Q2 SLR: station_precipitation

2024-12-06

Question 2

Exam Question:

In regional precipitation-frequency analysis, predicting the mean annual maximum precipitation at study sites using seasonal total precipitation as a predictor is a common approach. This allows researchers to estimate mean annual maximum precipitation in locations where seasonal total precipitation can be estimated.

PRISM climate normals provide spatially continuous maps of long-term averages for meteorological variables such as temperature and precipitation. Using PRISM as the basis for regression analysis enables predictions to be made almost anywhere. For this exercise, you will work with two variables from the dataset: seasonal total precipitation (predictor) and mean annual maximum precipitation (response).

- Perform a linear regression to evaluate whether seasonal total precipitation is a significant predictor of mean annual maximum precipitation.
- Evaluate the model's assumptions, including normality of residuals and homoscedasticity. If any assumptions appear to be violated, explore possible reasons for the violations and suggest ways to address them.
- Investigate whether any influential observations might be affecting the model fit by calculating and interpreting diagnostic metrics (e.g., Cook's Distance). Use these findings to propose or implement improvements to the model.
- Summarize your analysis in a clear, concise paragraph, including your regression results, assumption checks, and any modifications made to the model.

```
data <- read.csv('Q2-SLR-station precipitation.csv')
head(data)
```

```
##          name   n      l_1        t       t_3       t_4 Latitude Longitude
## 1 US1ORBN0013  85 4.949470 0.145362  0.149513  0.164376  44.6383 -123.5772
## 2 US1ORDG0044  44 5.656320 0.158076 -0.011484  0.028511  43.6918 -124.1246
## 3 US1ORLA0088  66 5.326953 0.130497  0.132726  0.121872  43.9087 -124.0884
## 4 US1ORLC0002  23 5.646357 0.198023  0.308209  0.142237  44.6210 -123.9370
## 5 US1ORLC0013 103 4.788742 0.182743  0.166362  0.116272  44.6773 -124.0592
## 6 US1ORTL0004  17 6.838586 0.181588 -0.161657 -0.010757  45.7235 -123.9391
##          Station_Na pm_wnt_ppt pm_wnt_tmp pm_elev  DISTCOAST RFA_Region
## 1      BLODGETT 1 N    1007.43     5.5800     252 33.7814441      ABCD1
## 2  REEDSPORT 0.8 SW    1115.48     8.0750      20  5.0414342      ABCD1
## 3    FLORENCE 5.4 S    1057.24     7.8875      43  5.6614377      ABCD1
## 4      TOLEDO 0.2 W    1009.03     7.7275      53  5.8014491      ABCD1
## 5     NEWPORT 4.2 N    1025.60     8.2275      46  0.1890756      ABCD1
## 6 MANZANITA 0.5 NNW    1279.41     7.8125      14  0.5914717      ABCD1
##   orig_reg
## 1        B
## 2        C
```

```
## 3        C
## 4        B
## 5        B
## 6        B
```

Set up regression

```
mod1 <- lm(l_1 ~ pm_wnt_ppt, data)
summary(mod1)
```

```
##
## Call:
## lm(formula = l_1 ~ pm_wnt_ppt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5414 -0.3310  0.0015  0.2888  2.0889
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 0.73793839 0.08262761   8.931 <0.0000000000000002 ***
## pm_wnt_ppt  0.00435654 0.00008177  53.275 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5558 on 293 degrees of freedom
## Multiple R-squared:  0.9064, Adjusted R-squared:  0.9061
## F-statistic:  2838 on 1 and 293 DF,  p-value: < 0.00000000000000022
```

Get standardized residuals

```
mod1.res <- rstandard(mod1)
shapiro.test(mod1.res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1.res
## W = 0.91601, p-value = 0.000000000008453
```

