West Creek Floodplain Jam Analysis

Update from Juli

2019-12-03

Correlation analysis for West Creek floodplain jams

In this code, I'm checking the data we collected at West Creek, calculating a few additional reach characteristics, and finding simple correlations between jams and measured valley characteristics. Beyond the simple correlations, multiple linear regressions were performed to determine more complex correlations between valley geometry, forest characteristics, and jams.

1. Upload jam Characteristics

Both jam characteristics and reach characteristics (below) were copied into a .csv from the Google Sheets data file in the West Creek Google Drive.

```
#upload jam data from .csv
jam_data <- read.csv('E:/WestCreek/data/WC_jam_orig.csv', header = TRUE) %>%
  tibble::add_row(reach = 32) %>%
  mutate_at('woodvolume_m3', ~replace(., is.na(.), 0))
```

2. Upload reach characteristics and join the two datasets by reach number

This creates one master file that we can use to calculate future reach characteristics.

```
#upload reach data from .csv
reach_data <- read.csv('E:/WestCreek/data/WC_reach_orig.csv',
                       header = TRUE)

#join reach and jam data by reach ID
master <- inner_join(jam_data, reach_data, by = 'reach')
```

3. Calculate more reach characteristics

Here, I'm going to calculate the total number of jams per reach, the number of organic jams per reach, the number of mixed/LW jams per reach, and the average jam size.

```
# count number of jams of each type per reach
count_jam_types <- master %>%
  group_by(reach, jam_type)%>%
  summarise(numberofjams = n()) %>%
  ungroup() %>%
  spread(jam_type, numberofjams) %>%
  mutate_if(is.integer, ~replace(., is.na(.), 0)) %>%
  mutate(numberofjam_LWmixed = LW + mixed) %>%
  mutate(numberofjam_organic = organic) %>%
  select(reach, numberofjam_LWmixed, numberofjam_organic)

#count total number of jams per reach
total_jams <- master %>%
  group_by(reach) %>%
  summarise(jam_total = n())
```

```
#edit reach with zero jams
total_jams$jam_total[total_jams$reach == 32] <- 0

#create reach df with all characteristics and add jam frequency
full_reach_ch <- master %>%
  mutate_at('woodvolume_m3', ~replace(., is.na(.), 0)) %>%
  group_by(reach) %>%
  summarise(jam_size_avg_m3 = mean(woodvolume_m3)) %>%
  inner_join(., count_jam_types, by = 'reach') %>%
  inner_join(., total_jams, by = 'reach') %>%
  inner_join(., reach_data, by = 'reach') %>%
  mutate(jam_per_m = jam_total/reach_length_m)
```

4. Normal distributions

In order to look at simple correlations, understanding whether the data are normal is important for determining correlation method. Here, I investigate normality using shapiro tests. Data are normal if $p > 0.05$.

Normality of variables in jam dataset

```
norm_test_jam <- jam_data %>%
  select_if(is.numeric) %>%
  map(shapiro.test)%>%
  map_df(pluck, 2) %>%
  gather(., variable, shapiro_p_value, reach:tot_blockage,
         factor_key = TRUE)
pander(norm_test_jam, caption ='Shapiro-Wilk Test, Jam Data')
```

Table 1: Shapiro-Wilk Test, Jam Data

variable	shapiro_p_value
reach	1.136e-10
GPS_ID	1.534e-08
latitude	5.074e-33
longitude	5.109e-33
length_m	3.575e-19
width_m	3.431e-21
height_m	6.996e-14
pct_wood_1	0.008605
pct_wood_2	0.01909
woodvolume_m3	1.448e-29
ht_above_water_surface_m	3.298e-14
ht_above_bf_m	2.035e-14
dist_from_channel_m	9.22e-12
bf_to_water_surface_height_m	2.368e-16
bf_width_m	6.288e-09
slope_HD_m	6.783e-12
slope_VD_m	2.774e-16
rangefinder_slope_.	4.766e-13
calculated_slope_.	1.15e-12
obstruction_index	5.202e-14
pin_DBH_cm_1	0.0001359

variable	shapiro_p_value
pin_DBH_cm_2	0.006229
pin_DBH_cm_3	0.03904
pins	7.242e-16
tot_blockage	2.789e-15

No variables in reach data are normal. For our analyses, we would like to have normally distributed response variables. To attempt fitting our data to a normal distribution, we can transform the response variable of interest in the jam dataset - wood volume of each jam.

Transformation of jam response variables

The response variable of interest at the jam level is wood volume. Here, I'm testing if a square root, log10, or natural log transformation will make it normal. First, I'm removing all jams where woodvolume = NA or 0, because these are not actually jams or they're not correctly measured.

```
norm_test <- jam_data %>%
  filter(!is.na(woodvolume_m3)) %>%
  filter(!(woodvolume_m3 == 0))

shapiro.test(sqrt(norm_test$woodvolume_m3)) #not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(norm_test$woodvolume_m3)
## W = 0.63402, p-value < 2.2e-16
```

```
shapiro.test(log10(norm_test$woodvolume_m3)) #not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log10(norm_test$woodvolume_m3)
## W = 0.98016, p-value = 0.001927
```

```
shapiro.test(log(norm_test$woodvolume_m3)) #not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(norm_test$woodvolume_m3)
## W = 0.98016, p-value = 0.001927
```

No transformation makes wood volume normal at the jam level.

Normality of variables in reach dataset

Normally distributed variables: reach, total number of jams, average slope, average valley width, floodplain area, basal area, jam frequency.

```
norm_test_reach <- full_reach_ch %>%
  select_if(is.numeric) %>%
  map(shapiro.test)%>% map_df(pluck, 2) %>%
  gather(., variable, shapiro_p_value, reach:jam_per_m,
         factor_key = TRUE)
pander(norm_test_reach, caption = 'Shapiro-Wilk test, reach data')
```

Table 2: Shapiro-Wilk test, reach data

variable	shapiro_p_value
reach	0.3277
jam_size_avg_m3	2.538e-05
numberofjam_LWmixed	0.02434
numberofjam_organic	0.007688
jam_total	0.03974
reach_length_m	8.627e-06
bankfull_width_ave_m	4.383e-06
horizontal_dist_m	0.5977
vertical_dist_m	0.1732
slope_ave_percent	0.311
val_width_ave_m	0.09624
RL_confinement	0.005218
RL_floodplain_area_ave_m2	0.08144
RL_floodplain_area_ave_ha	0.08144
basal_area_tally_ave	0.5474
basal_area_m2perha	0.5474
total_woodload_m3	0.003354
woodload_m3perm2	0.004123
woodload_m3perha	0.004123
jam_per_m	0.03906

Transformation of reach response variables

Total number of jams is normally distributed at the reach level, but total woodload and woodload per area are not. Here, I check to see if square root, log10, or natural log transformation will change that.

```
## testing for total woodload
shapiro.test(sqrt(full_reach_ch$total_woodload_m3)) #normal

## testing for woodload per area
shapiro.test(sqrt(full_reach_ch$woodload_m3perm2)) #normal
```

Square root transformation makes the remaining two response variables normal.

