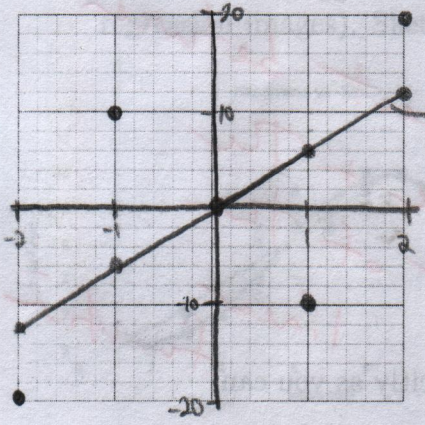
Directions: Show your work! Answers without justification will likely result in few points. Your written work also allows me the option of giving you partial credit in the event of an incorrect final answer (but good reasoning). Indicate clearly your answer to each problem (e.g., put a box around it). You **must skip** one problem (other than the first). Write "skip" clearly on the one you skip. **Good luck!**

Problem 1. (10 pts) (You can't skip this one!) Use the following data

x	-2	-1	0	1	2
y	-20	10	0	-10	20

a. (8 pts) to construct the simple linear regression model of best fit (find $y(x) = a + bx$). Show all work. You might sketch the data and your line in the grid provided.



$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad b = \frac{\sum x_i y_i}{\sum x_i^2}$$

$\bar{x} = 0$

$\bar{y} = 0$

$$b = \frac{(-2)(-20)+(-1)(10)+0+(1)(-10)+2(20)}{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2}$$

$$b = \frac{60}{10} = 6$$

$$a = \bar{y} - b\bar{x}$$

$$a = 0 - 6(0)$$

$$a = 0$$

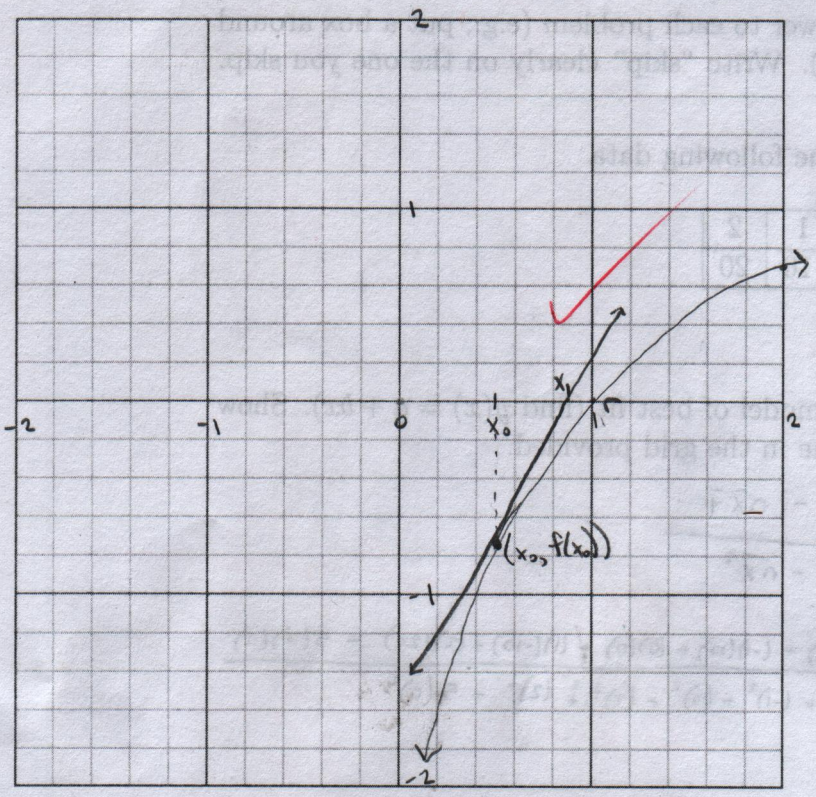
$$\boxed{y(x) = 6x}$$ ✓

b. (2 pts) Also compute R^2. Show all work.

