STAT 757 Assignment 2 Solutions

DUE 2/18/2018 11:59PM

AG Schissler

2/1/2018

2.8.1 [10 points]

```
## change the line below to your path!
my_data_path <- "~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data"
playbill <- read.csv(file.path(my_data_path,"playbill.csv"),header=TRUE)

# Figure 2.6 on page 38
plot(playbill$LastWeek,playbill$CurrentWeek,xlab="Gross box office results previous week",
ylab="Gross box office results current week")
```

2.8.1 Part A

```
lm_fit <- lm(formula = CurrentWeek ~ LastWeek, data = playbill)

## uncomment below to see the available functions for the lm fit object
## attributes(lm_fit)

lm_summary <- summary(lm_fit)
## attributes(lm_summary)

## 95% confidence interval using the output of lm
t_star <- qt(0.975, df = nrow(playbill) - 2)
B1_hat <- coefficients(lm_summary)["LastWeek", "Estimate"]
B1_se <- coefficients(lm_summary)["LastWeek", "Std. Error"]

margin_of_error <- t_star * B1_se
## technically half the interval's length (not the normal definition of margin of error)

print("A 95% confidence interval")
```

```
## [1] "A 95% confidence interval"
```

```
c(lower = B1_hat - margin_of_error, upper = B1_hat + margin_of_error)
```

```
##  lower  upper
## 0.9515 1.0127
```

```
## the easy way:
confint(lm_fit)
```

```
##                    2.5 %     97.5 %
## (Intercept) -14244.3274 27854.0994
## LastWeek         0.9515     1.0127
```

Since the interval contains the value 1, it's plausible that $\beta_1 = 1$.

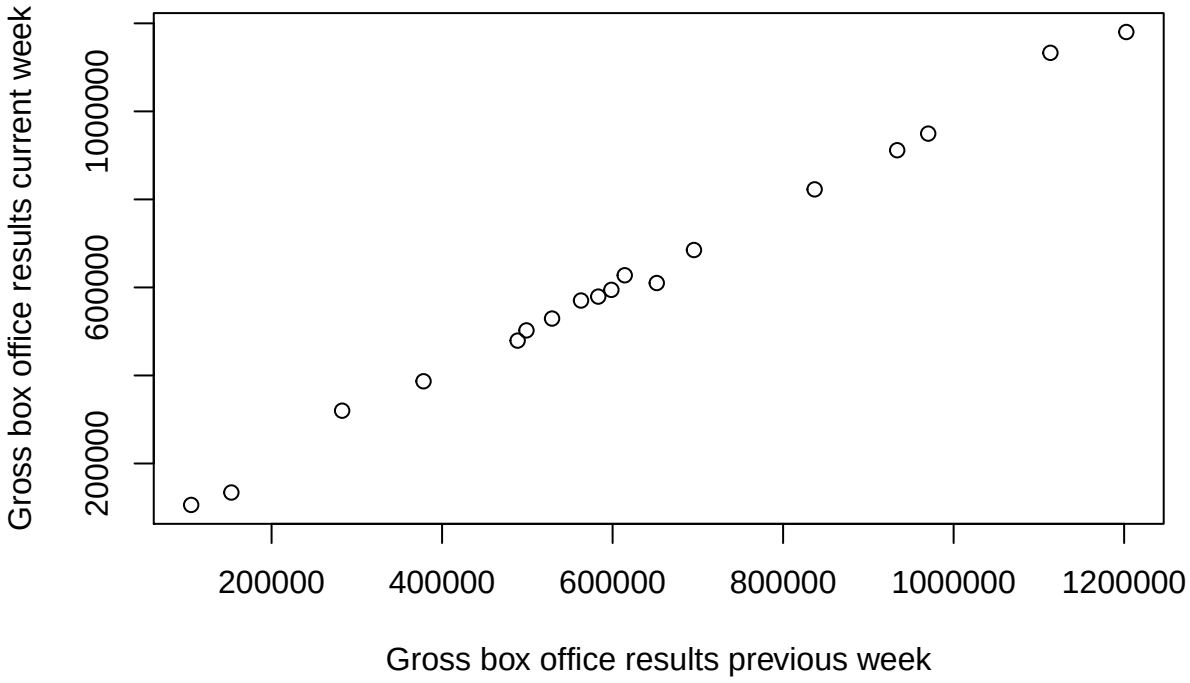

Figure 1: Note the striking linear relationship between box office results lagged by one week.

