MTH 326 - Spring 2022
Assignment #7
Due: Friday, April 1 2022 (11:59pm)

1. Suppose that we have a random sample of n observations from the density function

$$f(y \mid \theta) = \frac{y^2 e^{-y/\theta}}{2\theta^3} \quad \text{on support } y > 0$$

   (a) Determine the rejection region for the most powerful test of

$$\begin{aligned} &H_0 : \theta = \theta_0 \text{ versus} \\ &H_a : \theta = \theta_a, \end{aligned}$$

assuming $\theta_a > \theta_0$.

Solution: The likelihood function is

$$\begin{aligned} L(y \mid \theta) &= \prod_{i=1}^{n} \frac{{y_i}^2 e^{-y_i/\theta}}{2\theta^3} \\ &= \left(\frac{1}{2\theta^3}\right)^n \prod_{i=1}^{n} {y_i}^2 e^{-y_i/\theta} \\ &= \left(\frac{1}{2\theta^3}\right)^n \exp\left(-\frac{\sum y_i}{\theta}\right) \prod_{i=1}^{n} {y_i}^2 \end{aligned}$$

By the Neyman Pearson Lemma, there exists some k such that $\dfrac{L(\theta_0)}{L(\theta_a)} < k$ (strictly less than since $\theta_a > \theta_0$). And

$$\begin{aligned} \frac{L(\theta_0)}{L(\theta_a)} &= \frac{\left(\frac{1}{2{\theta_0}^3}\right)^n \exp\left(-\frac{\sum y_i}{\theta_0}\right) \prod_{i=1}^{n} {y_i}^2}{\left(\frac{1}{2{\theta_a}^3}\right)^n \exp\left(-\frac{\sum y_i}{\theta_a}\right) \prod_{i=1}^{n} {y_i}^2} \\ &= \frac{\frac{1}{{\theta_0}^{3n}} \exp\left(-\frac{\sum y_i}{\theta_0}\right)}{\frac{1}{{\theta_a}^{3n}} \exp\left(-\frac{\sum y_i}{\theta_a}\right)} \\ &= \left(\frac{\theta_a}{\theta_0}\right)^{3n} \cdot \exp\left(\frac{\sum y_i}{\theta_a} - \frac{\sum y_i}{\theta_0}\right) \\ &= \left(\frac{\theta_a}{\theta_0}\right)^{3n} \cdot \exp\left(\left(\frac{1}{\theta_a} - \frac{1}{\theta_0}\right) \sum_{i=1}^{n} y_i\right) \\ &< k \end{aligned}$$

To get this in terms of a statistic, we take the natural log of both sides. Thus,

$$\begin{aligned}
&\ln\left[\left(\frac{\theta_a}{\theta_0}\right)^{3n}\cdot\exp\left(\left(\frac{1}{\theta_a}-\frac{1}{\theta_0}\right)\sum_{i=1}^{n}y_i\right)\right] < \ln k\\
\iff &3n\left(\ln\theta_a-\ln\theta_0\right)+\left(\frac{1}{\theta_a}-\frac{1}{\theta_0}\right)\sum_{i=1}^{n}y_i < \ln k\\
\iff &\left(\frac{1}{\theta_a}-\frac{1}{\theta_0}\right)\sum_{i=1}^{n}y_i < \ln k-3n\left(\ln\theta_a-\ln\theta_0\right)\\
\iff &\sum_{i=1}^{n}y_i > \frac{\ln k-3n\left(\ln\theta_a-\ln\theta_0\right)}{\frac{1}{\theta_a}-\frac{1}{\theta_0}} := k' \qquad \text{(Divide by negative value)}
\end{aligned}$$

Let $S := \sum_{i=1}^{n} y_i$, then we reject H_0 if $S > k'$, with the best critical region

$$C := \left\{\vec{y} \mid S > k'\right\}$$

where k' is selected such that

$$\Pr\left(S > k' \mid H_0 : \mu = \mu_0\right) = \alpha$$

for some μ_0 and α.