5. Look at simple comparisons

Now that we have all of the dataframes and variables we need, we can perform some simple, univariate comparisons.

Starting with jam level variables

Comparing wood volume per jam to other jam variables. The table below contain the correlation coefficient for different variables with regard to wood volume (the cor() function in R).

The strongest correlations to wood volume (in m3) are distance from channel and height above bankfull.

Table 3: Jam Correlation (R values if significant)

	woodvolume_m3	
variable	r_value	p_value
reach		0.35
pct_wood_1		0.59
pct_wood_2		0.89
woodvolume_m3		1.00
ht_above_bf_m	-0.31	0.00
dist_from_channel_m	-0.31	0.00
bf_width_m	0.21	0.00
rangefinder_slope_.		0.17
obstruction_index	0.21	0.00
pin_DBH_cm_1	0.17	0.01
pin_DBH_cm_2	0.27	0.00
pin_DBH_cm_3	0.29	0.00
pins	0.19	0.00
tot_blockage	0.28	0.00

Moving on to reach level variables

See page below for table of correlation values. Significant values are bolded. Average jam size and woodload per unit area did not significantly correlate to any reach scale characteristics. These two variables were correlated to each other, which is not an interesting relationship.

Total woodload in m3 is signifcantly correlated to average slope and river left confinement ($r > 0.5$ and $p<0.05$).

The number of jams (both organic and LW/mixed) is significantly correlated to valley bottom width, river left confinement, floodplain area, and basal area ($r > 0.5$ and $p < 0.05$). Essentially, more space equals more jams.

Table 4: Reach Correlation Values (bolded if significant) (continued below)

	Avg_Jam_Size	Jams_LWmixed	Jams_org	Jams_total
Avg_Jam_Size	1	0.38	0.31	0.34
Jams_LWmixed	0.38	1	**0.89**	**0.99**
Jams_org	0.31	**0.89**	1	**0.92**
Jams_total	0.34	**0.99**	**0.92**	1
bf_width	0.23	0.25	0.32	0.22
Avg_slope	-0.41	**-0.62**	**-0.57**	**-0.63**
Valley_width	0.36	**0.77**	**0.87**	**0.79**
RL_conf	0.46	**0.69**	**0.73**	**0.71**
fp_area_m2	0.32	**0.79**	**0.89**	**0.81**
basal_tally	0.25	**0.85**	**0.94**	**0.89**
basal_m2perha	0.25	**0.85**	**0.94**	**0.89**
tot_woodload	**0.87**	**0.68**	**0.56**	**0.63**
woodload_m2	**0.91**	0.44	0.25	0.39
jam_per_m	0.3	**0.99**	**0.92**	**1**

Table 5: Table continues below

	bf_width	Avg_slope	Valley_width	RL_conf
Avg_Jam_Size	0.23	-0.41	0.36	0.46
Jams_LWmixed	0.25	**-0.62**	**0.77**	**0.69**
Jams_org	0.32	**-0.57**	**0.87**	**0.73**
Jams_total	0.22	**-0.63**	**0.79**	**0.71**
bf_width	1	-0.09	0.37	0.19
Avg_slope	-0.09	1	**-0.67**	**-0.65**
Valley_width	0.37	**-0.67**	1	**0.91**
RL_conf	0.19	**-0.65**	**0.91**	1
fp_area_m2	0.39	**-0.64**	**0.99**	**0.9**
basal_tally	0.24	**-0.56**	**0.79**	**0.68**
basal_m2perha	0.24	**-0.56**	**0.79**	**0.68**
tot_woodload	0.32	**-0.54**	**0.58**	**0.6**
woodload_m2	0.15	-0.34	0.19	0.32
jam_per_m	0.2	**-0.6**	**0.76**	**0.68**

Table 6: Table continues below

	fp_area_m2	basal_tally	basal_m2perha	tot_woodload
Avg_Jam_Size	0.32	0.25	0.25	**0.87**
Jams_LWmixed	**0.79**	**0.85**	**0.85**	**0.68**
Jams_org	**0.89**	**0.94**	**0.94**	**0.56**
Jams_total	**0.81**	**0.89**	**0.89**	**0.63**
bf_width	0.39	0.24	0.24	0.32
Avg_slope	**-0.64**	**-0.56**	**-0.56**	**-0.54**
Valley_width	**0.99**	**0.79**	**0.79**	**0.58**
RL_conf	**0.9**	**0.68**	**0.68**	**0.6**
fp_area_m2	1	**0.83**	**0.83**	**0.54**
basal_tally	**0.83**	1	**1**	0.46
basal_m2perha	**0.83**	**1**	1	0.46

	fp_area_m2	basal_tally	basal_m2perha	tot_woodload
tot_woodload	**0.54**	0.46	0.46	1
woodload_m2	0.16	0.2	0.2	**0.86**
jam_per_m	**0.79**	**0.89**	**0.89**	**0.6**

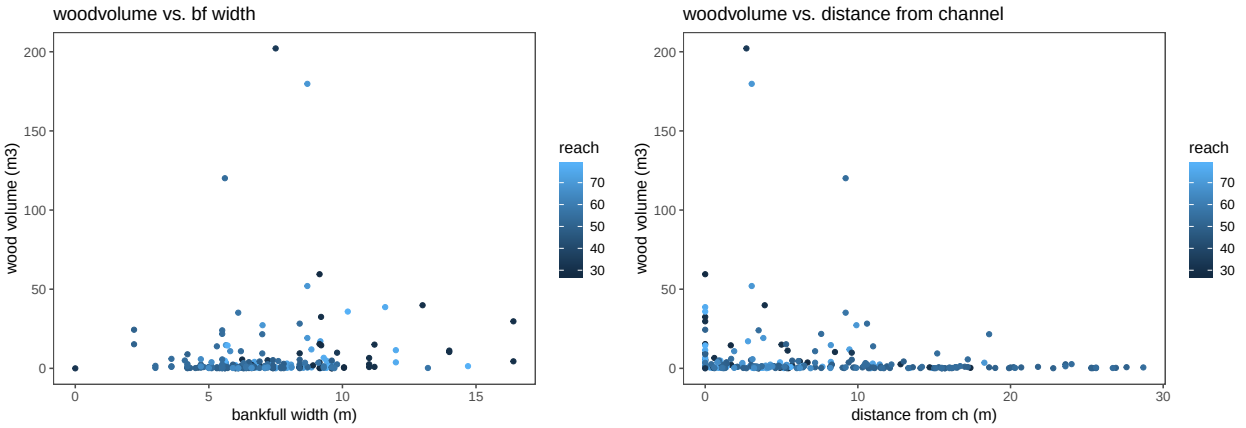
	woodload_m2	jam_per_m
Avg_Jam_Size	**0.91**	0.3
Jams_LWmixed	0.44	**0.99**
Jams_org	0.25	**0.92**
Jams_total	0.39	**1**
bf_width	0.15	0.2
Avg_slope	-0.34	**-0.6**
Valley_width	0.19	**0.76**
RL_conf	0.32	**0.68**
fp_area_m2	0.16	**0.79**
basal_tally	0.2	**0.89**
basal_m2perha	0.2	**0.89**
tot_woodload	**0.86**	**0.6**
woodload_m2	1	0.36
jam_per_m	0.36	1

6. Looking at non-linear relationships

Some relationships between variables of interest and predictors may be non-linear. Looking at plots could be a good idea for revealing these relationships.

