MA684 hw2
Jiayuan Shi

1.

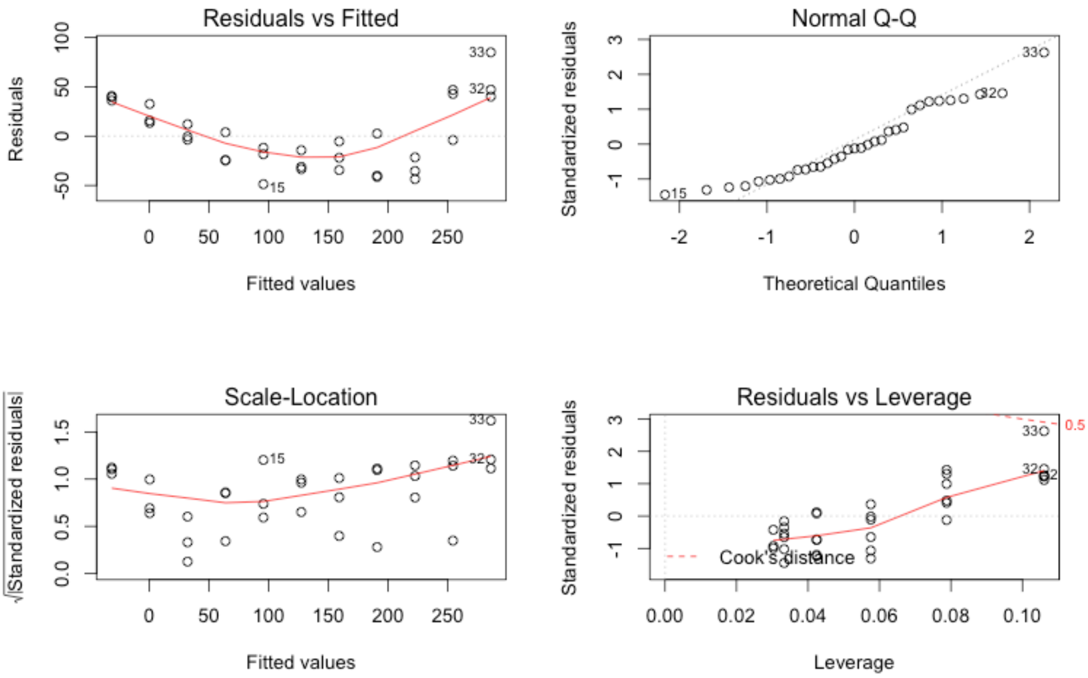
```
> distance <- read.csv("stoppingdistance.csv", header=TRUE)
> attach(distance)
> plot(dist~speed)
> speed
 [1] 15 15 15 20 20 20 25 25 25 30 30 30 35 35 35 40 40 40 45 45 45 50 50 50
[25] 55 55 55 60 60 60 65 65 65
> reg.results <- lm(dist~speed)
> summary(reg.results)
Call:
lm(formula = dist ~ speed)

Residuals:
   Min     1Q Median     3Q    Max
-48.73 -24.67  -3.98  32.63  84.97

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -126.729     16.197   -7.82  7.9e-09 ***
speed          6.350      0.377   16.86  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.2 on 31 degrees of freedom
Multiple R-squared:  0.902,  Adjusted R-squared:  0.899
F-statistic:  284 on 1 and 31 DF,  p-value: <2e-16

> par(mfrow=c(2,2))
> plot(reg.results)
```



The assumptions of the linear regression model are,
i) Independent, random sample from underlying population:
Yes, each data in the sample is independent from the others, and they are randomly selected.
ii) Linearity, the means of Y|X fall on a straight line:
Yes, from the plot, we can see the straight line $\mu_{Y/X} = \beta_0 + \beta_1 X$.
From R, we can get the estimate $\hat{Y} = -126.729 + 6.350X$, which is $\widehat{distance} = -126.729 + 6.350 * speed$
iii) Homoscedasticity, the variance of Y|X is the same for all X(variance of E_i is the same for all X):
Yes, from the above "Residuals vs Fitted" plot, variance of E_i is constant in the sample.
iv) Normality, the distribution of Y|X follows a normal distribution for all X:
Yes, we can see a straight line in the above "Normal Q-Q" plot, which means the distribution of Y|X follows a normal distribution for all X.
v) Existence, the model holds for valid values of X:
Yes, we can see the values of X through the command "speed" in R.