2.8.1 Part B

```
B0_hat <- coefficients(lm_summary)["(Intercept)", "Estimate"]
B0_se <- coefficients(lm_summary)["(Intercept)", "Std. Error"]
B0_null = 1e4

## build test statistic, see section 2.2.3
t_partb <- ( B0_hat - B0_null ) / B0_se
t_partb
```

```
## [1] -0.32179
```

```
## two-sided test
2*pt(q = t_partb, df = nrow(playbill) - 2)
```

```
## [1] 0.75178
```

There is little evidence to conclude that β_0 is different from 10000.

2.8.1 Part C

```
## y_hat vector is found by the predict function
(y_hat <- predict(object = lm_fit))
```

```
##       1       2       3       4       5       6       7       8       9
##  689781  496833  594655  526320  559681  284516  579532  156899  110609
##      10      11      12      13      14      15      16      17      18
##  828767  959611  646933  378265 1100362  610045  923894 1187793  486673
```

```r
## now predict a new value
x_star <- 400000
unname(coefficients(lm_fit)[1] + coefficients(lm_fit)[2]*x_star)
```

```
## [1] 399637
```

```r
## or use the predict function
x_star_data <- data.frame(LastWeek = 400000) ## change data structure make function work
(y_hat_star <- predict(object = lm_fit, newdata = x_star_data))
```

```
##      1
## 399637
```

```r
## compute a 95% prediction interval around the prediction for 40000
n <- nrow(playbill)
t_star <- qt(0.975, df = n - 2)

## make a prediction using Equations (2.15, 2.16, 2.17)
x_mean <- mean(playbill$LastWeek)
SXX <- sum( (playbill$LastWeek - x_mean)^2 )

## estimate sigma
RSS <- sum((playbill$CurrentWeek - y_hat)^2)
S <- sqrt( RSS / (n-2))
y_hat_star_se <- S * sqrt(1 + n^(-1) + ( (x_star - x_mean)^2 / SXX ) )

margin_of_error <- t_star * y_hat_star_se

print("A 95% prediction interval when x = 400000")
```

```
## [1] "A 95% prediction interval when x = 400000"
```

```r
c(lower = y_hat_star - margin_of_error, upper = y_hat_star + margin_of_error)
```

```
## lower.1 upper.1
##  359833  439442
```

```r
## or use the predict function
predict(object = lm_fit, x_star_data, interval = "prediction")
```

```
##      fit    lwr    upr
## 1 399637 359833 439442
```

```r
## now predict a new value
x_star <- 400000
unname(coefficients(lm_fit)[1] + coefficients(lm_fit)[2]*x_star)
```

```
## [1] 399637
```

```r
## or use the predict function
x_star_data <- data.frame(LastWeek = 400000) ## change data structure make function work
(y_hat_star <- predict(object = lm_fit, newdata = x_star_data))
```

```
##      1
## 399637
```

```r
## compute a 95% prediction interval around the prediction for 40000
n <- nrow(playbill)
t_star <- qt(0.975, df = n - 2)
```

```
## make a prediction using Equations (2.15, 2.16, 2.17)
x_mean <- mean(playbill$LastWeek)
SXX <- sum( (playbill$LastWeek - x_mean)^2 )

## estimate sigma
RSS <- sum((playbill$CurrentWeek - y_hat)^2)
S <- sqrt( RSS / (n-2))
y_hat_star_se <- S * sqrt(1 + n^(-1) + ( (x_star - x_mean)^2 / SXX ) )

margin_of_error <- t_star * y_hat_star_se

print("A 95% prediction interval when x = 400000")
```

```
## [1] "A 95% prediction interval when x = 400000"
```

```
c(lower = y_hat_star - margin_of_error, upper = y_hat_star + margin_of_error)
```

```
## lower.1 upper.1
##  359833  439442
```

```
## or use the predict function
predict(object = lm_fit, x_star_data, interval = "prediction")
```

```
##      fit    lwr    upr
## 1 399637 359833 439442
```

No, $450,000 is well outside the prediction interval.

2.8.1 Part D

Predicting the current week sales from the previous week is a somewhat reasonable strategy. Reformulating the question into terms of the model, predicting current sales using the exact previous sales is assuming that $\beta_0 = 0$ and $\beta_1 = 1$. From our investigation, we believe that β_0 could be 0 and β_1 could be one. However, if one continually predicts the current week from the last week then the obvious and common sense trend of decreasing sales is not being recognized. Indeed our $\hat{\beta}_1 < 1$, hinting at that logic. More data or a different model could produce a better prediction rule.