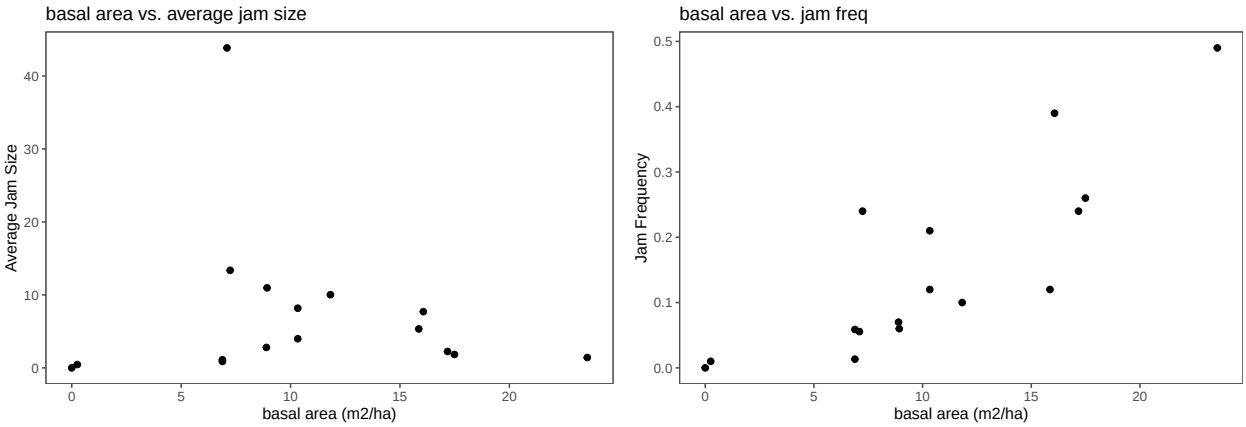
Jam scale plots of interest

We originally thought that there would be a predictable pattern in bankfull width and distance from channel to woodvolume per jam. However, plots reveal no strong pattern in these data. It is possible to see that woodvolume is greatest at intermediate bankfull widths and wood volume decreases with distance from channel, but there is substantial variability.



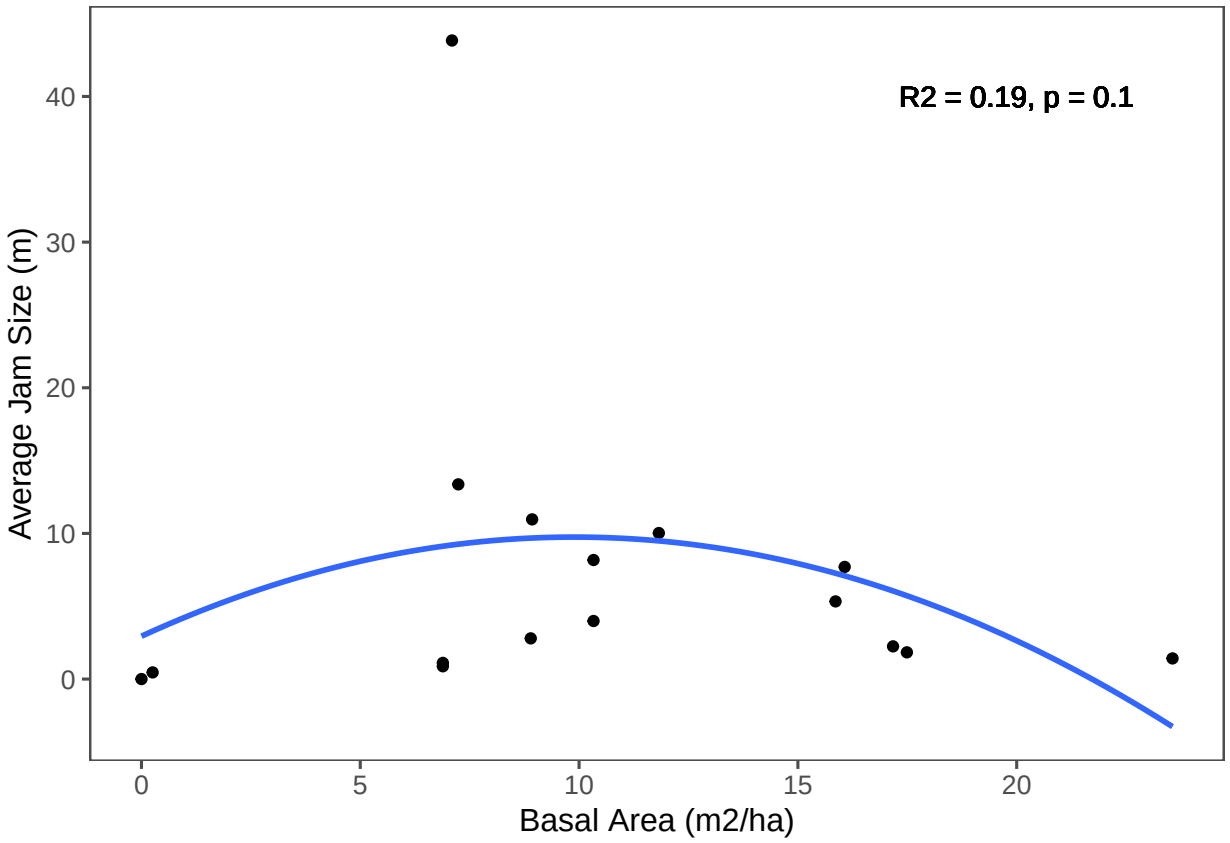
Reach scale plots of interest

At the reach scale, we thought average jam size and frequency would scale non-linearly with basal area (largest jams at some intermediate basal area). There appears to be a weak non-linear pattern between average jam size and basal area. Additionally, I plotted jam frequency vs. basal area - which ended up being a linear function with a significant R2.



Method for including non-linear relationship in regressions

Used in the models below for basal area or tally. A second degree polynomial might explain the relationship between jams and basal area. However, further testing reveals that this is not a significant correlation.



7. Multiple regression models

Now that normality and correlation has been checked, multiple linear regressions for a few variables of interest can be created. Here, non-significant variables have been removed from the models. Transformed response variables are used in all cases where transformation resulted in normally distributed data. The only significant model is the final model for number of jams per reach.

