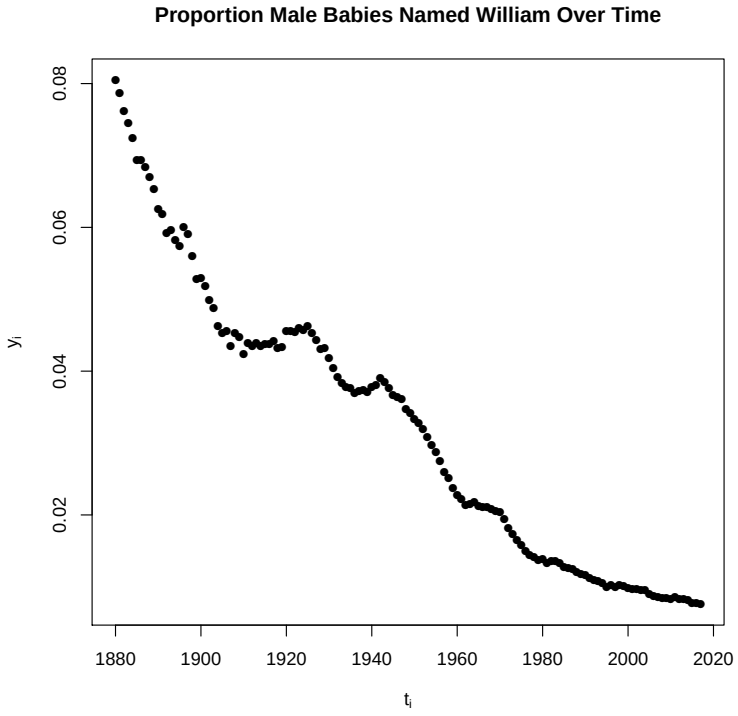
UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
August 2020

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. The number of points for each part of each question is listed inline with the part.

Question:	1	2	3	4	5	Total
Points:	24	25	17	15	19	100
Score:						

1. Consider the problem of modeling the proportion of male babies named William each year $\mathtt{y}_i$ as a function of time $\mathtt{t}_i$ from 1880 to 2017.



(a) (4 points) Consider the following model:

$$\mathtt{y}_i \overset{indep.}{\sim} \text{Normal}\left(\sum_{j=0}^{2} \beta_j \mathtt{t}_i^j, \sigma^2\right). \tag{1}$$

The least squares estimates of the parameters of this model fit to the data are provided in the R output printed below.

```
Call:
lm(formula = prop ~ year + I(year^2), data = wils)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0088088 -0.0019435 -0.0000959  0.0020842  0.0099837

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.878e+00  8.083e-01   9.747  < 2e-16 ***
year        -7.583e-03  8.299e-04  -9.138 8.41e-16 ***
I(year^2)    1.825e-06  2.129e-07   8.569 2.12e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.00355 on 135 degrees of freedom
Multiple R-squared:  0.967,Adjusted R-squared:  0.9665
F-statistic:  1977 on 2 and 135 DF,  p-value: < 2.2e-16
```

Give values of the least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and interpret them in words. Explain whether or not these describe contrasts that we might be interested in practice.

(b) (4 points) Write $\mathbb{E}\left[\mathtt{y}_{i+1} - \mathtt{y}_i\right]$ under the model given in (1) as a function of the time t_i and the parameters, $\beta_0, \beta_1, \beta_2$, and σ^2 (it is possible that not all of these parameters will appear in $\mathbb{E}\left[\mathtt{y}_{i+1} - \mathtt{y}_i\right]$). Use the fact that the data is observed annually, so $t_{i+1} = t_i + 1$, simplify as much as possible. Interpret $\mathbb{E}\left[\mathtt{y}_{i+1} - \mathtt{y}_i\right]$, and explain whether or not this describes a contrast that we might be interested in practice.

(c) (4 points) Now consider an alternative model:

$$\log\left(\mathtt{y}_i\right) \overset{indep.}{\sim} \text{Normal}\left(\sum_{j=0}^{1} \alpha_j \mathtt{t}_i^j, \omega^2\right). \tag{2}$$

