UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4900/9900 — Statistical methods and applications
Day of examination:	June 15th, 2021
Examination hours:	9.00 – 13:00
This problem set consists of 6 pages.	
Appendices:	None.
Permitted aids:	All printed, hand-written and internet-based resources. Calculator and R.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

From a large population-based case-control study on oral cancer conducted in the US (Day et al., 1993), the data related to the African American population (194 cases, `ccstatus` = 1, and 203 controls, `ccstatus` = 0) have been selected.
In a case-control study one samples (typically all) cases and (typically a small fraction) of the non-cases, referred to as controls, and then ascertains different covariates that the cases and controls have been exposed to. For our purposes, case-control data can be analysed in a similar way as data with binary outcomes from a complete population. In particular the logistic regression model holds on case-control data with the same regression parameters (except the intercept) as the logistic model in the full population seen in class.
The aim of the study was to evaluate the risk of oral cancer based on the variables `drinks` (number of one ounce ethanol-equivalent drinks consumed per week), `sex` (0 = female, 1 = male), `age` (in years) and `cigs` (number of cigarettes smoked per day).

a

Consider a new variable `smoker`, which assumes value 1 for the smokers (`cigs` ≥ 1) and 0 for not smokers (`cigs` = 0). Consider the following table, which includes the observed frequencies,

		`ccstatus` 0	1	Sum
`smoker`	0	69	22	91
	1	134	172	306
	Sum	203	194	397

Compute the proportion of smokers among the cases and the controls. Find 95% confidence intervals for both these proportions. Compare and comment.

(Continued on page 2.)

b

A way to test for the null hypothesis of no difference of experiencing oral cancer between smoker and no-smoker is to compare the observed values of the table in point (a) with the expected values under the null hypothesis.

		ccstatus		
		0	1	Sum
smoker	0	A	B	91
	1	C	D	306
	Sum	203	194	397

Compute the values for the numbers A, B, C and D in the table above (expected frequencies). Using the computed data, perform the statistical test, reporting the value of the χ^2 test statistics and the related p-value. Is the null hypothesis rejected or not rejected?

c

We now fit a logistic regression model using the dichotomized variable `smoker` as explanatory variable. The response is again `ccstatus`. Here is the output of R for this analysis:

```
Call:
glm(formula = ccstatus ~ smoker, family = binomial, data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1431     0.2448  -4.669 3.03e-06
smoker        1.3927     0.2706     ??? 2.65e-07
---
    Null deviance: 550.15  on 396  degrees of freedom
Residual deviance: 520.14  on 395  degrees of freedom
```

Fill in the output by adding the left-out value (z-value for the variable `smoker`). Moreover, comment on the result of this model: does being a smoker increase or decrease the risk of experiencing oral cancer? How much, in terms of log-odds?
Transform the latter estimate into an odds-ratio and interpret its value.

d

Consider now the original variable `cigs`, i.e. the number of cigarettes smoked per day. In this case, we consider it a continuous variable. Fitting again a logistic model in R, we obtain the following output (on the next page):

(Continued on page 3.)

```
Call:
glm(formula = ccstatus ~ cigs, family = binomial, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.909057   0.171766  -5.292 1.21e-07
cigs         0.053624   0.008614   6.225 4.81e-10
---
    Null deviance: 550.15  on 396  degrees of freedom
Residual deviance: 504.39  on 395  degrees of freedom
```

Comment on the result: What does the regression coefficient for this variable mean now?
Why do you think the residual deviance of this model is smaller than that of the model at point c?

e

Consider now the other three variables (`drinks`, `sex` and `age`) in the logistic model. The output is now

```
Call:
glm(formula = ccstatus ~ cigs + age + sex + drinks, family = binomial,
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.966071   0.620756  -3.167  0.00154
cigs         0.035480   0.009571   3.707  0.00021
age          0.006529   0.009960   0.656  0.51213
sex          0.594499   0.272752   2.180  0.02928
drinks       0.029623   0.004643   6.380 1.77e-10
---
    Null deviance: 550.15  on 396  degrees of freedom
Residual deviance: 443.84  on 392  degrees of freedom
```

What is the increase in terms of log-odds for an increasing number of cigarettes per day smoked estimated by this model?
Why did it change from the one obtained in model fitted in point (c)?

f

Consider now the model without `age`, which does not look significant in the model fitted in point (d). The R output is:

```
Call:
glm(formula = ccstatus ~ cigs + sex + drinks, family = binomial,
    data = data)
```

(Continued on page 4.)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.592919   0.238787  -6.671 2.54e-11
cigs         0.035536   0.009565   3.715 0.000203
sex          0.582183   0.271756   2.142 0.032169
drinks       0.029498   0.004638   6.360 2.01e-10
---
    Null deviance: 550.15  on 396  degrees of freedom
Residual deviance: 444.27  on 393  degrees of freedom
```

Perform a likelihood ratio test reporting the value of the test statistic G and the corresponding p-value. Which of the two models (with or without the covariate `age`) should be used in the study?

Problem 2

The UCI Machine Learning repository (Dua & Graff, 2019) contains a dataset with information about 194 models of cars in circulation in the United States in 1985. In this problem, we are interested in the distance in km travelled in the city with 1 litre of fuel (the response variable is `city.distance`), a measure of fuel-economy where high values correspond to low fuel consumption.

a

In order to evaluate the dependency of distance travelled on the car style (variable `bodystyle`, that can assume values 'cabriolet, 'hatchback', 'sedan' or 'wagon') and the location of the the drive wheels (variable `drive.wheels`, either 'in the front' or 'in the back'), a two-way ANOVA has been performed.