Model for wood volume at the jam scale

Total blockage and number of pins are the only significant ($p < 0.05$) predictors according to the full model. Therefore, it makes statistical sense to dredge other predictors from the model. **The model has a very low R2 value, but is significant.** Also, residuals are not normal, no homoscedacity.

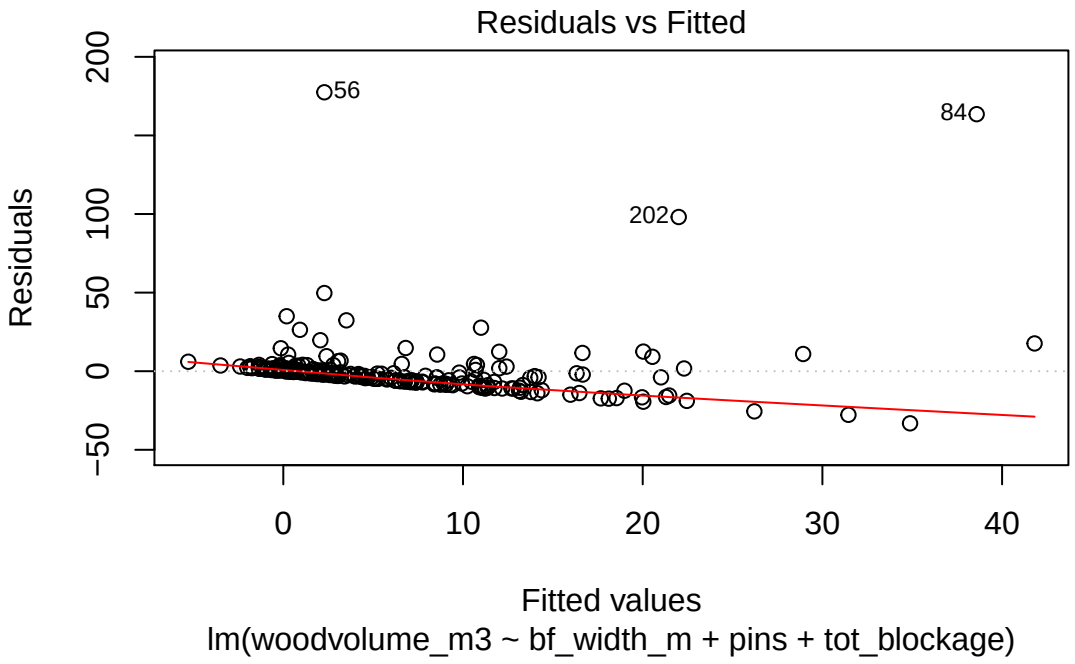
```
## create dataframe excluding reaches without jams
jam_vol_df <- jam_data %>%
  filter(!reach==32)

## full model for woodvolume at jam scale
jam_vol_mod_full <- lm(data = jam_vol_df, woodvolume_m3 ~
                         ht_above_bf_m +
                         dist_from_channel_m +
                         bf_width_m +
                         obstruction_index +
                         pins+
                         tot_blockage)
#summary(jam_vol_mod_full)

## dredge full model
options(na.action = "na.fail")
jam_dredge <- dredge(jam_vol_mod_full, extra = c("R^2"))

## final model
jam_final <- lm(data = jam_vol_df, woodvolume_m3 ~ bf_width_m + pins+tot_blockage)

## summary
s <- summary(jam_final)
pander(s, caption = 'Summary: Jam Wood Volume Model')
```

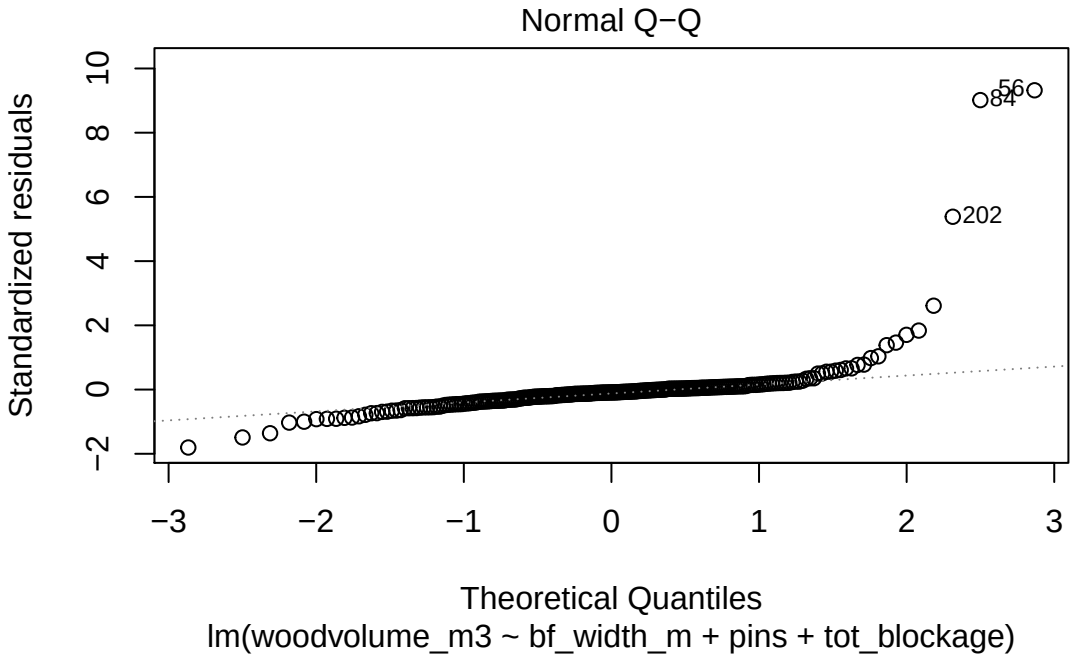
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.759	4.499	-1.058	0.2913
bf_width_m	0.8108	0.5624	1.442	0.1507
pins	-3.682	1.823	-2.019	0.04458
tot_blockage	0.3137	0.06842	4.585	7.359e-06

