Homework 2

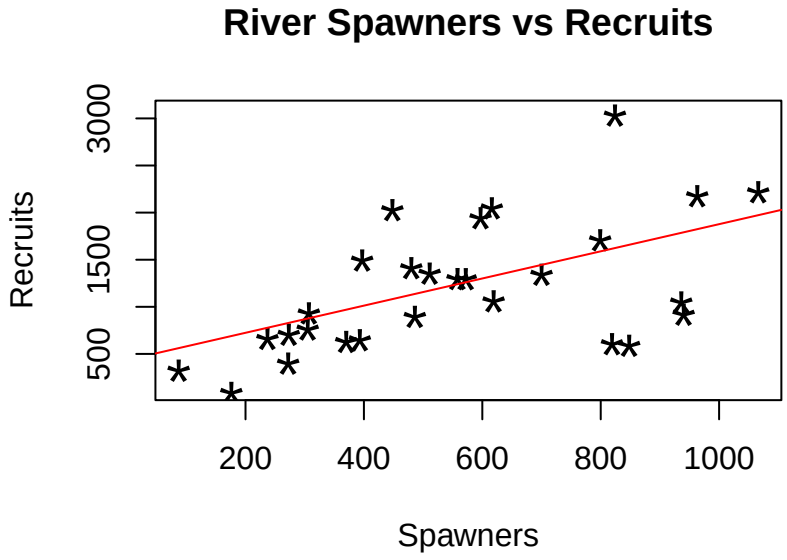
Skylar Liu

2024-01-20

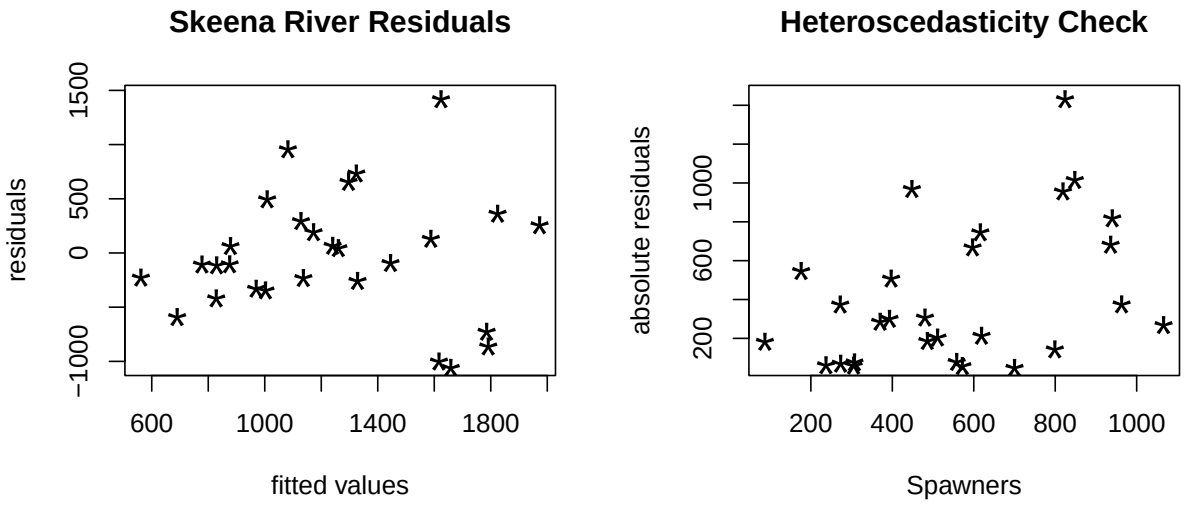
Homework #2, Stat 660, Spring 2024, Due Class #3, January 24

1. Perform a linear regression of Y on X.

```
# clear workspace
rm(list = ls())
# set the seed
set.seed(382957)

# import data and set X and Y variables
river = read.csv("~/660 - Flexible Regression/Homework/Homework2/Skeena_River.csv")
X = river$Spawners
Y = river$Recruits

# linear regression of Y on X
lin_recruits = lm(Y ~ X)

# plot X vs Y
plot(X, Y, type="p",pch='*',main="River Spawners vs Recruits",
     xlab="Spawners",ylab="Recruits",cex=2)
abline(lin_recruits, col="red")
```



```r
# plot the residuals
par(mfrow=c(1,2))
plot(fitted(lin_recruits),residuals(lin_recruits),xlab="fitted values",
     ylab="residuals",pch='*',cex=2,main="Skeena River Residuals")
plot(X,abs(residuals(lin_recruits)),xlab="Spawners",
     ylab="absolute residuals",pch='*',cex=2,main='Heteroscedasticity Check')
```



a. Do the residuals show heteroscedasticity?

