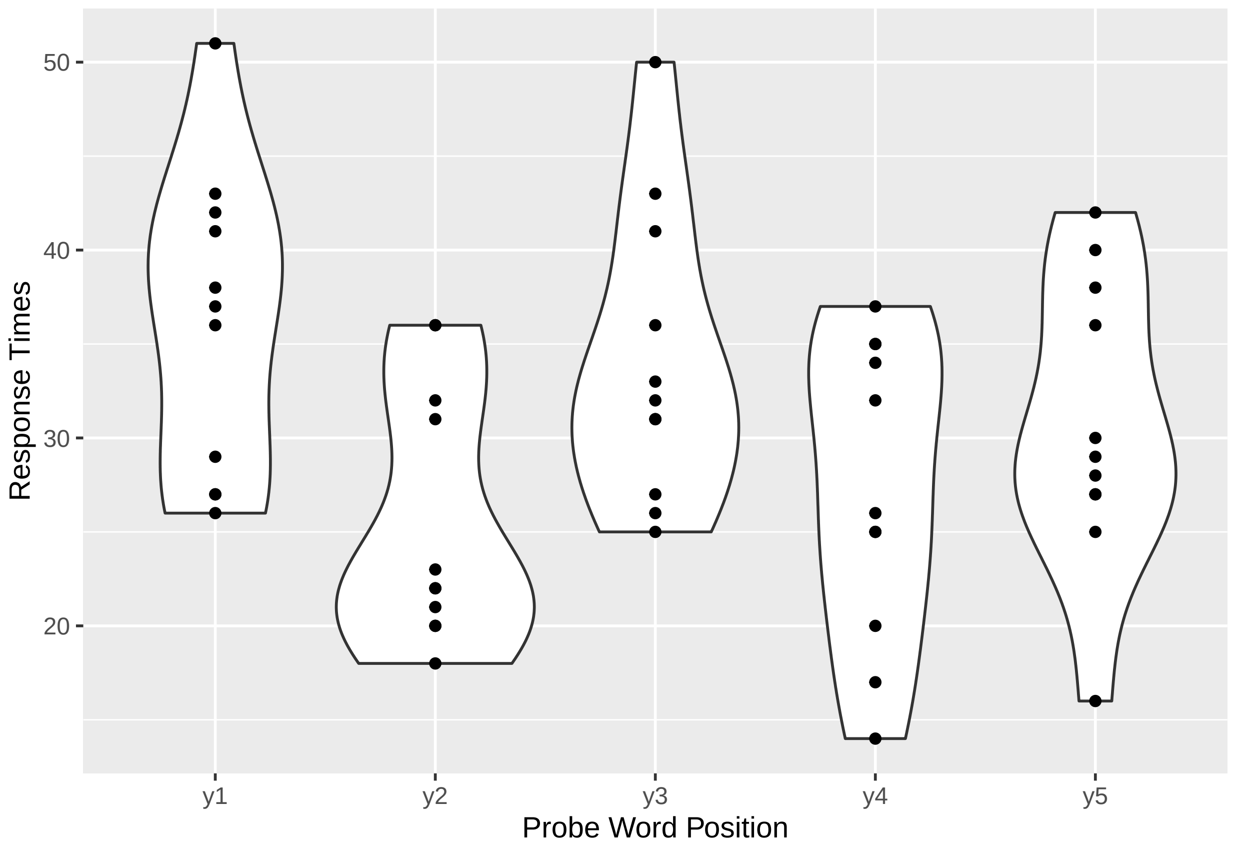
4/1/2022

-1- (30 points) The following table is also from Methods of Multivariate Analysis by Rencher. This is data from an experiment that measures response times to certain "probe words" at five positions in a sentence. The variables y_j are response times for the jth probe word, where $j = 1, ..., 5$. Analyze this dataset thoroughly in any way that you feel fit, present and interpret all results and justify everything you do.

