Rlab 3: Simple Linear Models

Eunjin Park

Oct 12th, 2022

DELETE ANYTHING FROM THIS TEMPLATE BELOW THAT IS NOT PART OF YOUR SOLUTION.

I. Simple Inference

(1) [4 pts] Which test can be used to check whether the total expected number of lobsters inside MPAs is different from the number outside MPAs under the assumption that the numbers of lobsters are approximately normal and the variances are the same inside and outside MPAs?

Which test can be used to check whether the expected number of lobsters inside MPAs is different from the number outside MPAs assuming only that the variances are the same inside and outside the MPA?

(a) t-test with equal variances (b) Welch test (c) Wilcoxon Mann-Whitney test

-> (a)

```
Lob = read.csv("6dd93320.csv")
t.test(Lob$Lob_total[Lob$Inside_Outside == "Inside"], var.equal = TRUE)
```

```
##
##  One Sample t-test
##
## data:  Lob$Lob_total[Lob$Inside_Outside == "Inside"]
## t = 7.7666, df = 81, p-value = 2.162e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   5.923308 10.003521
## sample estimates:
## mean of x
##  7.963415
```

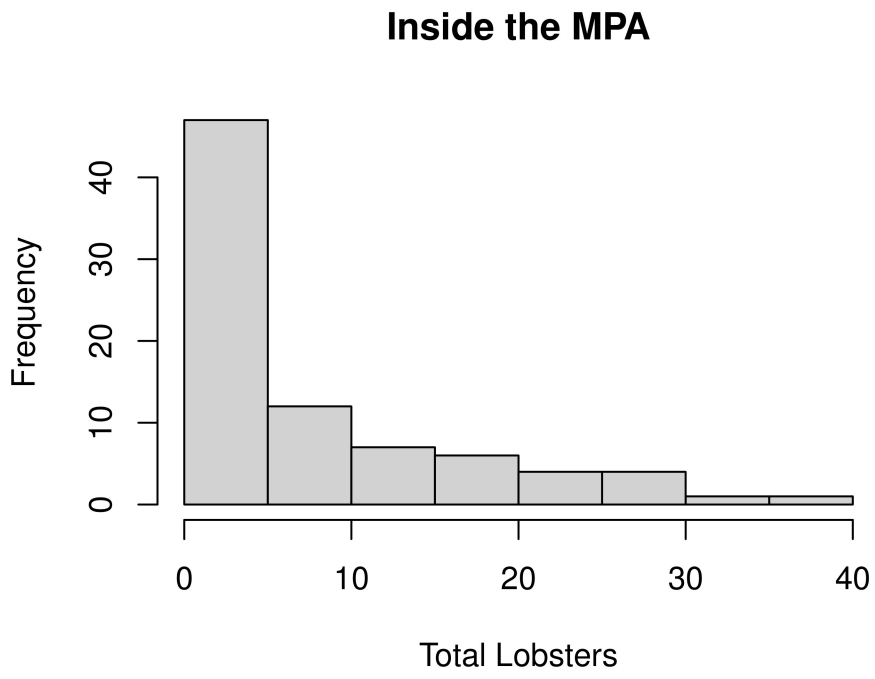
```
t.test(Lob$Lob_total[Lob$Inside_Outside == "Outside"], var.equal = TRUE)
```

```
##
##  One Sample t-test
##
## data:  Lob$Lob_total[Lob$Inside_Outside == "Outside"]
## t = 6.3628, df = 79, p-value = 1.2e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

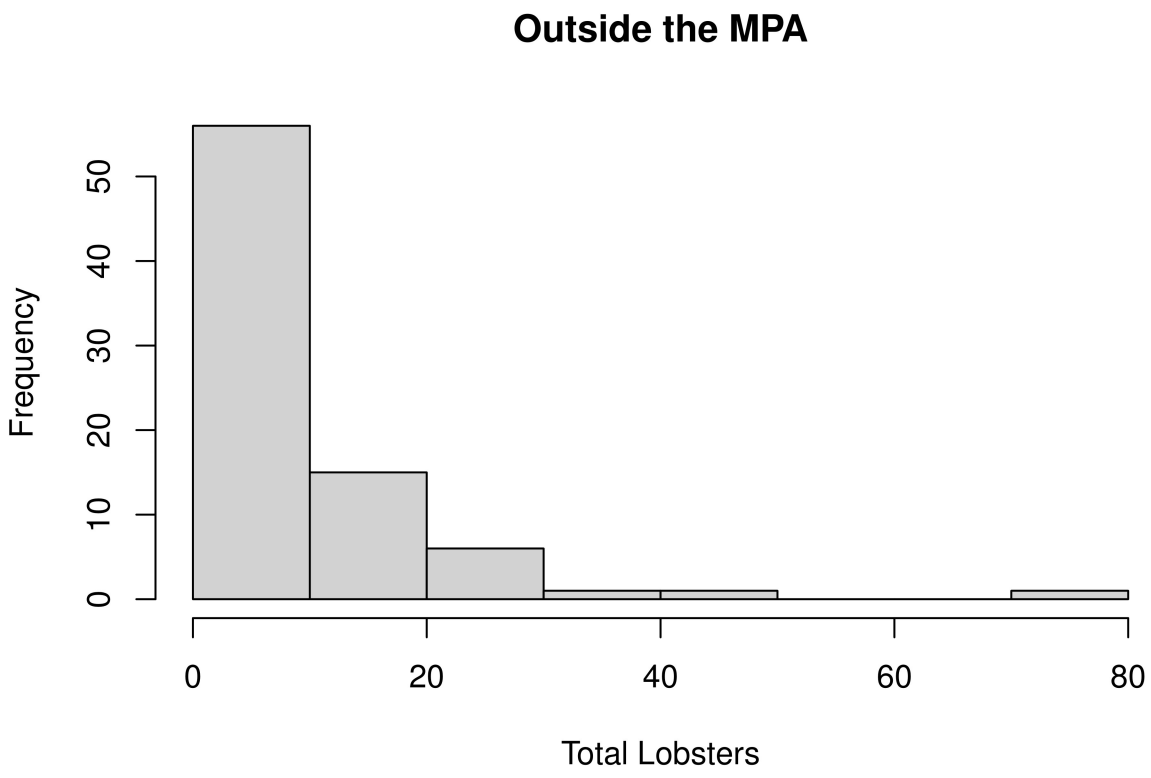
```
##   5.737893 10.962107
## sample estimates:
## mean of x
##      8.35
```

(2) [3 pts] Create histograms of the counts of lobsters inside and outside the MPA. If you put your histograms on separate plots, make sure to match the x and y axes so they are comparable. Based on your histograms, do you think the assumption of normality is reasonable? Why/why not? Do you think the assumption of equal variances is reasonable? Why/why not

