Homework 1

RABE 2.2

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

a. $\text{Cov}(Y, X)$ and $\text{Cov}(X, Y)$ can take values between $-\infty$ and $+\infty$

True, Cov can be positive or negative (see: regression) and has no max or min possible value

b. If $\text{Cov}(Y, X) = 0$ or $\text{Cov}(X, Y) = 0$ one can conclude that there is not relationship between Y and X

False, consider $x_1 = -1, x_2 = 0, x_3 = 1$ *and* $y_1 = y_3 = 1, y_2 = 0$. *Cor(X,Y)=0 but* $y = x^2$

c. The least squares line fitted to the points in the scatter plot made by the points $(\hat{Y}_i, Y_i)$ has a zero intercept and a unit slope

True. Say $Y_i = c\hat{Y}_i + d$ *and* $\hat{Y}_i = aX + b$. *It is approximately also true that* $Y_i = aX + b$, *so* $Y_i = c(aX + b) + d = aX + b$. *Hence* $caX = ax$ *and* $cb + d = b$, *thus* $c = 1$ *and* $d = 0$

RABE 2.3

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

a. $H_0 : \beta_1 = 15$ vs $H_A : \beta_1 \neq 15$

Check if 15 is in 99% CI:

```
computer.repair = read.csv('/home/peter/Desktop/Teaching/STAT_4400_5400/data/computer_repair.tsv',sep='
mod = lm(Minutes ~ Units, data=computer.repair)
confint(mod,level=.9)
```

```
##                   5 %     95 %
## (Intercept) -1.81810 10.14141
## Units       14.60875 16.40879
```

Conf int contains $\beta_1 = 15$ *so fail to reject* H_0.

b. $H_0 : \beta_1 = 15$ vs $H_A : \beta_1 > 15$

One-sided t-test, so first compute t-stat:

$$t = \frac{15.509 - 15}{.505} = 1.018$$

For one-sided test the cutoff t-value is $t_{12,0.1} = 1.782$ *(see eg. https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/more-significance-testing-videos/v/one-tailed-and-two-tailed-tests). Since* $t < t_{12,0.1}$ *fail to reject* H_0.

c. $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

p-value for this H_0 *given in table. Since 8.92e-13 < .1 reject* H_0

d. $H_0 : \beta_1 = 5$ vs $H_A : \beta_1 \neq 5$

Use confidence interval from (a). Since 5 not in CI reject H_0.

RABE 2.4

Using the regression output listed below ("Computer Repair Data Regression Table"), construct the 99% confidence interval for β_0

Two ways:

```
confint(mod,level=.99)
```

```
##                 0.5 %   99.5 %
## (Intercept) -6.086633 14.40994
## Units       13.966287 17.05126
```

Or:

$$CI = [\hat{\beta}_0 - t_{12,.005}\text{s.e.}(\hat{\beta}_0)\hat{\beta}_0 + t_{12,.995}\text{s.e.}(\hat{\beta}_0)]$$

Where the t-values are:

```
qt(.005,12)
```

```
## [1] -3.05454
```

```
qt(.995,12)
```

```
## [1] 3.05454
```

