Final Exam

MATH 60604A - Statistical Modelling
Fall 2019

Instructions:

- Be sure to show all of your work and justify your answers.
- Any time you carry out a statistical test, be sure to formally write out the underlying hypotheses, give the value of the test statistic and the corresponding p-value.
- Always make relevant conclusions in the context of the problem.

Breakdown of marks:

Question	Points
1	10
2	4
3	6
4	21
5	8
6	11
7	11
Total	**71**

Question 1

[10 points] State whether the following statements are true or false.

(a) In a simple linear regression model of the form $Y = \beta_0 + \beta_1 X_1 + \epsilon$, if $R^2 = 1$ then $\hat{\beta}_1 > 0$.

(b) In a logistic regression model, if a 95% confidence interval for the odds ratio corresponding to the variable X_j includes the value 0 then we can reject the null hypothesis $H_0 : \beta_j = 0$.

(c) When comparing two models which are identical in all ways except that one assumes an $AR(1)$ correlation structure for the errors while the other assumes an unstructured correlation on the errors, the value of $-2LL$ will always be higher for the model with the unstructured correlation (where $-2LL$ is an abbreviation for -2 times the log-likelihood evaluated at the maximum likelihood estimator $\hat{\theta}_{mle}$)).

(d) For a survival function $S(t)$, if $t_1 < t_2$ then $S(t_1) \leq S(t_2)$.

(e) In a linear regression model which includes both main effects and the interaction between two categorical variables, respectively with 5 and 2 levels, the model would include a total of 10 regression coefficients (β).

(f) In survival analysis, the log-rank test allows to formally test whether a categorical variable has a significant impact on the survival function.

(g) For a linear mixed model with both a random intercept and a random effect for X_1, the covariance structure on the observations, $Cov(Y_i)$, will be the same for all groups i.

(h) If $X_1 = 10 + 3X_2$, then it's impossible to uniquely estimate the regression coefficients in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

(i) In a generalized linear model with log link function for count data, using a Poisson distribution or a Negative Binomial distribution both lead to the same regression parameter estimates, $\hat{\beta}$.

(j) In the Cox proportional hazards model, the hazard function is time-invariant.

(k) Ignoring left censored observations will lead to an under-estimation of the survival function.

Question 2

[4 points] Suppose you are interested in modelling longitudinal data of the form

$$(Y_{ij}, X_{ij1}, \ldots, X_{ijp}),$$

for $i = 1, \ldots, m$ groups and three time points $j = 1, 2, 3$. You are considering various models with different correlation structures on the errors ϵ_i. Suppose that all of the models you are comparing include the same explanatory variables and that there are no random effects.

For each of the following comparisons, state whether or not the models can be formally compared using the likelihood ratio test (LRT). Whenever you can test the two models using the LRT, state the underlying hypotheses (H_0 and H_1) in terms of the covariance and/or correlation parameters.

(a) AR(1) vs. ARH(1)

(b) Compound symmetry (CS) vs. AR(1)

Question 3

[6 points] Recall that a Bernoulli random variable $Y \sim \mathcal{B}(\pi)$ takes on values $\{0, 1\}$ and has probability function

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}.$$

Suppose you observe a random sample of size 10, with 6 zeros and 4 ones:

$$y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 1, y_6 = 1, y_7 = 0, y_8 = 0, y_9 = 0, y_{10} = 0$$

(a) Write down the likelihood function.

(b) Write down the log-likelihood function.

(c) Compute the maximum likelihood estimator for π. Show all your work!

Question 4

[21 points] Data on the sales of houses in Saratoga County, New York were collected throughout the year 2006. In addition to the actual sale price, information on several other characteristics of the homes were collected. Consider the following variables:

`price`:	price of the house (in \$1,000's of US dollars)
`lotSize`:	size of the lot (in square feet)
`age`:	age of the house (in years)
`livingArea`:	living area (in square feet)
`bedrooms`:	number of bedrooms
`rooms`:	number of rooms
`fireplaces`:	number of fireplaces
`centralAir`:	indicator of whether the house has central air (`No` or `Yes`)
`fuel`:	type of fuel used for heating (`electric`, `gas`, `oil`)

A linear regression model was fit to the data including the variables `age`, `bedrooms`, `fireplaces`, `centralAir` and `fuel`. Some SAS output is given below.