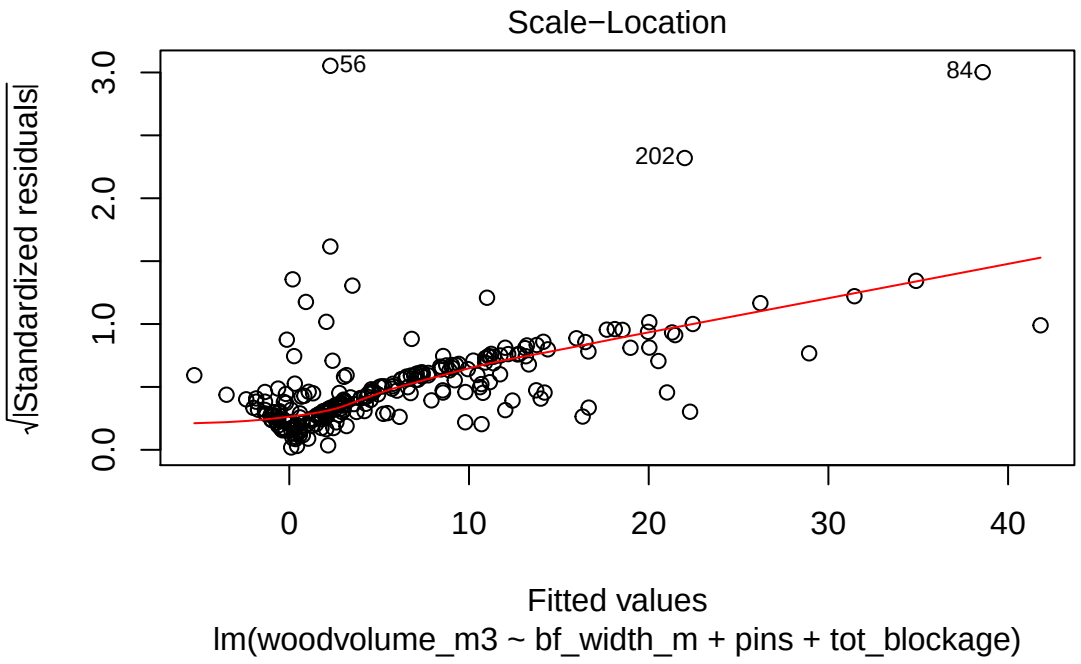
Table 9: Summary: Jam Wood Volume Model

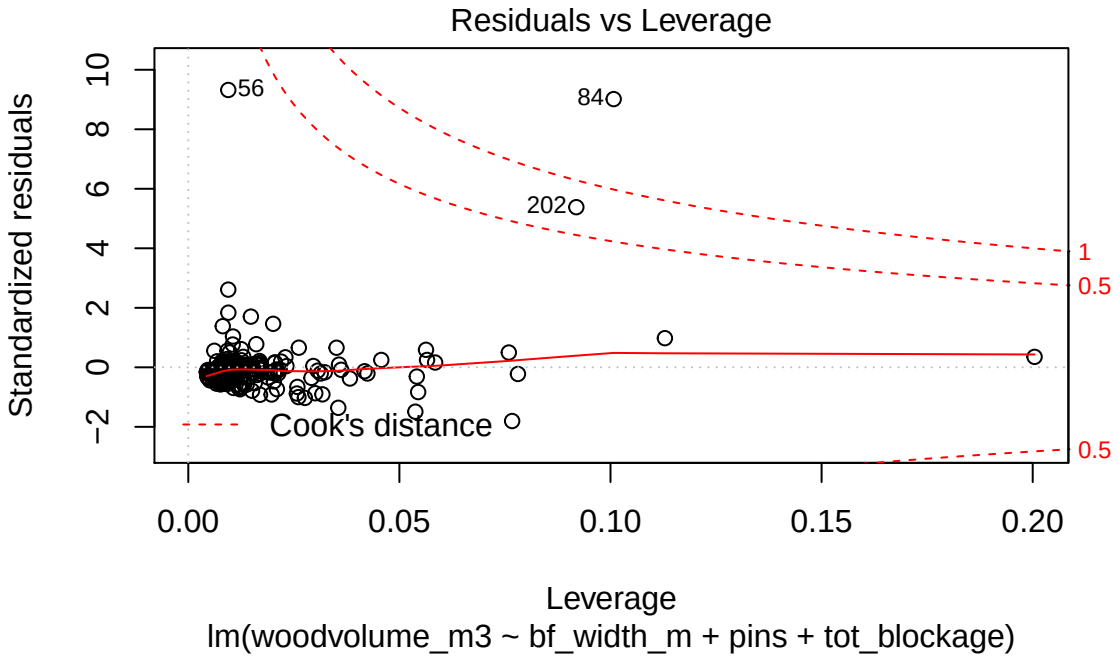
Observations	Residual Std. Error	R^2	Adjusted R^2
241	19.13	0.1324	0.1215

```
## residual plots
plot(jam_final)
```


Residuals vs Fitted
Residuals
Fitted values
lm(woodvolume_m3 ~ bf_width_m + pins + tot_blockage)
56
84
202


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(woodvolume_m3 ~ bf_width_m + pins + tot_blockage)
56
84
202


Scale−Location
√|Standardized residuals|
56
84
202
0.0
1.0
2.0
3.0
0
10
20
30
40
Fitted values
lm(woodvolume_m3 ~ bf_width_m + pins + tot_blockage)


Residuals vs Leverage
Standardized residuals
56
84
202
Cook's distance
1
0.5
0.5
−2
0
2
4
6
8
10
0.00
0.05
0.10
0.15
0.20
Leverage
lm(woodvolume_m3 ~ bf_width_m + pins + tot_blockage)

Model for woodload per area

No significant model.

```
## full model for woodload per area
woodload_mod_full <- lm(sqrt(woodload_m3perm2) ~ bankfull_width_ave_m +
                          jam_total +
                          RL_confinement +
                          poly(basal_area_tally_ave,2),
                          data = full_reach_ch)
#summary(woodload_mod_full)

## dredge full model
woodload_dredge <- dredge(woodload_mod_full, extra = c("R^2"))

## dredged model with highest AICc is not significant
woodload_final <- lm(sqrt(woodload_m3perm2) ~ jam_total, data = full_reach_ch)

## summary
s2 <- summary(woodload_final)
pander(s2, caption = 'Summary: Woodload per Area Model')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1817	0.05345	3.399	0.004322
jam_total	0.002667	0.002611	1.022	0.3242

Table 11: Summary: Woodload per Area Model

Observations	Residual Std. Error	R^2	Adjusted R^2
16	0.1448	0.06939	0.002922

Model for total woodload

No significant model.

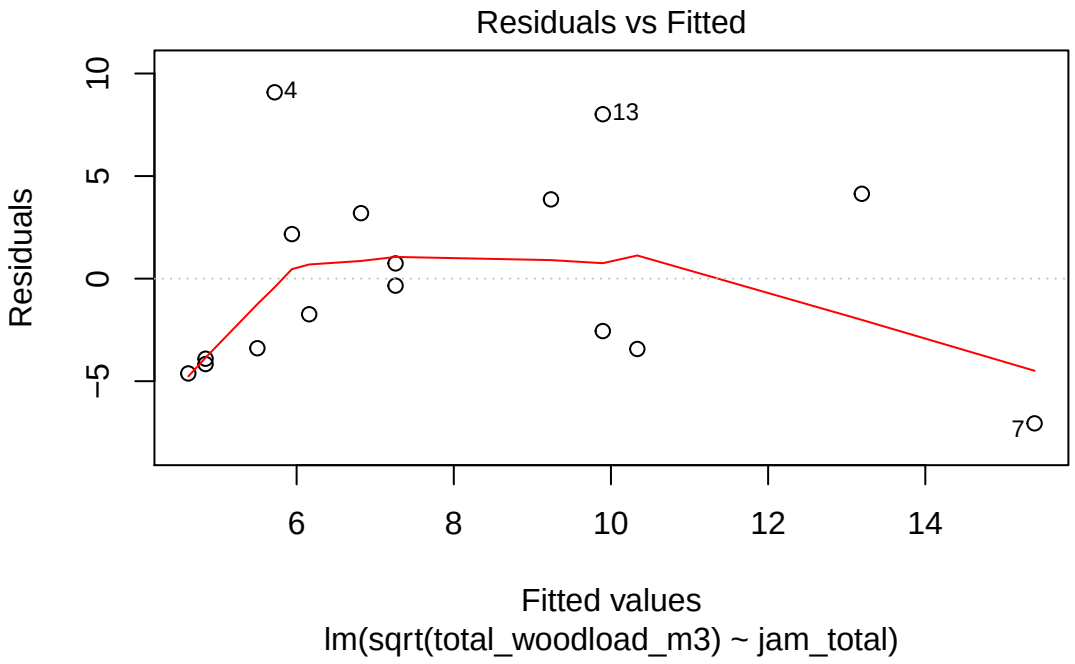
```
## full model for total woodload
tot_woodload_mod_full <- lm(sqrt(total_woodload_m3) ~
                              bankfull_width_ave_m +
                              jam_total +
                              RL_confinement +
                              poly(basal_area_tally_ave,2),
                            data = full_reach_ch)
#summary(tot_woodload_mod_full)

