UW TACOMA – MSBA – TBANLT – 540 A

Washington Bridge Data

Pooja sheth

Applied Regression – Final Project

2017

***Table of Contents***

[**Project Information**](#ProjectInformation)**2**

[**Dataset**](#Dataset)**4**

[**Cleaned Data set**](#CleanedDataset)**5**

[**Dataset Information**](#DatasetInformation)**6**

[**Analysis**](#Analysis)**7**

[**Descriptive Statistics**](#Descriptivestatistics)**7**

[**Regression**](#Regression)**9**

[**Graphs**](#Graphs)**13**

[**Interaction Variables**](#InteractionVariables)**17**

[**Partial F-Test**](#PartialFtest)**20**

[**Conclusion**](#Conclusion)**22**

**Project Information**

**Abstract**

Sufficiency rating is an overall rating of the bridge’s fitness for the duty that it performs. Sufficiency rating values are used in many development and improvement plans decisions taken by government. Different regression techniques have carried on the dataset for bridges in Washington state to find the factors affecting the sufficiency rating of the bridges.

**Introduction**

Washington State Department of Transportation (WSDOT) is assigning budget for the repair and seismic retrofit of bridges in Washington state.

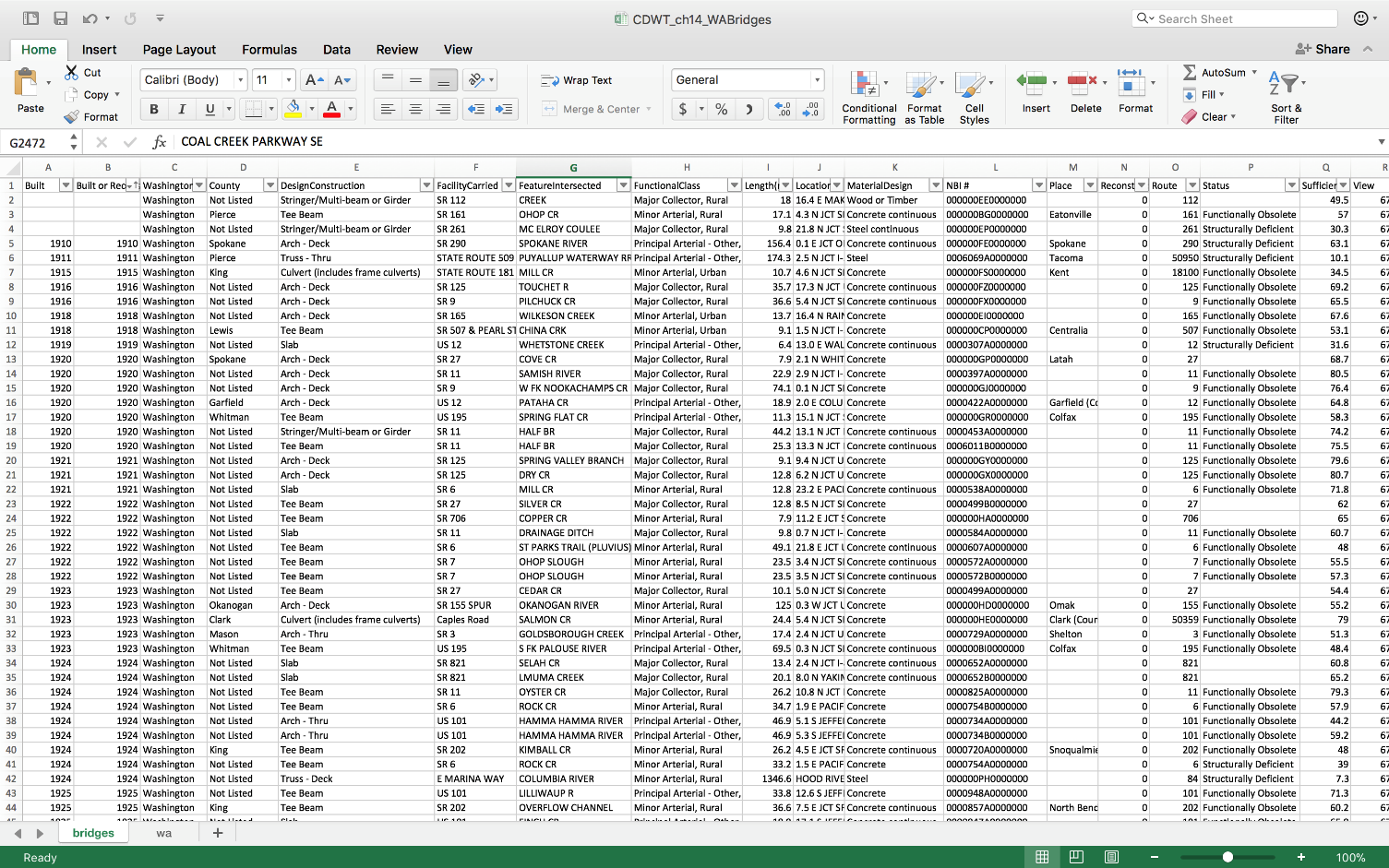
Since the budget is limited and there is not enough budget to cover the repair for all bridges that requires retrofit, board of engineers want to prioritize the bridges for receiving budget based on their serviceability and other critical criteria. One of these criteria is sufficiency rating of the bridge and we have chosen this criterion as our independent variable to investigate the effect of other variables on it. So, we would be identifying and evaluate the factors affecting the sufficiency rating variable.