Computer Repair Data Regression Table (RABE Table 2.9)

```
# data available here: http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P031.txt
computer.repair = read.csv('/home/peter/Desktop/Teaching/STAT_4400_5400/data/computer_repair.tsv',sep='
summary(lm(Minutes ~ Units, data=computer.repair))
```

```
##
## Call:
## lm(formula = Minutes ~ Units, data = computer.repair)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2318 -3.3415 -0.7143  4.7769  7.8033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.162      3.355    1.24    0.239
## Units         15.509      0.505   30.71 8.92e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.392 on 12 degrees of freedom
## Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864
## F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

RABE 2.12

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. Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between Daily and Sunday circulation? Do you think this is a plausible relationship?

b. Fit a regression line predicting Sunday circulation from Daily circulation
c. Obtain the 95% confidence itnervals for β_0 and β_1
d. Is there a significant relationship between Sunday circulation and Daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing, and your conclusion.
e. What proportion of the variability in Sunday circulation is accounted for by Daily circulation?
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)

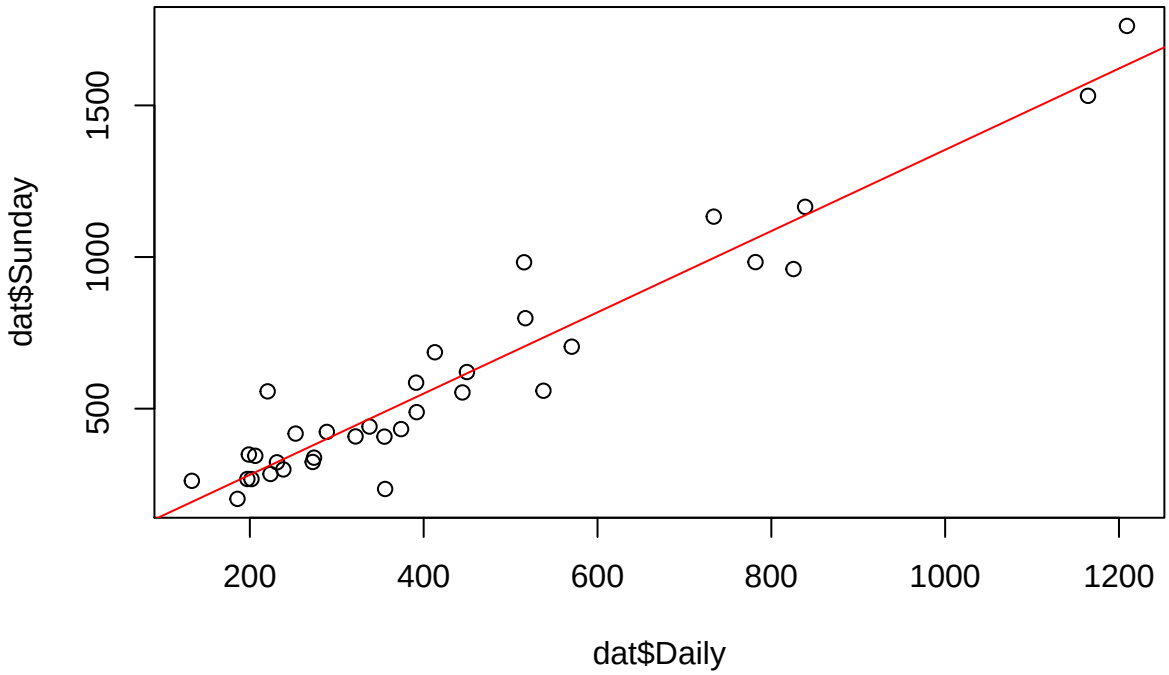
(h) Another newspaper being considered as a candidate for a Sunday edition has a daily circulation of 2,000,000. Provide an interval estimate for the predicted Sunday circulation for this paper. How does this interval compare with the one given in (g)? Do you think it is likely to be accurate?

```
library(tidyverse)
library(magrittr)

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')
```




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

```
##                   2.5 %    97.5 %
## (Intercept) -59.094743 86.766003
```

```
## Daily         1.195594   1.483836
```

```
# 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
predict(mod,interval='confidence',level=.95,newdata=data.frame(Daily=500))
```

```
##       fit      lwr     upr
## 1 683.693 644.1951 723.191
```

```
# g
predict(mod,interval='prediction',level=.95,newdata=data.frame(Daily=500))
```

```
##       fit      lwr      upr
## 1 683.693 457.3367 910.0493
```

```
# h
# as before
predict(mod,interval='prediction',level=.95,newdata=data.frame(Daily=2000))
```

```
##        fit      lwr      upr
## 1 2693.265 2373.463 3013.068
```

```
# probably too wide- way outside dataset so low information/hig
```

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. *