MODEL 1:

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		54157.07168	B	8401.937460	6.45	<.0001
age		-391.05629		73.746600	-5.30	<.0001
bedrooms		36920.61765		2594.982608	14.23	<.0001
fireplaces		25402.47689	B	4844.465577	5.24	<.0001
centralAir	Yes	24026.88175	B	7033.959561	3.42	0.0007
centralAir	No	0.00000	B	.	.	.
fuel	gas	26408.60328	B	5539.674670	4.77	<.0001
fuel	oil	19655.91746	B	7601.134798	2.59	0.0098
fuel	electric	0.00000	B	.	.	.
fireplace*centralAir	Yes	24946.29377	B	7709.147425	3.24	0.0012
fireplace*centralAir	No	0.00000	B	.	.	.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	187978759693	187978759693	28.12	<.0001
bedrooms	1	1.3532663E12	1.3532663E12	202.43	<.0001
fireplaces	1	603337338092	603337338092	90.25	<.0001
centralAir	1	78002644674	78002644674	11.67	0.0007
fuel	2	152166001289	76083000644	11.38	<.0001
fireplace*centralAir	1	70002508523	70002508523	10.47	0.0012

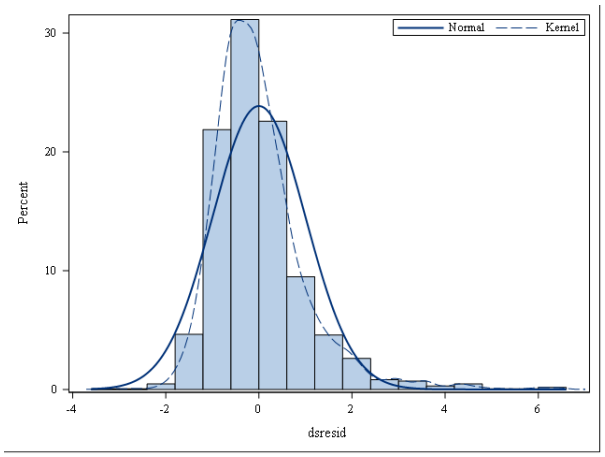
(a) Interpret the parameter for the variable `centralAir Yes`. **[3 points]**

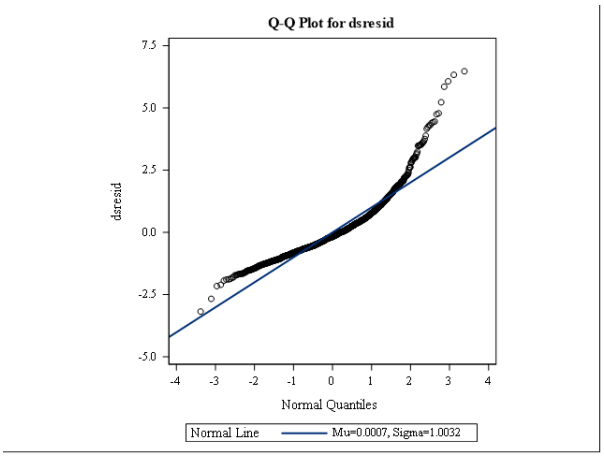
(b) Does the impact of having central air on the price of a home depend on the number of fireplaces in the house? Justify your answer. **[4 points]**

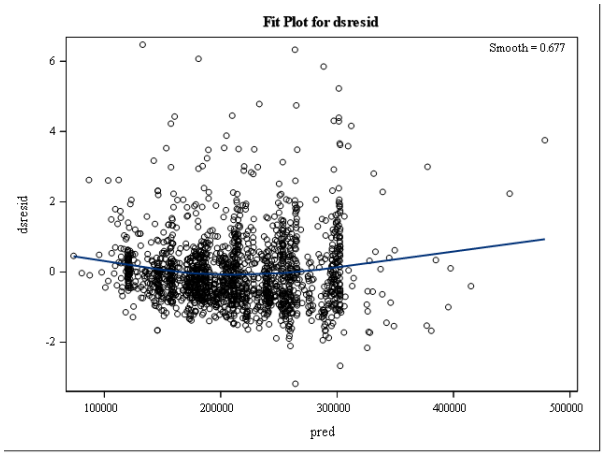
(c) Estimate the difference in the mean price of two homes that are the same in all ways, except that one heats with gas and the other with oil. **[2 points]**

(d) Does the type of fuel used for heating have a significant effect on the price of the house? Justify your answer. **[4 points]**

(e) Diagnostic graphs for the residuals (JSR) for this model are provided in the following figure. Based on these plots, what can you say about the model assumptions? Justify your answer. **[4 points]**







(f) A second model was fit to the data, this time including the explanatory variables `livingArea`, `bedrooms` and `rooms`. SAS output for this model is given below. Explain (in one or two sentences) why the estimated coefficient for `bedrooms` in Model 2 is so different from that in Model 1. **[2 points]**

MODEL 2:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	34421	6640.05108	5.18	<.0001	0
livingArea	1	119.05383	4.13160	28.82	<.0001	2.42048
bedrooms	1	-17330	2874.90771	-6.03	<.0001	2.03722
rooms	1	3305.52369	1126.48279	2.93	0.0034	2.51228

(g) Can you formally compare Models 1 and 2 using an F-test? If so, write out the underlying hypotheses (null and alternative) for the test. If not, explain why. **[2 points]**

Question 5

[8 points] A bank is interested in assessing the risk that customers will default on their credit card payments. In particular, they are interested in how a customer's income, student status and credit card balance impact the probability that they default on their debt. Consider the following variables:

`default:`	indicator for whether customer defaulted on their debt (0=no, 1=yes)
`balance:`	average balance on credit card (in \$100's)
`student:`	customer's student status (0=no, 1=yes)
`income:`	customer's income (in \$10,000's)

A logistic regression model was fit to the data to model the probability that a customer defaults, i.e. $P(\texttt{default} = 1)$. The results are summarized below.