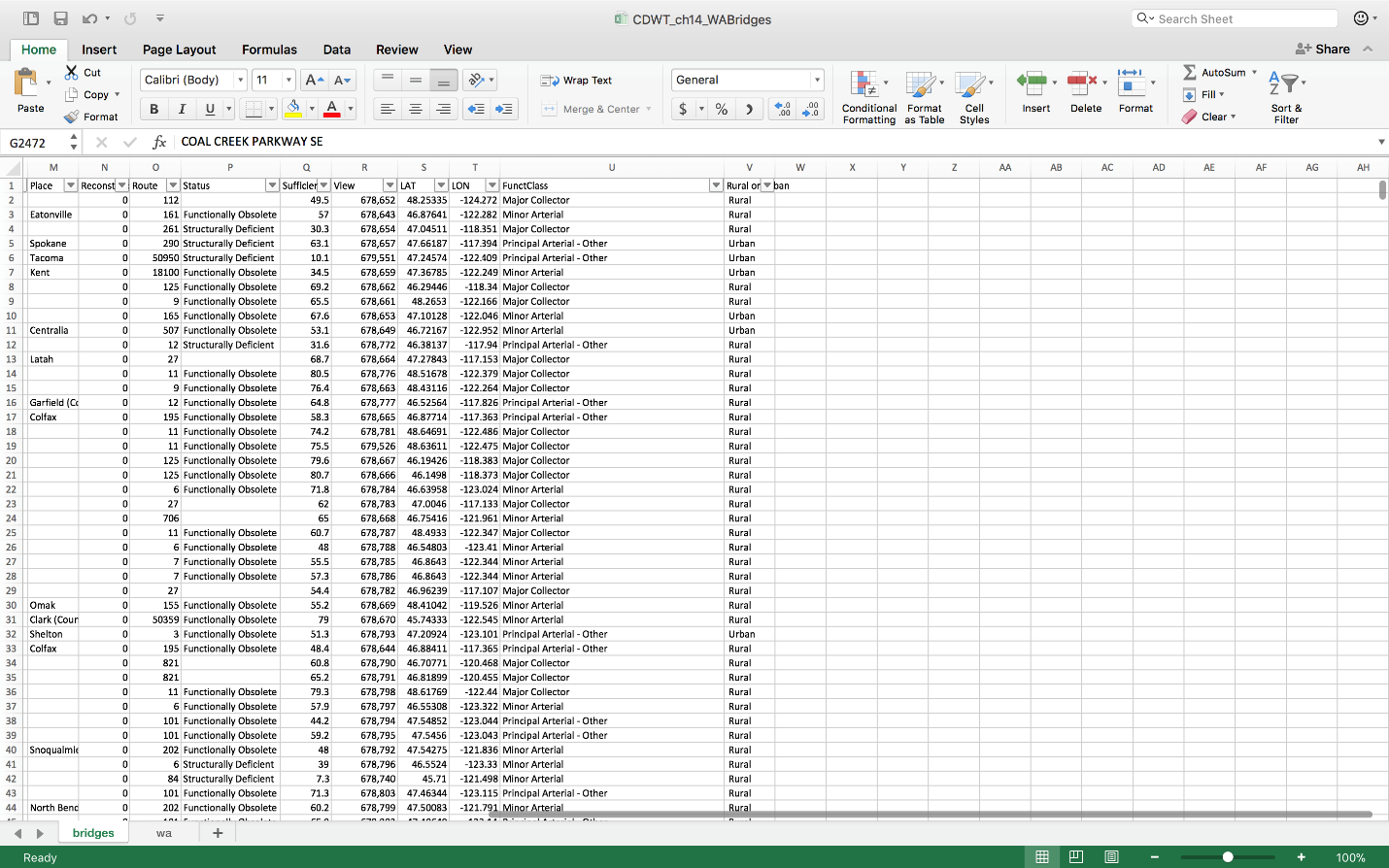
**Data Set**

Dataset is taken from “Washington State Department of Transportation” (WSDOT) and represent information about the structural specification, evaluation, and condition of bridges in Washington state.

