STAT 420 – Homework 2

1. Time Use (without R)

a. Find the equation of the least-squares regression line.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})^2$
16	30	–2	–2	4	4	4
12	52	–6	20	36	–120	400
25	7	7	–25	49	–175	625
19	32	1	0	1	0	0
21	9	3	–23	9	–69	529
15	56	–3	24	9	–72	576
18	38	0	6	0	0	36
126	224	0	0	108	–432	2170
				SXX	*SXY*	*SYY*

or …

x	y	x^2	xy	y^2
16	30	256	480	900
12	52	144	624	2704
25	7	625	175	49
19	32	361	608	1024
21	9	441	189	81
15	56	225	840	3136
18	38	324	684	1444
126	224	2376	3600	9338

$$\bar{x} = \frac{126}{7} = 18 \qquad \bar{y} = \frac{224}{7} = 32$$

$$SXX = 2376 - \frac{1}{7}(126)^2 = 108$$

$$SXY = 3600 - \frac{1}{7}(126)(224) = -432$$

$$SYY = 9338 - \frac{1}{7}(224)^2 = 2170$$

Putting it all together,…

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{-432}{108} = -4 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 32 - (-4)18 = 104$$

The least-squares regression line is $\hat{y} = 104 - 4x$.

b. Calculate the fitted values, $\hat{y}_i$.

c. Calculate the residuals, e_i. Does the sum of the residuals equal zero?

Fitted Values: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Residuals: $e = y - \hat{y} = y - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$

The sum of the residuals equals 0 as we would expect.

x	y	(b) $\hat{y}$	(c) e	e^2
16	30	40	– 10	100
12	52	56	– 4	16
25	7	4	3	9
19	32	28	4	16
21	9	20	– 11	121
15	56	44	12	144
18	38	32	6	36
			0	442 *RSS*

d. Give an estimate for σ, the standard deviation of the observations about the true regression line?

We need the residual sum of squares (*RSS*). It can be found by totaling the squares of the errors as seen in the table above. Or, recall that the total variation of *Y* comes from two sources: *SYY* = *SSReg* + *RSS*, or sometimes alternately written as *SST* = *SSR* + *SSE.*

$$SSR = \hat{\beta}_1^2 \cdot SXX = (-4)^2 \cdot 108 = 1728$$
$$RSS = SYY - SSR = 2170 - 1728 = 442$$

Often, the more common choice for estimating σ is the unbiased estimator,

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{442}{5}} = \sqrt{88.4} = 9.402\,.$$

Or, the maximum likelihood estimate of σ is another option,

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{442}{7}} = \sqrt{63.1} = 7.946\,.$$

e. What proportion of observed variation in TV viewing is explained by a straight-line relationship with physical activity?

This is answered by the coefficient of determination,

$$R^2 = \frac{SSR}{SYY} = 1 - \frac{RSS}{SYY} = 1 - \frac{442}{2170} = 0.796 = 79.6\%$$

f. Predict the number of TV viewing hours for a participant who engaged in 24 hours of physical activity in the same week.

The least-squares regression model predicts 8 hours of TV viewing in that week.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 104 - 4(24) = 8$$

2. Time Use (with R)

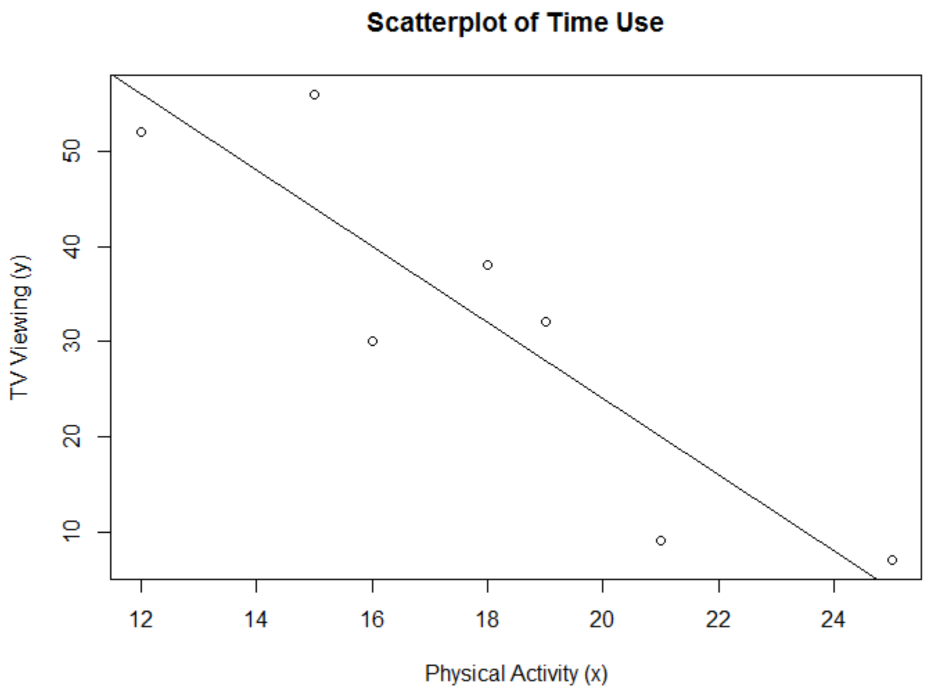
a. Find the equation of the least-squares regression line predicting the amount of TV watching when given the amount of physical activity in a given week.

```
> TVfit <- lm(y~x)
> TVfit

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
        104           -4
```