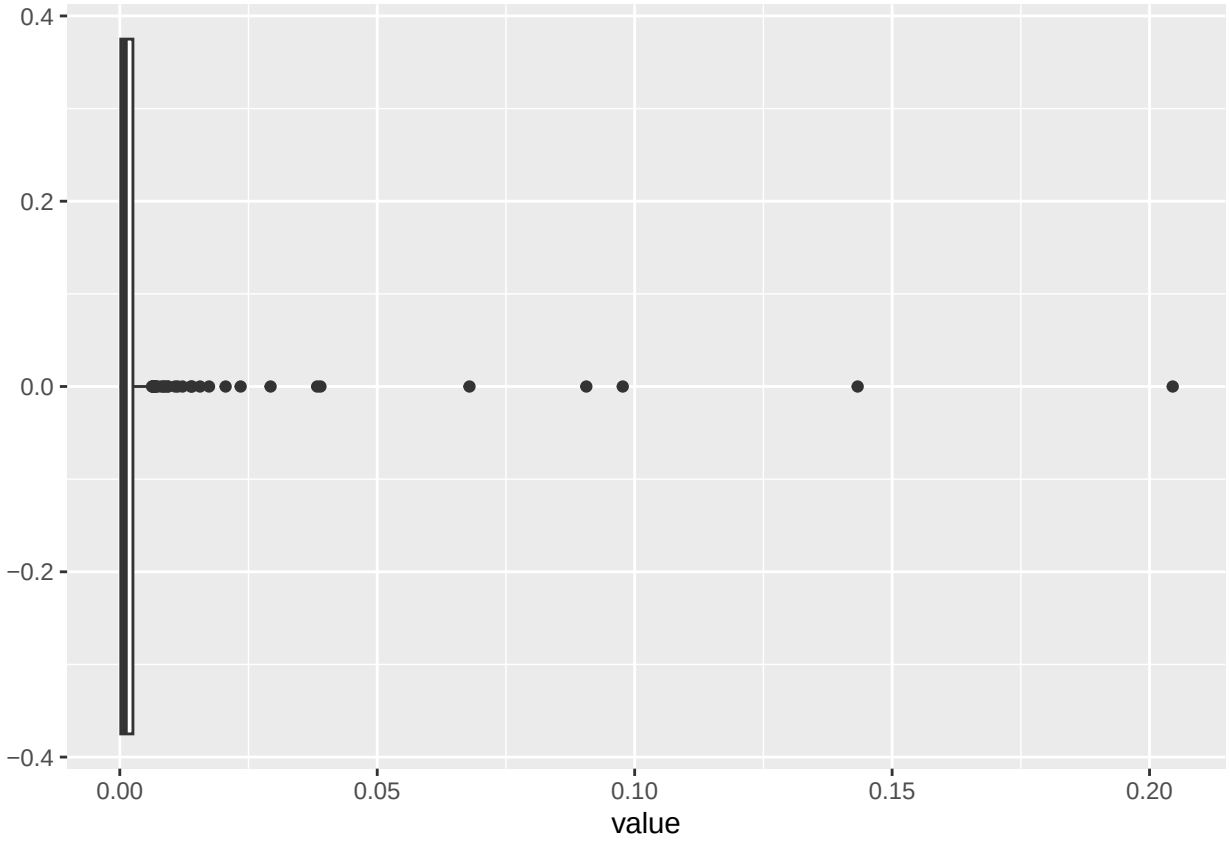
The Breusch-Pagan test to assess homoscedasticity

```
ncvTest(mod1)
```

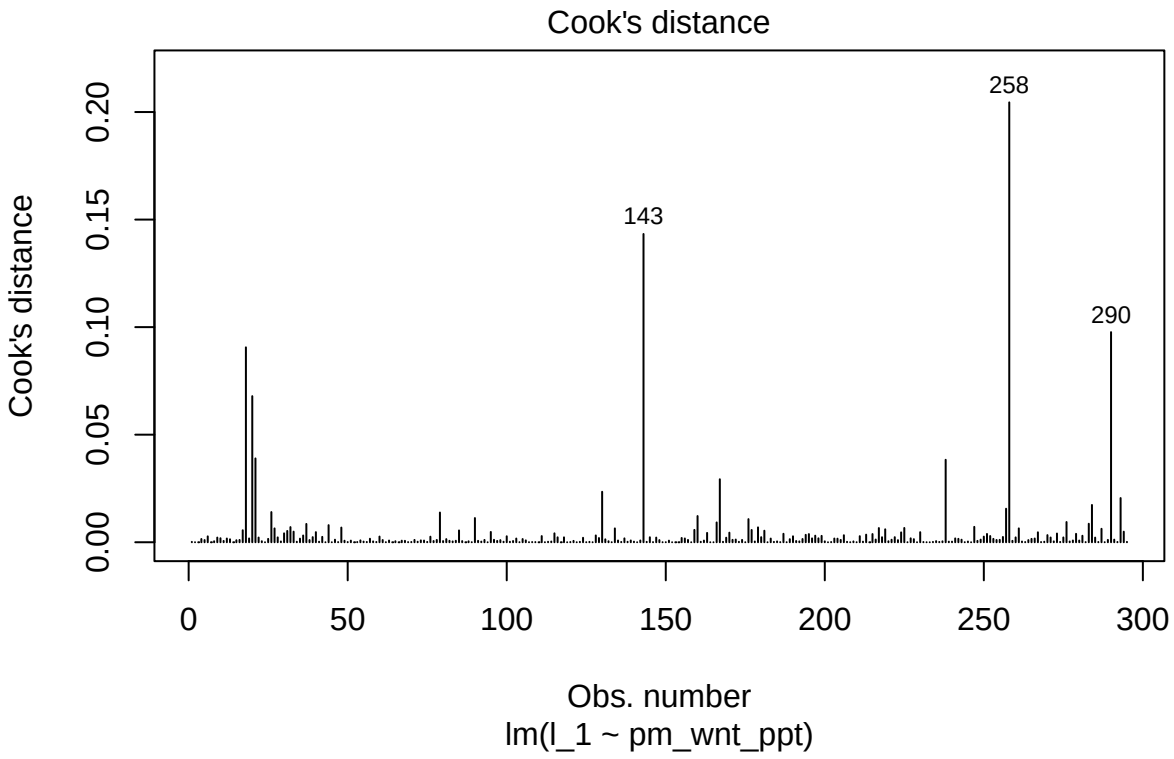
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 47.57819, Df = 1, p = 0.0000000000052853
```

#Cooks D

