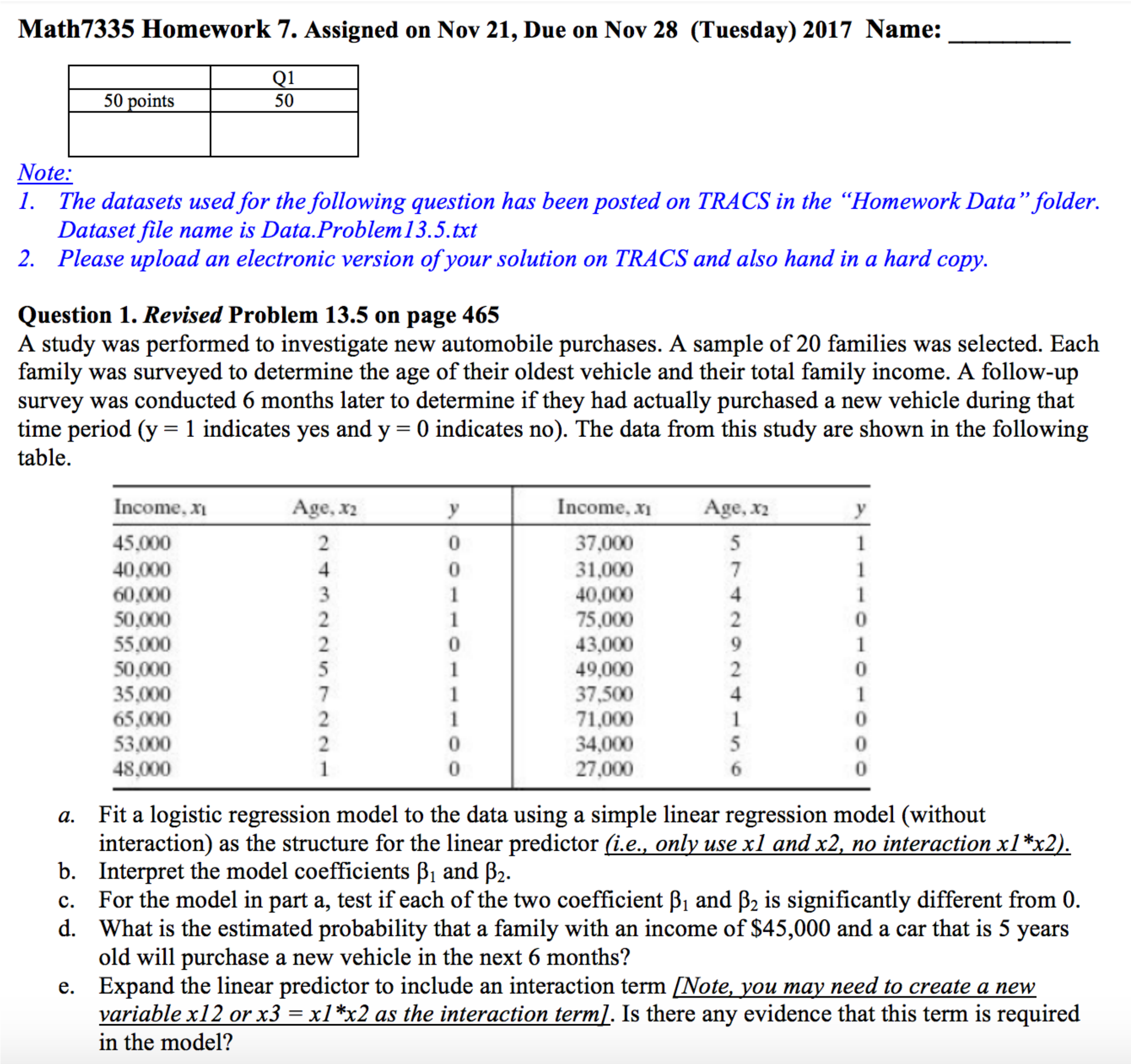

Math7335 Homework 7. Assigned on Nov 21, Due on Nov 28 (Tuesday) 2017 Name: __________

	Q1
50 points	50

Note:

1. *The datasets used for the following question has been posted on TRACS in the "Homework Data" folder. Dataset file name is Data.Problem13.5.txt*
2. *Please upload an electronic version of your solution on TRACS and also hand in a hard copy.*

Question 1. *Revised* Problem 13.5 on page 465

A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and their total family income. A follow-up survey was conducted 6 months later to determine if they had actually purchased a new vehicle during that time period (y = 1 indicates yes and y = 0 indicates no). The data from this study are shown in the following table.