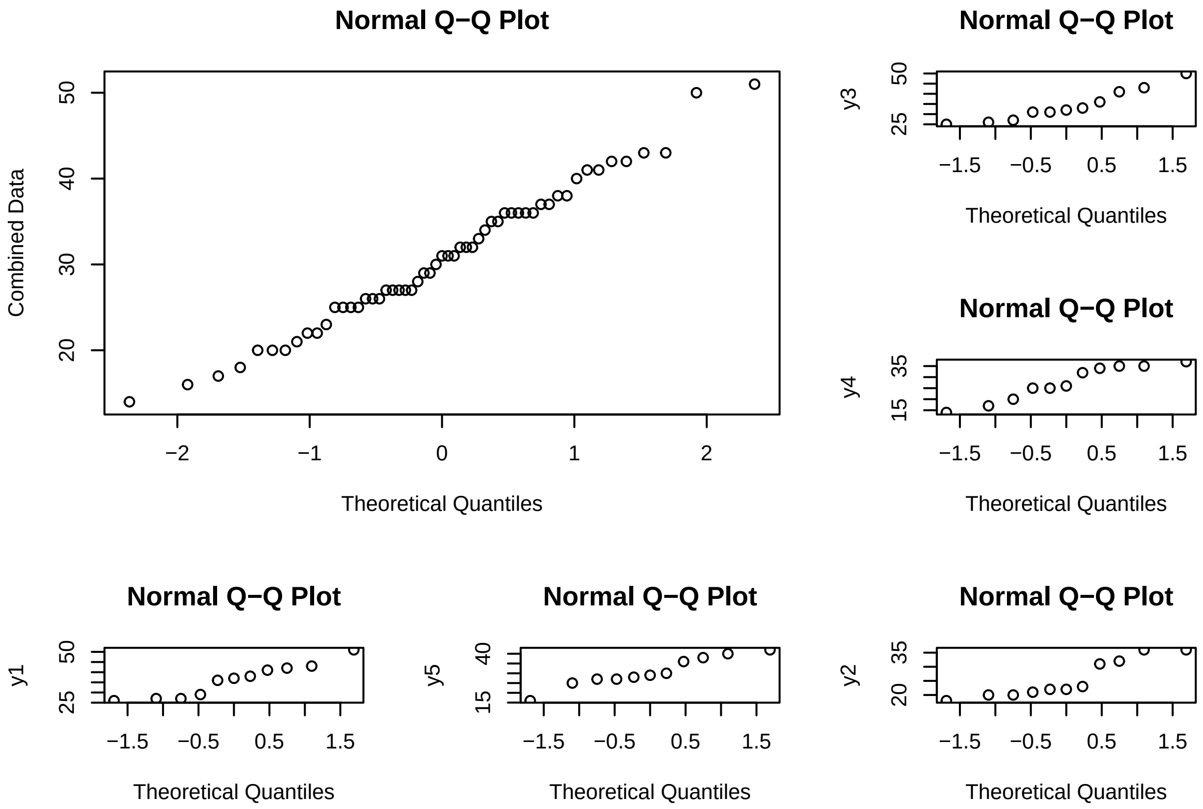
Table 1: Table 3.5 Response Times for Probe Word Positions

Subject_Number	y1	y2	y3	y4	y5
1	51	36	50	35	42
2	27	20	26	17	27
3	37	22	41	37	30
4	42	36	32	34	27
5	27	18	33	14	29
6	43	32	43	35	40
7	41	22	36	25	38
8	38	21	31	20	16
9	36	23	27	25	28
10	26	31	31	32	36
11	29	20	25	26	25

For this dataset I would want to find out if each probe word is substantively different and if so, which probe word location results in the fastest and slowest responses. So the first thing to look into is if the data associated with each probe word is significantly different than the others. One way to get some quick insights into the data would be to look at a plot of the distribution of each probe word location.



From this plot I see that position y_1 and y_3 seem to have a similar distribution. The other 3 positions have some overlap but they seem to be distinct. Another thing that would be good to check is the normality of the overall data set-that is can we model the times irrespective of word position as a normal distribution. A QQ plot can give some insight into the normality of a distribution. We can also look at the QQ plots of each variable individual



These QQ plots indicate that the data is close to normally distributed, and therefore a paired T-Test would be an appropriate way to check if these models have different means. Since there is so little data (only 11 points in each variable), I think it is wise to check to see if the largest pairwise difference in means is significant to start with.

```
##        mu1      mu2      mu3      mu4      mu5
## 1 36.09091 25.54545 34.09091 27.27273 30.72727
```

The biggest difference in means is between y_1 and y_2 so to start we use a T-Test

```
##
##  Paired t-test
##
## data:  p1data$y1 and p1data$y2
## t = 5.2804, df = 10, p-value = 0.0003575
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.095627 14.995283
## sample estimates:
## mean of the differences
##                10.54545
```

Since this test showed a difference with high significance we can move on to other tests. For example we can test if y_1 and y_3 are significantly different.

```
##
##  Paired t-test
##
## data:  p1data$y1 and p1data$y3
## t = 1.1991, df = 10, p-value = 0.2581
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.716265  5.716265
## sample estimates:
## mean of the differences
##                       2
```

As expected from the plots we get that y_1 and y_3 are not significantly different. The next thing to test is if all of the variables are distinct from the overall distribution of the times. I will test all 5 variables in a T-Test with a .01 confidence level rather than a .05 confidence level (to account for multiple testing) against the combined dataset of the other 4 variables.

```
##
##  Welch Two Sample t-test
##
## data:  p1data$y1 and otherTimes$times
## t = 2.4609, df = 15.307, p-value = 0.02619
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -1.296369 14.660005
## sample estimates:
## mean of x mean of y
##  36.09091  29.40909
```

```
##
##  Welch Two Sample t-test
##
## data:  p1data$y2 and otherTimes$times
## t = -2.7095, df = 18.284, p-value = 0.01422
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -13.3927859   0.3927859
## sample estimates:
## mean of x mean of y
##  25.54545  32.04545
```

```
##
##  Welch Two Sample t-test
##
## data:  p1data$y3 and otherTimes$times
## t = 1.567, df = 16.34, p-value = 0.1363
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -3.591389 11.955026
## sample estimates:
## mean of x mean of y
##  34.09091  29.90909
```

```
##
##  Welch Two Sample t-test
##
## data:  p1data$y4 and otherTimes$times
## t = -1.6069, df = 16.042, p-value = 0.1276
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -12.228446   3.546628
## sample estimates:
## mean of x mean of y
##  27.27273  31.61364
```

```
##
##  Welch Two Sample t-test
##
## data:  p1data$y5 and otherTimes$times
## t = -0.0085954, df = 17.042, p-value = 0.9932
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -7.683702  7.638248
## sample estimates:
## mean of x mean of y
##  30.72727  30.75000
```

