Homework 1

RABE 2.2 (3 pts)

1 point for each correct

Explain why you would or would not agree with each of the following statements:

a. **True**

b. **False**

c. **True**

RABE 2.3 (4 pts)

1 point for each correct

Using the regression output listed below ("Computer Repair Data Regression Table"), test the following hypotheses using $\alpha = 0.1$:

a. **Fail to reject**
b. **Fail to reject**
c. **Reject**
d. **Reject**

RABE 2.4

Ungraded

RABE 2.12 (3 pts)

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 news-papers concerning their daily and Sunday circulations (in thousands). The data are online here: http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P054.txt.

a. **Ungraded**
b. **Ungraded**
c. **Ungraded**
d. **Reject H_0 (1 point)**
e. **Ungraded**
f. Provide an interval estimate (based on 95% level) for the true average Sunday circulation of newspapers with Daily circulation of 500,000
g. The particular newspaper that is considering a Sunday a edition has a Daily circulation of 500,000. Provide an interval estiamte (based on the the 95% level) for the predicted Sunday circulation of this paper. How does this interval differ from that given in (f)

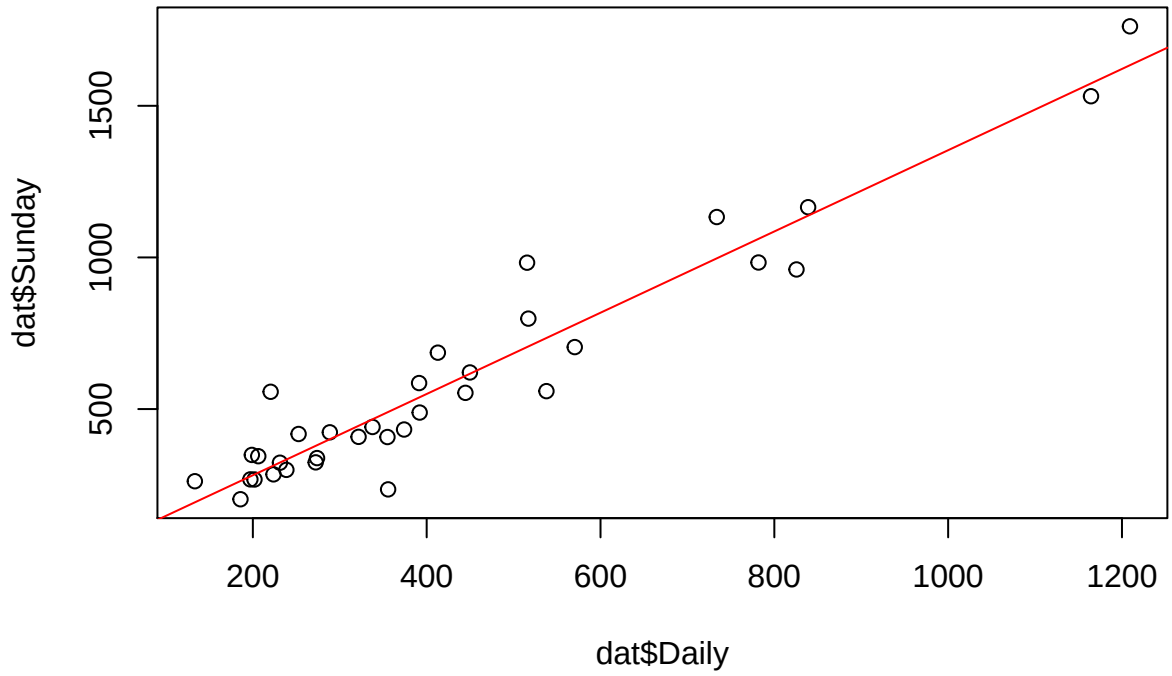
(h) Another newspaper being considered as a candidate for a Sunday edition has a daily circulation of 2,000,000. Provide an interval estimate for the predicted Sunday circulation for this paper. How does this interval compare with the one given in (g)? Do you think it is likely to be accurate?

```
library(tidyverse)
library(magrittr)
```

```r
dat = read.csv('../../data/newspaper.csv')

# a
plot(dat$Daily,dat$Sunday)
# yes linear relationship is plausible

# b
mod = lm(Sunday~Daily,data=dat)
plot(dat$Daily,dat$Sunday)
abline(mod,col='red')
```




```r
# c
confint(mod)
```

```
##                  2.5 %    97.5 %
## (Intercept) -59.094743 86.766003
## Daily         1.195594  1.483836
```

```r
# d
summary(mod)
```

```
##
## Call:
## lm(formula = Sunday ~ Daily, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.19  -55.57  -20.89   62.73  278.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.83563   35.80401   0.386    0.702
## Daily        1.33971    0.07075  18.935   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.4 on 32 degrees of freedom
## Multiple R-squared:  0.9181, Adjusted R-squared:  0.9155
## F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16
```

```
# H_0: beta_1=0, H_a: beta_1 =/= 0, alpha=.95
# p-value of t-stat is <2e-16, which is below alpha
# Yes there is a significant relationship


# e- about 92% (see above)

#f

# multiply upper and lower bounds of CI by 500K, add beta_0 to obtain prediction interval
beta0 = 13.83563
beta1.ci = as.numeric(confint(mod)[2,])

pred.int = beta0 + beta1.ci*5e6
pred.int
```

```
## [1] 5977982 7419193
```

```
# g
# previous prediction interval is for "true" Sunday, ie. \hat{y}_i
# now we want to predict y_i so add estimate of \hat{\sigma}
n = nrow(dat)
sigma.hat = sqrt(sum(mod$resid**2)/(n-2))

pred.int2 = pred.int + sigma.hat

# h
# as before

pred.int3 = beta0 + beta1.ci*2e6 + sigma.hat
pred.int3
```

```
## [1] 2391311 2967795
```

```
# probably too wide
```

