First name:

Last name:

Student ID number:

Statistical Modelling
Exam 23/07/2024

Exercise 1

The CPS1985 dataset consists of a random sample of 534 individuals from the 1985 census, with information on wages and other characteristics of the workers, including gender, age, number of years of education, years of work experience, and union membership. We wish to determine whether wages are related to these characteristics. Specifically, the covariates are

- EDUCATION: Number of years of education.
- SOUTH: Indicator variable for Southern Region (1=Lives in South, 0=Lives elsewhere).
- GENDER: Indicator variable for gender (1=Female, 0=Male).
- EXPERIENCE: Number of years of work experience.
- UNION: Indicator variable for union membership (1=Union member, 0=Not a member).
- WAGE: Wage (dollars per hour).
- AGE: Age (years).
- RACE: Race (1=Other, 2=Hispanic, 3=White).
- MARR: Marital Status (0=Unmarried, 1=Married)

Fitting a Gaussian linear model provides the following output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.4282	6.7940	-0.36	0.7209
EDUCATION	1.2699	1.1106	1.14	0.2534
SOUTH1	-0.7187	0.4297	-1.67	0.0951
GENDER1	-2.1837	0.3908	-5.59	0.0000
EXPERIENCE	0.4717	1.1106	0.42	0.6712
UNION1	1.4336	0.5087	?	?
AGE	-0.3711	1.1098	-0.33	0.7382
RACE2	0.7117	1.0120	0.70	0.4822
RACE3	?	0.5860	1.66	0.0970
MARR1	0.4563	0.4204	1.09	0.2782

Residual standard error: 4.412 on 524 degrees of freedom
Coefficient $R^2 = 0.2753$

*Residual standard error = $\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)}$

1. Define how the GENDER and RACE variables are encoded according to the output.
2. Write the statistical model corresponding to the analysis (model formulation and assumptions). Denote this model as "model A".

3. Complete the missing values in the table.

4. Explain the interpretation of the coefficients associated with the variables EDUCATION, RACE2, RACE3, and MARR1.

5. Perform a test of the overall significance of the model using a 5% significance level.

6. On the same dataset, it is then estimated a reduced model ("model B") that produces the following output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2689	0.4194	22.10	0.0000
EXPERIENCE	0.0428	0.0176	2.43	0.0153
GENDER1	-2.1960	0.4364	-5.03	0.0000

Residual standard error: 5.011 on 531 degrees of freedom
Coefficient $R^2 = 0.05275$

Write the statistical model corresponding to such output and perform a test to compare model A and model B. Which model do you prefer?

7. Can you use the R^2 coefficients to compare the two models? Explain.

8. Starting from model B, it is then introduced, as an additional covariate, the interaction between GENDER and EXPERIENCE. What is the purpose of estimating such a model? Derive and explain the interpretation of the coefficient associated with the variable EXPERIENCE:GENDER1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.6222	0.5053	17.06	0.0000
EXPERIENCE	0.0809	0.0242	3.34	0.0009
GENDER1	-0.7650	0.7646	-1.00	0.3176
EXPERIENCE:GENDER1	-0.0798	0.0351	-2.27	0.0233

Residual standard error: 4.992 on 530 degrees of freedom
Coefficient $R^2 = 0.06191$

Exercise 2

You have been given a sample dataset of 10,000 individuals from the insured population with the following characteristics:

- heart_disease: an indicator corresponding to whether an individual has heart disease (1 = yes, heart disease; 0 = no heart disease)
- coffee_drinker: an indicator corresponding to whether an individual drinks coffee regularly (1 = yes, coffee drinker; 0 = not a coffee drinker)
- fast_food_spend - a numerical variable corresponding to the annual spend of each individual on fast food
- income - a numerical variable corresponding to the individual's annual income

Fitting a logistic regression model provides the following output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	?	0.4923	-22.08	0.0000
coffee_drinker	-0.6468	0.2363	?	?
fast_food_spend	0.0023	0.0001	24.74	0.0000
income	0.0000	0.0000	0.37	0.7115

Null deviance: 2920.6
Residual deviance: 1571.5

1. Write the statistical model corresponding to the analysis (model formulation and assumptions).
2. Complete the missing values in the table.
3. What is the estimate of the probability of having a heart disease for an individual who regularly drinks coffee, spends 1000 in fast food and has an income of 20.000?
4. Consider a model that only includes the intercept ("model B"). What is the estimate of the intercept parameter in this case?
5. Perform a test to compare the full model with model B.