(b) Is the test you defined in part (a) uniformly most powerful for the alternative $\theta > \theta_0$? Briefly explain your answer.

Solution: The critical region C does not depend on a specific alternative θ_a with $\theta_a > \theta$. Therefore the test is a uniformly most powerful (UMP) test of size α.

The following table contains dietary data (calories and the content of fat, sodium, carbohydrate, and protein) in some standard hamburgers that can be found at local fast food restaurants.

	cal	fat (g)	sodium (mg)	carbs (g)	protein (g)
BK Jr.	310	18	390	27	13
Wendy's Jr.	250	11	420	25	13
McDonald's	250	9	480	31	12
Culvers	390	17	480	38	20
Steak-n-Shake	320	14	830	32	15
Sonic Jr.	330	16	610	32	15

2. We wish to explore if there is a relationship between fat and sodium. The conjecture is that leaner meat needs more salt to enhance flavor.

   (a) Compute the least squares regression line with response variable sodium content and input variable fat content. Clearly state sums of the intermediate calculations: $\bar{x}, \bar{y}, S_{xy}, S_{xx}$.

   Solution: Our x, (input) is the fat content, and our dependent y (response) is the sodium content. Then,

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{18+11+9+17+14+16}{6} = 14.1\bar{6} \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{390+420+480+480+830+610}{6} = 535 \\
S_{xy} &= (18-14.1\bar{6})(390-535) + (11-14.1\bar{6})(420-535) + (9-14.1\bar{6})(480-535) \\
&\quad + (17-14.1\bar{6})(480-535) + (14-14.1\bar{6})(830-535) + (16-14.1\bar{6})(610-535) \\
&= 25 \\
S_{yy} &= (390-535)^2 + (420-535)^2 + (480-535)^2 + (480-535)^2 + (830-535)^2 + (610-535)^2 \\
&= 132{,}950 \\
S_{xx} &= (18-14.1\bar{6})^2 + (11-14.1\bar{6})^2 + (9-14.1\bar{6})^2 + (17-14.1\bar{6})^2 + (14-14.1\bar{6})^2 + (16-14.1\bar{6})^2 \\
&= 62.8\bar{3}
\end{aligned}$$

Then

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{25}{62.8\bar{3}} = \frac{150}{377} \approx 0.39787798408488063660477453580901856763925729442970822281167108 75$$

and

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} = 535 - \frac{150}{377}\cdot 14.1\bar{6} = \frac{199570}{377} \approx 529.3633952254641909814323607427055702917771883289 1$$

Hence $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x = \dfrac{199570+150x}{377} \approx 529.363 + 0.398x$

(b) Calculate S^2. Again, state any necessary intermediary sums.

Solution: Using the values computed in (a),

$$\mathrm{SSE} = S_{yy} - \widehat{\beta}_1 \cdot S_{xy} = 132950 - \frac{150}{377} \cdot 25 = \frac{50{,}118{,}400}{377} \approx 132940.0530503979$$

and

$$S^2 = \frac{1}{n-2}\,\mathrm{SSE} = \frac{1}{6-2} \cdot \frac{50{,}118{,}400}{377} = \frac{12{,}529{,}600}{377} \approx 33235.013262599469496.$$

(c) Calculate the correlation coefficient ρ^2.

Solution: Solving for ρ^2,

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \widehat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}} \iff \rho^2 = \frac{S_{xx}}{S_{yy}} \widehat{\beta_1}^2.$$

Substituting in these values,

$$\begin{aligned} \rho^2 &= \frac{S_{xx}}{S_{yy}} \widehat{\beta_1}^2 \\ &= \frac{62.8\overline{3}}{132950} \left(\frac{150}{377}\right)^2 \\ &= \frac{75}{1{,}002{,}443} \\ &\approx 0.000074817221 \end{aligned}$$