```
hist(Lob$Lob_total[Lob$Inside_Outside == "Inside"], xlab = "Total Lobsters", main = "Inside the MPA")
```

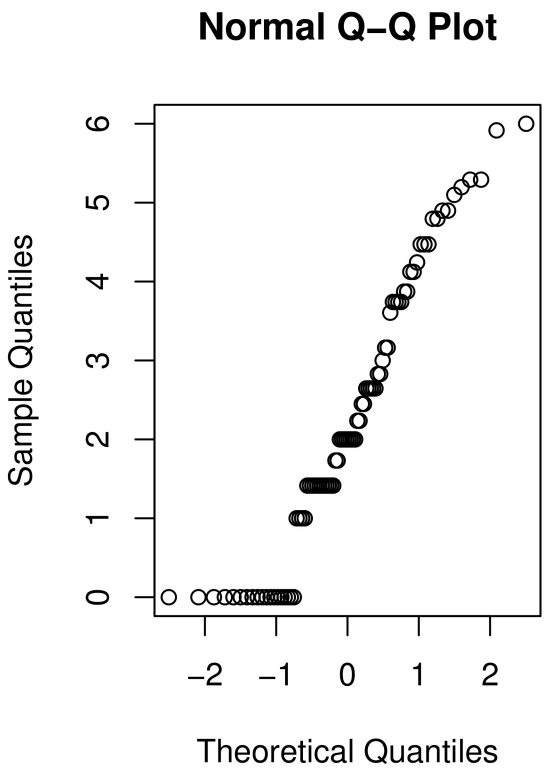


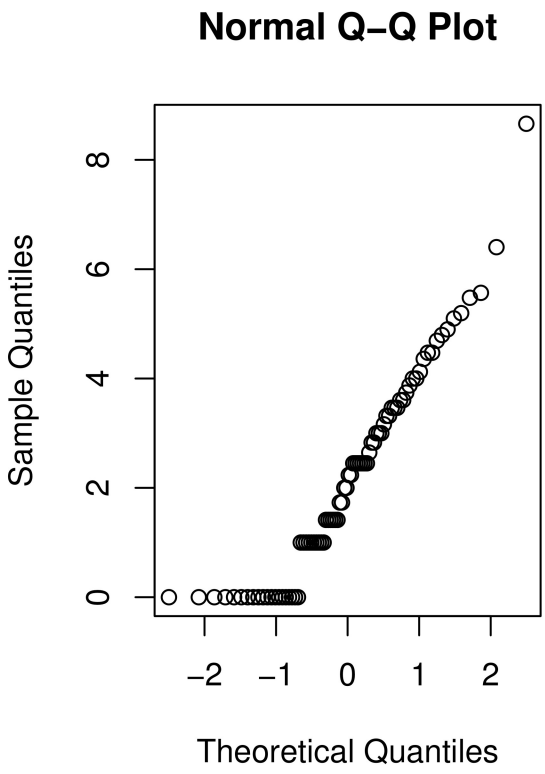