***Dataset***

Below are the snapshots of the master data

****

****

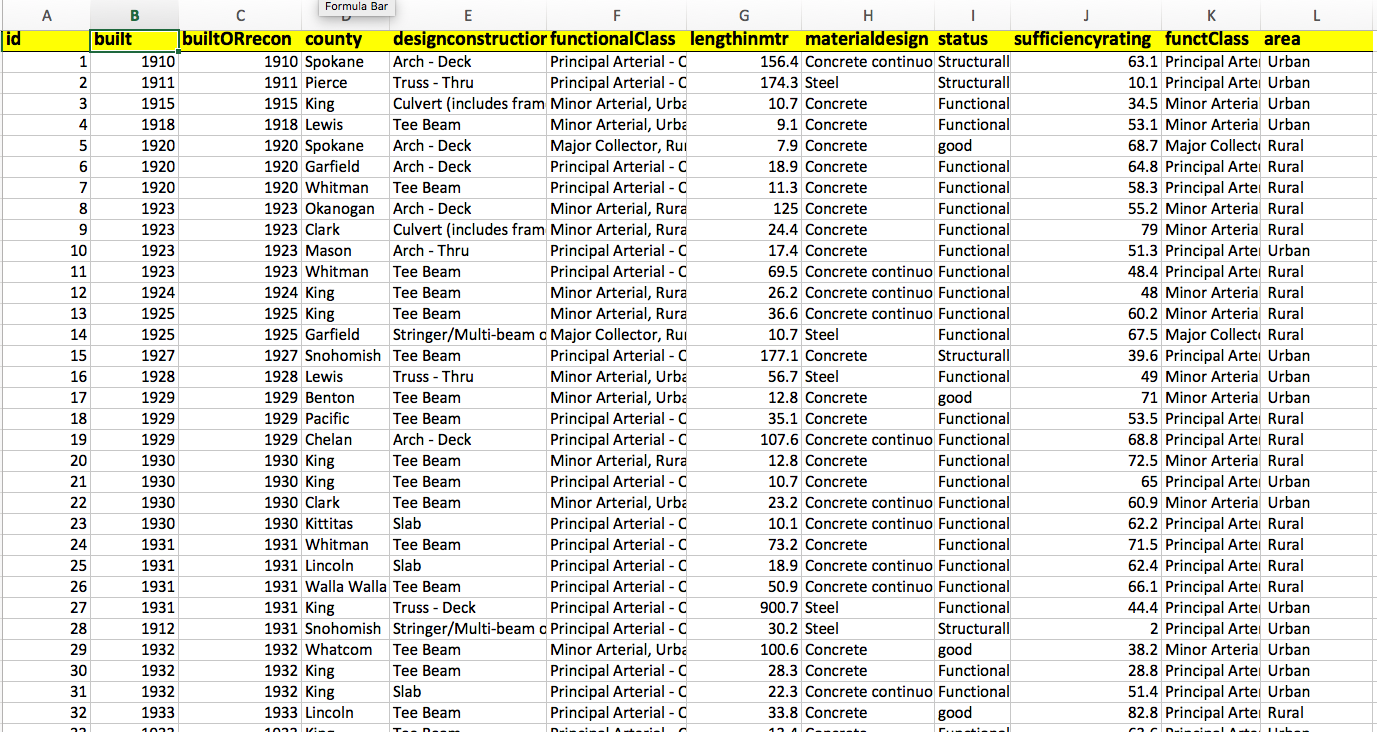
List of all variables included in master dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Built | Built or reconst | Washington | County | Design Construction |
| Facility Carried | Feature Intersected | Fucntional Class | Length (in mtr) | Location |
| MaterialDesign | NBI# | Place | Reconstruction | Route |
| Status | Sufficiency rate | View | LAT | LON |
| FunctClass | Area - Rural or Urban |  |  |  |

Data which are not affecting the structural efficiency of bridges are cleaned from data set.

***Cleaned Dataset***

Below is the snapshot of the cleaned data set



***Dataset Information***

**Variable:**

**Independent variable:**

* *Sufficiency Rating*: Sufficiency ratings are determined during the biennial bridge inspection and are intended to indicate a measure of the ability of a bridge to remain in service. Sufficiency ratings are used by the Federal Highway Administration to select the candidate bridges to receive budget for repair and retrofit.