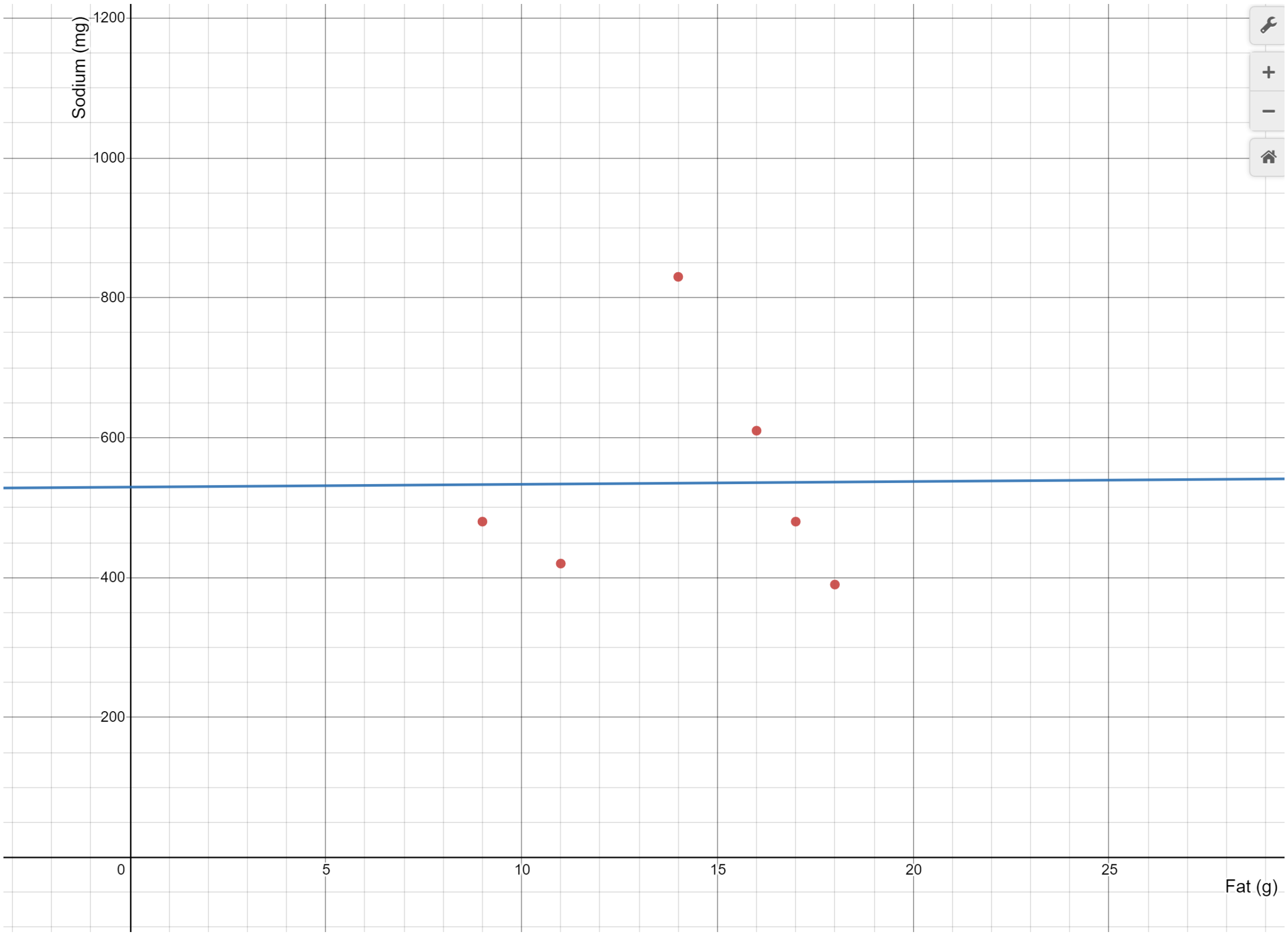
Hence $\rho^2 \approx 0.000075 = 7.5 \cdot 10^{-5}$.

(d) A good rule of thumb for the correlation coefficient in regards to best-fit lines is that if ρ^2 is greater than 0.70, then the line is a good model for the spread of the data. Is using a line a good model for this data?

