EXERCISE 7

10TH December 2024

Luca Danese – l.danese1@campus.unimib.it

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 results:

	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 $R^2 = 0.06191$

1.1)

The GENDER variable is categorical with 2 levels. It is encoded with 1 dummy variable:

$$\text{GENDER1}_i = \begin{cases} 1 & \text{if individual } i \text{ is a female} \\ 0 & \text{otherwise} \end{cases}$$

The RACE variable is categorical with 3 levels. It is encoded with 2 (3-1 to avoid collinearity) dummy variables.

Specifically, from the output we see that we have parameters associated with levels 2 and 3 (RACE2, RACE3), hence RACE = 1 is the baseline.

$$RACE2_i = \begin{cases} 1, \text{ if } RACE_i = 2 \text{ (individual is hispanic)} \\ 0, \text{ otherwise} \end{cases}$$

$$RACE3_i = \begin{cases} 1, \text{ if } RACE_i = 3 \text{ (individual is white)} \\ 0 \text{ otherwise} \end{cases}$$

1.2) The model is a Gaussian linear model where Y_i denotes the wage of individual i.

$$Y_i = \beta_1 + \beta_2 EDU_i + \beta_3 SOUTH1_i + \beta_4 GENDER1_i + \beta_5 EXPER_i + \\ + \beta_6 UNION1_i + \beta_7 AGE_i + \beta_8 RACE2_i + \beta_9 RACE3_i + \\ + \beta_{10} MARR1_i + \varepsilon_i \quad , \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ iid}$$

The assumptions of the model are the following:

i) normality, homoschedasticity, independence $\Rightarrow \varepsilon_i \sim N(0, \sigma^2)$ iid $i = 1, .., 534$

ii) linearity w.r.t. $\beta_1, ..., \beta_{10}$

iii) covariates are linearly independent

1.3)

- t-value of β_6 (UNION1)

$$t^{obs} = \frac{\hat{\beta}_6 - 0}{\sqrt{\hat{Var}(\hat{\beta}_6)}} = \frac{\hat{\beta}_6}{\hat{S.E.}(\hat{\beta}_6)} = \frac{1.4336}{0.5087} = 2.818$$

- p-value of β_6

$$\text{p-value} = P_{H_0}(|T| > |t^{obs}|) = 2 P_{H_0}(T > |t^{obs}|) = 2 P_{H_0}(T > 2.818)$$

$$= 2(1 - P_{H_0}(T < 2.818)) = 2(1 - 0.9975) = 0.005$$

$$T \sim t_{n-p=533}$$

- $\hat{\beta}_9$ (RACE3)

$$t^{obs} = \frac{\hat{\beta}_9}{\hat{S.E.}(\hat{\beta}_9)} \Rightarrow \hat{\beta}_9 = t^{obs} \cdot \hat{S.E.}(\hat{\beta}_9) = 1.66 \cdot 0.5860 = 0.9727$$

1.4)

- EDUCATION is a numeric variable. Hence β_2 represents the (additive) change in the expected wage for an additional year of education, keeping the other covariates fixed.
In other words, for every additional year of education, the mean wage increases of 1.26\$, with all other variables constant.

- RACE is categorical. I consider two individuals j and k such that:

RACE$_j$ = 1 and RACE$_k$ = 2, while all other covariates are equals.

$$\mu_j = \beta_1 + \beta_2 EDU_j + \ldots + \beta_7 AGE_j + \beta_8 \underbrace{RACE2_j}_{=0} + \beta_9 \underbrace{RACE3_j}_{=0} + \beta_{10} MARR_j$$

$$\mu_k = \beta_1 + \beta_2 EDU_k + \ldots + \beta_7 AGE_k + \beta_8 \underbrace{RACE2_k}_{=1} + \beta_9 \underbrace{RACE3_k}_{=0} + \beta_{10} MARR_k$$

$$\mu_k - \mu_j = \beta_8$$

$\Rightarrow$ β_8 represents the additive change in the mean hourly wage if I consider an individual in the hispanic population compared to an individual in the "other" population (keeping other covariates constant)

$\Rightarrow$ β_9 represents the additive change in the mean hourly wage if I consider an individual in the white population compared to an individual in the "other" population (keeping other covariates constant)

- MARR1 is binary

If I consider two individuals, identical for all covariates but the marital status, the married one has a mean hourly age of 0.4563\$ higher than the unmarried one.