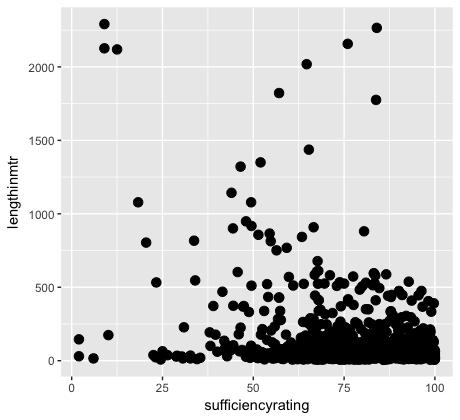
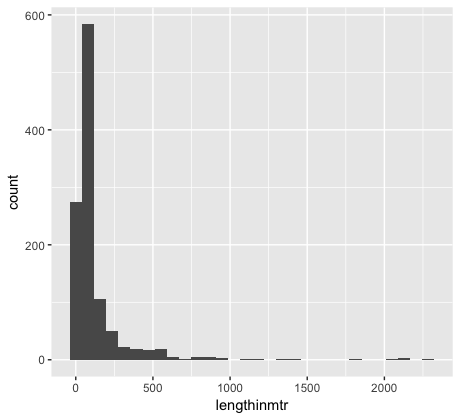
**Dependent Variables:**

1. *Year Build*: Year in which bridge was built
2. *Reconstructed*: Year in which bridge was repaired or reconstructed
3. *County*: County in which the bridge is located
4. *Design Construction*: Structural concept of the bridge
5. *Length(m):* Maximum length for the bridge span in meter
6. *Material Design:* Material for the structure of the bridge
7. *Status:* Structural status of the bridge
8. *Funct Class:* Functional Class of the bridge
9. *Area:* Type of location where the bridge is located Rural or Urban

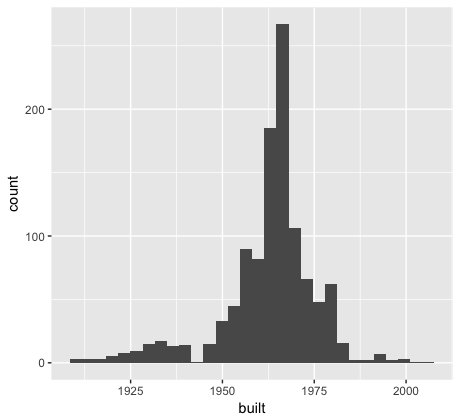
***Analysis***

**Analyze the Data using the techniques learned throughout class**

* + **Descriptive statistics**



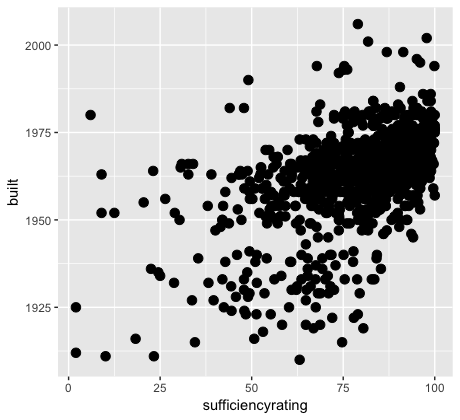
*For the most part, the bridges with length less than 750mtr sufficiency rating is higher and above 50%. There are few long bridges (>1000mtr) and still have good sufficiency ratings.*



*Histograms shows number of bridges constructed for the corresponding year.*

*1960’s saw the highest bridge construction compared to any other year.*

*Bridges constructed after 1960 observed sufficiency rating of 75%.*



* + **Regression**

**We performed linear regression along with area as a variable in the model too. But we got the same summary statistics as with this model below. And further when we carried out analysis, the variable wasn't much significant and area is rather a subset of the functional class values. i.e. urban and rural.**

summary(teamlm)

Call:

lm(formula = team$sufficiencyrating ~ built + countyFact + designconstructionFact +

functionalClassFact + lengthinmtr + statusFact, data = team)

Residuals:

Min 1Q Median 3Q Max

-35.292 -5.275 0.744 5.888 30.741

Coefficients:

