SS3859

Assignment 1

Yizhou Tang

250888541

1. Suppose that we have the following 9 observations for random variable X: x = c(12.21, 14.37, 17.18, 11.74, 13.84, 14.26, 15.42, 13.52, 17.97). Conduct a t-test for the true mean of X. Specifically, test H0 : μ = 16 vs H1 : μ < 16 at significance level α = 0.05. Give the test statistic (t value), the p-value, and your conclusion. (2 pt)

```
> #H0: u = 16, H1: u <16, alpha = 0.05
> # The test is one-sided.
> x = c(12.21, 14.37, 17.18, 11.74, 13.84, 14.26, 15.42, 13.52, 17.97)
>
> sample_mean = mean(x) # x_bar
> sample_sd = sd(x) # s
>
> t_stat = (sample_mean - 16)/(sample_sd/sqrt(9)) # (x_bar-mu0)/(s/sqrt(n))
> t_stat # test statistic
[1] -2.168426
>
> # calculating the p_value
> A = 1 - pt(abs(t_stat),df=8) # function "abs" gives the absolute value.
> p_val = A #One sided
> p_val # If p_val < alpha (0.05), reject H0
[1] 0.03098519
>
> cv = qt(0.95,8) # Gives t_value at which the p-value becomes 0.05
> abs(t_stat) > cv # TRUE means |t_stat| was greater than the critical value
-> Evidence against H0
[1] TRUE
```

2.

2. $\bar{x}=5$ $\quad\bar{y}=-90$

a) $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n} (x_i-\bar{x})^2} = -2022/102$

$\boxed{= -19.8235}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -90 + 19.8235(5)$

$\boxed{= 9.1175}$

b). $x=3$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(3)$

$= -50$

c) $\dfrac{47.13}{n-2} = 47.13/8 = 5.89125$

d) 95% C.I. for $E(Y|x=3)$

$\frac{\alpha}{2} = 0.025$

$-50 \pm (t_{\alpha/2,n-2})\, SE(\hat{Y})$

$= -50 \pm (2.306004)\left(\hat{\sigma}^2\left(1+\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)\right.$

$= -50 \pm 2.306004\left(5.89125^2\left(1+\frac{1}{10} + \frac{3-5}{102}\right)\right.$

$= -50 \pm 2.088372.$

3

3.

3

a) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot (\bar{x})$

$= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta} \bar{x}$

$= \bar{y}$

Since $\hat{y} = \bar{y}$, the fitted line passes through $(\bar{x}, \bar{y})$.

b) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

$SST = SSR + SSE + 2\sum (y_i \hat{y}_i - \hat{y}_i^2 - y_i \bar{y} + \hat{y}_i \bar{y})$

$= SSR + SSE + 2[\sum (y_i \hat{y}_i - \hat{y}_i^2) - \bar{y} \sum (y_i - \hat{y}_i)]$

We know ① $\sum (y_i - \hat{y}_i) = 0$
and ② $\sum x_i (y_i - \hat{y}_i) = 0$

$= SSR + SSE + 2[\sum (y_i \hat{y}_i - \hat{y}_i^2) - 0]$

$= SSR + SSE + 2 \sum (y_i \hat{y}_i - \hat{y}_i^2)$

$= SSR + SSE + 2 \sum \hat{y}_i (y_i - \hat{y}_i)$

$= SSR + SSE + 2\sum (\hat{\beta}_0 + \hat{\beta}_1 x_i)(y_i - \hat{y}_i)$

$= SSR + SSE + 2\sum \hat{\beta}_0 (y_i - \hat{y}_i) + \hat{\beta}_1 \sum x_i (y_i - \hat{y}_i)$

$= SSR + SSE + 2\hat{\beta}_0 \sum (y_i - \hat{y}_i) + 0$

$= SSR + SSE + 0$

$= SSR + SSE$

4.

a. Count the number of observations whose x1 are greater than 6. (1 pt)

```
> length(hw1_data$x1[hw1_data$x1>6])
[1] 26
```

b. Count the number of observations whose x1 are greater than 6 and x2 equal to H. (1 pt)

```
> temp <- hw1_data[hw1_data$x1>6,]
> length(temp$x2[temp$x2 == "H"])
[1] 23
```

c. Consider a subset A that contains all observations with x2 = H. Compute the mean, median and standard deviation of the x1 values in subset A. (1 pt)

```
> #median and standard deviation of the x1 values in subset A.
> A <- hw1_data[hw1_data$x2 == "H",]
> mean(A$x1) #mean
[1] 5.832919
> median(A$x1) #median
[1] 5.684439
```

```
> sd(A$x1) #standard deviation
[1] 1.790704
```