Solution: Based on this correlation coefficient this line is an egregious model of the data.

(e) Sketch a scatterplot of the data and draw the best-fit line and interpret the picture in context of your answer in (d).

Solution: Look at this photograph:



Now, the line does not look as terrible. It could be around 1200 on the y-intercept and about 20 on the x-intercept and go between the set of points to be a bit more accurate. If $(11, 420)$ is an outlier, then a quadratic regression could be useful with a maximum around $(14, 800)$. The right cluster looks very linear, but there is some weird shear between the 2nd and 3rd points (from left to right). That, or there is just no correlation between these with any regression curve and it is just a coincidence that the right cluster is linear.

3. American culture is focused on fat intake as corresponding to a high-calorie diet.

   (a) Compute the least squares regression line with response variable calorie count and input variable fat content. Clearly state sums of the intermediate calculations: $\bar{x}, \bar{y}, S_{xy}, S_{xx}$.

   Solution: Our x represents the fat content, and y is the calorie count. Hence,

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{18+11+9+17+14+16}{6} = 14.1\overline{6} \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{310+250+250+390+320+330}{6} = 308.\overline{3} \\
S_{xy} &= (18-14.1\overline{6})(310-308.\overline{3}) + (11-14.1\overline{6})(250-308.\overline{3}) + (9-14.1\overline{6})(250-308.\overline{3}) \\
&\quad + (17-14.1\overline{6})(390-308.\overline{3}) + (14-14.1\overline{6})(320-308.\overline{3}) + (16-14.1\overline{6})(330-308.\overline{3}) \\
&= 761.\overline{6} \\
S_{yy} &= (310-308.\overline{3})(310-308.\overline{3}) + (250-308.\overline{3})(250-308.\overline{3}) + (250-308.\overline{3})(250-308.\overline{3}) \\
&\quad + (390-308.\overline{3})(390-308.\overline{3}) + (320-308.\overline{3})(320-308.\overline{3}) + (330-308.\overline{3})(330-308.\overline{3}) \\
&= 14083.\overline{3} \\
S_{xx} &= (18-14.1\overline{6})^2 + (11-14.1\overline{6})^2 + (9-14.1\overline{6})^2 + (17-14.1\overline{6})^2 + (14-14.1\overline{6})^2 + (16-14.1\overline{6})^2 \\
&= 62.8\overline{3}
\end{aligned}
$$

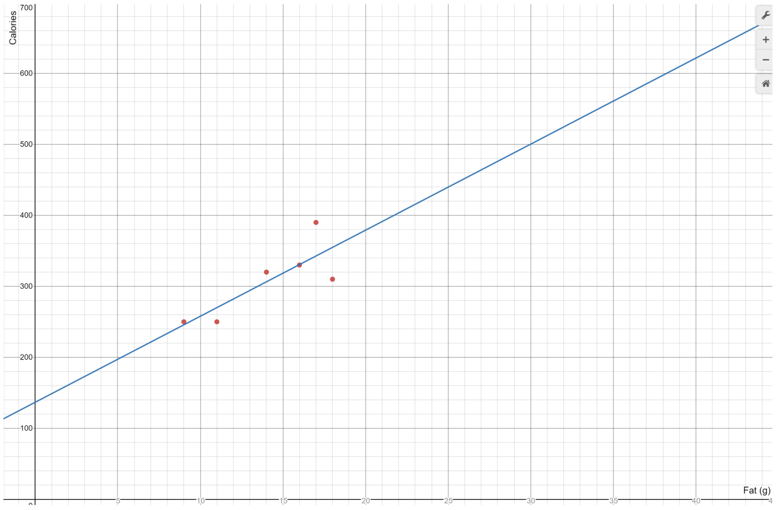
Then

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{761.\overline{6}}{62.8\overline{3}} = \frac{4570}{377} \approx 12.122015915119363395225464190981432360742705570291777188328912466$$

and

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} = 308.\overline{3} - \frac{4570}{377} \cdot 14.1\overline{6} = \frac{51{,}500}{377} \approx 136.604774535809018567639257294429708222811671087$$

Hence $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x = \dfrac{51500 + 4570x}{377} \approx 136.60477 + 12.12202x.$



(b) Suppose a new burger on the market is known to have 650 calories. What is a good estimate for how much fat is in the burger?