```
Analysis of Variance Table

Response: city.distance
                       Df  Sum Sq Mean Sq F value Pr(>F)
bodystyle               D   69.14   23.05      F1     P1 **
drive.wheels            1  448.71  448.71      F2     P2 ***
bodystyle:drive.wheels  3   13.83    4.61      F3     P3
Residuals             186 1015.67    5.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fill in the left-out values in the table (D = degrees of freedom for `bodystyle`, and the values of the F test statistics, F1, F2 and F3 with the related p-values P1, P2 and P3).
What can we conclude from these results? In particular, what does it means that the result of the test on the interaction is not statistically significant?

(Continued on page 5.)

b

Let us ignore for the rest of the analysis the variables `bodystyle` and `drive.wheels`. Other variables that may affect the outcome of interest are the engine size (`engine.size`, in cubic decimetre) and the kind of fuel (`fuel`, with possible values 0 = 'diesel' and 1 = 'gas'). The resulting linear Gaussian model is

```
Call:
lm(formula = city.distance ~ engine.size + fuel, data = auto)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.0349     0.6285  30.286  < 2e-16
engine.size  -2.7672     0.2027 -13.652  < 2e-16
fuel         -2.7587     0.4635  -5.952 1.25e-08
---
Residual standard error: 1.959 on 191 degrees of freedom
Multiple R-squared:  0.5263,Adjusted R-squared:  0.5214
F-statistic: 106.1 on 2 and 191 DF,  p-value: < 2.2e-16
```

Comment on the results, in particular focus on the the interpretation of the three regression coefficient estimates.
Also provide a 95% confidence interval for the regression coefficients.

c

For a new car with a 2.3 cubic decimetre diesel engine, we would like to predict the amount of km that can be travelled with 1 litre of fuel by using the model fitted in point c. Below you find an output showing:

- the predicted value with the lower and upper limits of the 95% prediction interval;
- the fitted value with the lower and upper limits of the 95% confidence interval;

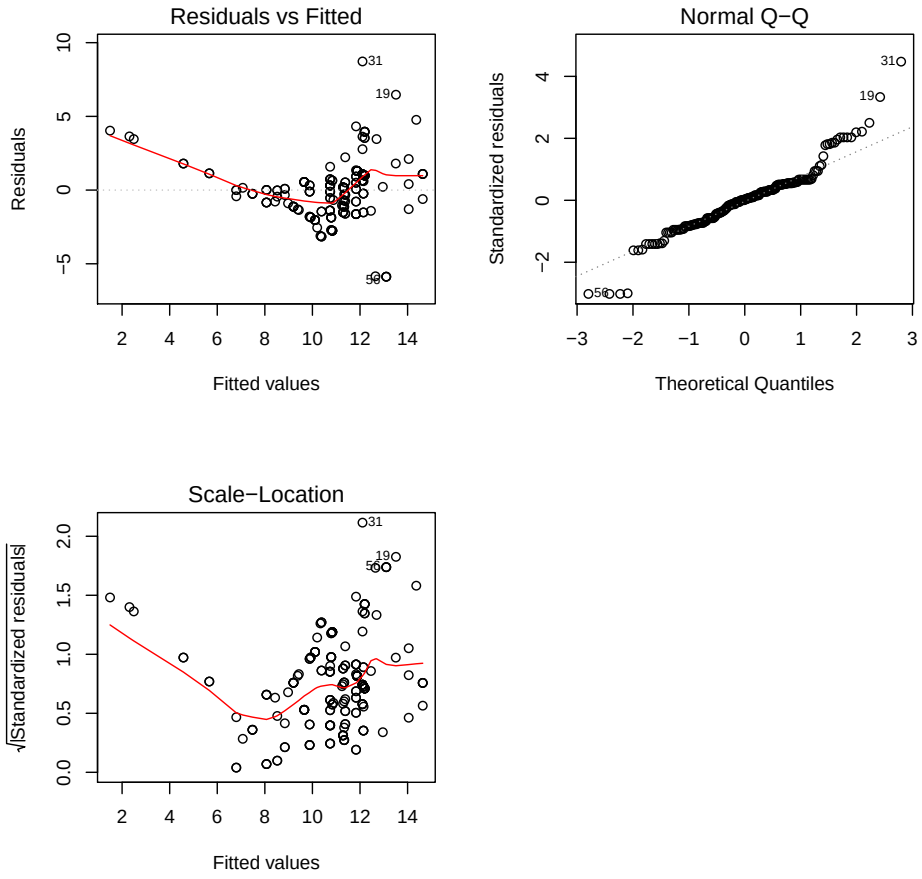
```
x.new <- data.frame(engine.size = 2.3, fuel = 'diesel')
predict(model, newdata = x.new, interval = 'prediction',
        level = 0.95)
      fit      lwr      upr
 12.67043 8.710924 16.62993
predict(model, newdata = x.new, interval = 'confidence',
        level = 0.95)
      fit      lwr      upr
 12.67043 11.80588 13.53498
```

Show how the predicted value is calculated, proving the formula used to compute it.
Moreover, focus on the difference between the prediction interval and the confidence interval showed in the R output. Interpret the two intervals and comment on their difference.

(Continued on page 6.)

d

Consider the following diagnostic plots,



Comment on what you see, describing what it is displayed in the three plots and what are the deviations of the model from the multiple linear regression assumptions.

References

Day, G. L., Blot, W. J., Austin, D. F., Bernstein, L., Greenberg, R. S., Preston-Martin, S., Schoenberg, J. B., Winn, D. M., McLaughlin, J. K. & Fraumeni Jr, J. F. (1993). Racial differences in risk of oral and pharyngeal cancer: alcohol, tobacco, and other determinants. *JNCI: Journal of the National Cancer Institute* **85**, 465–473.

Dua, D. & Graff, C. (2019). UCI machine learning repository.

THE END