STAT 516 hw 1

Solutions

Chp 4 Ex 8

It is of interest to test H_0: $\mu \leq 129$ versus H_1: $\mu > 129$. The R code below computes the test statistic and the critical value at $\alpha = 0.01$ for testing these hypotheses.

```
bp <- c(115,134,131,143,130,154,119,137,155,130,110,138)
mu0 <- 129
n <- length(bp)
xbar <- mean(bp)
s <- sd(bp)
Tstat <- (xbar - mu0)/( s / sqrt(n))
alpha <- 0.01
tcrit <- qt(1-alpha,n - 1)
```

The value of the test statistic is $T_{\text{stat}} = 0.993901$, which does not exceed the critical value $t_{11,0.01} = 2.7180792$. So we fail to reject H_0. There is insufficient evidence to claim that the average blood pressure in this community exceeds 129.

The next chunk of R code compute the p-value.

```
pval <- 1 - pt(Tstat,n-1)
```

The p-value is 0.1708158, so we would fail to reject H_0 for all significance levels less than 0.1708158.

We can obtain these results with the `t.test()` function as follows:

```
t.test(bp,mu = mu0,alternative="greater")
```

```
    One Sample t-test
```

```
data:  bp
t = 0.9939, df = 11, p-value = 0.1708
alternative hypothesis: true mean is greater than 129
95 percent confidence interval:
 125.7724      Inf
sample estimates:
mean of x
      133
```

Chp 4 Ex 19

The following R code reads in the data builds the 95% confidence interval.

```
hl <- c(2.50,2.20,1.60,1.30,
        1.20,1.60,2.20,2.20,
        2.60,1.00,1.50,3.15,
        1.44,1.26,1.98,1.98,
        1.87,2.31,1.4,
        2.48,2.80,0.69)

n <- length(hl)
xbar <- mean(hl)
s <- sd(hl)
alpha <- 0.05
lo95 <- xbar - qt(1-alpha/2,n-1) * s / sqrt(n)
up95 <- xbar + qt(1-alpha/2,n-1) * s / sqrt(n)
```

A 95% confidence interval for the true mean half life of Amikacin is $(1.5961643, 2.1547448)$.

This can also be obtained with the `t.test()` function.

```
t.test(hl)
```

```
    One Sample t-test

data:  hl
t = 13.965, df = 21, p-value = 4.237e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
 1.596164 2.154745
sample estimates:
mean of x
 1.875455
```

Chp 4 Ex 20

```
alpha <- 0.10
lo90 <- xbar - qt(1-alpha/2,n-1) * s / sqrt(n)
up90 <- xbar + qt(1-alpha/2,n-1) * s / sqrt(n)
```

A 90% confidence interval for the true mean half life of Amikacin is $(1.6443603, 2.1065488)$.

This can also be obtained with the `t.test()` function.

```
t.test(hl, conf.level = 0.90)
```

```

    One Sample t-test

data:  hl
t = 13.965, df = 21, p-value = 4.237e-12
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 1.644360 2.106549
sample estimates:
mean of x
 1.875455
```

Chp 5 Ex 10

The following R code reads in the data.

```
before <- c(12,16,10,17,12,15)
after <- c(15,16,15,18,14,17)
d <- after - before
```

a)

Let μ be the mean difference (after minus before) in the employees' self-ratings of their knowledge. We wish to test H_0: $\mu \leq 0$ versus H_1: $\mu > 0$.

```
alpha <- 0.05
mu0 <- 0
n <- length(d)
xbar <- mean(d)
s <- sd(d)
Tstat <- (xbar - mu0) / (s / sqrt(n))
pval <- 1 - pt(Tstat,n-1)
```

The p-value is 0.0137146, which is quite small (less than the commonly used significance level $\alpha = 0.05$), so there is fairly strong evidence that the employees believe they have gained knowledge from the seminar.

b)

It could be that the employees who did not return a follow-up rating did so because they did not learn anything from the seminar; in this case, if they had reported their follow-up ratings, the complete data may *not* have carried as much evidence in favor of the effectiveness of the seminar. On the other hand, if those who did not return a follow-up rating learned more from the seminar than those who did return a follow-up rating, the complete data would have carried even stronger evidence in favor of the seminar. When data is missing like this, one must investigate whether the missing values would have contributed in such a way as to change the outcome of the study.