Estimate Std. Error t value

(Intercept) -741.746785 61.011691 -12.157

built 0.418979 0.031520 13.292

countyFactAsotin 13.551770 7.281297 1.861

countyFactBenton 3.365584 5.084942 0.662

countyFactChelan -1.195080 5.458143 -0.219

countyFactClallam -7.281910 5.813670 -1.253

countyFactClark 0.665235 4.694218 0.142

countyFactColumbia 5.905702 6.268443 0.942

countyFactCowlitz 2.870498 4.805807 0.597

countyFactDouglas 4.838226 6.136229 0.788

countyFactFerry -11.995989 5.486032 -2.187

countyFactFranklin 6.020182 5.138114 1.172

countyFactGarfield 6.085496 8.364897 0.728

countyFactGrant 2.559380 5.399305 0.474

countyFactGrays Harbor 0.150143 4.859751 0.031

countyFactJefferson -16.569527 7.595187 -2.182

countyFactKing 2.421142 4.482419 0.540

countyFactKitsap 3.428185 4.937262 0.694

countyFactKittitas 2.667663 5.310730 0.502

countyFactKlickitat 8.704964 8.247056 1.056

countyFactLewis 4.141601 4.993481 0.829

countyFactLincoln 5.540332 5.086143 1.089

countyFactMason -4.229185 6.672186 -0.634

countyFactOkanogan 1.806414 5.242851 0.345

countyFactPacific -7.066729 6.368566 -1.110

countyFactPend Oreille 12.060244 8.453524 1.427

countyFactPierce 2.464938 4.559320 0.541

countyFactSan Juan -1.723639 10.740036 -0.160

countyFactSkagit 5.289949 4.898877 1.080

countyFactSkamania -16.011021 6.837939 -2.341

countyFactSnohomish 1.373943 4.551920 0.302

countyFactSpokane 6.172206 4.759663 1.297

countyFactThurston 2.662973 4.922660 0.541

countyFactWalla Walla 2.321512 5.247278 0.442

countyFactWhatcom -1.125924 4.723144 -0.238

countyFactWhitman 5.172517 5.122161 1.010

countyFactYakima 1.827432 4.649127 0.393

designconstructionFactArch -16.825169 5.606594 -3.001

designconstructionFactBox Beam or Girders - Multiple

-5.427418 3.379845 -1.606

designconstructionFactBox Beam or Girders - Single or Spread

-6.442268 3.374419 -1.909

designconstructionFactCulvert (includes frame culverts)

-9.712738 4.345183 -2.235

designconstructionFactFrame (except frame culverts)

-7.785690 10.472246 -0.743

designconstructionFactGirder and Floorbeam System

-10.319077 4.171211 -2.474

designconstructionFactMovable - Bascule

-21.965009 5.537293 -3.967

designconstructionFactMovable - Lift

-22.580726 4.309557 -5.240

designconstructionFactMovable - Swing

-16.307817 5.849627 -2.788

designconstructionFactOther

-17.819289 6.936644 -2.569

designconstructionFactSegmental Box Girder

-14.764034 8.148320 -1.812

designconstructionFactSlab

-6.733746 3.293439 -2.045

designconstructionFactStringer/Multi-beam or Girder

-8.724298 3.307212 -2.638

designconstructionFactSuspension

2.539497 10.578496 0.240

designconstructionFactTee Beam

-9.998659 3.294220 -3.035

designconstructionFactTruss - Deck

-12.276599 6.706944 -1.830

designconstructionFactTruss - Thru

-22.346334 3.673875 -6.082

functionalClassFactLocal, Rural

-4.494435 2.859927 -1.572

functionalClassFactLocal, Urban

5.354478 3.927971 1.363

functionalClassFactMajor Collector, Rural

-1.154346 2.372401 -0.487

functionalClassFactMinor Arterial, Rural

2.254481 2.745265 0.821

functionalClassFactMinor Arterial, Urban

-0.907346 2.165365 -0.419

functionalClassFactMinor Collector, Rural

3.383747 3.174817 1.066

functionalClassFactPrincipal Arterial - Interstate, Rural

5.389839 2.725963 1.977

functionalClassFactPrincipal Arterial - Interstate, Urban

3.616247 1.962938 1.842

functionalClassFactPrincipal Arterial - Other Freeways, Urban

3.232058 2.016532 1.603

functionalClassFactPrincipal Arterial - Other, Rural

2.950438 2.619319 1.126

functionalClassFactPrincipal Arterial - Other, Urban

-0.162571 2.084493 -0.078

lengthinmtr -0.014258 0.001494 -9.545

statusFactgood 9.950203 0.689371 14.434

statusFactStructurally Deficient

-19.508822 1.548878 -12.595

Pr(>|t|)

(Intercept) < 0.0000000000000002 \*\*\*

built < 0.0000000000000002 \*\*\*

countyFactAsotin 0.06300 .