		p						
distribution		0.90	0.95	0.975	0.99	0.995	0.9975	0.999
standard Normal	z_p	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902
t with 1 df	$t_{1,p}$	3.0777	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088
t with 2 df	$t_{2,p}$	1.8856	2.9200	4.3027	6.9646	9.9248	14.0890	22.3271
t with 7 df	$t_{7,p}$	1.4149	1.8946	2.3646	2.9980	3.4995	4.0293	4.7853
t with 8 df	$t_{8,p}$	1.3968	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008
t with 9 df	$t_{9,p}$	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968
t with 10 df	$t_{10,p}$	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437
t with 524 df	$t_{524,p}$	1.2832	1.6478	1.9645	2.3335	2.5852	2.8190	3.1059
t with 526 df	$t_{526,p}$	1.2832	1.6478	1.9645	2.3335	2.5852	2.8189	3.1058
t with 527 df	$t_{527,p}$	1.2832	1.6478	1.9645	2.3334	2.5852	2.8189	3.1058
t with 532 df	$t_{532,p}$	1.2831	1.6477	1.9644	2.3334	2.5851	2.8188	3.1056
t with 533 df	$t_{533,p}$	1.2831	1.6477	1.9644	2.3334	2.5851	2.8188	3.1056
t with 534 df	$t_{534,p}$	1.2831	1.6477	1.9644	2.3334	2.5851	2.8187	3.1056
χ^2 with 1 df	$\chi^2_{1,p}$	2.7055	3.8415	5.0239	6.6349	7.8794	9.1406	10.8276
χ^2 with 2 df	$\chi^2_{2,p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
χ^2 with 3 df	$\chi^2_{3,p}$	6.2514	7.8147	9.3484	11.3449	12.8382	14.3203	16.2662
χ^2 with 4 df	$\chi^2_{4,p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668
χ^2 with 5 df	$\chi^2_{5,p}$	9.2364	11.0705	12.8325	15.0863	16.7496	18.3856	20.5150

Table 1: Some quantiles of Gaussian, t, and χ^2 distribution: $p = \mathbb{P}(X \leq q_p)$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, for the t and the χ^2, each row corresponds to different degrees of freedom (df).

		p						
distribution		0.90	0.95	0.975	0.99	0.995	0.9975	0.999
F with $(6, 524)$ df	$f_{6,524;p}$	1.7854	2.1159	2.4324	2.8365	3.1345	3.4277	3.8095
F with $(6, 534)$ df	$f_{6,534;p}$	1.7852	2.1155	2.4319	2.8358	3.1337	3.4267	3.8083
F with $(7, 524)$ df	$f_{7,524;p}$	1.7282	2.0270	2.3117	2.6735	2.9394	3.2003	3.5393
F with $(7, 534)$ df	$f_{7,534;p}$	1.7280	2.0267	2.3112	2.6728	2.9386	3.1993	3.5380
F with $(8, 524)$ df	$f_{8,524;p}$	1.6820	1.9561	2.2161	2.5453	2.7865	3.0227	3.3289
F with $(8, 534)$ df	$f_{8,534;p}$	1.6817	1.9557	2.2156	2.5446	2.7857	3.0217	3.3277
F with $(9, 524)$ df	$f_{9,524;p}$	1.6435	1.8977	2.1380	2.4412	2.6628	2.8794	3.1598
F with $(9, 534)$ df	$f_{9,534;p}$	1.6433	1.8974	2.1375	2.4406	2.6621	2.8785	3.1586
F with $(10, 524)$ df	$f_{10,524;p}$	1.6109	1.8488	2.0728	2.3548	2.5604	2.7611	3.0204
F with $(10, 534)$ df	$f_{10,534;p}$	1.6107	1.8484	2.0724	2.3542	2.5597	2.7601	3.0192
F with $(524, 6)$ df	$f_{524,6;p}$	2.7268	3.6771	4.8619	6.9005	8.9074	11.4322	15.7996
F with $(524, 7)$ df	$f_{524,7;p}$	2.4759	3.2385	4.1554	5.6698	7.1031	8.8462	11.7451
F with $(524, 8)$ df	$f_{524,8;p}$	2.2980	2.9367	3.6835	4.8789	5.9769	7.2785	9.3795
F with $(524, 9)$ df	$f_{524,9;p}$	2.1650	2.7161	3.3465	4.3307	5.2135	6.2391	7.8563
F with $(524, 10)$ df	$f_{524,10;p}$	2.0615	2.5477	3.0937	3.9292	4.6643	5.5044	6.8045
F with $(534, 6)$ df	$f_{534,6;p}$	2.7267	3.6770	4.8616	6.9001	8.9069	11.4315	15.7985
F with $(534, 7)$ df	$f_{534,7;p}$	2.4758	3.2383	4.1551	5.6694	7.1026	8.8456	11.7442
F with $(534, 8)$ df	$f_{534,8;p}$	2.2979	2.9365	3.6833	4.8786	5.9764	7.2778	9.3786
F with $(534, 9)$ df	$f_{534,9;p}$	2.1649	2.7160	3.3462	4.3303	5.2130	6.2385	7.8555
F with $(534, 10)$ df	$f_{534,10;p}$	2.0614	2.5475	3.0935	3.9288	4.6638	5.5038	6.8037

Table 2: Some quantiles of the F distribution: $p = \mathbb{P}(X \leq f_{df_1,df_2;p})$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, each row corresponds to different degrees of freedom (df).