d. The sample mean of x1 is 4.435. Can we argue that the true mean of x1 differs from 4? Conduct a t-test at significance level α = 0.05. Give the test statistic (t-value), p-value and your conclusion. (1 pt)

```
> sample_mean = 4.435 # x_bar
> sample_sd = sd(hw1_data$x1) # s
> n = length(hw1_data$x1)#n
> t_stat = (sample_mean - 4)/(sample_sd/sqrt(n)) # (x_bar-mu0)/(s/sqrt(n))
> t_stat # test statistic
[1] 1.719877
>
> # calculating the p_value
> A = 1 - pt(abs(t_stat),df=n-1) # function "abs" gives the absolute value.
> p_val = 2*A #Two sided
> p_val
[1] 0.08857947
> # Conclusion: Since p_val > alpha (0.05), failed to reject H0
```

e. Consider the statement: "Given that x2 equals to H, the true mean of x1 is larger than 4." Is this statement convincing? Use a t-test (α = 0.05) to support your answer. (2 pt)

```
> #Hypothesis:
> #H0: mu = 4, H1: mu >4
>
> x = hw1_data[hw1_data$x2 == "H",]
> x = x$x1
> sample_mean = mean(x) # x_bar
> sample_sd = sd(x) # s
> n = length(x)#n
> t_stat = (sample_mean - 4)/(sample_sd/sqrt(n)) # (x_bar-mu0)/(s/sqrt(n))
> t_stat # test statistic
[1] 7.727821
>
> cv = qt(0.95,n-1) #Critical value
> cv
[1] 1.672522
> abs(t_stat) > cv
[1] TRUE
> # Conclusion: Since it is TRUE, there is evidence against H0, null hypothes
is is rejected,
> # Hence, the gievn statement is convicing.
```

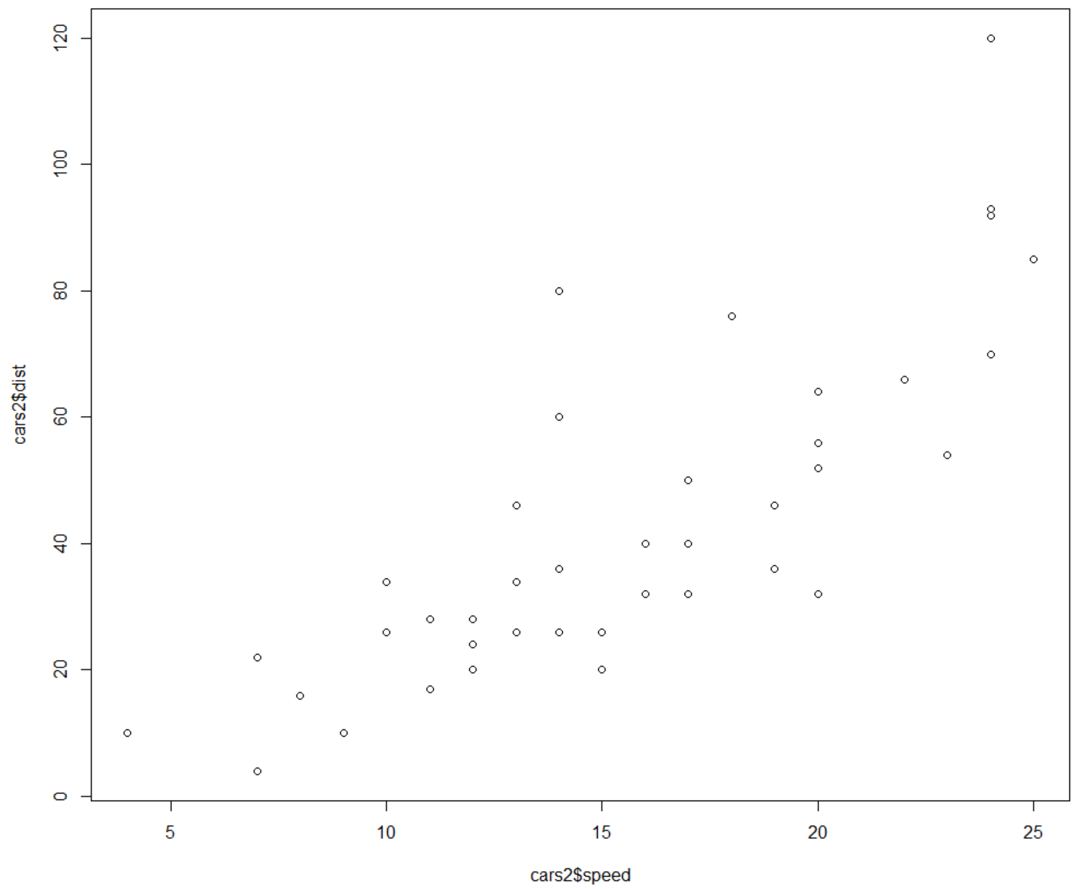
5. For question 5, you will use a subset of the cars data. Run the following R codes and use the cars2 data set to answer the questions. Include your R codes and output for the following questions.

```
> set.seed (50)
> idx = sample(nrow(cars),40,replace=FALSE)
> cars2 = cars[idx,]
>
```

```
> y = cars2$dist
> x = cars2$speed
> n = 40
```