2.

```
> distance <- read.csv("stoppingdistance.csv", header=TRUE)
> attach(distance)
> plot(dist~speed)
> speed
 [1] 15 15 15 20 20 20 25 25 25 30 30 30 35 35 35 40 40 40 45 45 45 50 50 50
[25] 55 55 55 60 60 60 65 65 65
> reg.results <- lm(dist~speed)
> summary(reg.results)
Call:
```

lm(formula = dist ~ speed)

Residuals:
Min 1Q Median 3Q Max
-48.73 -24.67 -3.98 32.63 84.97

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-126.729	16.197	-7.82	7.9e-09	***
speed	6.350	0.377	16.86	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.2 on 31 degrees of freedom
Multiple R-squared: 0.902, Adjusted R-squared: 0.899
F-statistic: 284 on 1 and 31 DF, p-value: <2e-16

$$\widehat{distance} = \mathrm{b0} + \mathrm{b1} * speed$$

From the above results, we can get the estimate b0= -126.729, and b1=6.350.

$$\widehat{distance} = -126.729 + 6.350 * speed$$

2a. Yes, there is a significant association between speed and stopping distance. From the results of "summary(reg.results)", we can see that the p-value is very small(< 2e-16), so we can reject the null hypothesis that $\beta = 0$(no association between speed and stopping distance), which means there is a significant association between speed and stopping distance.

2b.

$$\Delta speed = 60 - 50 = 10$$
$$\Delta distance = 10 * 6.35 = 63.5$$

(95%) Confidence interval for β_1 :
b1±t*(*crit)*se(b1)=6.350±*2.04*0.377* =(5.581, 7.119)
95% CI for this increase in distance $= (5.581, 7.119) * 10 = (55.81, 71.19)$
So the stopping distance expected to increase from 55.81 to 71.19 mph with 95% CI.

3a.

	Parameter Estimate	Standard Error	t-value (df)	p-value	95% CI
Intercept	115.44	14.73	---	---	---
Maternal Age	-0.49	0.56	-0.875	0.389	(-1.64 , 0.658)

$t_{obs} = \frac{b_1}{se(b_1)}$ =-0.49/0.56=-0.875, df=28, p-value=0.389
95% Confidence interval for the slope (t(crit) = 2.05)
-0.49± 2.05 (0.56) or (-1.64 , 0.658)
H0: $\beta_1 = 0$ (No association between maternal age and a child's cognitive ability)
The p-value is 0.389 (>0.05), so we fail to reject the null hypothesis that $\beta = 0$(no association between speed and stopping distance), which means there is not a significant association between maternal age and a child's cognitive ability.

With 95% confidence, we can say the slope for maternal age is between -1.64 and 0.658.

The slope is not big enough, which shows a weak association between maternal age and a child's cognitive ability. In addition, the p-value is 0.389 (>0.05), also showing that there is not a significant association between the two.

3b.
Predicted IQ for a child born to a 20 year old mother:
$\hat{y}_{X0}$ =115.44-0.49*20=105.64

$$\hat{y}_{X0} \pm t_{crit} s_{Y|X}\sqrt{1+\frac{1}{n}+\frac{(x0\text{-}\bar{x})^2}{(n\text{-}1)s_X^2}}=105.64\pm \mathit{2.05{*}13.5}\sqrt{1+\frac{1}{30}+\frac{(20\text{-}27)^2}{(30\text{-}1)\mathit{{*}4.5{\char94}2}}}=$$

(76.394, 134.89)

We are 95% confident that the IQ for a child born to a 20 year old mother is between 76.394 and 134.89.