$\bar{y} = 0$

$$R^2 = \frac{\sum(\hat{y_i} - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{(-12)^2+(-6)^2+(0)^2+(6)^2+(12)^2}{(-20)^2+(10)^2+(0)^2+(-10)^2+(20)^2} = \frac{360}{1000} = \boxed{0.36 \text{ or } 36\%}$$ ✓

Problem 2. (10 pts) We like linear functions, operations, etc. in mathematics, because they tend to make our lives easy. Illustrate **graphically** how Newton's method is an example of using a linear method to solve a non-linear problem in the case of finding the root of $\ln(x)$ from a starting point of $x_0 = \frac{1}{2}$. I suggest a "plot window" of $[0, 2] \times [-2, 2]$, using the entire grid below.



$f(x) = \ln(x)$ $\quad x_0 = \frac{1}{2}$

$f'(x) = \frac{1}{x}$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$= \frac{1}{2} - \frac{f(\frac{1}{2})}{f'(\frac{1}{2})}$$

$$= \frac{1}{2} - \frac{-0.6931}{\frac{1}{1/2}}$$

$$= 0.84657$$

No need to go beyond the first step of Newton's method. Describe as explicitly as you can

a. (3 pts) the non-linear problem,

We are trying to approximate the solution to $f(x) = \ln(x) = 0$. This is a non-linear equation, but we are going to use linear ones to approximate its root.

b. (4 pts) the associated linear problem, and

We use linear tangent lines at "guesses" for r and find roots for these tangent lines instead.

c. (3 pts) describe how we make use of the linear problem to solve the non-linear problem.

Iteratively, we take the initial guess, find the tangent line at that point on the original function, and find the root of the tangent line. This root is now our next guess, and the process is repeated.

10

Problem 3. (10 pts) I have issues with some regressions: I hope that you do, too!

a. Comment on the linear regression model featured in this graphic, and sketch in a model you might prefer (justify).

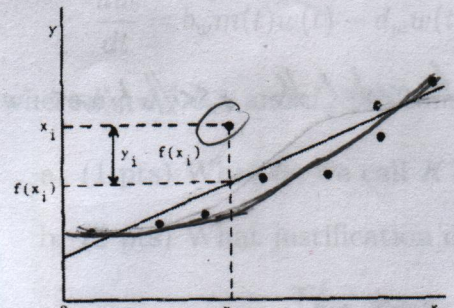
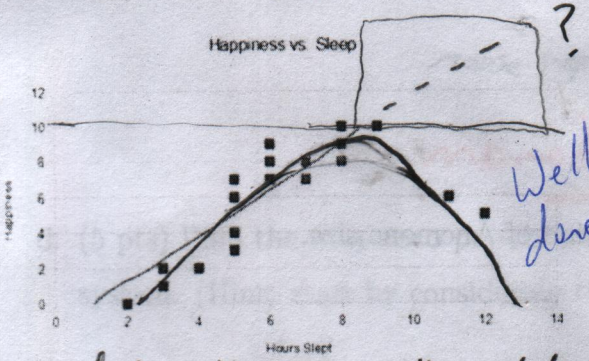


Figure 5. Deviations

I'd prefer a quadratic model, seeing as the values increase slowly at first, and then slowly increase in value. I'd say the point circled is an outlier to this model, and I would investigate how this value came about - was it entered incorrectly?

✓ good

b. Comment on the linear regression model featured in this graphic (from "study.com" – and it's not the squiggliness of the "line" that bothers me, or that it was the second hit googling regression images). Sketch in a model you might prefer (justify).



?

The model as presented doesn't take the data for hours >10, which is really disconcerting, because if the line were continued, the difference between the "line" and the data points would be >5! Also, if the scale is from 0 to 10, then the model wouldn't make sense, as it projects the happiness to be >10. I like a quadratic model to fit this data. Not only are the residuals minimized with a quadratic model, but the model also captures the idea that a lot of sleep (too much) is bad for happiness.