Yes, the residuals appear to show heteroscedasticity.

b. How did you decide?

I first plotted the fitted values vs. their residuals. As the fitted values increase, the variance of the residuals also increase showing an unequal scatter of the residuals. I then plotted the absolute residuals which increase as X increases, showing an increase in variance.

2. Is there a statistically significant relationship between Y and X, ignoring possible heteroscedasticity?'

```r
summary(lin_recruits)
```

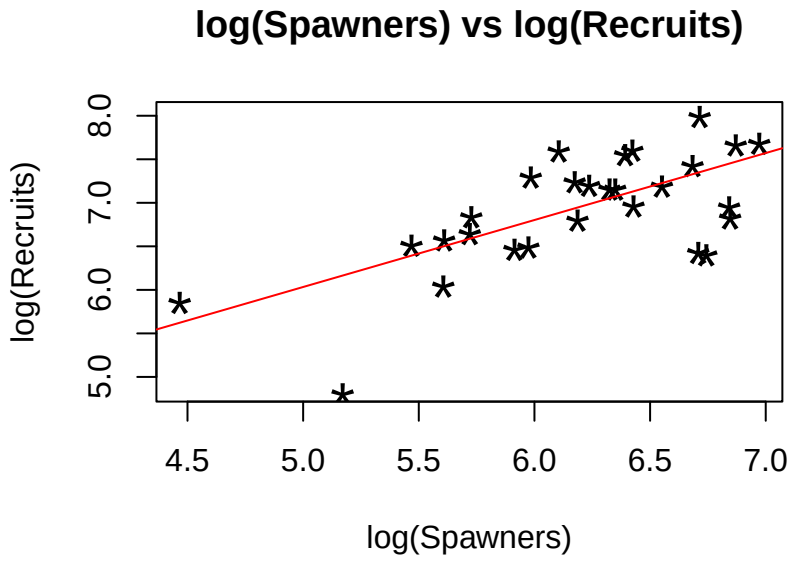
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1031.4  -305.2   -69.9   294.0  1447.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 436.0809   260.2775   1.675  0.10583
## X             1.4414     0.4233   3.405  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 582.8 on 26 degrees of freedom
## Multiple R-squared:  0.3084, Adjusted R-squared:  0.2818
## F-statistic:  11.6 on 1 and 26 DF,  p-value: 0.002156
```

The p-value for X on Y is 0.002156, which is statistically significant at the 0.05 level.

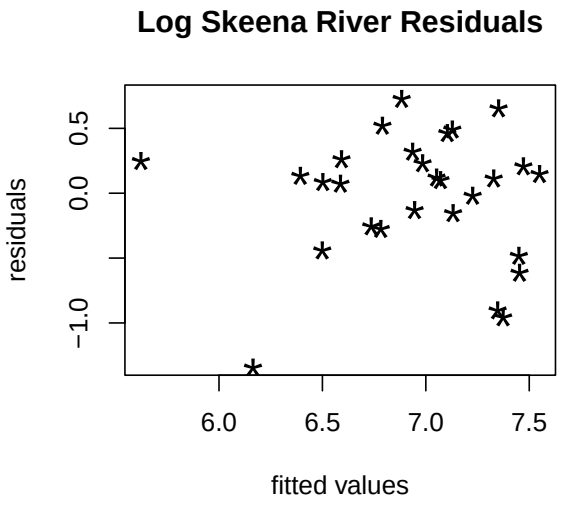
3. Try regressing log(Y) on log(X). Which points are obviously 1951 and 1955? It is hard to tell which is which, but the two are obvious, to me (but I know what they are, so I am an oracle in this one).

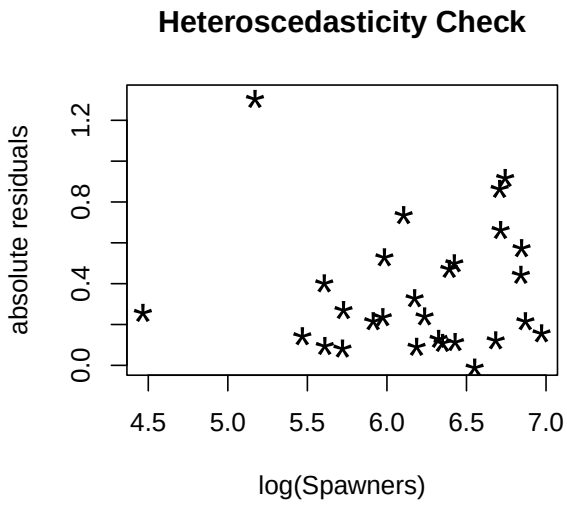
```
# linear regression of log(Y) on log(X)
log_recruits = lm(log(Y) ~ log(X))