a. Make a scatterplot that shows the relationship between x and Y . From the plot, do you find any relationship between speed and dist? (1 pt)

```
> plot(cars2$speed,cars2$dist)
> # According to the plot, it appears that there is a linear relationship bet
ween speed and dist
```



b. Assume that there is a linear relationship between x and Y . That is, Yi = β0 + β1xi + i where i ∼ N(0, σ2). Obtain the LS estimates for β0, β1 and an unbiased estimate for σ 2 . (1 pt)

```
> #b)
> cars_lm = lm(dist~speed,data=cars2)
> fitted_y = cars_lm$fitted.values
> summary(cars_lm)

Call:
lm(formula = dist ~ speed, data = cars2)
```

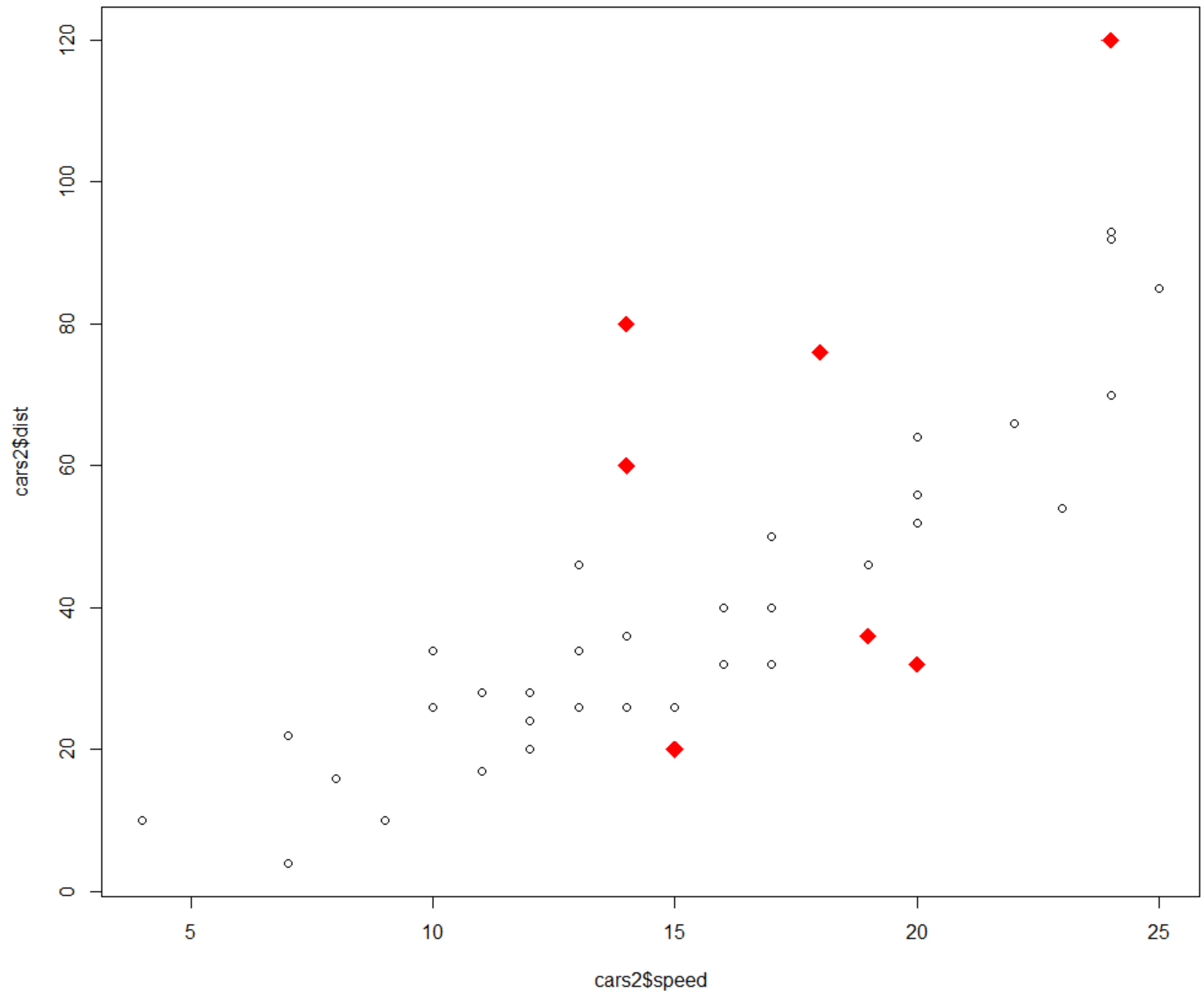
```
Residuals:
    Min      1Q  Median      3Q     Max 
-28.403  -8.904  -3.285   6.818  44.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.2369     7.7336  -2.229   0.0318 *  
speed         3.8820     0.4698   8.264 5.15e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.85 on 38 degrees of freedom
Multiple R-squared:  0.6425,	Adjusted R-squared:  0.6331 
F-statistic: 68.29 on 1 and 38 DF,  p-value: 5.152e-10

> #B0 =  -17.2369
> #B1 = 3.8820
> 
> #Unbiased sigma^2 estimator from LSE
> est_var = sum((y-fitted_y) ^2)/(n-2)
> est_var
[1] 251.0771
> #Unbiased estimate of sigma^2 is 173.4714
```