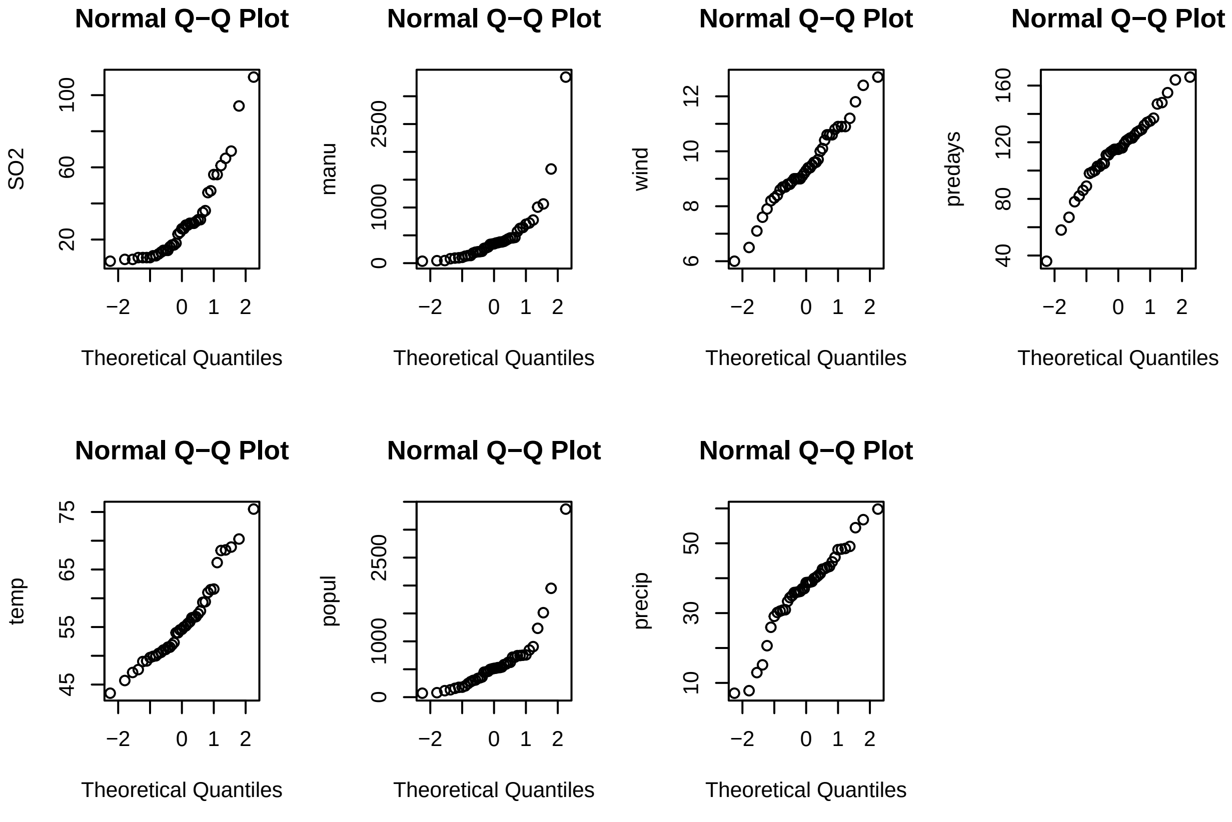
For this test none of the word positions were significantly different than the rest of the data. We know from other tests that there is a difference between the extremes, but this shows that between arbitrary probe word positions there is not a huge difference in response times. This dataset is too small to get any truly meaningful insights but it does suggest that there may be some benefit to placing a probe word in positions y_1 or y_3 rather than y_2 with respect to response time.

-2- (30 points) In R use the commands install.packages("HSAUR3") followed by data("USairpollution", package="HSAUR3") to download air pollution information for 41 cities in the United States. You can find documentation on this data set at https: //www.rdocumentation.org/packages/HSAUR3/versions/1.0-8/topics/USairpollution Plot normal qq-plots [qqnorm in R] on each of the variables S02, temp, manu, popul, wind, precip and predays and present them in a grid. What do you find? Use S02 and temp either together or individually as response variables and the rest as predictor variables and perform multivariate linear regression. Analyze the dataset completely including any diagnostics that you feel appropriate.