plot(log(X), log(Y), type="p",pch='*',main="log(Spawners) vs log(Recruits)",
     xlab="log(Spawners)",ylab="log(Recruits)",cex=2)
abline(log_recruits, col="red")
```



The 1951 and 1955 points appear to be the two outliers in the bottom left quadrant, since they had the smallest X and Y values from the data.

```
# plot the residuals
par(mfrow=c(1,2))
plot(fitted(log_recruits),residuals(log_recruits),xlab="fitted values",
     ylab="residuals",pch='*',cex=2,main="Log Skeena River Residuals")
plot(log(X),abs(residuals(log_recruits)),xlab="log(Spawners)",
     ylab="absolute residuals",pch='*',cex=2,main='Heteroscedasticity Check')
```



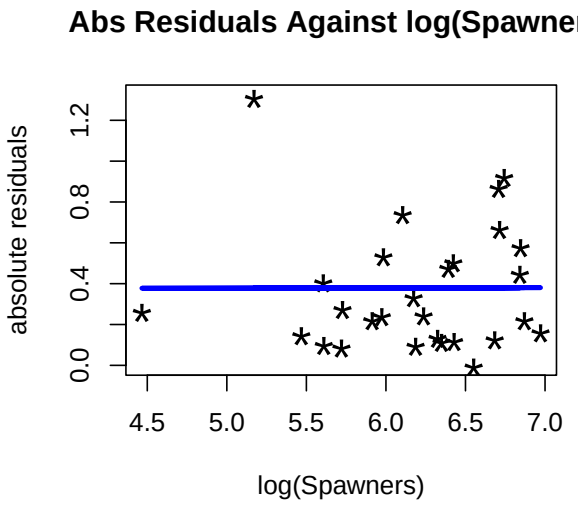


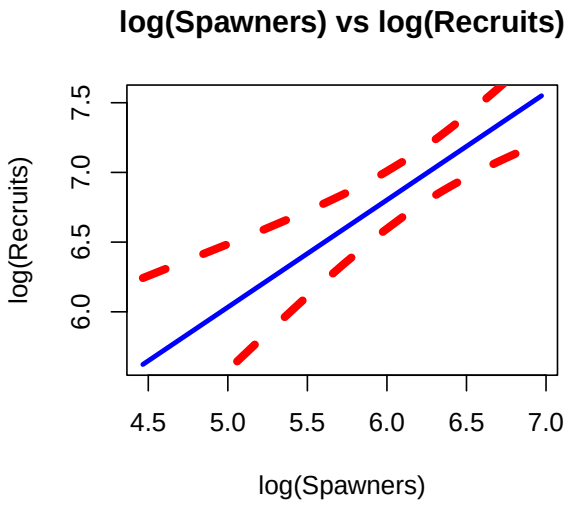
4. Does the log-log regression show heteroscedasticity?

   When including the 1951 and 1955 points, heteroscedasticity is questionable since these outliers make the variance appear to be relatively constant. Removing these two point would create heteroscedasticity, since variance would then increase as X increases.

5. Plot the fitted lines of the log-log regression of the absolute residuals against log(X).

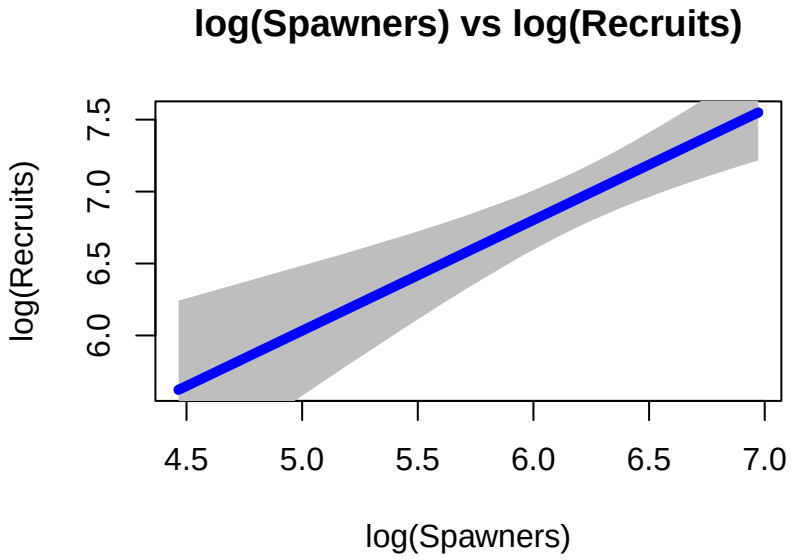
```
# log-log regression of the absolute residuals against log(X)
abs_recruits = lm(abs(residuals(log_recruits))~log(X))