Income, x_1	Age, x_2	y	Income, x_1	Age, x_2	y
45,000	2	0	37,000	5	1
40,000	4	0	31,000	7	1
60,000	3	1	40,000	4	1
50,000	2	1	75,000	2	0
55,000	2	0	43,000	9	1
50,000	5	1	49,000	2	0
35,000	7	1	37,500	4	1
65,000	2	1	71,000	1	0
53,000	2	0	34,000	5	0
48,000	1	0	27,000	6	0

a. Fit a logistic regression model to the data using a simple linear regression model (without interaction) as the structure for the linear predictor *(i.e., only use x1 and x2, no interaction x1*x2).*

b. Interpret the model coefficients β_1 and β_2.

c. For the model in part a, test if each of the two coefficient β_1 and β_2 is significantly different from 0.

d. What is the estimated probability that a family with an income of $45,000 and a car that is 5 years old will purchase a new vehicle in the next 6 months?

e. Expand the linear predictor to include an interaction term *[Note, you may need to create a new variable x12 or x3 = x1*x2 as the interaction term]*. Is there any evidence that this term is required in the model?

Question 1

Q1 Part (a)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.0471	4.6742	-1.51	0.1316
x1	0.0001	0.0001	1.16	0.2466
x2	0.9879	0.5274	1.87	0.0610

```
    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 21.082  on 17  degrees of freedom
AIC: 27.082

Number of Fisher Scoring iterations: 5
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			19	27.73
x1	1	0.73	18	26.99
x2	1	5.91	17	21.08

Q1 Part (b)

Since every one unit increase in x1 results in a 0.0001 increase in ln($\frac{p}{1-p}$), it looks like the income of the family not very relevant, where as the age of the car (x2) is highly relevant.

Q1 Part (c)

We will test if $\hat{\beta}_1$ and $\hat{\beta}_2$ are statistically significant (i.e. testing $H_0 : \hat{\beta}_1 = 0$ and $\hat{\beta}_2 = 0$)
Our Wald statistics (given by the z value in the summary) are:
$\hat{\beta}_1$: 1.16
$\hat{\beta}_2$: 1.87
We will test at $\alpha = 0.05$
$Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = -1.96$
$\hat{\beta}_1 : |1.16| > 1.96 \longrightarrow$ Not true.
$\hat{\beta}_2 : |1.87| > 1.96 \longrightarrow$ Not true.
Thus,
we fail to reject H_0.

Q1 Part (d)

```
new <- data.frame(x1 = 45000, x2 = 5)
predict(glm.1, new)
```

gives us 1.214124
So,

$$\ln(p/(1-p)) = 1.214124$$

$$p/(1-p) = e^{1.214124}$$

$$p = e^{1.214124}(1-p)$$

$$p = e^{1.214124} - pe^{1.214124}$$

$$(1+e^{1.214124})p = e^{1.214124}$$

$$p = \frac{e^{1.214124}}{(1+e^{1.214124})}$$

$$p(y=1 \mid x1 = 45,000,\ x2 = 5) = 0.77102783$$

They have a 77.1% chance of buying a new vehicle in the next 6 months.

Q1 Part (e)

```
glm.2 <- glm(y ~ x1 + x2 + (x1)*(x2), family=binomial(link=logit), data=q1data)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3144	6.3940	0.05	0.9608
x1	-0.0001	0.0001	-1.00	0.3177
x2	-2.4617	2.0815	-1.18	0.2369
x1:x2	0.0001	0.0001	1.61	0.1074

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 16.551  on 16  degrees of freedom
AIC: 24.551

Number of Fisher Scoring iterations: 6
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			19	27.73
x1	1	0.73	18	26.99
x2	1	5.91	17	21.08
x1:x2	1	4.53	16	16.55

Yes - both parameters for x1 and x2 switched signs when the interaction term was added, indicating a possible dependency. The interaction term also has a p value of 0.1074, which itself is not a strong p value, but out of the 4 parameters, it is the most significant.