Table 2: First Rows of US Air Pollution data

	SO2	temp	manu	popul	wind	precip	predays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55.0	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148

Looking at the QQ plots:



These plots show that most of these variables are close to normally distributed. Temperature, wind speed, amount of precipitation, and days with precipitation are all reasonably close to normally distributed considering this is a real dataset with a relatively low number of observations. Population and number of manufacturing enterprises seem to be mostly normal except for a couple serious outliers and SO2 does not seem to be fit very well by a normal distribution.

Next we should check for highly correlated data. The correlation matrix shows:

Table 3: Correlation Matrix

	SO2	temp	manu	popul	wind	precip	predays
SO2	1.0000000	-0.4336002	0.6447687	0.4937796	0.0946905	0.0542943	0.3695636
temp	-0.4336002	1.0000000	-0.1900422	-0.0626781	-0.3497396	0.3862534	-0.4302421
manu	0.6447687	-0.1900422	1.0000000	0.9552693	0.2379468	-0.0324169	0.1318293
popul	0.4937796	-0.0626781	0.9552693	1.0000000	0.2126438	-0.0261187	0.0420832
wind	0.0946905	-0.3497396	0.2379468	0.2126438	1.0000000	-0.0129944	0.1641056
precip	0.0542943	0.3862534	-0.0324169	-0.0261187	-0.0129944	1.0000000	0.4960967
predays	0.3695636	-0.4302421	0.1318293	0.0420832	0.1641056	0.4960967	1.0000000

Population and number of manufacturing enterprises are highly correlated (corr(popul,manu) $\approx .955$). This suggests that we should only use one of these two for our regression. No other pair of variables has an absolute correlation above 0.5, so we can not immediately rule out any other variables for colinearity.

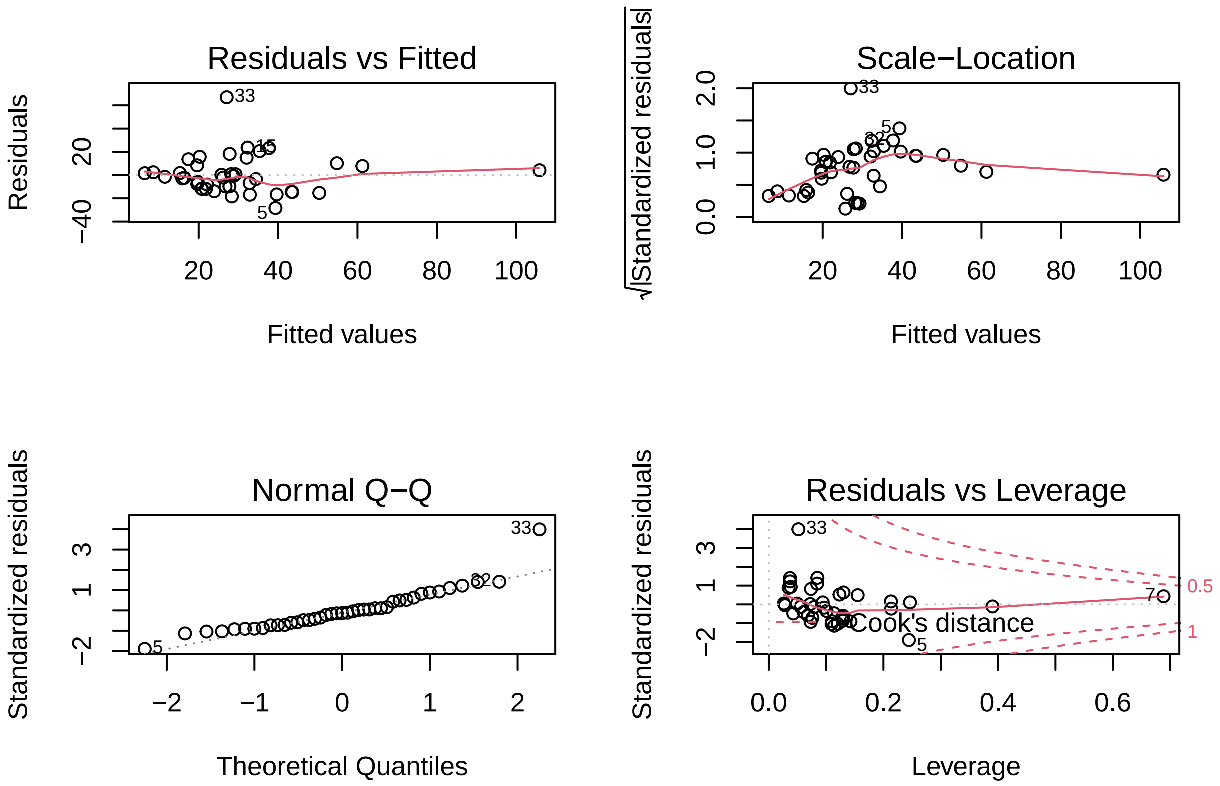
I will do two separate linear regressions using SO2 and temperature as response variables with the other variables as predictor variables. For the regression with SO2 as the response variable I will include number of manufacturing enterprises, wind speed, amount of precipitation and, number of days with precipitation as predictor variables. Population is not used in this regression because it has a high correlation with manufacturing and has a worse correlation with SO2 concentration. For the regression for Temperature, the same predictor variables are used execpt that population is used over manufacturing because neither has a very good correlation with temperature and population has less colinearity with the other predictor variables.