## dredge full model
tot_woodload_dredge <- dredge(tot_woodload_mod_full, extra = c("R^2"))


## dredged model with highest AICc is slightly significant
tot_woodload_final <- lm(sqrt(total_woodload_m3) ~ jam_total,
                         data = full_reach_ch)

## summary
s3 <- summary(tot_woodload_final)
pander(s3, caption = 'Summary: Total Woodload Model')
```

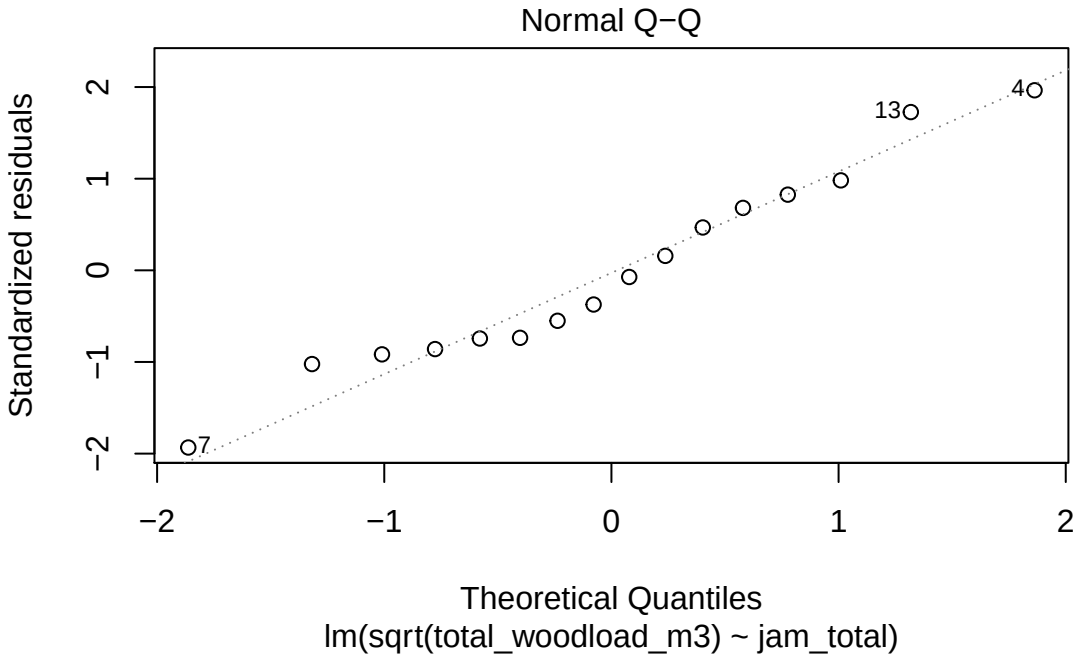
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.622	1.795	2.575	0.022
jam_total	0.2198	0.08765	2.507	0.02511

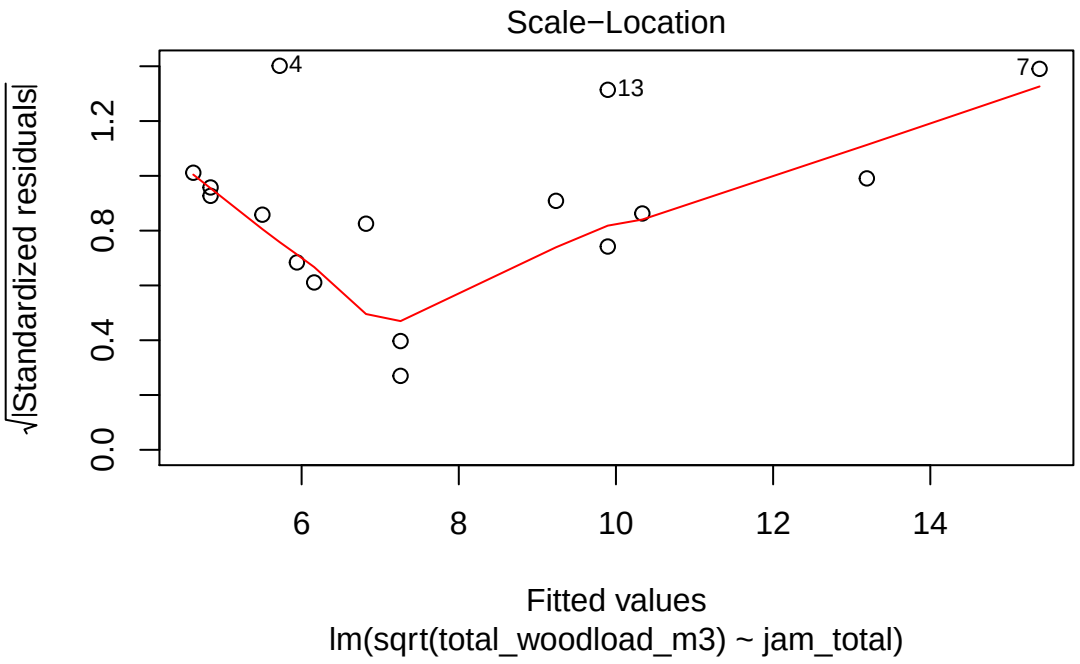
Table 13: Summary: Total Woodload Model

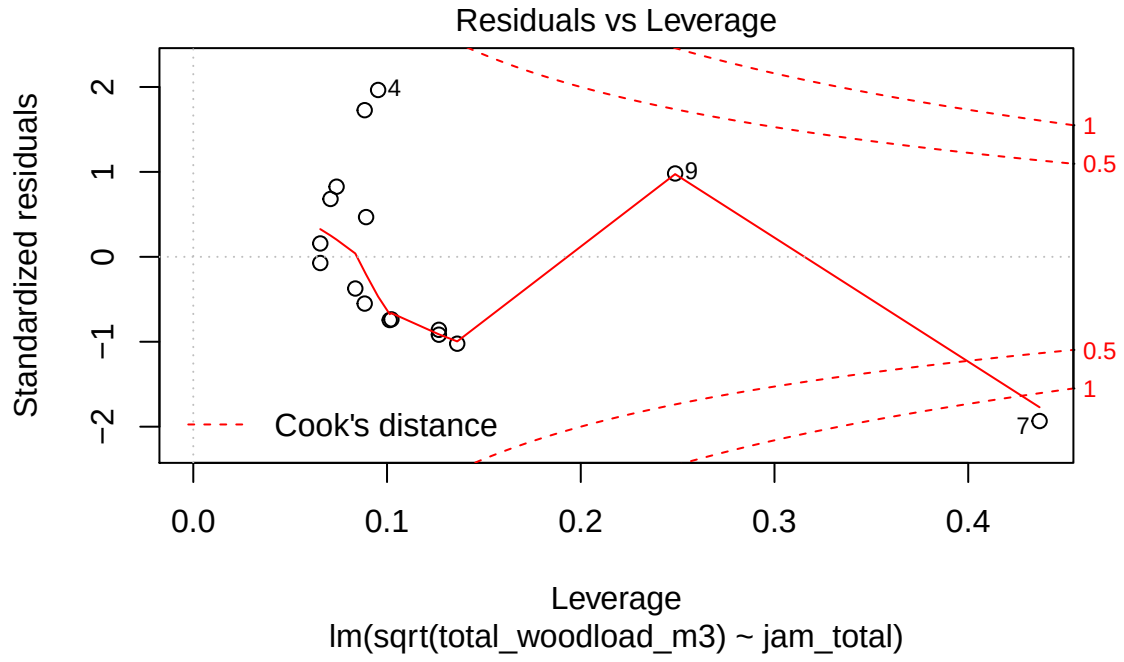
Observations	Residual Std. Error	R^2	Adjusted R^2
16	4.862	0.3099	0.2606