countyFactBenton 0.50820

countyFactChelan 0.82673

countyFactClallam 0.21065

countyFactClark 0.88733

countyFactColumbia 0.34634

countyFactCowlitz 0.55044

countyFactDouglas 0.43060

countyFactFerry 0.02899 \*

countyFactFranklin 0.24159

countyFactGarfield 0.46708

countyFactGrant 0.63558

countyFactGrays Harbor 0.97536

countyFactJefferson 0.02936 \*

countyFactKing 0.58921

countyFactKitsap 0.48762

countyFactKittitas 0.61555

countyFactKlickitat 0.29143

countyFactLewis 0.40707

countyFactLincoln 0.27627

countyFactMason 0.52631

countyFactOkanogan 0.73050

countyFactPacific 0.26741

countyFactPend Oreille 0.15398

countyFactPierce 0.58887

countyFactSan Juan 0.87253

countyFactSkagit 0.28047

countyFactSkamania 0.01939 \*

countyFactSnohomish 0.76284

countyFactSpokane 0.19499

countyFactThurston 0.58865

countyFactWalla Walla 0.65827

countyFactWhatcom 0.81163

countyFactWhitman 0.31281

countyFactYakima 0.69435

designconstructionFactArch - Thru

0.00275 \*\*

designconstructionFactBox Beam or Girders - Multiple

0.10861

designconstructionFactBox Beam or Girders - Single or Spread

0.05652 .

designconstructionFactCulvert (includes frame culverts)

0.02561 \*

designconstructionFactFrame (except frame culverts)

0.45737

designconstructionFactGirder and Floorbeam System

0.01352 \*

designconstructionFactMovable - Bascule

0.00007780114 \*\*\*

designconstructionFactMovable - Lift

0.00000019430 \*\*\*

designconstructionFactMovable - Swing

0.00540 \*\*

designconstructionFactOther

0.01034 \*

designconstructionFactSegmental Box Girder

0.07029 .

designconstructionFactSlab 0.04114 \*

designconstructionFactStringer/Multi-beam or Girder

0.00846 \*\*

designconstructionFactSuspension

0.81033

designconstructionFactTee Beam

0.00246 \*\*

designconstructionFactTruss - Deck

0.06747 .

designconstructionFactTruss - Thru

0.00000000165 \*\*\*

functionalClassFactLocal, Rural

0.11636

functionalClassFactLocal, Urban

0.17312

functionalClassFactMajor Collector, Rural

0.62666

functionalClassFactMinor Arterial, Rural

0.41170

functionalClassFactMinor Arterial, Urban

0.67528

functionalClassFactMinor Collector, Rural

0.28675

functionalClassFactPrincipal Arterial - Interstate, Rural

0.04828 \*

functionalClassFactPrincipal Arterial - Interstate, Urban

0.06572 .

functionalClassFactPrincipal Arterial - Other Freeways, Urban

0.10928

functionalClassFactPrincipal Arterial - Other, Rural

0.26025

functionalClassFactPrincipal Arterial - Other, Urban

0.93785

lengthinmtr < 0.0000000000000002 \*\*\*

statusFactgood < 0.0000000000000002 \*\*\*

statusFactStructurally Deficient

< 0.0000000000000002 \*\*\*

---

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

Residual standard error: 9.707 on 1051 degrees of freedom

Multiple R-squared: 0.6555, Adjusted R-squared: 0.6335

F-statistic: 29.84 on 67 and 1051 DF, p-value: < 0.00000000000000022

From the above regression output, the factors which are ***significant*** in ***contributing to sufficiency******rating*** are **Built, design construction fact, length in meter, status Fact** are highly significant.

The factors which are *not that significant* in this model are **County** and **functional class**.

* + **Graphs**

**A screenshot of a cell phone

Description generated with very high confidence**

**Interpretation:**

The regression line is very close to the fitted line. This is a good graph of area residuals. The deviation of values from fit line is not too large. The regression line gives us the idea of pattern of residual movement. However, as the area variable is a subset of functionalclass this wasn't considered in the model.

**A close up of a logo

Description generated with high confidence**

**Interpretation:**

The regression line is very close to the fitted line. This is a good graph of county residuals. The deviation of values from fit line is very less. The regression line gives us the idea of pattern of residual movement. There are few outliers, but it can be said that the county is useful in deciding the sufficiency rating.

**A screenshot of a cell phone

Description generated with high confidence**

**Interpretation:**

In the above graph the regression line is almost overlapping the fitted line. This is a good graph of design construction residuals. The deviation of values from fit line is very less. The regression line gives us the idea of pattern of residual movement. There are few outliers, but it can be said that the Design Construction is useful in deciding the sufficiency rating.

**A screenshot of a cell phone

Description generated with very high confidence**

**Interpretation:**

The regression line in the graph of functional class residuals is somewhat curved in the middle, however if we look at the deviation from the fitted line, it is very less. This is a good graph of functional class residuals. The regression line gives us the idea of pattern of residual movement. There are few outliers, but it can be said that the Functional Class is useful in deciding the sufficiency rating.

**A close up of a piece of paper

Description generated with high confidence**

**Interpretation:**

This graph of length in meters residuals is cloud shaped with outliers at the start. However, the regression line is close to the fitted line with one intersection. The deviation of the residual values is not that high. Therefore, Length in Meters is a useful factor in determining the sufficiency rating.