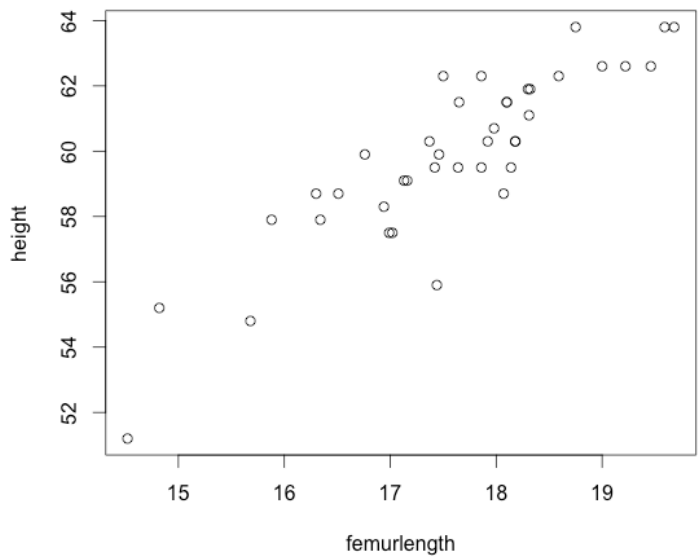
3c.
Predicted mean IQ for children born of 30 year old mothers:
$\hat{y}_{X0}$ =115.44-0.49*30=100.74

$$\hat{y}_{X0} \pm t_{crit} s_{Y|X}\sqrt{\frac{1}{n}+\frac{(x0\text{-}\bar{x})^2}{(n\text{-}1)s_X^2}}=100.74\pm \mathit{2.05{*}13.5}\sqrt{\frac{1}{30}+\frac{(30\text{-}27)^2}{(30\text{-}1)\mathit{{*}4.5{\char94}2}}}=$$

(94.635, 106.84)

We are 95% confident that the mean IQ for children born of 30 year old mothers is between 94.635 and 106.84.

4a.
> csi <- read.csv("CSI femur stature inches.csv", header=TRUE)
> attach(csi)
> plot(height~femurlength



4b.

```
> mean(femurlength)
[1] 17.604
> sd(femurlength)
[1] 1.1652
```

Mean +/- sd for femur length: 17.604+/-1.1652

```
> mean(height)
[1] 59.892
> sd(height)
[1] 2.6374
```

Mean +/- sd for height: 59.892+/-2.6374

4c.

```
> reg.results <- lm(height~femurlength)
> summary(reg.results)

Call:
lm(formula = height ~ femurlength)

Residuals:
   Min     1Q Median     3Q    Max
-3.667 -0.740  0.015  0.678  2.614

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.848      3.083    8.06  9.5e-10 ***
femurlength    1.991      0.175   11.39  8.1e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.27 on 38 degrees of freedom
Multiple R-squared:  0.773,  Adjusted R-squared:  0.768
F-statistic:  130 on 1 and 38 DF,  p-value: 8.1e-14

> confint(reg.results,"femurlength")
            2.5 % 97.5 %
femurlength 1.637 2.3446
```

Linear regression predicting height (inches) from femur length (inches)

	Parameter Estimate	Standard Error	t-value (df)	p-value	95% CI
Intercept	24.848	3.083	8.06	9.5e-10	---
Femur Length	1.991	0.175	11.39	8.1e-14	(1.637, 2.3446)

4d. $\widehat{height} = 24.848 + 1.991 * femurlength$

For each inch increase in the femur length, on average, the height expected to increase 1.991 inch.

4e. s(y|x)=$\sqrt{s^2_{Y|X}}$=$\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-2}}$=1.27

s(y|x) is called the 'standard error of the estimate', and so the standard deviation of femur length, for the sub-set of subjects with a particular height value is 1.27 inches.

4f.
Predicted height for a person with a femur of 19.00 inches:
```
> predict(reg.results,data.frame(femurlength=19),interval="predict")
    fit   lwr    upr
1 62.673 60.02 65.325
```
We are 95% confident that the height for a person with a femur of 19.00 inches is between 60.02 and 65.325.

4g.
Predicted mean height for people with a femur of 19.00 inches:
```
> predict(reg.results,data.frame(femurlength=19),interval="confidence")
    fit    lwr    upr
1 62.673 62.032 63.313
```
We are 95% confident that the mean height for people with a femur of 19.00 inches is between 62.032 and 63.313.