Looking closely at the SO2 regression

```
##
## Call:
## lm(formula = p2data$SO2 ~ p2data$manu + p2data$wind + p2data$precip +
##     p2data$predays)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.348 -11.726  -2.159   7.759  66.938
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.186097  20.971790   0.343   0.7339
## p2data$manu     0.025872   0.005015   5.159 9.25e-06 ***
## p2data$wind    -1.860634   1.987045  -0.936   0.3553
## p2data$precip  -0.208905   0.268933  -0.777   0.4424
## p2data$predays  0.317237   0.121581   2.609   0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.21 on 36 degrees of freedom
## Multiple R-squared:  0.5164, Adjusted R-squared:  0.4627
## F-statistic: 9.612 on 4 and 36 DF,  p-value: 2.151e-05
```

Firstly the F-statistic has a very low p-value so this regression models the data well. The only predictor variable that has a significant t value is manufacturing, which suggests that the most significant predictor for SO2 concentration is the amount of manufacturing in a city.

Looking at some diagnostic plots:



We see that this regression has a linear relationship with normally distributed errors from the residuals vs fitted values and QQ plots. While there are a couple outliers none of them are high leverage, so their inclusion does not greatly skew the model. The scale location plot indicates that there may be some heteroskedasticity but only slightly.

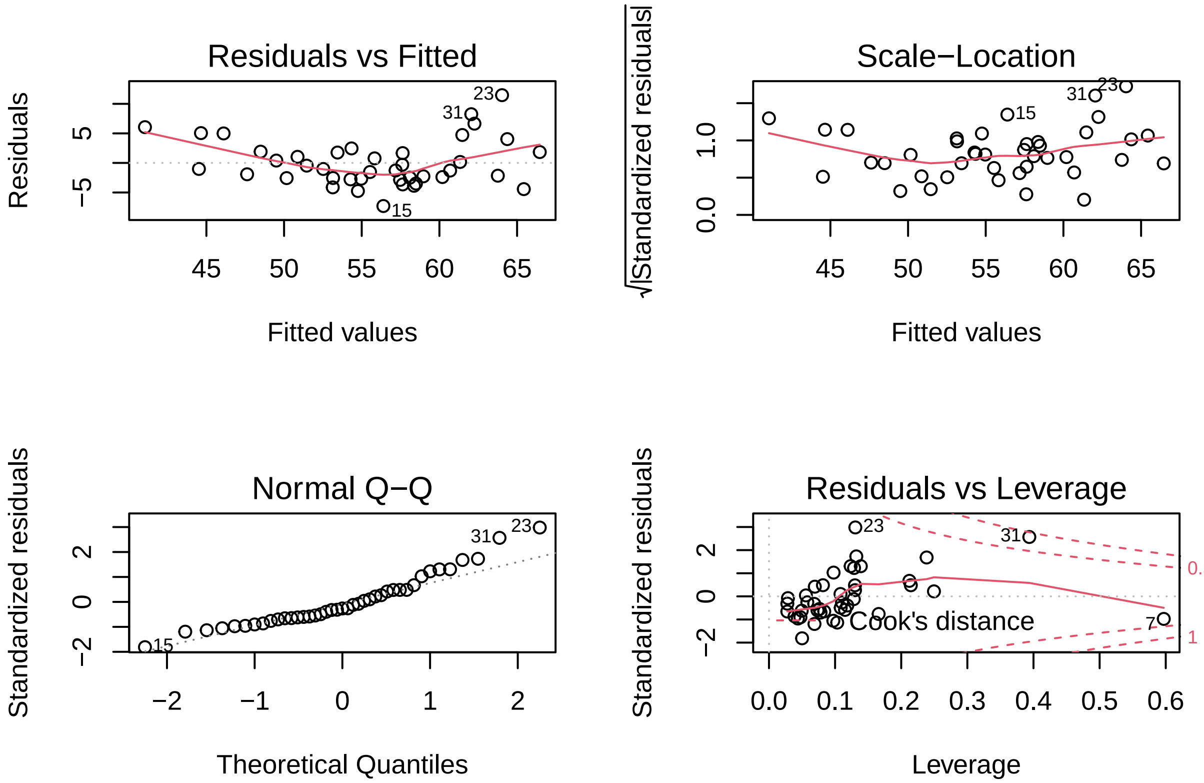
Looking at the Temperature regression:

```
##
## Call:
## lm(formula = p2data$temp ~ p2data$popul + p2data$wind + p2data$precip +
##     p2data$predays)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -7.293 -2.566 -1.022  1.839 11.467
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    72.7602537  4.9967063  14.562  < 2e-16 ***
## p2data$popul    0.0004611  0.0011534   0.400    0.692
## p2data$wind    -1.1127241  0.4757523  -2.339    0.025 *
## p2data$precip   0.4729201  0.0642470   7.361 1.11e-08 ***
## p2data$predays -0.2120920  0.0289145  -7.335 1.20e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.126 on 36 degrees of freedom
## Multiple R-squared:  0.7068, Adjusted R-squared:  0.6742
## F-statistic: 21.69 on 4 and 36 DF,  p-value: 3.527e-09
```