Chp 7 Ex 7

The following R code reads in the data set.

```
heatcost <- read.table(file = "Data Tables 4th edition/Chapter 7/datatab_7_19.prn",
                       header = TRUE)
head(heatcost)
```

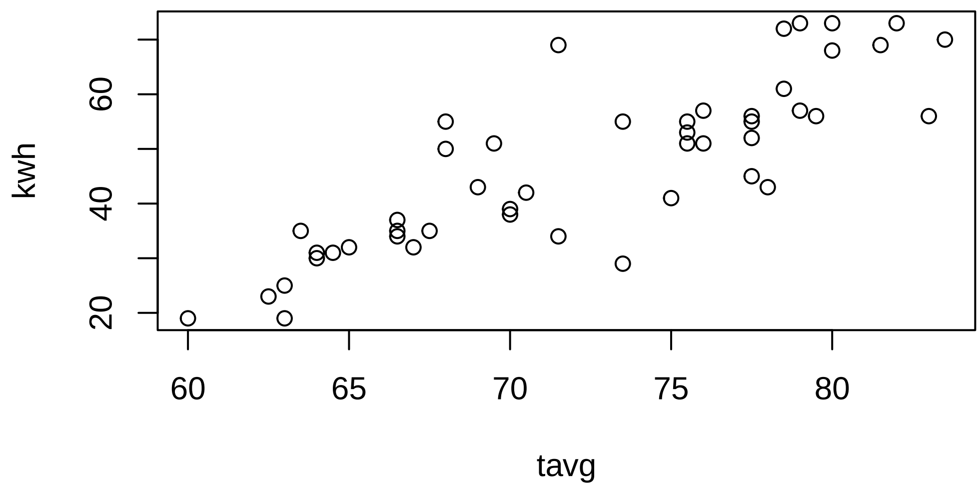
```
  mo day tavg kwh
1  9  19 77.5  45
2  9  20 80.0  73
3  9  21 78.0  43
```

```
4   9   22 78.5   61
5   9   23 77.5   52
6   9   24 83.0   56
```

a)

The following R code makes a scatterplot of power consumption versus temperature.

```
plot(kwh ~ tavg, data = heatcost)
```

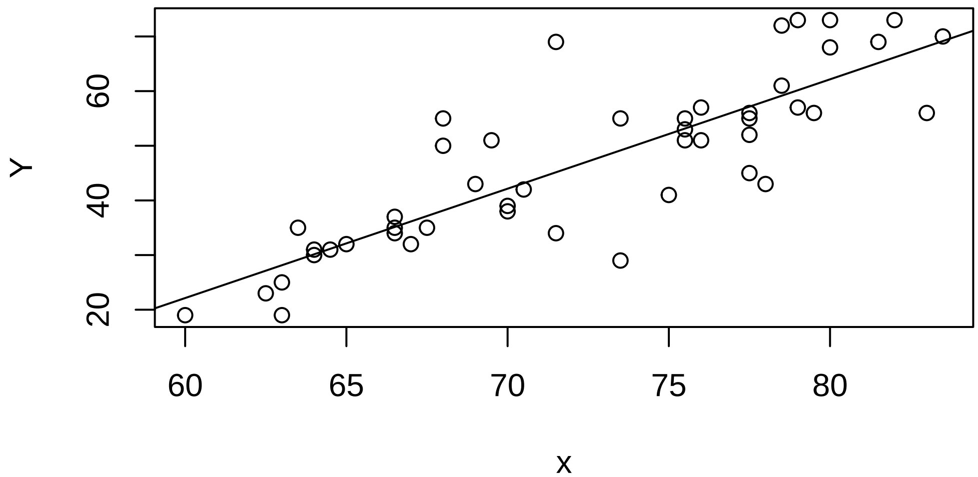


b)

We now fit a simple linear regression model to estimate the linear effect of the temperature on power consumption.

```
x <- heatcost$tavg
Y <- heatcost$kwh
b1 <- cor(x,Y) * sd(Y) / sd(x)
b0 <- mean(Y) - b1 * mean(x)

plot(Y~x)
abline(b0,b1)
```



The estimated regression function is

$$\text{Kwh} = -97.9238928 + 2.0010054 \cdot \text{Tavg},$$

according to which an increase in the temperature by 1 degree is associated with an increase in power consumption by 2.0010054 kilowatt hours. Of course, this is only an estimate.