The least-squares regression line is $\hat{y} = 104 - 4x$.

b. Create a scatterplot of the data and add the least-squares regression line to it.

```
> plot(x, y, main="Scatterplot of Time Use",
+      xlab="Physical Activity (x)",
+      ylab="TV Viewing (y)")
> abline(TVfit$coefficients)
```



c. Find the fitted values, $\hat{y}_i$.

```
> TVfit$fitted
 1  2  3  4  5  6  7
40 56  4 28 20 44 32
```

d. Find the residuals, e_i. Does the sum of the residuals equal zero?

```
> TVfit$residuals
  1   2   3   4   5   6   7
-10  -4   3   4 -11  12   6
> sum(TVfit$residuals)
[1] -4.440892e-16
```

Aside from some approximation and rounding error in R, yes, the residuals sum to 0.

e. Give an estimate for σ, the standard deviation of the observations about the true regression line?

```
> summary(TVfit)$sigma
[1] 9.402127
```

f. What proportion of observed variation in TV viewing is explained by a straight-line relationship with physical activity?

```
> summary(TVfit)$r.squared
[1] 0.7963134
```

g. Predict the number of TV viewing hours for a participant who engaged in 24 hours of physical activity in the same week.

```
> predict(TVfit, data.frame(x=24))
8
```

3. Modeling without an intercept

To derive the formula for the slope of the least-squares regression line with an intercept at the origin, we want to minimize $f(\beta)=\sum_{i=1}^{n}(y_i-\beta x_i)^2$.

Take the first derivative: $f'(\beta)=\sum_{i=1}^{n}2(y_i-\beta x_i)(-x_i)=-2\sum_{i=1}^{n}x_i y_i+2\beta\sum_{i=1}^{n}x_i^2$

To find the extreme points, set the first derivative equal to 0 and solve.

$$f'(\beta)=0 \quad \Rightarrow \quad \hat{\beta}=\sum_{i=1}^{n}x_i y_i \Big/ \sum_{i=1}^{n}x_i^2$$

To make sure that the estimate provides a minimum for $f(\beta)$, use the second derivative test. Since $f''(\beta)=2\sum_{i=1}^{n}x_i^2$ always has to be greater than 0, then $f(\beta)$ has a minimum at $\hat{\beta}$.

4. Concert Venue

a. Find the least-squares estimate, $\hat{\beta}$.

x	y	x^2	xy
3.6	28	12.96	100.8
4.2	24	17.64	100.8
5.4	32	29.16	172.8
3	13.6	9	40.8
4.8	36	23.04	172.8
6	44	36	264
27	177.6	127.8	852

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \bigg/ \sum_{i=1}^{n} x_i^2 = \frac{852}{127.8} = 6.67$$

b. Calculate the fitted values, $\hat{y}_i$.

c. Calculate the residuals, e_i. Does the sum of the residuals equal zero?

Fitted Values: $\hat{y} = \hat{\beta}x$

Residuals: $e = y - \hat{y} = y - \hat{\beta}x$

		(b)	(c)
x	y	$\hat{y}$	e
3.6	28	24	4
4.2	24	28	–4
5.4	32	36	–4
3	13.6	20	–6.4
4.8	36	32	4
6	44	40	4
			–2.4

Note that the residuals do not add up to zero. Without the intercept (β_0) and the vector of all 1's in the model, the vector $\boldsymbol{e}$ of the residuals does not have to be orthogonal to it, so the residuals do not have to add up to zero.

d. Find the least-squares estimate, $\hat{\beta}$.

```
> venue <- lm(y ~ 0+x)
> venue$coef
       x
6.666667
```

e. Create a scatterplot of the data and add the least-squares regression line to it.

```
> plot(x, y, xlim=c(0,6), ylim=c(0,45),
+   main="Scatterplot of Concert Venue Revenue",
+   xlab="Number of Patrons, in thousands (x)",
+   ylab="Revenue, in thousands of dollars (y)")
> abline(a=0,b=venue$coeff)
```