This regression has a much smaller p-value for the F-statistic and aside from population all of the predictor variables have small p-values associated with their t-statistic. This model has a significantly better R-Squared value than the previous regression so it has more explanatory power.

Looking at the same diagnostics:



We still have a reasonable QQ plot and the residuals vs fitted plot still shows a linear relationship. This model has a much flatter scale location plot which indicates that this model does not have much heteroskedasticity, but there is also a high leverage outlier in this model.

These models seem to suggest that unlike for SO2 concentrations, which can be predicted reasonably without much consideration given to natural conditions such as wind speed or precipitation, average temperature is much more affected by the natural conditions of the area. Therefore, if you want to predict the temperature it is important not to neglect the conditions of the area in your model to get good predictions.

-4- (30 points) Lab mice were trained for 1, 2, or 3 days to jump down from a pedestal and catch a piece of cheese. The data from the experiment is in mice.csv. Analyze the dataset thoroughly including any diagnostics that you feel appropriate.

Table 4: First Rows of Mice data

Jumped	Weight_Ounce	Length_Inch	TrainingDays
0	1.0	7.1	1
0	1.1	8.5	1
0	1.1	7.6	1
1	0.7	7.1	2
1	1.0	8.0	1
1	0.9	8.3	3

For this dataset the question is how does training mice affect the likelihood that a mouse jumps in the experiment. So a natural starting point would be to check if training the mice has any noticeable change in the proportion of mice that jump.

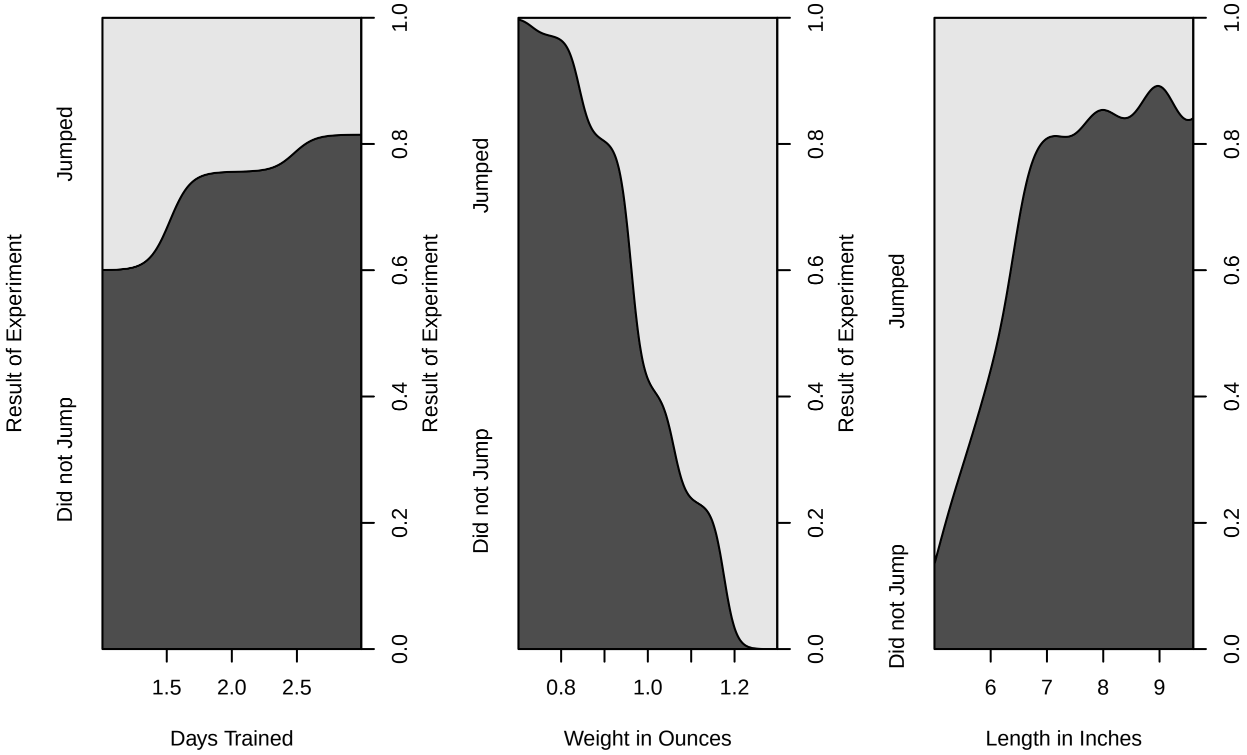
Table 5: Mice Jumping by Days Trained

TrainingDays	Count	NumberJumped	JumpProp
Total	150	108	0.72
1	55	33	0.6
2	41	31	0.75609756097561
3	54	44	0.814814814814815