```
hist(Lob$Lob_total[Lob$Inside_Outside == "Outside"], xlab = "Total Lobsters", main = "Outside the MPA")
```



(3)

```
par(mfrow=c(1,2))
qqnorm(sqrt(Lob$Lob_total[Lob$Inside_Outside == "Inside"]))
qqnorm(sqrt(Lob$Lob_total[Lob$Inside_Outside == "Outside"]))
```





(4) [3 pts] Repeat your test from (1) on the transformed counts of lobsters and report your p-value (PEC).

```
t.test(sqrt(Lob$Lob_total[Lob$Inside_Outside == "Inside"]),var.equal = F)
```

```
##
##  One Sample t-test
##
## data:  sqrt(Lob$Lob_total[Lob$Inside_Outside == "Inside"])
## t = 11.532, df = 81, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.840850 2.608497
## sample estimates:
## mean of x
##  2.224674
```

```
t.test(sqrt(Lob$Lob_total[Lob$Inside_Outside == "Outside"]),var.equal = F)
```

```
##
##  One Sample t-test
##
## data:  sqrt(Lob$Lob_total[Lob$Inside_Outside == "Outside"])
## t = 10.451, df = 79, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
##  1.782009 2.620471
## sample estimates:
## mean of x
##   2.20124
```

(5) [4 pts] Which test can be used to check whether the expected number of lobsters inside MPAs is different from the number outside MPAs assuming only that the variances are the same inside and outside the MPA?

(a) t-test with equal variances (b) Welch test (c) Wilcoxon Mann-Whitney test

->(B)

Use R to conduct your chosen test on the un-transformed counts of lobsters and report the p-value (PEC).

```
t.test(sqrt(Lob$Lob_total[Lob$Inside_Outside == "Inside"]),
       sqrt(Lob$Lob_total[Lob$Inside_Outside == "Outside"]),
       var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  sqrt(Lob$Lob_total[Lob$Inside_Outside == "Inside"]) and sqrt(Lob$Lob_total[Lob$Inside_Outside
## t = 0.082048, df = 158.41, p-value = 0.9347
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5406647  0.5875325
## sample estimates:
## mean of x mean of y
##  2.224674  2.201240
```

II. Simple Linear regression

(6) [3 pts] PEC that fits a simple linear regression model with the total number of observed lobsters as the response and the percent cover of Surfgrass as the predictor. Give a 95% confidence interval for the effect of Surfgrass.