Well done!

c. Reflect a little on the linear regression model(s) featured in this graphic: the dashed curve is

oblem 3. (10 pts) I have issues with some regressions: I hope that you do, too!

a. Comment on the linear regression model featured in this graphic, and sketch in a model you might prefer (justify).

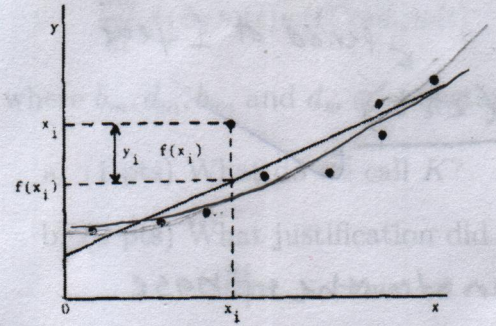


Figure 5. Deviations

The residuals will not be randomly scattered around zero. There are far more points below the linear regression line than above it.

The model I drew has curvature (possibly quadratic or exponential) which allows for the data points to fall more evenly around it.

good

Problem 3. (10 pts) I have issues with some regressions: I hope that you do, too!

a. Comment on the linear regression model featured in this graphic, and sketch in a model you might prefer (justify).

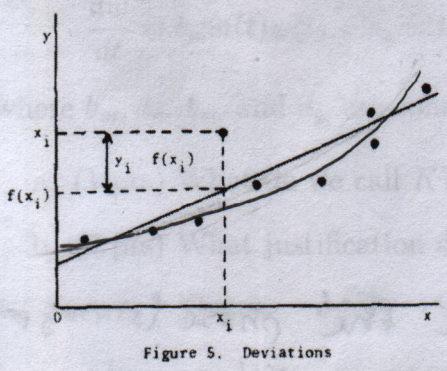
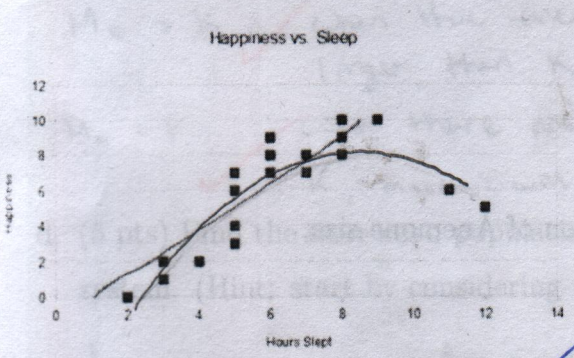


Figure 5. Deviations

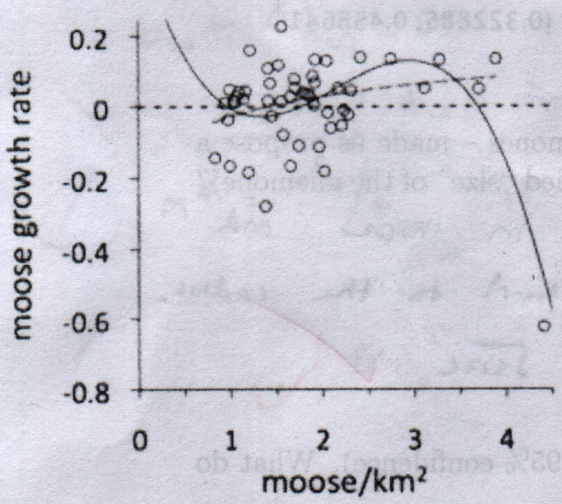
A lot of the points in this plot fall below the regression line - particularly in the "middle" of the x-values. This suggests lack-of-fit, and the residual plot would probably show a big smiley face. I would suggest a quadratic model to capture the curve in the data a little better.

b. Comment on the linear regression model featured in this graphic (from "study.com" – and it's not the squiggliness of the "line" that bothers me, or that it was the second hit googling regression images). Sketch in a model you might prefer (justify).