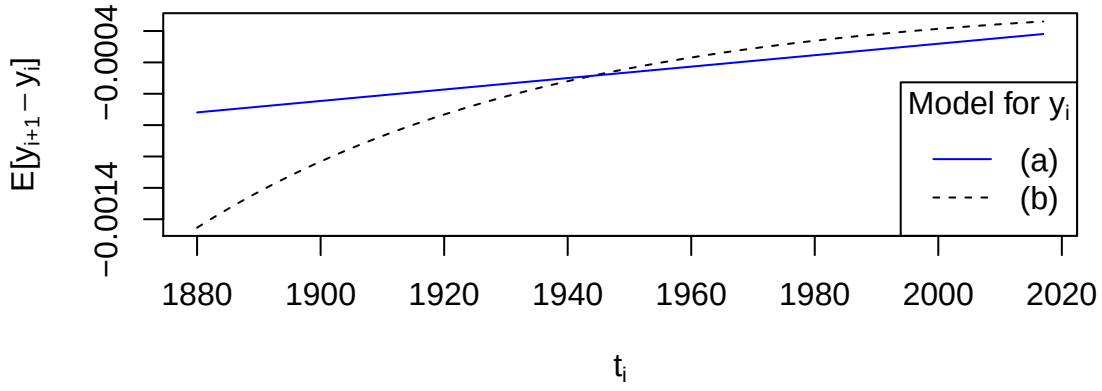
The least squares estimates of the parameters of this model fit to the data are provided in the R output printed below.

```
Call:
lm(formula = I(log(prop)) ~ year, data = wils)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20320 -0.10619 -0.06137  0.11937  0.29412

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.7528543  0.5801098   51.29   <2e-16 ***
year        -0.0171389  0.0002977  -57.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1393 on 136 degrees of freedom
Multiple R-squared:  0.9606,Adjusted R-squared:  0.9603
F-statistic:  3315 on 1 and 136 DF,  p-value: < 2.2e-16
```

Give the value of the least squares estimate $\hat{\alpha}_1$ and interpret it in words. Explain whether or not it describes a contrast that we might be interested in practice.

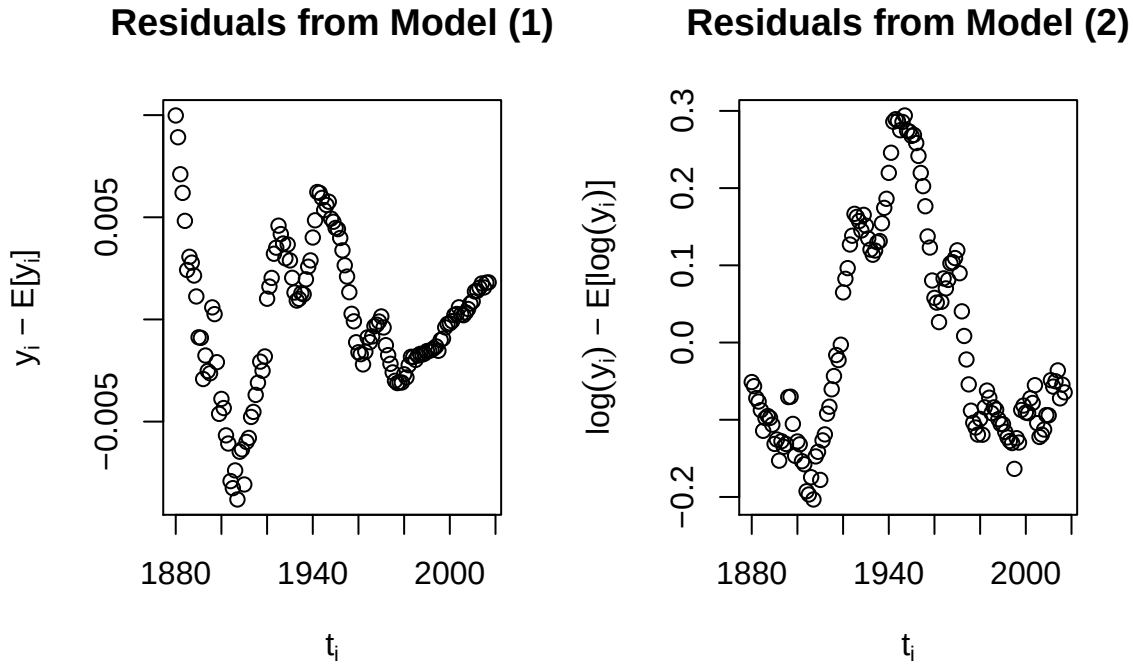
(d) (3 points) Write $\mathbb{E}\left[\mathtt{y}_{i+1} - \mathtt{y}_i\right]$ under the model given in (2) using the mean derived in the previous part, using the fact that $\mathbb{E}\left[\mathtt{y}_i\right] = \exp\left\{\sum_{j=0}^{1} \alpha_j \mathtt{t}_i^j + \frac{\omega^2}{2}\right\}$. The expression will be a function of the time t_i and the parameters, α_0, α_1, and ω^2. Simplify as much as possible. Note that the data is observed annually, so $t_{i+1} = t_i + 1$. Explain how $\mathbb{E}\left[\mathtt{y}_{i+1} - \mathtt{y}_i\right]$ compares under the models given in (1) and (2), with respect to how they depend on time t_i.