Solution: Since the calorie count is the y value of our regression line, we substitute it in for y. Thus,

$$\begin{aligned} y = 650 &= \frac{51500 + 4570x}{377} \\ \iff 245050 &= 51500 + 4570x \\ \iff 193550 &= 4570x \\ \iff x &= \frac{19355}{457} \approx 42.352 \end{aligned}$$

Hence the expected value for the grams of fat in a burger with 650 calories is approximately 42.352 grams.

(c) Find an 80% confidence interval for the slope of the regression line.

Solution: For an 80% C.I., then $\alpha = 0.2$ and $t_{\alpha/2} = t_{0.1}$. We want to use the formula $I \equiv \widehat{\beta}_i \pm t_{\alpha/2}(\text{df}) \cdot S\sqrt{c_{i\,i}}$. Since the slope coefficient is β_1 then $i = 1$.

$t_{0.1}(4) = 1.533$ by Table 5.

$\sqrt{c_{i\,i}} = \sqrt{\dfrac{1}{S_{xx}}} = \sqrt{\dfrac{6}{377}}$ by part (3a).

$\text{SSE} = S_{yy} - \widehat{\beta}_1 \cdot S_{xy} = \dfrac{42250}{3} - \dfrac{4570}{377} \cdot \dfrac{2285}{3} = \dfrac{1{,}828{,}600}{377} \approx 4850.39788$ from (3a).

$S = \sqrt{\dfrac{1}{n-2}\,\text{SSE}} = \sqrt{\dfrac{1}{6-2}\dfrac{1{,}828{,}600}{377}} = \dfrac{5\sqrt{6{,}893{,}822}}{377} \approx 34.822399$ by above.

Substituting in the respective values, then

$$\begin{aligned} \widehat{\beta}_1 \pm t_{\alpha/2}(\text{df}) \cdot S\sqrt{c_{1\,1}} &= \frac{4570}{377} \pm 1.533 \cdot \frac{5\sqrt{6{,}893{,}822}}{377} \cdot \sqrt{\frac{6}{377}} \\ &= \frac{4570}{377} \pm \frac{15.33\sqrt{27429}}{377} \\ &= \frac{4570 \pm 15.33\sqrt{27429}}{377} \end{aligned}$$

$$\text{C.I.} = \left(\frac{4570 - 15.33\sqrt{27429}}{377},\quad \frac{4570 + 15.33\sqrt{27429}}{377}\right)$$

$$\text{C.I.} \approx (5.387509175,\ 18.85652265)$$

(d) Is there statistical evidence that the slope of the regression line is greater than 10? Run a hypotheses test at $\alpha = 0.05$.

Solution: Let $H_0 : \beta_1 = 10$ and $H_a : \beta_1 > 10$. Then our one-sided t-value is $t_\alpha(\text{df}) = t_{0.05}(4) = 2.132$.

Then the $\mathcal{T}$ statistic is defined as

$$\begin{aligned}\mathcal{T} &= \frac{\widehat{\beta}_1 - \beta_1}{S\sqrt{c_{11}}} \\ &= \frac{\dfrac{4570}{377} - 10}{\dfrac{5\sqrt{6893822}}{377}\sqrt{\dfrac{6}{377}}} \\ &= \frac{377\left(\dfrac{4570}{377} - 10\right)}{10\sqrt{27429}} \\ &= \frac{80\sqrt{27429}}{27429} \\ &\approx 0.483042.\end{aligned}$$

Then $\Pr(\mathcal{T} < t) \approx 0.327157$ by webassign technology (t-distribution, right tail, $\mathcal{T} = 0.4830$, df $= 4$). Since $p > \alpha$ then we accept the null hypothesis. Hence, there is **not** enough statistical evidence that the slope is greater than 10.