Exercise 2.8.2 [10 points]

```
## change the line below to your path!
my_data_path <- "~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data"
indicators_dat <- read.table(file.path(my_data_path,"indicators_v2.txt"), header=T, sep = "\t")

# Figure 2.6 on page 38
plot(indicators_dat$LoanPaymentsOverdue,indicators_dat$PriceChange, xlab="Percentage of mortgage loans
ylab="Percentage change in average price from Jul 2006 to Jul 2007.")
```

2.8.2 Part A

```
lm_fit <- lm(formula = PriceChange ~ LoanPaymentsOverdue, data = indicators_dat)
## uncomment below to see the available functions for the lm fit object
```

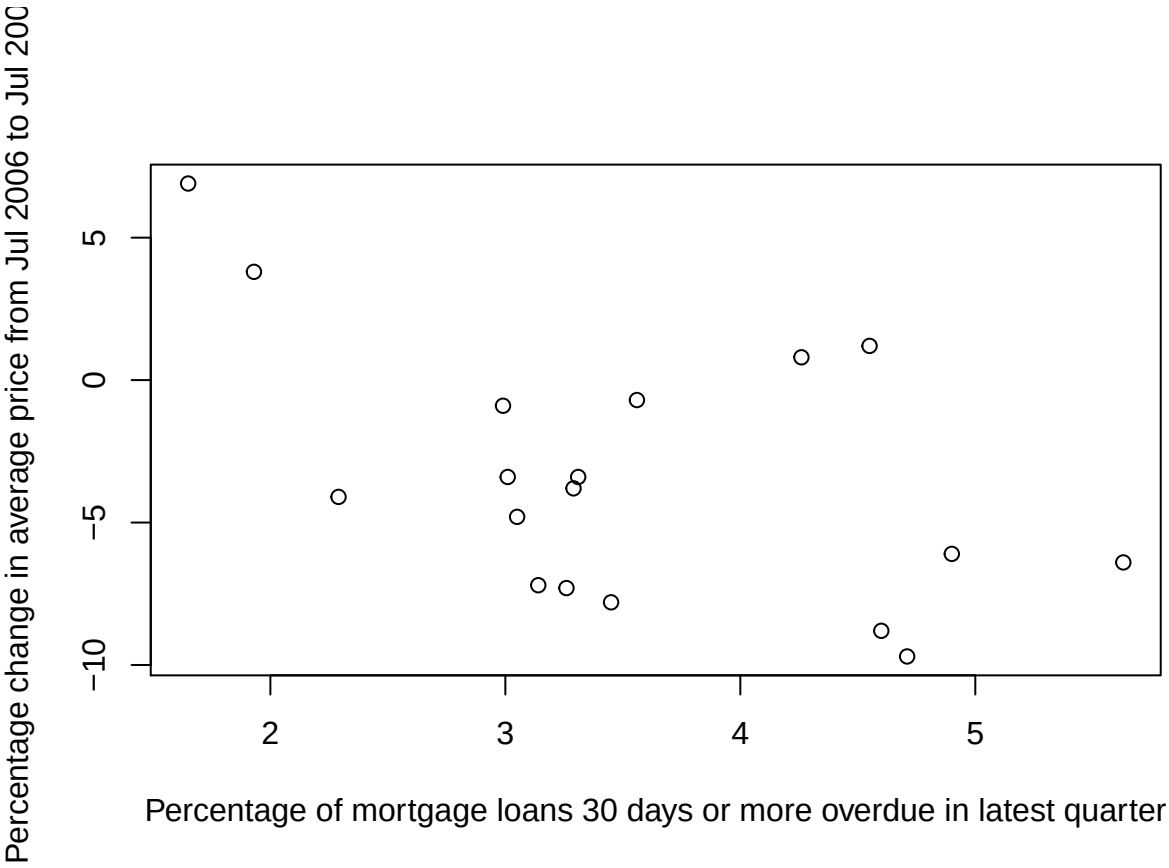

Figure 2: The relationship between the variables is noisy and perhaps only weakly associated.

```
## attributes(lm_fit)
lm_summary <- summary(lm_fit)
## attributes(lm_summary)
## the easy way:
confint(lm_fit)
```

```
##                         2.5 %   97.5 %
## (Intercept)         -2.5321 11.56110
## LoanPaymentsOverdue -4.1635 -0.33359
```

Since the 95% confidence interval lies entirely the negative side of 0, there is statistically significant evidence that $\beta_1 \not\geq 0$.

2.8.2 Part B

```
x_new <- data.frame(LoanPaymentsOverdue = 4)
predict(object = lm_fit, x_new, interval = "confidence")
```

```
##       fit     lwr     upr
## 1 -4.4796 -6.6488 -2.3103
```

Note that the question is referring to the parameter $E(Y|X = 4)$, not the random variable $\hat{y}$. Regression parameters estimates are less variable than the prediction estimates associated with the random variable. See `?predict.lm` for more reading. This more narrow interval does not contain 0 and so the average value of $Y|X = 4$ is likely to be less than 0.