RABE 2.13

Let $y_1, y_2, ..., y_n$ be a sample drawn from a normal population with unknown mean μ and unknown variance σ^2 ($y_i \sim N(\mu, \sigma^2)$). One way to estimate μ is to fit the linear model:

$$y_i = \mu + \epsilon; i = 1, 2, ...n$$

And use the least squares method estimator (LSE), that is to chose μ such that it minimizes the sum of squares $\sum_{i=1}^{n}(y_i - \mu)^2$. Another way is to use least absolute value estimator (LAVE), that is, to minimize the sum of the absolute values: $\sum_{i=1}^{n}|y_i - \mu|$

a. Show that the LSE of μ is the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Compute LSE by taking derivative $\frac{d}{d\mu}$ and setting equal to 0:

$$\frac{d}{d\mu}\sum_{i=1}^{n}(y_i-\mu)^2=0$$
$$-\sum_{i=1}^{n}2(y_i-\hat{\mu})=0$$
$$-2\sum_{i=1}^{n}(y_i-\hat{\mu})=0$$
$$\sum_{i=1}^{n}(y_i-\hat{\mu})=0$$
$$(\sum_{i=1}^{n}y_i)-n\hat{\mu}=0$$
$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

b. Show that the LAV of μ is the sample median

Since:

$$|y_i-\mu|=\begin{cases}y_i-\mu, y_i>\mu\\ \mu-y_i, y_i<\mu\end{cases}$$

Thus:

$$\frac{d}{d\mu}|y_i-\mu|=\begin{cases}-1, y_i>\mu\\ 1, y_i<\mu\end{cases}$$

Let $\mathbb{1}_{y_i<\mu}$ denote the function which is 1 when $y_i<\mu$ and -1 otherwise. Therefore:

$$\frac{d}{d\mu}\sum_{i=1}^{n}|y_i-\mu|=0$$
$$\sum_{i=1}^{n}\mathbb{1}_{y_i<\mu}=0$$

Since the left hand side is a big sum of 1s and -1s, the only way it can be equal to 0 is if there are are equal numbers of 1s and -1s. This only occurs when equal numbers of $y_i<\mu$ and $y_i>\mu$, ie. when μ is the median.

c. State one advantage and one disadvantage each of the LSE and the LAVE

- Median is not as sensitive to outliers, mean minimizes expected square error loss. *