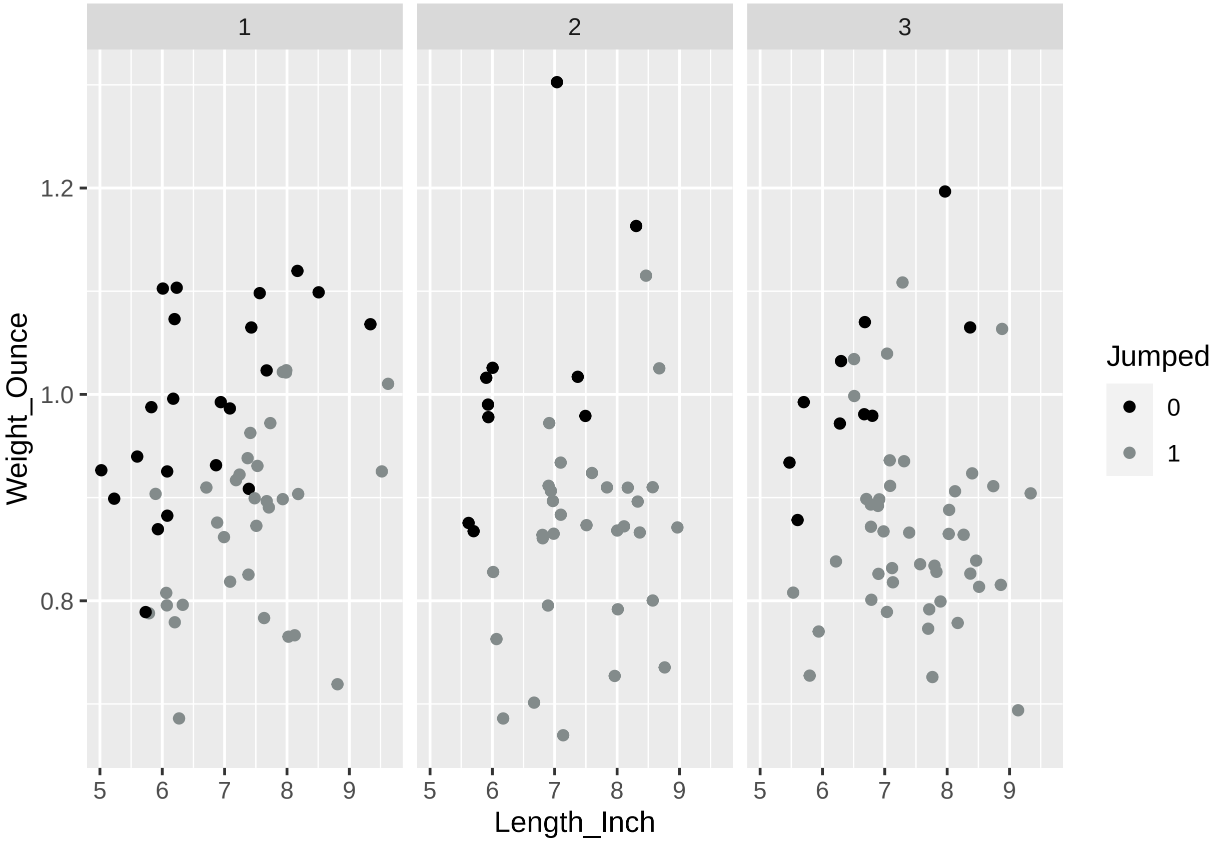
So as an initial check the number of days trained does seem to impact a mouse's probability of jumping. To be more sure that this is a significant effect I am going to run a simple logistic regression.

```
##
## Call:
## glm(formula = Jumped ~ TrainingDaysNumeric, family = "binomial",
##     data = p4data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8658  -1.3715   0.6211   0.7927   0.9950
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.1055     0.4475  -0.236   0.8136
## TrainingDaysNumeric   0.5510     0.2239   2.461   0.0139 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 177.89  on 149  degrees of freedom
## Residual deviance: 171.53  on 148  degrees of freedom
## AIC: 175.53
##
## Number of Fisher Scoring iterations: 4
```

Based on the z-value of this regression there is a significant difference as a result of training days. So, training a mouse for extra days does have a significant affect. This model has a relativity small coefficient and the residuals are not very close to zero. Since this model is not particularly good, the next natural thing to check is if it is the best predictor. One way to get some intuition for this would be to look at the conditional density plots for each predictor variable in the data.