c. Using the estimates, calculate the residuals e4, e7 and e10 (i.e. residuals for the observations at the 4th, 7th and 10th rows of the cars2 data). (1 pt)

```
> #Residuals e4,e7, e10:
> e4 = y[4] - fitted_y[4]
> e7 = y[7] - fitted_y[7]
> e10 =  y[10] - fitted_y[10]
> e4
      37 
-10.5208 
> e7
      31 
1.243172 
> e10
       5 
2.181033 
```

d. Find the residuals whose absolute values are greater than 20. Indicate those residuals in the scatterplot with different a color and shape. (2 pt)

```
> #d)
> residuals = y - fitted_y
> #residuals whose absolute values are greater than 20:
> residuals[abs(residuals) > 20]
       36        22        24        23        34        49        39 
-20.52080  22.88913 -20.99286  42.88913  23.36119  44.06928 -28.40278 
> idx = abs(residuals)>20
> 
> plot(cars2$speed,cars2$dist)
> points(cars2$speed[idx],cars2$dist[idx], col = "red",pch = 18,cex=2)
```



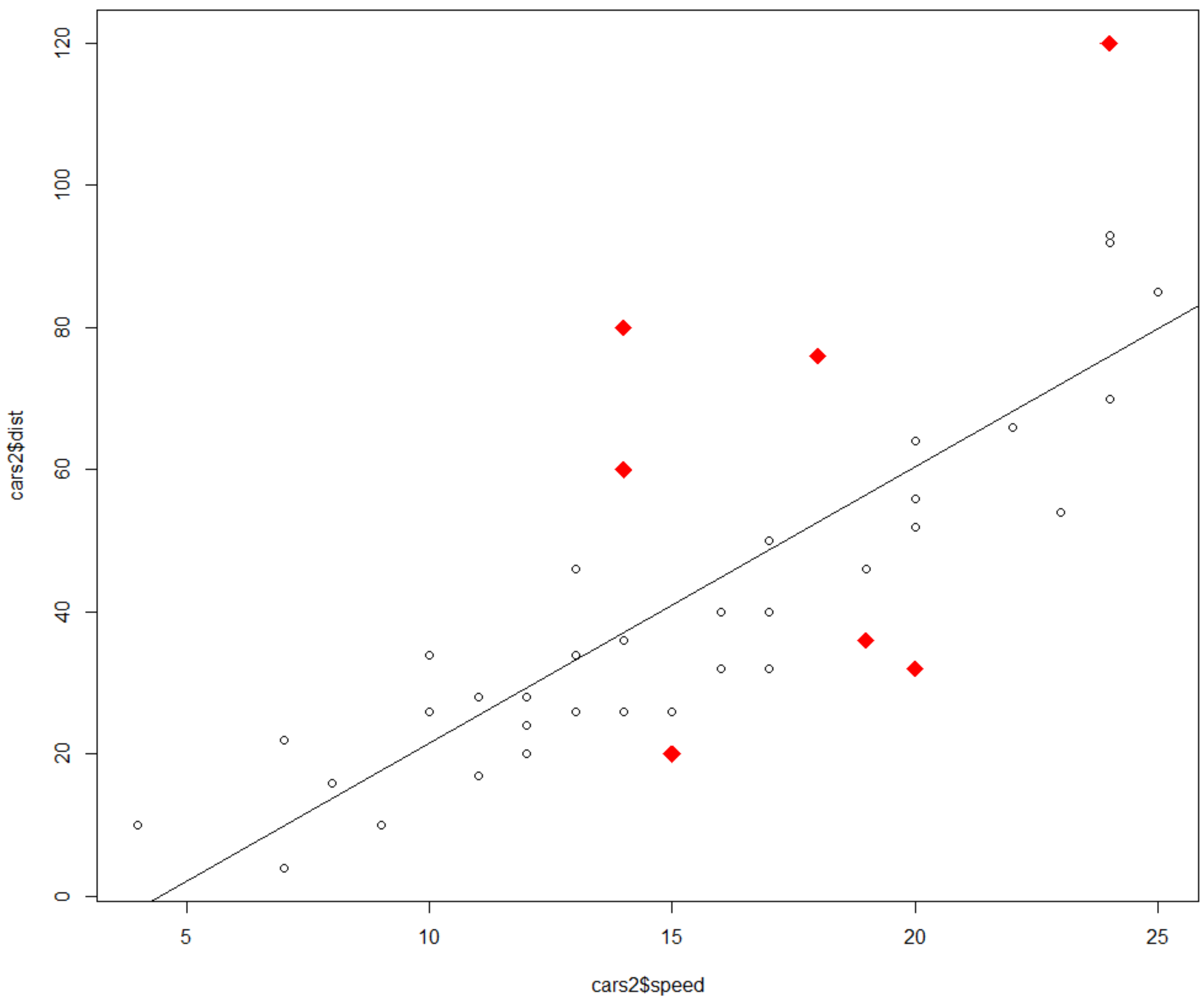
e. Calculate the sum of the residuals (i.e. Pn i=1 ei). (1 pt)

```
> #e)
> sum(y-fitted_y)
[1] 0
> #sum of residuals = 0, as expected
>
```

f. Report the fitted model. Add the fitted regression line to the current scatterplot. Predict the distance taken to stop when the speed of the car is 17. (1 pt)

```
> #Report the fitted model:
> summary(cars_lm)

Call:
lm(formula = dist ~ speed, data = cars2)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-28.403  -8.904  -3.285   6.818  44.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.2369     7.7336  -2.229   0.0318 *  
speed         3.8820     0.4698   8.264 5.15e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.85 on 38 degrees of freedom
Multiple R-squared:  0.6425,	Adjusted R-squared:  0.6331 