```
fit = lm(Lob_total ~ Surfgrass, data = Lob)
summary(fit)
```

```
##
## Call:
## lm(formula = Lob_total ~ Surfgrass, data = Lob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.248  -7.355  -3.855   2.645  57.011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.35514    0.85323   8.620 6.16e-15 ***
## Surfgrass    0.14820    0.05039   2.941  0.00375 **
```

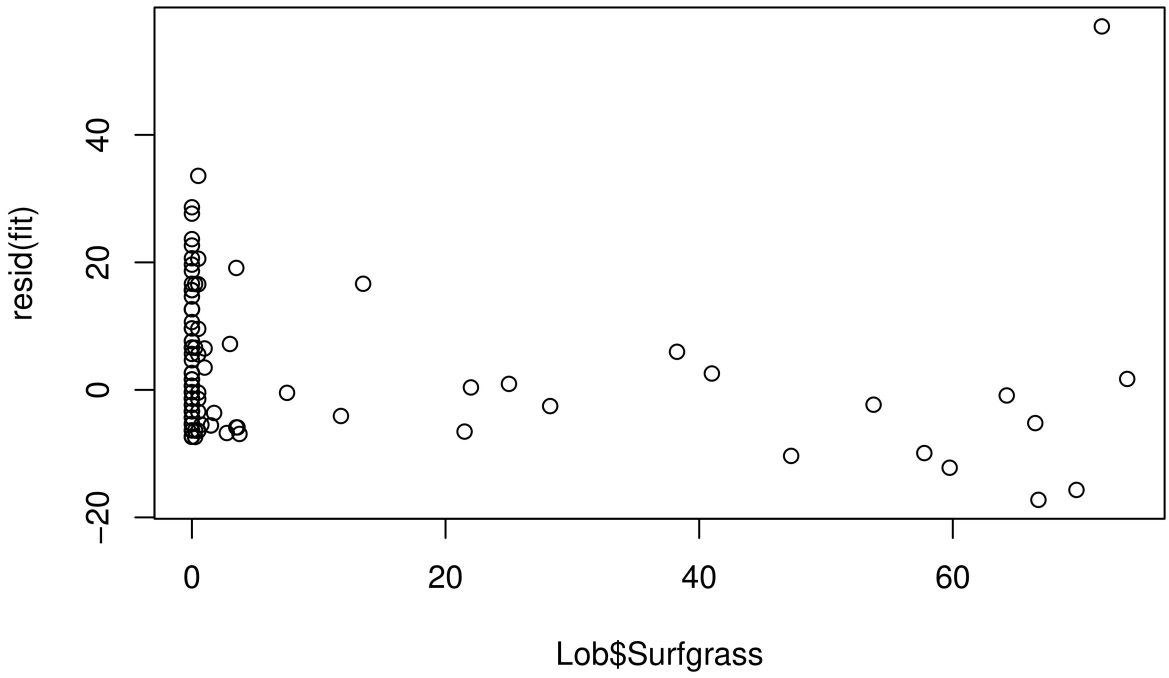
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 160 degrees of freedom
## Multiple R-squared:  0.0513, Adjusted R-squared:  0.04537
## F-statistic: 8.651 on 1 and 160 DF,  p-value: 0.003753
```

```
confint(fit, 'Surfgrass', level=0.95)
```

```
##                2.5 %    97.5 %
## Surfgrass 0.04869419 0.2477065
```

(7) [3 pts] Make a plot of the residuals as a function of percent cover of Surfgrass (PEC). Do you see evidence of heteroskedasticity? Explain your reasoning.