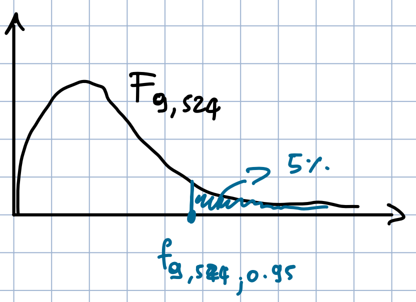
1.5)

$$\begin{cases} H_0: \beta_2 = \beta_3 = \ldots = \beta_{10} = 0 \\ H_1: \text{at least one } \beta_j \text{ is} \neq 0 \quad (j = 2, \ldots, 10) \end{cases}$$

We use the following test statistic:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p}{p-1} = \frac{R^2}{1-R^2} \cdot \frac{524}{9} \quad \text{where } F \overset{H_0}{\sim} F_{9,524}$$

$$f^{obs} = \frac{0.2753}{1-0.2753} \cdot \frac{524}{9} = 22.47$$



$f_{9,524;0.95} = 1.8977$

$\Rightarrow R = (1.8977, +\infty)$

$\Rightarrow$ we reject H_0

1.6) The model is

$$Y_i = \gamma_1 + \gamma_2 \, EXPER_i + \gamma_3 \, GENDER1_i + \varepsilon_i \, , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ iid}$$

This model is nested to model A, hence I can compare them through a test

$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0 \\ H_1: \text{at least one is} \neq 0 \end{cases}$$

We have the following test statistic:

$$F = \frac{SSE_B - SSE_A}{SSE_A} \cdot \frac{n-p_A}{p_A - p_B} \overset{H_0}{\sim} F_{7,524}$$

To compute f^{obs} we need first to compute SSE_A and SSE_B.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSE_A = (\text{residual s.e.})^2 \cdot (n-10) = 4.412^2 \cdot 524 = 10\,200.05$$

$$SSE_B = (\text{residual s.e.})^2 \cdot (n-3) = 5.04^2 \cdot 531 = 13\,333.47$$

$$f^{obs} = \frac{13\,333 - 10\,200}{10\,200} \cdot \frac{524}{7} = 22.99$$

$\Rightarrow$ we reject H_0. We prefer model A

1.7)

No, because the R^2 always increases (or stay the same) when I add covariate. We should use R^2 adjusted.

1.8)

The inclusion of the interaction allows studying if the effect on the mean wage of an additional year of experience is different for men and women.

The model is:

$$Y_i = \xi_1 + \xi_2 \, EXPER_i + \xi_3 \, GENDER1_i + \xi_4 \, EXPER:GENDER_i + \varepsilon_i \,, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ iid},$$

where

$$EXPER:GENDER_i = \begin{cases} EXPER_i & \text{if } GENDER_i = 1 \\ 0 & \text{if } GENDER_i = 0 \end{cases}$$

If I consider a man the expected wage is:

$$\mathbb{E}[Y_i] = \xi_1 + \xi_2 \, EXPER_i$$

If I consider a woman the expected wage is

$$\mathbb{E}[Y_i] = \xi_1 + \xi_2 \, EXPER_i + \xi_3 + \xi_4 \, EXPER_i$$
$$= (\xi_1 + \xi_3) + (\xi_2 + \xi_4) \, EXPER_i$$

Hence ξ_4 is the change in the effect of an additional year of experience on the mean wage due to being a woman (compared to being a man)

In other terms: an additional year of experience leads to an increase of 0.0809 \$ in the mean wage for a man, while it leads to an increase of $(0.0809 - 0.0798) = 0.011$ \$ for a woman.

distribution		0.90	0.95	0.975	0.99	0.995	0.9975	0.999
		p						
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).

distribution		0.90	0.95	0.975	0.99	0.995	0.9975	0.999
		p						
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).

EXERCISE 2

Consider an experiment to study the resistance to the tension of a machine component. The dataset studies how many breaks occurred during 54 replications of the experiment for two types of material (A and B) and different levels of tension (L = low; M = medium; H = high).

To study such relationship we fit a Poisson regression model. The output of the model is the following:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.6920	0.0454	81.30	0.0000
material B	-0.2060	0.0516	-3.99	0.0001
tension M	-0.3213	0.0603	-5.33	0.0000
tension H	-0.5185	0.0640	-8.11	0.0000

Null deviance: 297.37 on 53 degrees of freedom
Residual deviance: 210.39 on 50 degrees of freedom

1. Write the model formulation and assumptions.
2. Derive and explain the interpretation of the coefficient associated with the variable "material B".
3. A second model ("model B") assumes that the type of material and the level of tension do not have an impact on the number of breaks. Specify the model and perform a test to compare the model fitted in point (a) with model B.