The linear model appears to do very well until hours slept gets larger. I would check to see if it were worth dropping the large x-points, or focusing the regression only on the beginning. If I had to model all the data, I would suggest a quadratic, or perhaps a piecewise function of some sort, we need to adjust for those 2 points

c. Reflect a little on the linear regression model(s) featured in this graphic: the dashed curve is one regression, whereas the solid curve (a cubic) is called "the most parsimonious relationship between moose abundance and population growth rate". What concerns might you have?



I am concerned about the cubic model because it appears to only be used because of one large outlier. I don't think one point should have that much control over the model shape & structure.

The dashed regression appears to fit the data better, however the residuals are going to be non-constant. Variability about the fit is much larger for small values of moose/km² than for larger values.

10

Item 3. (10 pts) I have issues with some regressions: I hope that you do, too!

a. Comment on the linear regression model featured in this graphic, and sketch in a model you might prefer (justify).

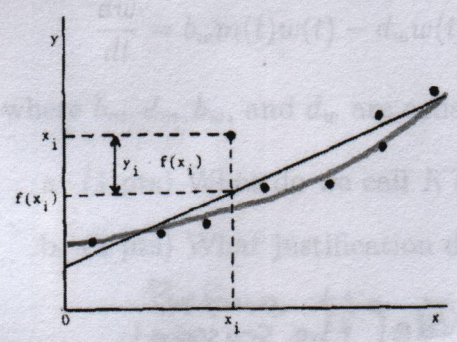
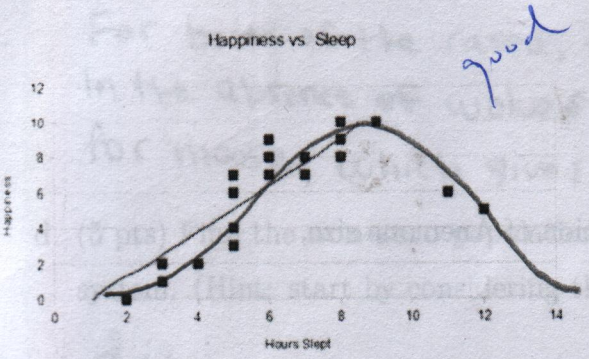


Figure 5. Deviations

This data definitely has some curvature to it, so a linear model may not be right. If we were to look at the residuals for the fitted model, we would see a curve, as the residuals on the extreme values are negative, and all of the data near X_i have positive residuals.

b. Comment on the linear regression model featured in this graphic (from "study.com" – and it's not the squiggliness of the "line" that bothers me, or that it was the second hit googling regression images). Sketch in a model you might prefer (justify).

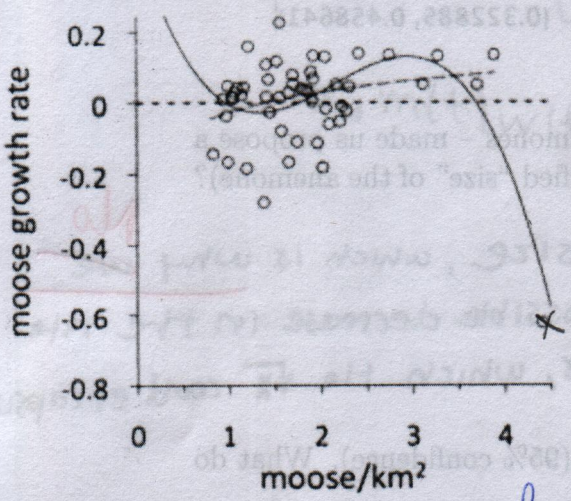


good

This linear model seems to completely disregard the two data points near 11 hours slept. If they did not include these two points then the model would appear to be fine. But since it is included a model with some curvature is needed. To me, it appears that the normal distribution seems to fit well. It is also because you can't have negative happiness, I think.

c. Reflect a little on the linear regression model(s) featured in this graphic: the dashed curve is one regression, whereas the solid curve (a cubic) is called "the most parsimonious relationship between moose abundance and population growth rate". What concerns might you have?

Let's hope not!