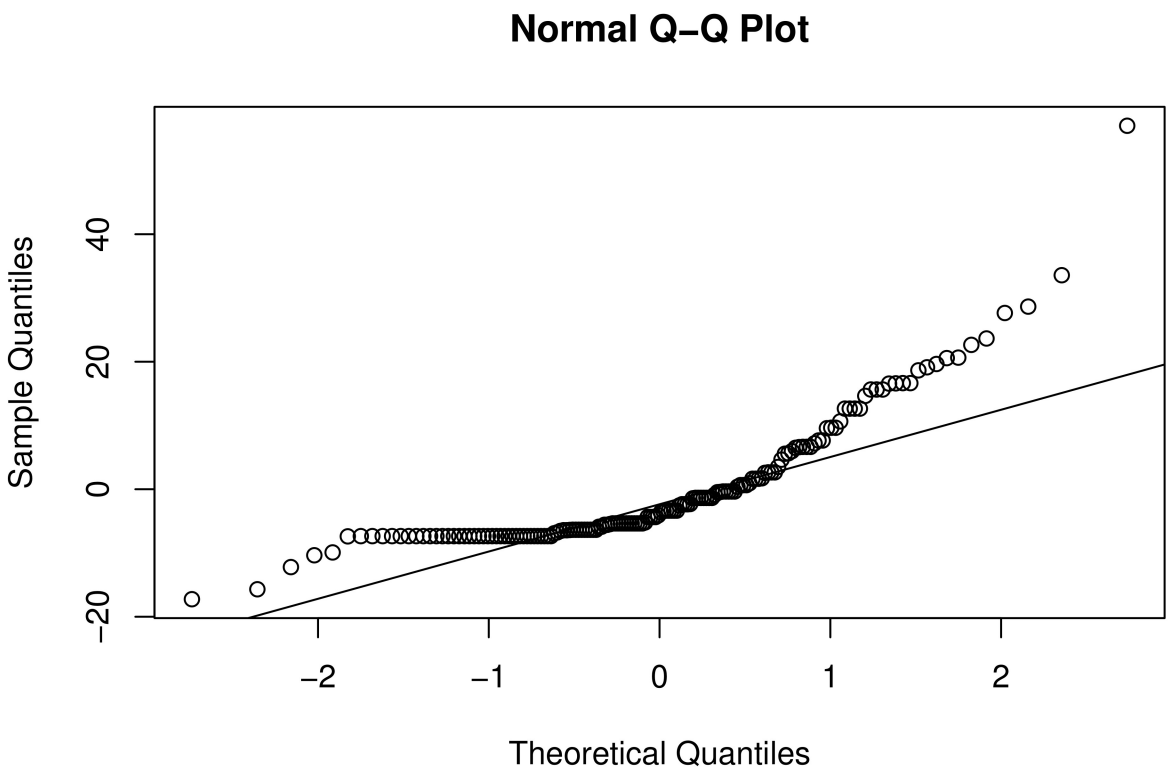
```
plot(resid(fit) ~Lob$Surfgrass)
```



-> Yes it is heterokedasticity pretty much. Since residuals are very small for low Lob$Surfgrass while there is great variation in the si

(8) [3 pts] Make a QQ-plot of the residuals (PEC). Do you see evidence of non-normal residuals? Explain your reasoning.

```
qqnorm(resid(fit))
qqline(resid(fit))
```



(9) [3 pts] Setting aside any potential issues related to the assumptions for linear regression, use your model for total lobster counts as a function of Surfgrass cover to make a 95% confidence interval for the expected number of lobsters observed at sites with 50% Surfgrass cover (PEC).

```
fit = lm(Lob_total ~ Surfgrass, data = Lob)
predict(fit, newdata = data.frame(Surfgrass=50), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 14.76515 10.04778 19.48253
```

(10) [4 pts] Fit another simple regression model with the Inside_Outside variable as the sole predictor (PEC). Compare the p-value for the effect of being Outside an MPA to your value from (1). What do you notice?

```
fitt = lm(Lob_total ~ Inside_Outside, data = Lob)
summary(fitt)
```

```
##
## Call:
## lm(formula = Lob_total ~ Inside_Outside, data = Lob)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.350 -7.350 -3.963  4.650 66.650
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.9634     1.1670   6.824 1.73e-10 ***
## Inside_OutsideOutside     0.3866     1.6606   0.233    0.816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.57 on 160 degrees of freedom
## Multiple R-squared:  0.0003386,  Adjusted R-squared:  -0.005909
## F-statistic: 0.05419 on 1 and 160 DF,  p-value: 0.8162
```

(11) [4 pts] Fit another simple regression model with MPA as the sole predictor. Use a parameterization where each regression coeffcient corresponds to the expected number of lobster observed at each MPA (PEC).Report 95% confidence intervals for the expected number of lobsters observed at all five MPAs. Do you notice anything concerning about any of the confidence intervals?

```
fit_MPA <- lm(Lob_total ~ MPA, data = Lob)
predict(fit_MPA, newdata = data.frame(MPA= c("Cabrillo State Marine Reserve", "Laguna Beach State Marin
```

```
##          fit       lwr       upr
## 1 10.3421053  7.035096 13.649114
## 2  6.6363636  3.087656 10.185071
## 3  0.6428571 -4.805469  6.091184
## 4  9.7631579  6.456149 13.070167
## 5  8.4358974  5.171562 11.700233
```

-> Regarding Cabrillo State Marine Reserve: 7.0351< < 13.649 Laguna Beach State Marine Reserve: 3.08766< < 10.1851 Point Vicente State Marine Conservation Area : -4.805< < 6.0911 South La Jolla State Marine Reserve: 6.456< < 13.070 Swami's State Marine Conservation Area: 5.1716< < 11.700

(12) [3 pts] How well are the assumptions of homoskedasticity, independence of residuals, and normality of residuals met? Explain your reasoning. You can choose to use figures as part of your explanation if it is helpful.

->A condition in which the variance of the residual, or error term, in a regression model is constant, so assumption of homoscedasticity is satisfied.