(e) (3 points) A plot of $\mathbb{E}\left[\mathrm{y}_{i+1} - \mathrm{y}_i\right]$ computed from the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}^2$, $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\omega}^2$ is given below.



For curves (a) and (b), indicate whether or not they correspond to estimates under the model given in (1) or (2) and explain your reasoning.

(f) (3 points) Based on the available information provided up to this point as well as the residual plots provided below, indicate whether or not you would prefer the model given in (1) or the model given in (2) for analysis of this data, and explain your reasoning.



(g) (3 points) Identify and describe an assumption that is common to the models given

in (1) and (2) and appears to be violated. Cite whatever available information leads you to this conclusion, and explain how you expect such a violation might affect the results.

2. Consider the set of diagnostic plots below. Each represents a plot of relationships between the residuals r_i and the univariate predictor x_i or fitted value $\hat{y}_i$ of a simple linear regression model.

   (a) (4 points) There are four key assumptions required for standard inference using linear regression. List these 4 assumptions.

   (b) (4 points) For each of the assumptions in the previous part, can you detect violations of this assumption using residual plots like the ones in these figures? If not, give one example of a diagnostic tool you could use to diagnose violations of this assumption.

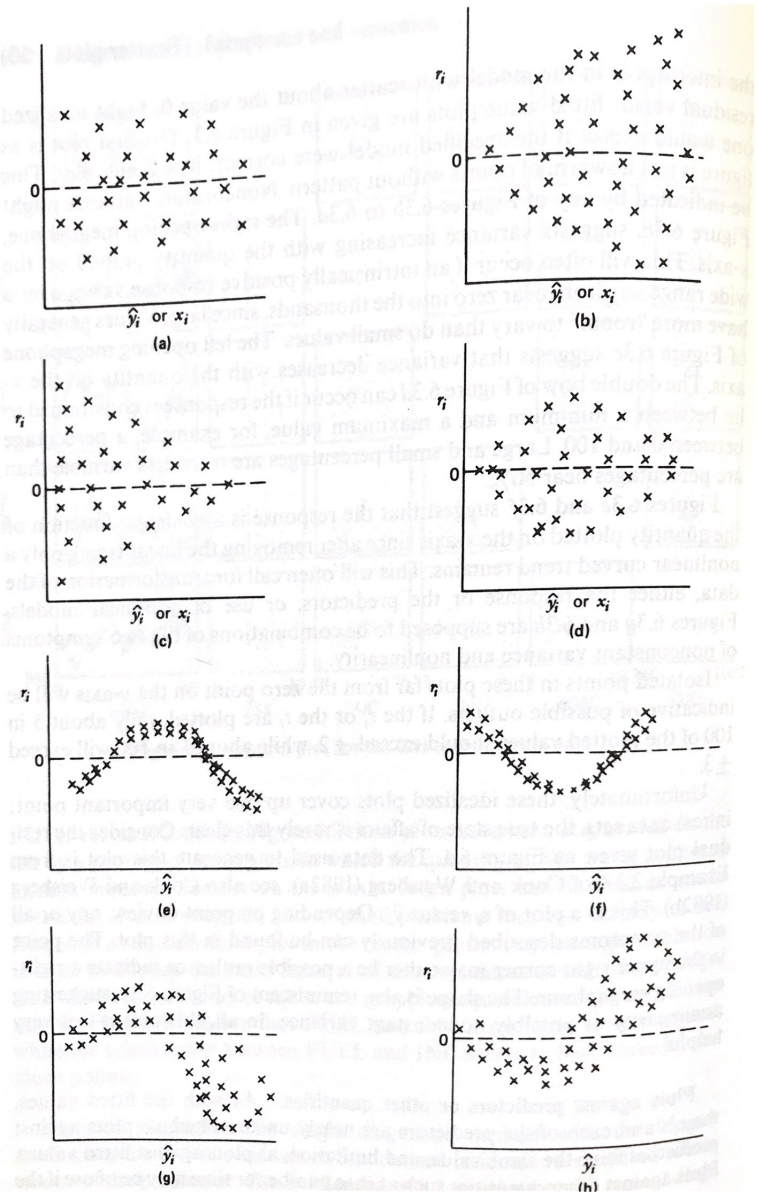
   (c) (2 points) Are there any plots here that do not show violation of any assumptions of linear regression? Which plot(s)?

   (d) (4 points) Are there any plots here that show violation of more than one assumption? Which plot(s)? Which assumptions do they violate?