**A close up of a piece of paper

Description generated with high confidence**

**Interpretation:**

This cloud shaped graph of year built residuals shows that the regression line is close to the fitted line, however, it bends downwards towards the end. The deviation of the values is minimal at the start, but it increases at the end. This graph has outliers. Year built can be used in determining the sufficiency rating, however, there are other more significant factors than year built.

**A screenshot of a cell phone

Description generated with very high confidence**

**Interpretation:**

This residual graph for the status vs sufficiency rating is a good graph. This regression line is close to the fitted line at the start with minimum deviation, however, it almost coincides with the fitted line later. Status variable can be used in a model to determine sufficiency rating.

It is observed that countyfact, functionclassfact are variables with high collinearity and designconstructionfact has high potential collinearity. So, we further create and test the interaction variables, partial F test to come up a parsimonious model.

vif(lm(formula = team$sufficiencyrating ~ built + countyFact + designconstructionFact + lengthinmtr+ statusFact + functionalClassFact, data = team))

GVIF Df GVIF^(1/(2\*Df))

built 1.942890 1 1.393876

countyFact 26.668127 35 1.048024

designconstructionFact 8.818955 17 1.066121

lengthinmtr 1.404107 1 1.184950

statusFact 1.493385 2 1.105460

functionalClassFact 11.175448 11 1.11595

***Interaction Variables***

Further, by creating the interaction variables adjusted R-squared value can be checked.

|  |  |  |
| --- | --- | --- |
| Interaction Variable | Adjusted R-squared Value | Interpretation |
| CountyFact\*designConstructionFact | 0.6517 | This first variable gave the adjusted R-squared value of 0.6517. This value can be compared to other model’s adjusted R-squared values further. |
| CountyFact\*built | 0.6606 | Insertion of interaction variable county and year built is significant as the adjusted R-squared value is greater than before. |
| lengthinmtr\*functionalClassFact | 0.6643 | Interaction variable of length in meter and functional class increased the adjusted R-squared value but by very small number. Model with this interaction variable is good, but not that significant. |
| statusFact\*functionalClassFact | 0.6797 | Addition of interaction variable of Status and Functional Class is significant as the adjusted R-squared value is increased. |
| statusFact\*built | 0.6798 | This interaction variable is not that significant in the model, as the adjusted R-squared value is increased has a very less difference. However, this model till now has the highest adjusted R-squared value and can be further tested. |
| statusFact\*countyFact | 0.6878 | This is the best Adjusted R-square value. It can be said that interaction variable of status and county is significant. |

**The parsimonious model till now is:**

* Interaction of County and Design Construction
* Interaction of County and Year Built
* Interaction of Length in Meters and Functional Class
* Interaction of Status and Functional Class
* Interaction of Status and County
* Interaction of Status and Year Built

**Whereas, parsimonious model from stepAIC() is:**

* Status
* County
* Length in Meters
* Functional Class
* Design Construction
* Year Built
* Interaction of Status and County
* Interaction of Length in Meters and Functional Class
* Interaction of Status and Functional Class
* Interaction of County and Year Built

Call:

lm(formula = team$sufficiencyrating ~ statusFact + countyFact +

lengthinmtr + functionalClassFact + designconstructionFact +

built + statusFact:countyFact + lengthinmtr:functionalClassFact +

statusFact:functionalClassFact + countyFact:built, data = team)

AIC=5123.16

**After running the linear regression on above parsimonious model from stepAIC(), the Adjusted R-squared value is 0.6717.**

**Regression on stepAIC() model:**

> summary(A)

Call:

lm(formula = team$sufficiencyrating ~ statusFact + countyFact +

lengthinmtr + functionalClassFact + designconstructionFact +

built + statusFact:countyFact + lengthinmtr:functionalClassFact +

statusFact:functionalClassFact + countyFact:built, data = team)

Residuals:

Min 1Q Median 3Q Max

-35.034 -4.437 0.435 5.478 27.499

---

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

.

.

.

Residual standard error: 9.187 on 944 degrees of freedom

Multiple R-squared: 0.7228, Adjusted R-squared: 0.6717

F-statistic: 14.15 on 174 and 944 DF, p-value: < 2.2e-16

**This Adjusted R-squared value is less than the manually derived model.**

**Regression on Manually derived model:**

> summary(suffratingfit)

Call:

lm(formula = team$sufficiencyrating ~ (statusFact \* countyFact) +

(lengthinmtr \* functionalClassFact) + (statusFact \* functionalClassFact) +

(countyFact \* designconstructionFact) + (countyFact \* built) +

(statusFact \* built), data = team)

Residuals:

Min 1Q Median 3Q Max

-38.078 -3.856 0.000 4.908 24.862

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

Residual standard error: 8.959 on 827 degrees of freedom

Multiple R-squared: 0.769, Adjusted R-squared: 0.6878

F-statistic: 9.463 on 291 and 827 DF, p-value: < 2.2e-16

**Therefore, it can be said that manually derived model is better than the model derived from stepAIC().**

***Partial F-test***

Further the Partial F-test can be performed on the above derived model.