2.1)

Y_i = number of breaks

$X_{i1} = \begin{cases} 1 \text{ if material}_i = B \\ 0 \text{ if material}_i = A \end{cases}$ $X_{i2} = \begin{cases} 1 \text{ if tension}_i = M \\ 0 \text{ otherwise} \end{cases}$ $X_{i3} = \begin{cases} 1 \text{ if tension} = H \\ 0 \text{ otherwise} \end{cases}$

model: $Y_i \sim \text{Poisson}(\mu_i)$ indep. $i = 1, \dots, 54$

$\eta_i = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \beta_4 X_{i3}$

$\log(\mu_i) = \eta_i \iff \mu_i = e^{\eta_i}$

2.2)

consider two experiments with same tension and different material

exp. i : material A

exp. j : material B

$$\log \mu_i = \beta_1 + \beta_2 \underset{0}{\underset{\shortparallel}{x_{i1}}} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

$$\log \mu_j = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

$$\Rightarrow \log \mu_j - \log \mu_i = \beta_2$$

$\Rightarrow$ β_2 is the difference in the log. of the expected number of breaks if I consider material B instead of material A, for fixed level of tension.

2.3)

model B

$Y_i \sim Pois(\mu_i) \quad i = 1, \dots, n$

$\log(\mu_i) = \beta_1 \quad \Rightarrow \mu_i = \mu$ for every i

To compare model A and model B we run the following test

$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: \text{at least one is} \neq 0 \end{cases}$$

We use the LRT:

$$W = 2\left[\hat{\ell}(\text{model A}) - \hat{\ell}(\text{model B})\right] \overset{H_0}{\sim} \chi^2_3$$

Since B is the null model:

$$W^{obs} = D(\text{null}) - D(\text{model A}) = 297.37 - 210.39 = 96.98$$

Since $w^{obs} > \chi^2_{3,1-\alpha}$ for every α we reject H_0.

		p						
		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 18 d.o.f	$t_{18,p}$	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105
t with 17 d.o.f	$t_{17,p}$	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458
t with 16 d.o.f	$t_{16,p}$	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862
t with 15 d.o.f	$t_{15,p}$	1.3406	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328
t with 14 d.o.f	$t_{14,p}$	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874
χ^2 with 1 d.o.f	$\chi^2_{1,p}$	2.7055	3.8415	5.0239	6.6349	7.8794	9.1406	10.8276
χ^2 with 2 d.o.f	$\chi^2_{2,p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
χ^2 with 3 d.o.f	$\chi^2_{3,p}$	6.2514	7.8147	9.3484	11.3449	12.8382	14.3203	16.2662
χ^2 with 4 d.o.f	$\chi^2_{4,p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668

Table 1: Some quantiles of Gaussian, Student's T and chi-squared distribution: $p = \mathbb{P}(X \leq q_p)$. Columns correspond to probabilities p. Rows correspond to different distributions, in particular, for the T and χ^2, each row corresponds to different degrees of freedom (d.o.f.).

	p						
	0.9000	0.9500	0.9750	0.9900	0.9950	0.9975	0.9990
$f_{1,18;p}$	3.0070	4.4139	5.9781	8.2854	10.2181	12.3208	15.3793
$f_{2,18;p}$	2.6239	3.5546	4.5597	6.0129	7.2148	8.5130	10.3899
$f_{3,18;p}$	2.4160	3.1599	3.9539	5.0919	6.0278	7.0351	8.4875
$f_{1,17;p}$	3.0262	4.4513	6.0420	8.3997	10.3842	12.5525	15.7222
$f_{2,17;p}$	2.6446	3.5915	4.6189	6.1121	7.3536	8.7006	10.6584
$f_{3,17;p}$	2.4374	3.1968	4.0112	5.1850	6.1556	7.2053	8.7269
$f_{1,16;p}$	3.0481	4.4940	6.1151	8.5310	10.5755	12.8201	16.1202
$f_{2,16;p}$	2.6682	3.6337	4.6867	6.2262	7.5138	8.9179	10.9710
$f_{3,16;p}$	2.4618	3.2389	4.0768	5.2922	6.3034	7.4027	9.0059
$f_{1,15;p}$	3.0732	4.5431	6.1995	8.6831	10.7980	13.1328	16.5874
$f_{2,15;p}$	2.6952	3.6823	4.7650	6.3589	7.7008	9.1726	11.3391
$f_{3,15;p}$	2.4898	3.2874	4.1528	5.4170	6.4760	7.6343	9.3353
$f_{1,14;p}$	3.1022	4.6001	6.2979	8.8616	11.0602	13.5026	17.1434
$f_{2,14;p}$	2.7265	3.7389	4.8567	6.5149	7.9216	9.4748	11.7789
$f_{3,14;p}$	2.5222	3.3439	4.2417	5.5639	6.6804	7.9097	9.7294

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).