   (e) (3 points) What would you do to fix the violations in the previous part? If you identified more than one plot in the previous part, you may focus on only one of the plots in the previous part.

   (f) (4 points) A researcher suggests replacing dependent variable Y with $\log(Y)$ to remedy the violation in one of these plots. Which plot or plots would you expect to fix this way? Why?

   (g) (4 points) One researcher suggests replacing predictor X with X^2 to remedy the violation in another of these plots. Another researcher suggests keeping X in the model and adding X^2 as a second predictor (keeping X in the model) instead. Which plot or plots do you think might be fixed with one of these strategies? Which of these two strategies would you prefer? Why? Are there any special considerations or additional adjustments you would recommend? If you identified more than one plot, you may focus on only one of the plots to discuss in detail.


0
$\hat{y}_i$ or x_i
(a)
r_i
0
$\hat{y}_i$ or x_i
(b)
r_i
0
$\hat{y}_i$ or x_i
(c)
r_i
0
$\hat{y}_i$ or x_i
(d)
r_i
0
$\hat{y}_i$
(e)
r_i
0
$\hat{y}_i$
(f)
r_i
0
$\hat{y}_i$
(g)
r_i
0
$\hat{y}_i$
(h)

3. Consider N independent binary random variables $Y_1, Y_2, \ldots Y_N$ such that

$$P(Y_i = 1) = \pi_i \quad \text{and} \quad P(Y_i = 0) = 1 - \pi_i.$$

The probability mass function of Y_i can be written as $\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$. This is an exponential family distribution with natural parameter $\log\left(\frac{\pi_i}{1-\pi_i}\right)$. Note that the function $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is called the *logit* of π_i.

(a) (2 points) Show that $E(Y_i) = \pi_i$.

(b) (3 points) If the link function is defined as

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = x^T\beta$$

show that this is equivalent to modeling the probability π as

$$\pi = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}.$$

(c) (3 points) Consider the case when when $x^T\beta = \beta_0 + \beta_1 x$. Sketch a graph of π against x in this case, taking $\beta_0 = 0$ and $\beta_1 = k$, a positive constant. Label both coordinates of the point where $\pi = .5$, and the values 0 and 1 on the vertical axis.

(d) Now add a dotted line to the previous plot representing the relationship between π and x with $\beta_0 = 0$ and $\beta_1 = 2k$.

(e) (2 points) Now add a dashed line to the previous plot representing the relationship between π and x with $\beta_0 = -1$ and $\beta_1 = k$.

(f) (2 points) Finally, add another solid line to the previous plot representing the relationship between π and x with $\beta_0 = 0$ and $\beta_1 = -k$.

(g) (5 points) Suppose one of these lines represents data on the effectiveness of an insecticide (a chemical used to kill insects), where x is the dose of an insecticide and π is the probability of an insect dying. State which of the lines from parts (d)-(f) most plausibly represents this relationship and describe why. Interpret the resulting model in the context of the insecticide.

4. In this problem, you'll write a function to multiply two matrices to compute $\hat{y} = X\beta$.

   (a) (10 points) In R or python write a function that takes two parameters `A` and `B`. `A` and `B` are each lists of lists of size $n \times n$. For example, in python you may have `A = B = [ [1, 2, 3], [4, 5, 6], [7, 8, 9] ]`. The function should return the matrix product.

   Assume that the sizes of the matrices are compatible with matrix multiplication. You may only use scalar product operations inside the function and you may not employ functions in special modules or libraries.

   (b) (5 points) Provide an analysis of the time complexity of the code that you wrote in the previous part to answer the question: how does the time complexity scale with the number of elements in the matrices `A` and `B`? Use Big O notation to express your results.

5. Suppose you are consulting with a researcher who wants to use cross-validation to assess the generalization performance of their prediction algorithm (regression model). They want to know many groupings or folds they should use for their cross-validation analysis. The sample size is $N = 100$ and the number of groupings or folds is k. Recall, the first fold is treated as a validation set, and the model is fit on the remaining $k - 1$ folds. This process is repeated for all folds.

   (a) (5 points) What is the quantity that the cross-validation is aiming to estimate? You can describe this using words or equations.

   (b) (7 points) Let $k = N$. Describe the (1) bias and (2) variance of the resulting cross-validation estimate from the previous part. Is the bias expected to be high or low and why? Is the variance expected to be high or low and why?

   (c) (7 points) Let $k = 2$. Describe the (1) bias and (2) variance of the resulting cross-validation estimate from the previous part. Is the bias expected to be high or low and why? Is the variance expected to be high or low and why?