The cubic model is fitting the data set with a very possible outlier (it appears to be greater than 3 standard deviations away from the mean). Without that data point in there, the cubic model would not nearly fit the data as well as the linear model. Unless biologists have certain information that would lead them to believe that the point is needed and accurate, there is no need to account for it in the model, which is what the linear model does.

Good work

10

Problem 4. (10 pts)

a. For the Keeling data, with data given monthly (but assume time t measured in years),

i. (2 pts) what precise pair of trigonometric functions did we use (along with a quadratic trend model) to model **the oscillations** in the data?

$$y(t) = a + bt + ct^2 + d\sin(2\pi t) + e\cos(2\pi t)$$

← period of 1 year

ii. (2 pts) Why did we use **a pair** of such functions?

If we only use one of the functions the phase will be fixed and will not be able to change where the oscillations start. + amplitude

iii. (2 pts) Given the pair $a\sin(x)$ and $b\cos(x)$, what is the **amplitude** of the combined oscillation?

$$\sqrt{a^2+b^2}$$

b. For the sea anemone data, Duration of Response time as a function of Anemone size,

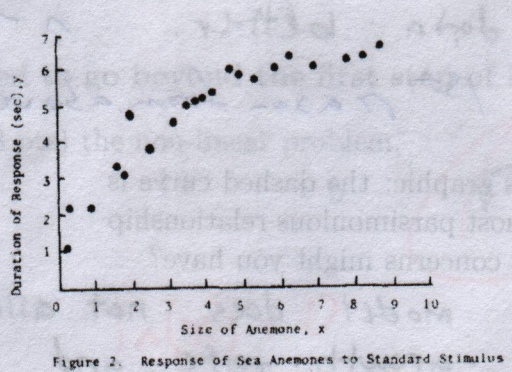


Figure 2. Response of Sea Anemones to Standard Stimulus

FittedModel[$2.94754\,x^{0.390763}$]

	Estimate	Standard Error	Confidence Interval
a	2.94754	0.15844	{2.61592, 3.27916}
b	0.390763	0.0324306	{0.322885, 0.458641}

sea anemone

i. (2 pts) what **structural** idea – that is, characteristic of anemones – made us propose a model of the form $Duration(x) = a\sqrt{x}$ (based on the unspecified "size" of the anemone)?

The data looks like a square root curve.

If response time is proportional to r, then

Area $= \pi r^2$

$$R \sim \sqrt{A} = \sqrt{\pi r^2} = r$$

ii. (2 pts) Power model regression produced the results shown (95% confidence). What do the results suggest about our proposed model?

If our model ($b = 1/2$) was appropriate then we'd expect $1/2$ to be within the C.I. for b. It is not contained within, so our model is not appropriate.

9.5

Problem 5. (10 pts) One change we made to the equations for the Moose/Wolf compartmental model was to replace one Lotka-Volterra birth process by a logistic growth process: that is,

$$\begin{aligned} \frac{dm}{dt} &= b_m m(t) - d_m m(t) w(t) \\ \frac{dw}{dt} &= b_w m(t) w(t) - d_w w(t) \end{aligned} \longrightarrow \begin{aligned} \frac{dm}{dt} &= b_m m(t)\left(1 - \frac{m(t)}{K}\right) - d_m m(t) w(t) \\ \frac{dw}{dt} &= b_w m(t) w(t) - d_w w(t) \end{aligned}$$

where b_m, d_m, b_w, and d_w are constants.

a. (1 pts) What do we call K? Carrying capacity ✓

b. (2 pts) What justification did we have for making that change?

When there are no wolves, the moose will grow exponentially, so we must introduce carrying capacity to prevent this. ✓

c. (2 pts) Describe the dynamics of the moose population in the absence of wolves. Consider both cases: initial population values below or above K.

