Regression HW 2

Jun Lu

Fall 2023

Total points: 55

1. Work on Lab 1 (not collected with homework).

2. Study two-sided p-value and one-sided p-value conversion.

Copier maintenance

The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1.1) is appropriate.

```
# Load data
cm <- read.table("CH01PR20.txt", header=F)
colnames(cm) <- c("time", "copiers")
head(cm, 3)  # check the first 3 observations.
```

```
##   time copiers
## 1   20       2
## 2   60       4
## 3   46       3
```

```
tail(cm, 2)  # check the last 2 observations.
```

```
##    time copiers
## 44   61       4
## 45   77       5
```

```
# For tidyverse users
library(tidyverse)
cm <- read.table("./CH01PR20.txt", header=F)
cm %>%
  rename("time" = "V1",
         "copiers" = "V2") -> cm
```

3. Problem 2.5(a, b, c, d) (30 pts.)

2.5.a. CI for the slope (β_1) (10 pts,5 for fomula-based, 5 for software)

From the previous `summary()` output, $b_1 = 15.0352$, $se(b_1) = 0.4831$.

The t-critical value is $t_{(1-\alpha/2, df=n-2)}$. At 90% level (ie $\alpha = 0.1$), n = 45, the t-critical value is (we need 0.1/2 because R uses a one-sided calculator).

```
qt(1 - (0.1/2), df = 45-2)
```

```
## [1] 1.681071
```

The 90% confidence interval for the slope (β_1) is

$$\begin{aligned}&(\widehat{\beta}_1 - t_{crit} \times se(\widehat{\beta}_1),\ \widehat{\beta}_1 + t_{crit} \times se(\widehat{\beta}_1)\\ =&(15.0352 - (1.6811)(0.4831),\ 15.0352 + (1.6811)(0.4831))\\ =&(14.22,\ 15.85)\end{aligned}$$

With 90% confidence, the mean service time change is between 14.22min and 15.85min, when number of copiers increases by 1.

The result matches what's calculated from software.

```
confint(cm.SLR, level = 0.9)
```

```
##                   5 %      95 %
## (Intercept) -5.29378  4.133467
## copiers     14.22314 15.847352
```

2.5.b. Conduct a t test to determine whether or not there is a linear association between X and Y here; control the a risk at 0.10. State the alternatives, decision rule, and conclusion. What is the P-value of your test? (10 pts,5 for fomula-based, 5 for software)

- $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
- test statistic: $t_{obs} = (b_1 - 0)/se(b_1) = (15.035 - 0)/0.483 = 31.13$
- $p - value = 2 \times P(t_{(df=n-2)} > |t_{obs}|) = 2 \times P(t_{df=43} > 31.13) \approx 0$. Since $p - value < \alpha = 0.1$, we reject H_0, and conclude H_a.
- Alternatively, one can use decision rules: when $|t_{obs}| > t_{(1-\alpha/2;n-2)} = 1.68$, reject H_0. Since $|31.31| > t_{(1-\alpha/2;n-2)} = 1.68$. Conclude H_a
- Conclusion: There is significant linear regression association between "number of copiers" and "service time."
- The output form the `summary()` shows that the p-value for the slope of "copiers" is less than $2 \times 10^{-16} \approx 0$. Our formula-based computation is consistent with the software's calculation.

```
tobs <- (15.035 - 0)/0.483
data.frame("t_statistic" = tobs, "p_value"= 2*(1-pt(abs(tobs), df=43)))
```

```
##   t_statistic p_value
## 1    31.12836       0
```

Pr.2.5.c. Are your results in parts (a) and (b) consistent? Explain. (5 pts)

- From part (a), the 90% confidence interval for the slope (β_1) is (14.22, 15.85).
- Yes, the result in parts(a) and part(b) are consistent. A two-sided significance test (at α significance level) rejects the null hypothesis, when the hypothesized value (in this case, 0) falls outside the corresponding confidence interval (at $(1 - \alpha)$ confidence level) for β.

Pr.2.5.d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test? (5 pts)

- $H_0 : \beta_1 = 14$, $H_a : \beta_1 > 14$
- test statistic: $t_{obs} = (b_1 - 14)/se(b_1) = (15.035 - 14)/0.483 = 2.14$
- $p - value = P(t_{(df=n-2)} > t_{obs}) = P(t_{df=43} > 2.14) = 0.0189$. Since $p - value < \alpha = 0.05$, we reject H_0, and conclude H_a.
- Conclusion: There is significant evidence to believe that it took more than 14 minutes for each additional copier. We conclude that the Tri-City standard is not being satisfied.

```
tobs <- (15.035 - 14)/0.483
data.frame("t_statistic" = tobs, "p_value"= 1-pt(tobs, df=43))
```

```
##   t_statistic    p_value
## 1    2.142857 0.01891305
```

4. Problem 2.14 (a, b) (10 pts.)

Pr.2.14.a. Obtain a 90 percent confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval. (5 pts)

```
predict(cm.SLR, newdata=data.frame(copiers = 6), se.fit = T,
        interval = "confidence", leve=0.90)
```

```
## $fit
##        fit      lwr     upr
## 1 89.63133 87.28387 91.9788
##
## $se.fit
## [1] 1.396411
##
## $df
## [1] 43
##
## $residual.scale
## [1] 8.913508
```

- From the above output, at $x = 6$, $\hat{y} = 89.6313 = (-0.5802 + 15.0352 \cdot 6)$, $se(\hat{y}_{mean}) = 1.3964$.
- The t-critical value is $t_{(1-\alpha/2,df=n-2)}$. At 90% level (ie $\alpha = 0.1$), n = 45, the t-critical value is 1.6811.

  ```
  qt(1 - (0.1/2), df = 45-2)
  ```

  ```
  ## [1] 1.681071
  ```

- The 90% confidence interval for the mean service time, when x=6, is

$$\begin{aligned}&(\hat{y}_{mean} - t_{crit} \times se(\hat{y}_{mean}),\ \hat{y}_{mean} + t_{crit} \times se(\hat{y}_{mean}))\\ =&(89.6313 - (1.6811)(1.3964),\ 89.6313 + (1.6811)(1.3964))\\ =&(87.28,\ 91.98)\end{aligned}$$

- We can be 90% confident the mean for the service time will be between 87.28 and 91.98 minutes for 6 copiers.