F-statistic: 68.29 on 1 and 38 DF,  p-value: 5.152e-10

> #Add a fitted regression line
> abline(x,fitted_y)


Call:
line(x, fitted_y)

Coefficients:
[1]  -17.237    3.882  

> #Predict the distance when the speed is 17
> predict(cars_lm,newdata=data.frame(speed=17))
       1 
48.75683 
> 
> #Distance is estimated to be 48.7571 when sped is 17
```



g. State the goodness of fit for the fitted model. What percentage of the variation in the response variable is explained by the fitted model? (1 pt)

```
> #g)
> #To measure goodness of fit, we can use R^2:
> summary(cars_lm)$r.squared
[1] 0.6424875
> #According to R^2,64% of the variation is explained by the model
>
```

h. Consider the statement: “If someone is driving at 100mph, according to the fitted model, the distance taken to stop will be exactly 370.9615ft.” Give a brief (reasonable) criticism of the statement. (2 pt)

```
> # This statement is assuming that the model is able to predict every outcom
e with residual = 0; which is very unlikely to happen. Include a confidence i
```

```
nterval for E(Y|x = 100) could be a way to provide a better description for t
he model.
```

i. Construct a 90% confidence interval for β1. (2 pt)

```
> #Confidence interval for the beta parameters
> confint(cars_lm,level = 0.90)
                   5 %      95 %
(Intercept) -30.275356 -4.198463
speed         3.089992  4.673977
> #Confidence for B1 is (3.235501  4.629317)
```

j. Construct a 95% confidence interval for E(Y |x = 15). (1 pt)

```
> # Confidence interval for the mean response at speed=15
> predict(cars_lm,newdata=data.frame(speed=15),interval="confidence",level=0.95)
       fit      lwr      upr
1 40.99286 35.89159 46.09413
> #Confidence interval for E(Y|x = 15) is (35.89159,46.09413)

>
```