```
## residual plots
plot(tot_woodload_final)
```


Residuals vs Fitted
Residuals
Fitted values
lm(sqrt(total_woodload_m3) ~ jam_total)
4
13
7


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(total_woodload_m3) ~ jam_total)
4
13
7


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
4
13
7
6
8
10
12
14
Fitted values
lm(sqrt(total_woodload_m3) ~ jam_total)


Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
4
9
7
1
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
Leverage
lm(sqrt(total_woodload_m3) ~ jam_total)

Model for total number of jams

Bankfull width and RL_confinement are the only significant model variables. This is a significant model (r2 = 0.74, p <<0.01).

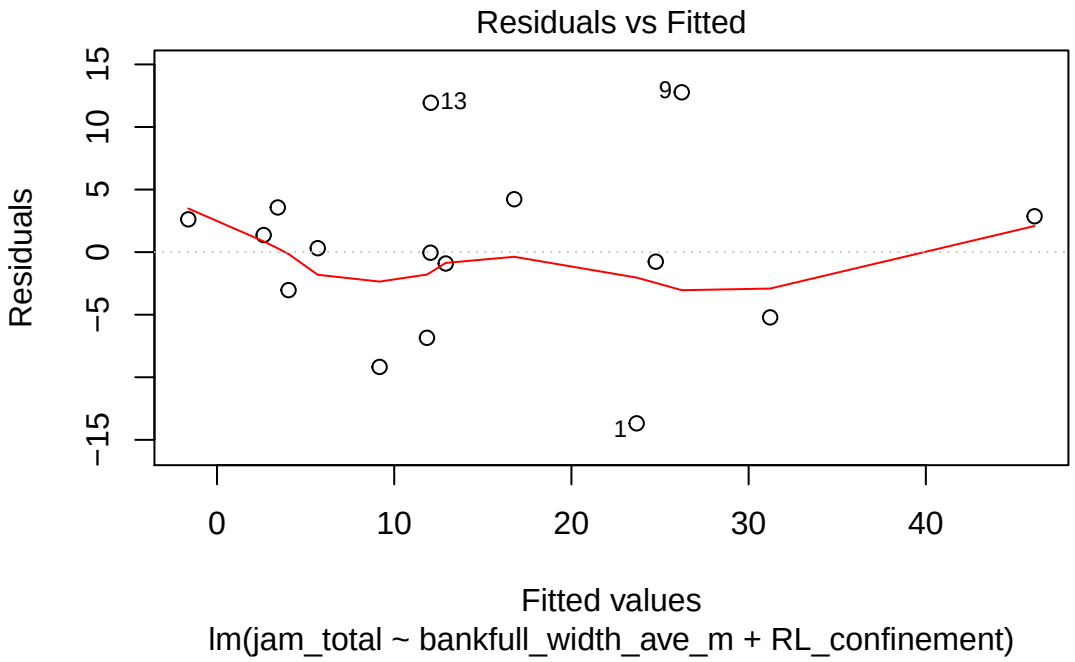
```
## full model for number of jams
number_model_full <- lm(jam_total ~ bankfull_width_ave_m +
                          RL_confinement +
                          poly(basal_area_tally_ave, 2),
                        data = full_reach_ch)
#summary(number_model_full)

## dredge full model
number_dredge <- dredge(number_model_full, extra = c("R^2"))

## dredged model for number of jams per reach is significant!!
number_final <- lm(jam_total ~ bankfull_width_ave_m + RL_confinement,
                         data = full_reach_ch)