Exercise 2.8.3 [10 points]

```
## change the line below to your path!
my_data_path <- "~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data"
invoices_dat <- read.table(file.path(my_data_path,"invoices.txt"),header=TRUE)

# Figure 2.6 on page 38
plot(invoices_dat$Invoices,invoices_dat$Time ,xlab="Number of Invoices for the day",
ylab="Processing time in hours")
```

2.8.3 Part A

```
lm_fit <- lm(formula = Time ~ Invoices, data = invoices_dat)
## uncomment below to see the available functions for the lm fit object
## attributes(lm_fit)
(lm_summary <- summary(lm_fit))
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5952 -0.2785  0.0349  0.1935  0.5308
```

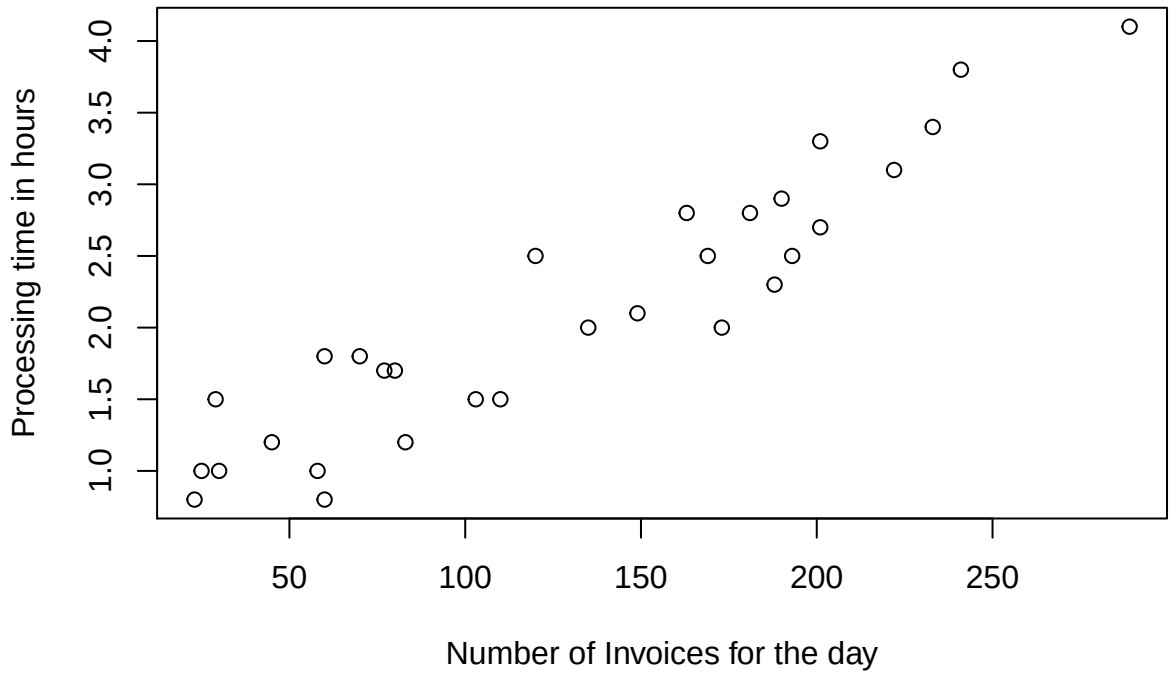

Figure 3: Fairly linear relationship between x and y.

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.641710   0.122271    5.25  1.4e-05
## Invoices    0.011292   0.000818   13.80  5.2e-14
##
## Residual standard error: 0.33 on 28 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.867
## F-statistic:  190 on 1 and 28 DF,  p-value: 5.17e-14
```

```
## attributes(lm_summary)
## the easy way:
confint(lm_fit)[1,]
```

```
##   2.5 %  97.5 %
## 0.39125 0.89217
```

2.8.3 Part B

```
B1_hat <- coefficients(lm_summary)["Invoices", "Estimate"]
B1_se <- coefficients(lm_summary)["Invoices", "Std. Error"]
B1_null = 0.01

## build test statistic, see section 2.2.3
t_partb <- ( B1_hat - B1_null ) / B1_se
t_partb
```

```
## [1] 1.5783
```

```
## two-sided test (careful with the direction here)
2*pt(q = t_partb, df = nrow(invoices_dat) - 2, lower.tail = F)
```