The following code constructs a 95% confidence interval for the slope parameter β_1:

```
# estimate sigma
n <- length(Y)
Yhat <- b0 + x * b1
ehat <- Y - Yhat
sghat <- sqrt(sum(ehat^2) / (n-2))

# construct confidence interval
alpha <- 0.05
Sxx <- sum((x - mean(x))^2)
tval <- qt(1-alpha/2,n-2)
lo <- b1 - tval * sghat / sqrt(Sxx)
up <- b1 + tval * sghat / sqrt(Sxx)
```

The 95% confidence interval for β_1 is $(1.6170855, 2.3849253)$.

To check whether the temperature has a linear effect on power consumption, we would test the hypothesis H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. The following R code computes the p-value for testing this set of hypotheses.

```r
Tstat <- b1 / (sghat / sqrt(Sxx))
pval <- 2*(1 - pt(abs(Tstat),n-2))
```

We obtain the p-value is 1.110223×10^{-13}.

All of these results can be obtained using the `lm()` function, as shown below:

```r
lm_out <- lm(kwh ~ tavg, data = heatcost)
summary(lm_out)
```

```
Call:
lm(formula = kwh ~ tavg, data = heatcost)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.150  -3.647  -0.152   3.348  23.852 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -97.9239    13.8616  -7.064 8.18e-09 ***
tavg          2.0010     0.1906  10.498 1.11e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.388 on 45 degrees of freedom
Multiple R-squared:  0.7101,    Adjusted R-squared:  0.7036 
F-statistic: 110.2 on 1 and 45 DF,  p-value: 1.111e-13
```

A 95% confidence for the true regression coefficient β_1 describing the relationship between power consumption and temperature can be retrieved from the fitted model with the `confint()` function.

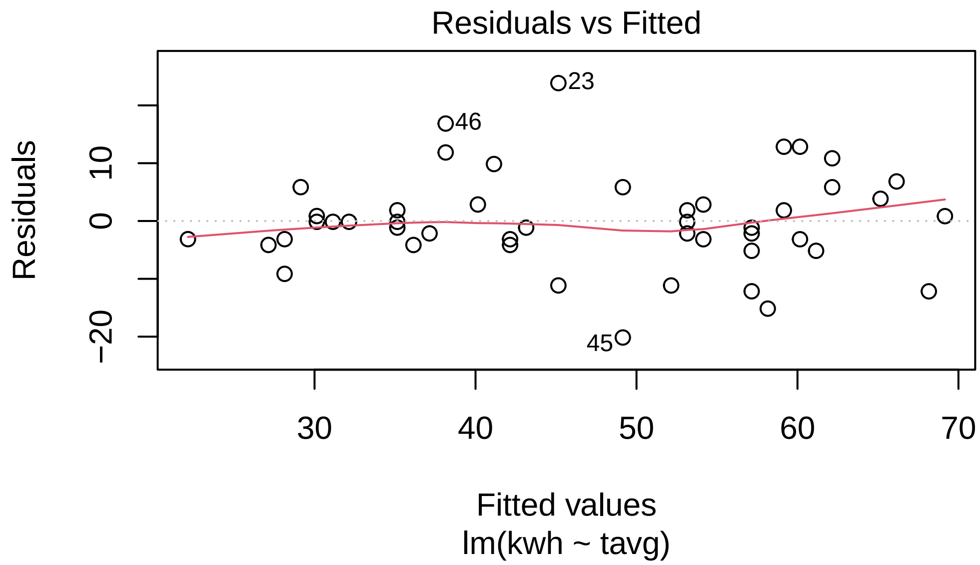
```r
confint(lm_out)
```

```
                  2.5 %     97.5 %
(Intercept) -125.842544 -70.005242
tavg           1.617086   2.384925
```

c)

The following code produces a residuals versus fitted values plot for the fitted model.

```
plot(lm_out,which = 1)
```



There is no evidence in the residuals versus fitted values plot of a nonlinear relationship between power consumption and temperature, as the residuals appear to be centered around zero across the entire range of fitted values. There is some evidence that the variance of the power consumption is smaller at smaller fitted values, as the plot shows a decrease in variability of the residuals toward the left, but it does not appear to be a very strong pattern. The assumptions of the linear model appear to be approximately satisfied.