# Statistical Modelling
# Exam preparation

January 11, 2024

## Exercise 1

The `mtcars` dataset comprises fuel consumption (`mpg`: Miles/(US) gallon) and 8 aspects of automobile design and performance for 32 automobiles. Specifically, the covariates are

- `wt`: Weight (1000 lbs)
- `am`: Transmission (0 = automatic, 1 = manual)
- `cyl`: Number of cylinders
- `disp`: Displacement (cu.in.)
- `hp`: Gross horsepower
- `drat`: Rear axle ratio
- `qsec`: 1/4 mile time
- `vs`: Engine (0 = V-shaped, 1 = straight)

Fitting a Gaussian linear model outputs the following summary

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---:|---:|---:|---:|---:|
| `(Intercept)` | 15.5731 | 16.3817 | 0.951 | 0.3517 |
| `wt` | -3.9437 | 1.2874 | -3.063 | 0.0055 |
| `am = 1` | 2.7937 | 1.8682 | 1.495 | 0.1484 |
| `cyl` | -0.2786 | 0.9348 | -0.298 | 0.7683 |
| `disp` | 0.0147 | 0.0120 | 1.223 | 0.2338 |
| `hp` | -0.0214 | 0.0162 | ?? | 0.1995 |
| `drat` | 0.8151 | 1.5101 | 0.540 | 0.5946 |
| `qsec` | ?? | 0.6587 | 1.229 | 0.2314 |
| `vs = 1` | 0.3684 | 2.0116 | 0.183 | ?? |

Residual standard error $\sqrt{\sum_{i=1}^{32}(y_i - \hat{y}_i)^2/23} = 2.544$
Coefficient $R^2 = 0.8678$

a) Write the statistical model corresponding to the analysis (quantities and assumptions). Denote this model as "model A".

b) Write the parameter space and sample space.

c) Complete the missing values in the table.
For "Pr(>|t|)" of `vs1`, write the meaning of the missing value and how to obtain it.

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

e) On the same dataset, it is then estimated a reduced model ("model B") that only includes the variables `wt` and `am`. The software estimates the following quantities:
Residual standard error = 3.098
Coefficient $R^2 = 0.7528$
What procedure would you use to compare model A and model B? Following your chosen procedure, which model do you prefer?

f) Starting from model B, it is then introduced as additional covariate the interaction between `wt` and `am`. Explain the resulting model and how you interpret the parameters.

## Exercise 2

Given a set of $n = 30$ observations, consider fitting the model $Y_i \sim \text{Bernoulli}(\pi_i)$ where $\text{logit}(\pi_i) = \beta_1 + \beta_2 x_i$, with $x_i$ is a dummy variable taking value 1 for the first 10 observations and 0 otherwise. Fitting this model returns the following output

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| `(Intercept)` | 1.3863 | 0.5590 | 2.480 | 0.01314 |
| `x` | -2.0794 | 0.7826 | -2.657 | 0.00788 |

| | |
|---|---|
| Null deviance | 47.111 |
| Residual deviance | 39.112 |

a) Write the likelihood, log-likelihood and score functions for $(\beta_1, \beta_2)$. Write the fitted model.

b) Compute the estimate of the probability $\hat{\pi}$ for $x = 0$ and $x = 1$. Obtain the odds for $x = 0$ and $x = 1$ and interpret them. Give an estimate of the odds ratio and interpret it.

c) Test the hypothesis $H_0 : \beta_2 = -1$ vs $H_1 : \beta_2 < -1$.

d) What are the two quantities "Null deviance" and "Residual deviance"?

|  |  | $p$ |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  |  | 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 21 df | $t_{21,p}$ | 1.3232 | 1.7207 | 2.0796 | 2.5176 | 2.8314 | 3.1352 | 3.5272 |
| $t$ with 22 df | $t_{22,p}$ | 1.3212 | 1.7171 | 2.0739 | 2.5083 | 2.8188 | 3.1188 | 3.5050 |
| $t$ with 23 df | $t_{23,p}$ | 1.3195 | 1.7139 | 2.0687 | 2.4999 | 2.8073 | 3.1040 | 3.4850 |
| $t$ with 31 df | $t_{31,p}$ | 1.3095 | 1.6955 | 2.0395 | 2.4528 | 2.7440 | 3.0221 | 3.3749 |
| $t$ with 32 df | $t_{32,p}$ | 1.3086 | 1.6939 | 2.0369 | 2.4487 | 2.7385 | 3.0149 | 3.3653 |
| $t$ with 33 df | $t_{33,p}$ | 1.3077 | 1.6924 | 2.0345 | 2.4448 | 2.7333 | 3.0082 | 3.3563 |

Table 1: Some quantiles of Gaussian and Student's t distribution: $p = \mathbb{P}(X \leq q_p)$. Columns correspond to probabilities $p$. Rows correspond to different distributions, in particular, for the t, each row corresponds to different degrees of freedom (df).

|  | $p$ |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 | 0.9975 | 0.999 |
| $f_{6,23;p}$ | 2.0472 | 2.5277 | 3.0232 | 3.7102 | 4.2591 | 4.8366 | 5.6486 |
| $f_{7,23;p}$ | 1.9949 | 2.4422 | 2.9023 | 3.5390 | 4.0469 | 4.5807 | 5.3308 |
| $f_{8,23;p}$ | 1.9531 | 2.3748 | 2.8077 | 3.4057 | 3.8822 | 4.3826 | 5.0853 |
| $f_{9,23;p}$ | 1.9189 | 2.3201 | 2.7313 | 3.2986 | 3.7502 | 4.2243 | 4.8896 |
| $f_{6,32;p}$ | 1.9668 | 2.3991 | 2.8356 | 3.4269 | 3.8886 | 4.3653 | 5.0211 |
| $f_{7,32;p}$ | 1.9132 | 2.3127 | 2.7150 | 3.2583 | 3.6819 | 4.1185 | 4.7186 |
| $f_{8,32;p}$ | 1.8702 | 2.2444 | 2.6202 | 3.1267 | 3.5210 | 3.9271 | 4.4846 |
| $f_{9,32;p}$ | 1.8348 | 2.1888 | 2.5434 | 3.0208 | 3.3919 | 3.7738 | 4.2977 |
| $f_{23,6;p}$ | 2.8223 | 3.8486 | 5.1284 | 7.3309 | 9.4992 | 12.2271 | 16.9460 |
| $f_{23,7;p}$ | 2.5796 | 3.4179 | 4.4263 | 6.0921 | 7.6688 | 9.5865 | 12.7758 |
| $f_{23,8;p}$ | 2.4086 | 3.1229 | 3.9587 | 5.2967 | 6.5260 | 7.9832 | 10.3357 |
| $f_{23,9;p}$ | 2.2816 | 2.9084 | 3.6257 | 4.7463 | 5.7516 | 6.9197 | 8.7618 |
| $f_{32,6;p}$ | 2.7953 | 3.7998 | 5.0521 | 7.2073 | 9.3290 | 11.9983 | 16.6155 |
| $f_{32,7;p}$ | 2.5504 | 3.3670 | 4.3491 | 5.9712 | 7.5066 | 9.3740 | 12.4795 |
| $f_{32,8;p}$ | 2.3777 | 3.0703 | 3.8806 | 5.1776 | 6.3691 | 7.7816 | 10.0616 |
| $f_{32,9;p}$ | 2.2491 | 2.8543 | 3.5468 | 4.6282 | 5.5984 | 6.7255 | 8.5031 |

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).