# residual plot
par(mfrow=c(1,2))
plot(log(X),abs(residuals(log_recruits)),xlab="log(Spawners)",
     ylab="absolute residuals",pch='*',cex=2,main='Abs Residuals Against log(Spawners)')
# add fitted lines
lines(log(X),fitted(abs_recruits),lwd=3,col="blue")

# confidence interval plot
n = length(X)
ord = order(X)
x = X[ord]
y = Y[ord]
pred = predict(lm(log(y)~log(x)),newdata=as.data.frame(log(X)),se.fit=TRUE)
plot(log(x),pred$fit,type="l",col="blue",lwd=3,xlab="log(Spawners)",ylab="log(Recruits)",
     main="log(Spawners) vs log(Recruits)")
t = qt(0.975,df=n-2)
lines(log(x),pred$fit+t*pred$se.fit,lty=2,lwd=5,col="red")
lines(log(x),pred$fit-t*pred$se.fit,lty=2,lwd=5,col="red")
```





6. On page 36 of the book, there is a command polygon, which allows you to form a confidence interval which is shaded, unlike what I did in class. Give it a try, and report your plot. Do not obsess about it: we will do lots of these things, and the instructions are deliberately (on my part) vague.

```
# shaded confidence interval plot
plot(log(x),pred$fit,type="l",col="blue",lwd=5,xlab="log(Spawners)",ylab="log(Recruits)",
     main="log(Spawners) vs log(Recruits)")
t = qt(0.975,df=n-2)
upperCI = pred$fit+t*pred$se.fit
lowerCI = pred$fit-t*pred$se.fit
polygon(x=c(log(x), rev(log(x))), y=c(upperCI, rev(lowerCI)), col="gray", border=NA)
lines(log(x),pred$fit,col="blue",lwd=5)
```



7. Please tell me what you think about the line plots of CI versus the polygon plots of CI. No wrong answer. I find the line plots look better in published papers, but if you were presenting a CI for your job, what would you choose, and why? Remember, what looks great on a computer screen may well not look so great in a document, or an overhead projector, so your answer should be based on what You would use.

   I would prefer to use the line plot of the CI. To me, the dashed lines present the idea that the confidence interval is not 100% certain like the shaded region suggests. Although the polygon plot CI may look better in a presentation, I think it would be more important to ensure that the audience does not assumem 100% certainty.