- The results from formula-based calculation and the software are the same.

Pr.2.14.b. Obtain a 90 percent prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be? (5 pts)

```
predict(cm.SLR, newdata=data.frame(copiers = 6), se.fit = T,
        interval = "prediction", leve=0.90)
```

```
## $fit
##        fit      lwr      upr
## 1 89.63133 74.46433 104.7983
##
## $se.fit
## [1] 1.396411
##
## $df
## [1] 43
##
## $residual.scale
## [1] 8.913508
```

- From the above output, at $x = 6$, $\hat{y} = 89.6313$, $se(\hat{y}_{mean}) = 1.3964$, $\hat{\sigma} = 8.9135$. We need to compute the standard error of the individual prediction.

$$se(\hat{y}_{new}) = \sqrt{[se(\hat{y}_{mean})]^2 + \hat{\sigma}^2} = \sqrt{(1.3964)^2 + (8.9135)^2} = 9.0222$$

- The t-critical value is $t_{(1-\alpha/2, df=n-2)}$. At 90% level (ie $\alpha = 0.1$), n = 45, the t-critical value is 1.6811.
- The 90% confidence interval for the next (i.e., an individual case) service time, when x=6, is

$$\begin{aligned}
&(\hat{y}_{new} - t_{crit} \times se(\hat{y}_{new}),\ \hat{y}_{new} + t_{crit} \times se(\hat{y}_{new})) \\
=&(89.6313 - (1.6811)(9.0222),\ 89.6313 + (1.6811)(9.0222)) \\
=&(74.46,\ 104.80)
\end{aligned}$$

- We can be 90% confident that the service time on the next call for 6 copiers is between 74.46 and 104.80 minutes.
- The results from formula-based calculation and the software are the same.
- At the same x-value, the prediction interval of an individual response is wider than the confidence interval of the mean response. When predicting an individual response, we encounter variability in the estimation of the regression line and variability of the random error of the response. That's why the standard error of the individual response is greater than the standard error of the mean response (at the same x-value).

5. Plot confidence band and prediction band (5 pts)

```
plot( time ~ copiers, data=cm)
abline(cm.SLR$coef)
newx<-seq(0, 10, by=0.1)
cm.CI<-predict(cm.SLR, newdata=data.frame(copiers = newx), interval="confidence", leve=0.95)
cm.PI<-predict(cm.SLR, newdata=data.frame(copiers=newx), interval="prediction", leve=0.95)
lines(newx, cm.CI[,2], lty=2, col=2)
lines(newx, cm.CI[,3], lty=2, col=2)
```

```
lines(newx, cm.PI[,2], lty=3, col=3, lwd=3)
lines(newx, cm.PI[,3], lty=3, col=3, lwd=3)

# If you know tidyverse
# library(tidyverse)

cm.PI <- predict(cm.SLR, interval="prediction", level=0.95)

cm.new <- cbind(cm, cm.PI)
cm.new %>%
  ggplot(aes(x=copiers, y=time)) +
  geom_point() +
  geom_smooth(method=lm, se=TRUE, level=0.95) +
  geom_line(aes(y=lwr), color = "darkgreen", linetype = "dashed") +
  geom_line(aes(y=upr), color = "darkgreen", linetype = "dashed") +
  theme_bw() +
  ggtitle("Data, confidence band and prediction band. (95% level)")
```

6. Problem 2.24 (b, ~~c, d~~) Copier maintenance (~~15~~ 5 pts)

Pr.2.24.b. Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced; use $\alpha = 0.10$. State the alternatives, decision rule, and conclusion.

```
anova(cm.SLR)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## copiers    1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $SSE = 3416$, $SSR = 76960$, SST is not the in R's output (SPSS has it), but can be easily calculated as $SST = SSR + SSE = 76960 + 3416 = 80376$.
- $MSR = SSR/df_R = 76960/1 = 76960$. $MSE = SSE/df_E = 3416/43 = 79.4419$. $F_{obs} = MSR/MSE = 76960/79.4419 = 968.76$. The results matches the output (small rounding error).
- $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
- Test statistic: $F_{obs} = \frac{MSR}{MSE} = \frac{76960}{79.4419} = 968.76$
- p-value: $p - value = P(F_{(df_1=1,\ df_2=n-2)} > F_{obs}) = P(F_{(df_1=2,\ df_2=45-2)} > 968.76) \approx 0$. Note that the p-value is also provided in the software output. Since p-value $< \alpha = 0.1$, reject H_0.

  ```
  1-pf(968.76, df1=1, df2=43)
  ```

  ```
  ## [1] 0
  ```

- Alternatively, use rejection rule: reject H_0 if $F_{obs} > F_{crit}$, where $F_{crit} = F_{(1-\alpha,\ df1=1,\ df2=n-2)} = 2.826$. Reject H_0.

  ```
  qf(1-0.1, df1=1, df2=45-2)
  ```

  ```
  ## [1] 2.825999
  ```

- We conclude that there is significant linear association between the copier number and total minutes for service.

Pr.2.24.c,d. are moved to the next assignment.

7. Read output from another software. 5 pts.

—— This is the end of Homework 2. ——