Parameter Estimates

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-10.8690	0.4923	9996	-22.08	<.0001
balance	0.5737	0.02319	9996	24.74	<.0001
student	-0.6468	0.2363	9996	-2.74	0.0062
income	0.03033	0.08203	9996	0.37	0.7115

(a) Write an expression for the fitted model on the log-odds scale. **[2 points]**

(b) Interpret the intercept in this model. **[3 points]**

(c) What is the estimated probability that a non-student with an income of \$90,000 and an average balance of \$1,000 on their credit card will default? (Careful with the scales of the variables `balance` and `income` in the model!) **[3 points]**

Question 6

[**11 points**] An insurance company is interested in understanding the impact of a health care reform on the number of doctor visits. They collected data from patients and recorded the number doctor visits over the course of a 3-month period, both before and after the reform. They also collected information on other relevant variables, all of which are listed below.

`numvisit:`	number of visits to the doctor over a 3-month period
`reform`	indicator variable (1= after the reform, 0= before the reform)
`badh:`	health status indicator (1=bad health, 0=good health)
`age:`	patient's age
`educ:`	patient's education level (categorized into 3 levels)

The SAS output below shows the results from a Poisson generalized linear model with log link function for the response variable `numvisit`.

Parameter Estimates

Effect	educ	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.5019	0.05186	2221	9.68	<.0001
reform		-0.1388	0.02656	2221	-5.22	<.0001
badh		1.1283	0.03024	2221	37.30	<.0001
age		0.005787	0.001247	2221	4.64	<.0001
educ	1	0.05120	0.03588	2221	1.43	0.1537
educ	2	0.1598	0.03135	2221	5.10	<.0001
educ	3	0	.	.	.	.

(a) Based on this model, what is the estimated mean number of doctor visits for a 30 year old patient who is in good health and has an education of level 3, before the reform is put in place? What about after the reform? [**4 points**]

(b) Interpret the coefficient for the variable `reform`. [**3 points**]

(c) Why is it not necessary to include an offset term in this model? [**2 points**]

(d) The same model was fit again, this time using a Negative Binomial distribution. Relevant output comparing the Negative Binomial model with the above Poisson model are given below. (Note that $-2LL$ represents -2 times the log-likelihood evaluated at the maximum likelihood estimator). Which model do you think is more appropriate for the data? Justify your answer. [**2 points**]

Model	−2LL	AIC	BIC
Poisson	11878.23	11890.23	11924.48
Negative Binomial	9125.98	9139.98	9179.94

Question 7

[**11 points**] A researcher wants to evaluate students' progress in a reading test. To do this, the test is administered at age 11 and again at age 16. Data was collected on a total of 4 059 students from 65 different schools. The dataset includes the following variables:

`score16`:	standardized reading test score at age 16
`school`:	school identification number
`score11`:	standardized reading test score at age 11
`schgender`:	categorical variable for the type of school, with levels `mixed` (both girls and boys), `girls` (girls only), and `boys` (boys only)

A mixed effects linear regression model was fit to the data using `score16` as the response variable, with a random intercept (specific to each school). In this model, no correlation structure is assumed for the errors, that is, the errors are assumed to be independent. The SAS code and corresponding select output for this model are given below.

```
proc mixed data=mydata.exam22 method=reml covtest;
class school schgender;
model score16 = schgender score11  /solution cl;
random intercept / subject=school  solution;
run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	school	0.08463	0.01761	4.81	<.0001
Residual		0.5659	0.01267	44.67	<.0001

Solution for Fixed Effects

Effect	schgender	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept		-0.08725	0.05241	62	-1.66	0.1010	0.05	-0.1920	0.01751
schgender	boys	0.09687	0.1116	3993	0.87	0.3855	0.05	-0.1220	0.3157
schgender	girls	0.2452	0.08713	3993	2.81	0.0049	0.05	0.07440	0.4161
schgender	mixed	0	.	.	.	.	.	.	.
score11		0.5636	0.01246	3993	45.22	<.0001	0.05	0.5391	0.5880

Solution for Random Effects

Effect	school	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	0.4533	0.09701	3993	4.67	<.0001
Intercept	2	0.3588	0.1142	3993	3.14	0.0017
Intercept	3	0.5785	0.1088	3993	5.32	<.0001
Intercept	4	0.1006	0.09455	3993	1.06	0.2872
Intercept	5	0.3125	0.1246	3993	2.51	0.0122

(a) What type of covariance/correlation structure does this model imply for the observations Y_i? **[1 point]**

(b) According to this model, what is the estimated correlation between two observations from the same school? What about for two observations from different schools? **[4 points]**

(c) What is the estimated marginal mean reading score at age 16 for a student from a mixed gender school with a reading score at age 11 of 1? **[3 points]**

(d) What is the predicted reading score at age 16 for a student from school #1 (which is an all girls school) with a reading score at age 11 of 0? **[3 points]**