First, the above full model is compared with the partial model that does not contain interaction variable (Status \* Year Built), in Anova().

**Code:**

A<-lm(formula = team$sufficiencyrating ~ (statusFact\*countyFact)+(lengthinmtr\*functionalClassFact) + (statusFact\*functionalClassFact )+(countyFact\*designconstructionFact)+(countyFact\*built) + (statusFact\*built), data = team)

B<-lm(formula = team$sufficiencyrating ~ (statusFact\*countyFact)+(lengthinmtr\*functionalClassFact) + (statusFact\*functionalClassFact )+(countyFact\*designconstructionFact)+(countyFact\*built) , data = team)

anova(A,B)

Analysis of Variance Table

Model 1: team$sufficiencyrating ~ (statusFact \* countyFact) + (lengthinmtr \*

functionalClassFact) + (statusFact \* functionalClassFact) +

(countyFact \* designconstructionFact) + (countyFact \* built) +

(statusFact \* built)

Model 2: team$sufficiencyrating ~ (statusFact \* countyFact) + (lengthinmtr \*

functionalClassFact) + (statusFact \* functionalClassFact) +

(countyFact \* designconstructionFact) + (countyFact \* built)

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

1 827 66383

2 829 66694 -2 -310.25 1.9326 0.1454

Since, the p-value is 0.1454 which is greater than 0.05, we cannot reject the null hypothesis at 5% level of significance. This means that the interaction variable (Status \* Year Built) does not contribute significantly to decide the sufficiency rating of the bridge over the other variables in the model.

Partial F-test is performed further on the interaction variable (county \* Design construction).

**Code:**

A<-lm(formula = team$sufficiencyrating ~ (statusFact\*countyFact)+(lengthinmtr\*functionalClassFact) + (statusFact\*functionalClassFact )+(countyFact\*designconstructionFact)+(countyFact\*built) + (statusFact\*built), data = team)

C<-lm(formula = team$sufficiencyrating ~ (statusFact\*countyFact)+(lengthinmtr\*functionalClassFact) + (statusFact\*functionalClassFact )+(countyFact\*built) , data = team)

anova(A,C)

Analysis of Variance Table

Model 1: team$sufficiencyrating ~ (statusFact \* countyFact) + (lengthinmtr \*

functionalClassFact) + (statusFact \* functionalClassFact) +

(countyFact \* designconstructionFact) + (countyFact \* built) +

(statusFact \* built)

Model 2: team$sufficiencyrating ~ (statusFact \* countyFact) + (lengthinmtr \*

functionalClassFact) + (statusFact \* functionalClassFact) +

(countyFact \* built)

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

1 827 66383

2 961 88694 -134 -22311 2.0742 6.905e-10 \*\*\*

---

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

Since, the p-value is less than 0.05, we reject the null hypothesis at 5 % level of significance. This means that the interaction variable (county \* Design Construction) is contributing significantly to decide the sufficiency rating of the bridge. Therefore, most parsimonious model till now is -

team$sufficiencyrating ~ (statusFact \* countyFact) + (lengthinmtr \*

functionalClassFact) + (statusFact \* functionalClassFact) +

(countyFact \* designconstructionFact) + (countyFact \* built) +

(statusFact \* built)

***Conclusion***

***Plan***:

We performed the liner regression techniques, plotting of scatter plots, histograms to understand the nature of variables and which of the variables has more impact on the sufficiency rating and also checked for collinearity to select the variables from the dataset for the initial model. Later we removed the insignificant variable out of the model and performed variable interaction to see which variable were significant enough to be used in the parsimonious model.

***Results:***

After applying all the regression techniques, we observed that variables **Built, design construction fact, length in meter, status Fact** are highly significant along with the interaction variable (**county \* Design Construction**).

***Limitations of study and conclusion:***

The analysis applied to this data set is limited to specific structural type and construction method of the bridge which are mentioned in our data set. For instance, this method of analysis can’t be used for other bridges with different type of material like suspension bridge, cable bridge or temporary bridges with other method of constructions.

Also, we tried predicting the sufficiency rating of bridges but each bridge data was unique and didn’t have data over the years which constrained our prediction analysis.