```
## [1] 0.12574
```

There is some evidence to believe that β_1 is different from 0.01 hours/invoice, but not at conventional significance levels.

2.8.3 Part C

```
x_star_dat <- data.frame(Invoices = 130)
predict(lm_fit, newdata = x_star_dat, interval = "prediction", level = 0.95)
```

```
##      fit    lwr    upr
## 1 2.1096 1.4229 2.7963
```

The predicted value of the number of hours is 2.1 hours with a 95% prediction interval between 1.4 and 2.8 hours.

Exercise 2.8.4 [10 points]

2.8.4 Part A

We proceed by minimizing the residual sum of squares (RSS).

$$\begin{aligned} 0 \equiv \frac{\partial RSS}{\partial b} &= \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - bx_i)^2 \\ &= -2\sum_{i=1}^{n}(y_i - bx_i) * x_i \\ &= -2[\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2] \Rightarrow \\ b\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Which gives

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{1}$$

Since the function is quadratic in b this is a global minimum.

2.8.4 Part B

First note that $\hat{\beta} = \sum_{i=1}^{n} c_i y_i$ where $c_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$. Conditional on the observed X this c_i is a constant.

(i)

Then in a similar fashion to Sheather (2009) Section 2.7.1, we proceed as follows:

$$\begin{aligned}
E(\hat{\beta}|X) &= E[\sum_{i=1}^{n} c_i y_i | X = x_i] \\
&= \sum_{i=1}^{n} c_i E[y_i | X = x_i] \\
&= \sum_{i=1}^{n} c_i E[\beta x_i + \epsilon_i] \\
&= \sum_{i=1}^{n} c_i [E(\beta x_i) + E(\epsilon_i)] \\
&= \sum_{i=1}^{n} c_i [\beta x_i + 0] \\
&= \beta \sum_{i=1}^{n} c_i x_i \\
&= \beta \sum_{i=1}^{n} \frac{x_i x_i}{\sum_{i=1}^{n} x_i^2} \\
&= \beta
\end{aligned}$$

(ii)

$$\begin{aligned}
Var(\hat{\beta}|X) &= Var[\sum_{i=1}^{n} c_i y_i | X = x_i] \\
&= \sum_{i=1}^{n} c_i^2 Var[y_i | X = x_i] \\
&= \sum_{i=1}^{n} c_i^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} c_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)^2 \\
&= \sigma^2 \sum_{i=1}^{n} \frac{x_i^2}{(\sum_{i=1}^{n} x_i^2)^2} \\
&= \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i^2)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

(iii)

Using parts (i) and (ii) above and since $\hat{\beta}|X$ is a linear combination of normal randoms variables by construction (namely, the $Y_i|X_i$), $\hat{\beta}|X$ must be normally distributed with the indicated mean and standard deviation.

Exercise 2.8.5 [10 points]

Statement (d) is true. Total variation in Y, denoted SST or SYY, can be decomposed into residual sum of squares RSS and regression sum of squares SS_{reg} and each component of the variation competes with one another. So it would make sense that the directionality of the statements must contradict. To see the direction, the sum of squared residuals in Model 1 are small (differences between the regression line and the points) while deviations from $\bar{y}$ (imagine a horizontal line around y = 5) that contribute to SS_{reg} are large. Exactly the opposite is true for Model 2.

Exercise 2.8.6 [10 points]

Part a

$$\begin{aligned}(y_i - \hat{y}_i) &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i \\ &= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

Part b

$$\begin{aligned}(\hat{y}_i - \bar{y}) &= (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y} \\ &= [(\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i] - \bar{y} \\ &= \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

Part c

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} ([(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})][\hat{\beta}_1 (x_i - \bar{x})]) \\ &= \hat{\beta}_1 \sum_{i=1}^{n} [(y_i - \bar{y})(x_i - \bar{x})] - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{SXY}{SXX} * SXY - \left(\frac{SXY}{SXX}\right)^2 * SXX \\ &= \frac{(SXY)^2 - (SXY)^2}{SXX} = 0\end{aligned}$$

This shows that $SST = RSS + SS_{reg}$ by simply expanding SST and seeing the result in part c (the cross term) is zero.

Exercise 2.8.7 [10 points]

The confidence interval shown illustrates the variation in inferences on the regression parameters at every observed x value. This is different from variation that can be observed for the random variable Y (condition on $X = x$). The regression paramaters are the expected value for Y and as such the estimation of confidence intervals can be very narrow when there are many data points. Nothing in the theory guarantees any points to be in the mean's CI.

References

Sheather, Simon. 2009. *A Modern Approach to Regression with R*. Springer Science & Business Media.