If the initial population is below K, then $\frac{dm}{dt}$ will be positive, so the moose population will increase, but not past the carrying capacity. ✓

If the initial population is above K, then $\frac{dm}{dt}$ will be negative, so the population will decrease.

but not past the carrying capacity.

d. (5 pts) Find the **non-zero** populations of wolves and moose that are **equilibria** for this new system. (Hint: start by considering the $\frac{dw}{dt}$ equation, which didn't change.)

$$\frac{dw}{dt} = b_w wm - d_w w = 0$$

$$= w(b_w m - d_w) = 0$$

$$b_w m - d_w = 0$$

$$\boxed{m^* = \frac{d_w}{b_w}}$$

✓

$$\frac{dm}{dt} = b_m m\left(1 - \frac{m}{K}\right) - d_m m w = 0$$

$$= m\left(b_m\left(1 - \frac{m}{K}\right) - d_m w\right) = 0$$

$$b_m\left(1 - \frac{d_w}{K b_w}\right) - d_m w = 0$$

$$d_m w = b_m\left(1 - \frac{d_w}{K b_w}\right)$$

$$\boxed{w^* = \frac{b_m}{d_m}\left(1 - \frac{d_w}{K b_w}\right)}$$

✓

Well done!

Problem 6. (10 pts)

My bad!

a. (6 pts) Demonstrate that one can estimate the coefficients in the model $y(t) = ax^b$ by linearization: that is, that one can estimate the coefficients a and b of the non-linear model from data $\{x_i, y_i\}_{i=1,\ldots,n}$ using **linear** regression on transformed data, then back-transforming the parameters from the appropriate linear model, α and β.

Identify how you would transform the data for use in the linear regression. Natural Log!

$y(x) = ax^b$

$\ln y(x) = \ln(ax^b)$

$\ln y(x) = \ln a + \ln x^b$

$\ln y(x) = \ln a + b \ln x$

$\ln y(x) = \alpha + \beta \ln x$

Linear regression to get parameters

exponentiate

$e^{\ln y(x)} = e^{\alpha + \beta \ln x}$

$y(x) = e^{\alpha} e^{\beta \ln x} = e^{\alpha} \cdot x^{\beta}$

b. i. (2 pts) How do you back-transform to obtain estimates for a and b?

$a = e^{\alpha}$ b/c $\alpha = \ln a$

$b = \beta$ b/c $b = \beta$

$y(x) = e^{\alpha} x^{\beta}$

ii. (2 pts) What important change occurs in the confidence intervals obtained from linear regression if you back-transform them in the same way you back-transform the parameters? How do they differ from traditional CIs for linear regression?

These intervals are no longer symmetric around the estimate. This is because the error structure changes to multiplicative rather than additive.

10

roblem 7. (10 pts) In class we demonstrated that the simple linear regression line must pass through the center of mass of the data, $(\overline{x}, \overline{y})$. Thus there is really only one "degree of freedom" for the line – the slope m. Use a little univariate calculus to **derive** the value of m, by minimizing

$$f(m) = \sum_{i=1}^{N} (y_i - \overline{y} - m(x_i - \overline{x}))^2$$

$$f'(m) = \sum_{i=1}^{N} \left[(y_i - \bar{y} - m(x_i - \bar{x}))^2 \right]'$$

$$= \sum_{i=1}^{N} 2(y_i - \bar{y} - m(x_i - \bar{x})) \cdot (x_i - \bar{x})$$

$$0 = 2 \sum_{i=1}^{N} (y_i - \bar{y} - m(x_i - \bar{x}))(x_i - \bar{x})$$

$$0 = \sum_{i=1}^{N} y_i (x_i - \bar{x}) - \sum_{i=1}^{N} \bar{y}(x_i - \bar{x}) - \sum_{i=1}^{N} m(x_i - \bar{x})^2$$

$$m \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} y_i (x_i - \bar{x}) - \sum_{i=1}^{N} \bar{y}(x_i - \bar{x})$$

$$m = \frac{\sum_{i=1}^{N} y_i (x_i - \bar{x}) - \sum_{i=1}^{N} \bar{y}(x_i - \bar{x})}{m \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

-.5 rats! otherwise perfect!