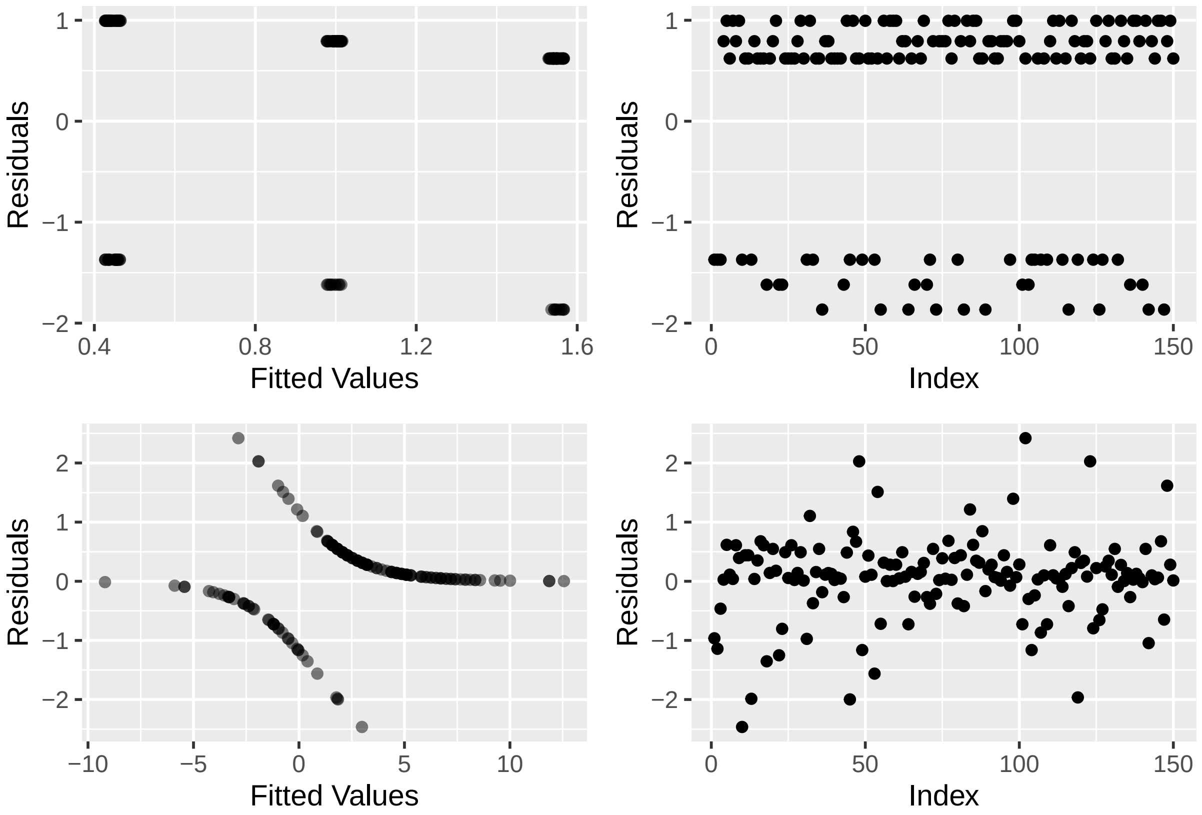


This set of plots seems to indicate that both the weight and length of a mouse may be better predictors of the mouse's probability to jump than the number of days trained. The pair of plots showing weight and length also suggest that the heavier and shorter a mouse is the less likely it is to see that mouse jump. It is worthwhile to see how the length and weight of mice are distributed, and see the effects of a logistic regression with all the variables included.



```
##
## Call:
## glm(formula = Jumped ~ TrainingDaysNumeric + Weight_Ounce + Length_Inch,
##     family = "binomial", data = p4data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.27911  -0.07723   0.03872   0.27739   2.05161
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          10.5947     3.7596   2.818  0.00483 **
## TrainingDaysNumeric   1.2859     0.4654   2.763  0.00573 **
## Weight_Ounce        -33.9015     6.5746  -5.156 2.52e-07 ***
## Length_Inch           2.8582     0.5977   4.782 1.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 177.886  on 149  degrees of freedom
## Residual deviance:  62.729  on 146  degrees of freedom
## AIC: 70.729
##
## Number of Fisher Scoring iterations: 7
```

Comparing this regression to the simple training days only regression we see that this model has deviance residuals that are almost zero and the AIC score is much better. We can also see that while having more days of training is still a significant predictor it has both the largest p-value and the smallest coefficient. This model shows that the best predictor of a mouse jumping in the experiment is the weight of a mouse. The residual deviance plots of a logistic model only considering the number of days trained and a logistic model with only the length and weight confirm that just using the number of training days is not a good way of modeling whether or not a mouse will jump.



These results seem to indicate that while training a mouse can increase the chances it will jump in an experiment the amount of training is less significant than the length and weight of a mouse. If we need mice to jump in an experiment it is a better idea to get long and light mice rather than spend more time training. If for some reason this is not possible, then in order to get more mice to jump-especially smaller and heavier mice-more training is required. For mice that are less inclined to jump, three days of training is insufficient to see much progress.