## summary
s4 <- summary(number_final)
pander(s4, caption = 'Summary: Jam Count Model')
```

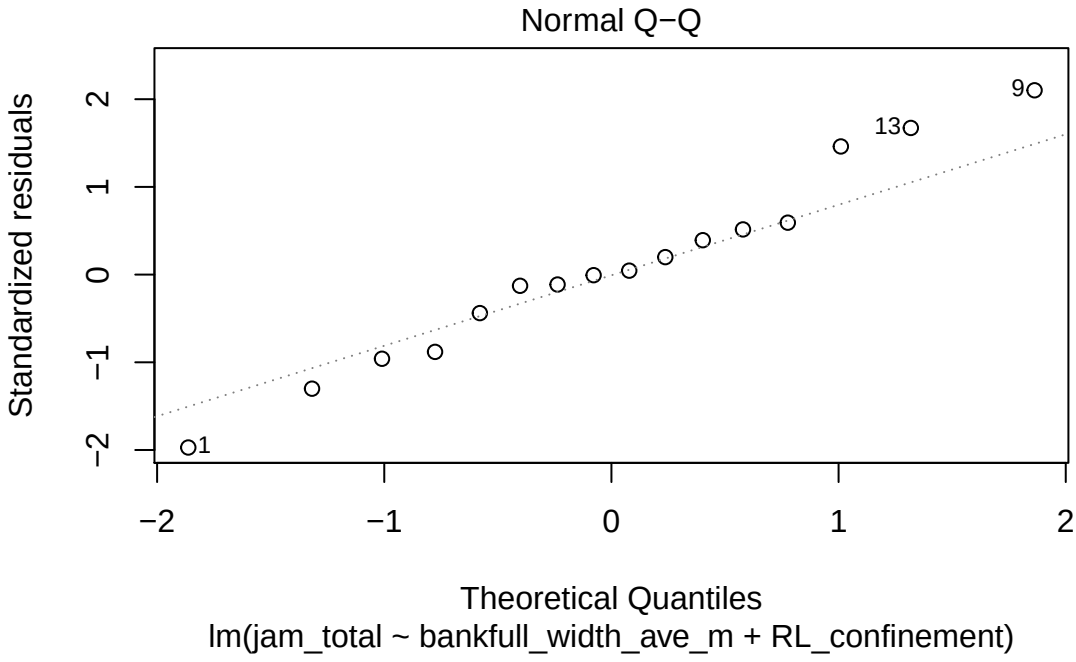
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.08	4.525	-2.67	0.01926
bankfull_width_ave_m	1.748	0.335	5.217	0.000166
RL_confinement	6.649	1.36	4.888	0.0002964

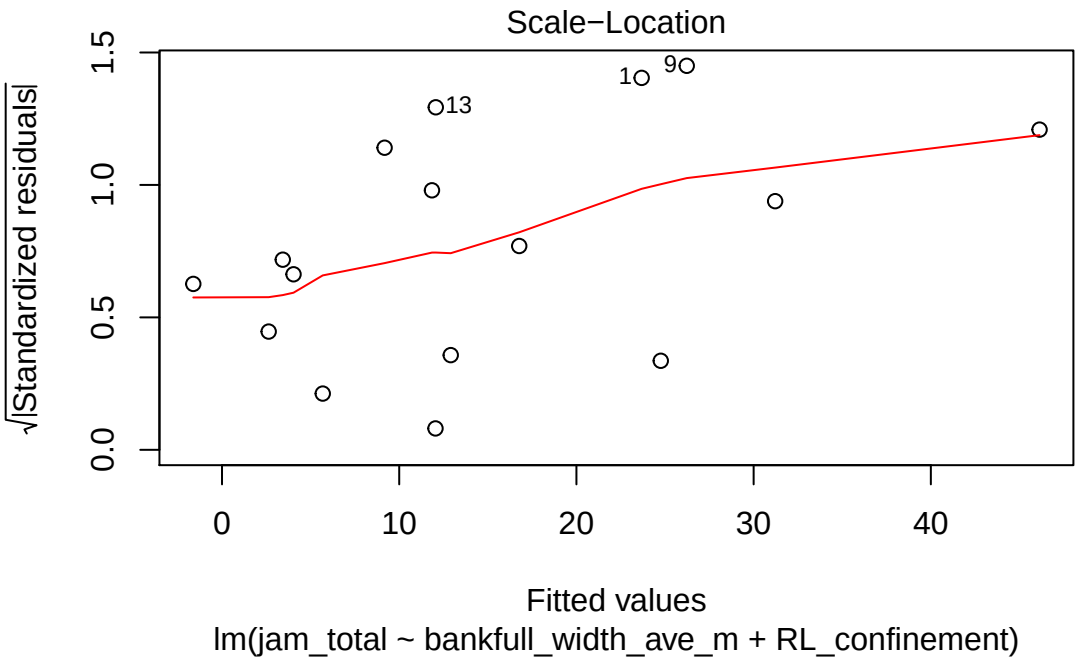
Table 15: Summary: Jam Count Model

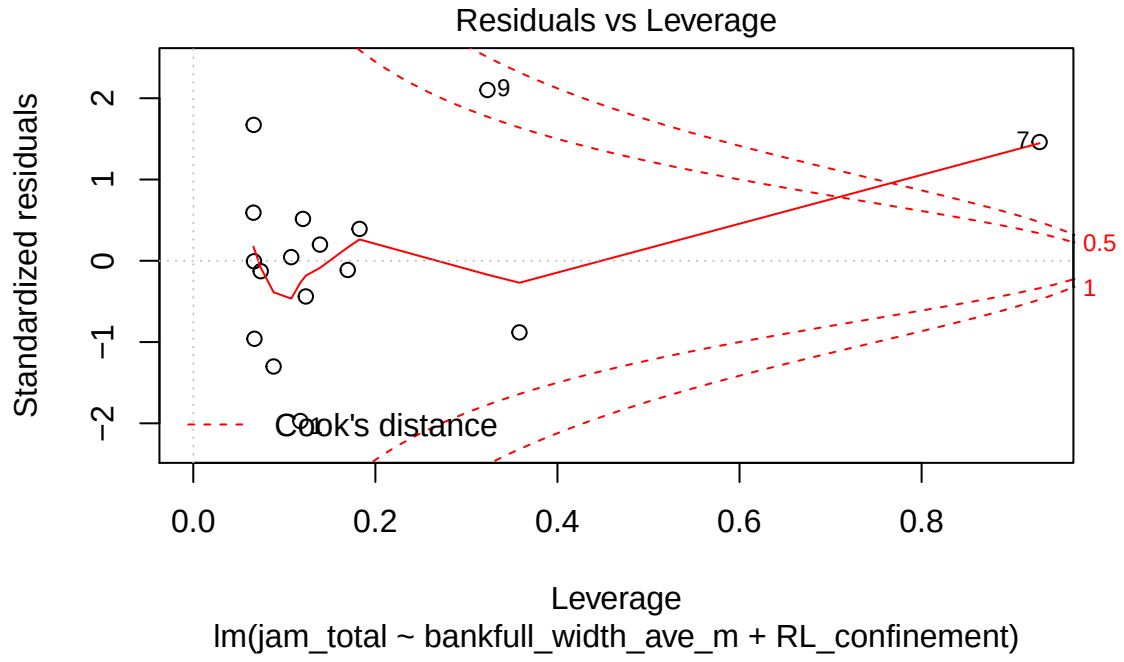
Observations	Residual Std. Error	R^2	Adjusted R^2
16	7.387	0.7695	0.734

```
## residual plots
plot(number_final)
```


Residuals vs Fitted
Residuals
Fitted values
lm(jam_total ~ bankfull_width_ave_m + RL_confinement)
13
9
1


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(jam_total ~ bankfull_width_ave_m + RL_confinement)
9
13
1





8. Conclusions

While there is a weak non-linear relationship between jam characteristics and basal area, the strongest modeled relationship was between total number of jams and bankfull width + RL confinment. Geometric characteristics are more significant than forest characteristics in multiple regressions.