```
ggplot(as_tibble(cooks.distance(mod1)), aes(value)) + geom_boxplot()
```

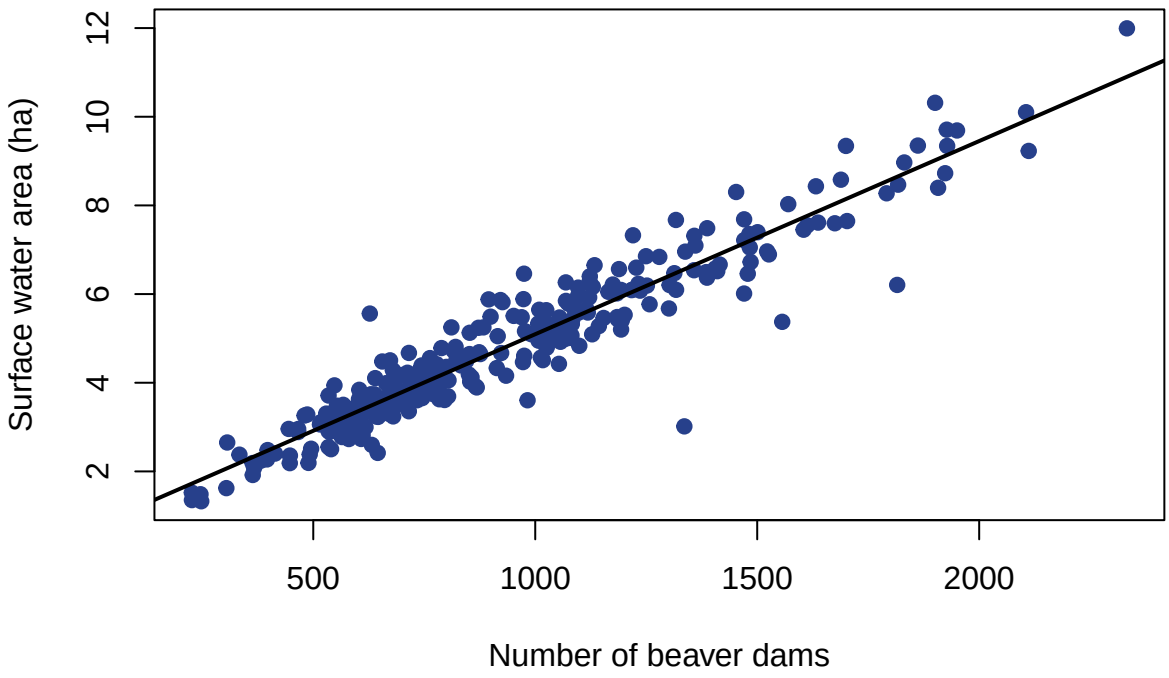



```
plot(mod1,which=4)
```



Make plot of beaver dams and surface water

```
plot(l_1 ~ pm_wnt_ppt, data,
     pch = 19,
     col = "royalblue4",
     ylab = "Surface water area (ha)",
     xlab =  "Number of beaver dams")
#add regression line
#make line width thicker
abline(mod1, lwd=2)
```



The linear regression model indicates that seasonal total precipitation (pm_wnt_ppt) is a highly significant predictor of mean annual maximum precipitation ($p < 0.001$), with an adjusted R2R2 of 0.9061, suggesting the model explains a large portion of the variability in the response. However, diagnostic checks reveal violations of assumptions:

- Residual Normality: The Shapiro-Wilk test is significant ($W=0.916, p<0.001$), indicating non-normal residuals.
- Heteroscedasticity: The non-constant variance test is significant ($p<0.001$), showing heteroscedasticity.
- Influential Points: Cook's Distance analysis suggests potential influence from observations 143, 258, and 290, which warrant further investigation.

These issues suggest the need for model improvements, such as addressing outliers or transforming variables.

References:

- https://www.hec.usace.army.mil/confluence/sspdocs/ssptutorialsguides/r-based-statistics-tutorials/linear-regression-using-r