Coursework 1: STAT 570

Philip Pham

October 4, 2018

1. The data we analyze are from a 1970s study that investigated insurance redlining on $n = 47$ zipcodes. Information on who was being refused homeowners is not available so instead we take as response the number of FAIR plan policies written and renewed in Chicago by zip code over the period December 1977 to May 1978. The FAIR plan was offered by the city of Chicago as a default policy to homeowners who had been rejected by the voluntary market. The data we will analyze are named `chredlin` and are in the `faraway` package. The variable `involact` are the number of new FAIR plan policies and renewals per 100 housing units.

   We will consider five covariates for modeling the response: racial composition in percent minority (`race` x_{i1}), fires per 100 housing units (`fire` x_{i2}), theft per 1000 population (`theft` x_{i3}), percent of housing units built before 1939 (`age` x_{i4}), log median family income in thousands of dollars (`lincome` x_{i5}‘), $i = 1, \ldots, 47$.

   We will examine the model with the main effects due to race, fire, theft, age and log(income).

   We let Y_i represent `involact`, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{i5})$, the covariates, for individual i, $i = 1, 2, \ldots, 47$. We fit the model

$$y_i = \beta_0 + \sum_{j=1}^{5} x_{ij}\beta_j + \epsilon_i \tag{1}$$

   for $i = 1, \ldots, n$ using least squares.

   (a) Provide informative plots to illustrate what we might expect to learn from the model in Equation 1.

   Solution: See Figure 1 and the corresponding code in `chredlin_explore.ipynb`. `fire`, `race`, and `age` appear to be positively correlated with `involact`. `income` appears to be negatively correlated.
   Zipcodes in the northern `side` of Chicago have a lower minority population and higher income. `involact` is smaller in these northern zipcodes, too.

   (b) Give interpretations of the parameters β_j, $j = 1, \ldots, 5$.

   Solution: Fitting such a model, we get the estimates in Table 1 for β_j.
   The percent of minorities (`race`) and frequency of fires (`fire`) are positively correlated with the number of FAIR plan policies. `involact` is the number of FAIR plans per 100 housing units. Thus, every percent increase in racial minorities means about 1 FAIR plan, and for every fire per 100 housing units, there are 3 FAIR plans.

	estimate	std_error	t-statistic	p-value
(intercept)	-1.185540	1.100255	-1.077514	0.287550
race	0.009502	0.002490	3.816831	0.000449
fire	0.039856	0.008766	4.546588	0.000048
theft	-0.010295	0.002818	-3.653264	0.000728
age	0.008336	0.002744	3.037749	0.004134
log_income	0.345762	0.400123	0.864137	0.392540

Table 1: The result of fitting the model described in Equation 1. The procedure for obtaining the estimates and test statistics is described in Part 1c.

`age` seems to have postive effect on `involact`, while `theft` has a negative effect.
`log_income` doesn't seem to tell us anything new: it's correlated with other covariates, and its effect is mainly due to chance.

(c) Reproduce every number in the handout using matrix and arithmetic operations.

Solution: Let us assume that $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. The log-likelihood of this model is

$$\begin{aligned}\sum_{i=1}^{n} \log \mathbb{P}\left(y_i \mid x_i, \beta, \sigma^2\right) &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i^\intercal \beta\right)^2 \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|y - X\beta\right\|_2^2, \end{aligned} \tag{2}$$

where we 0-index β and the columns of X, so each row of X is $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{i5})$.

Estimating $\hat{\beta}$

To maximize Equation 2, we choose $\hat{\beta}$ such that $X\hat{\beta}$ is the projection of y onto the hyperplane spanned by the columns of X. Thus, we must have that $X^\intercal\left(y - X\hat{\beta}\right) = 0$ since the residuals will orthogonal to the columns of X if $X\hat{\beta}$ is the projection that minimizes the squared error. Solving for $\hat{\beta}$, we have that

$$\hat{\beta} = \left(X^\intercal X\right)^{-1} X^\intercal y. \tag{3}$$

The results of apply Equation 3 can be seen in the first column of Table 1.

Estimating $\hat{\sigma}^2$

Let us derive an unbiased estimator for residual standard error. Consider the residual random vector.

$$R = y - X\hat{\beta} \tag{4}$$

As stated earlier, the residuals are orthogonal to hyperplane spanned by the columns of X, so they must lie in some orthonormal hyperplane of $N - p$

vectors, where $p = \dim(\beta)$. Thus, residuals are y projected down to this space.
Let $w_1, \ldots, w_{n-p}$ be an orthonormal basis of this space. Let W be matrix with these basis vectors as the columns.
We have that

$$\begin{aligned} R &= y - X\hat{\beta} \\ &= W\left(W^\intercal y\right) \\ &= W\left(W^\intercal\left(X\beta + \sigma\epsilon\right)\right) \\ &= W\left(W^\intercal X\right)\beta + \sigma W\left(W^\intercal\epsilon\right) \\ &= \sigma W\left(W^\intercal\epsilon\right). \end{aligned} \tag{5}$$

Now, $W^\intercal\epsilon \sim \mathcal{N}\left(0, I_{n-p}\right)$. To see this, note that the ith entry is $\sum_{j=1}^{n} w_{ij}\epsilon_j \sim \mathcal{N}\left(0,1\right)$, and for $i \neq i'$,

$$\begin{aligned} \operatorname{Cov}\left(\left(W^\intercal\epsilon\right)_i, \left(W^\intercal\epsilon\right)_{i'}\right) &= \mathbb{E}\left[\left(\sum_{j=1} w_{ij}\epsilon_j\right)\left(\sum_{k=1} w_{i'k}\epsilon_k\right)\right] \\ &= \sum_{j=1}^{n}\mathbb{E}\left[w_{ij}w_{i'j}\epsilon_j^2\right] + 2\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}\mathbb{E}\left[w_{ij}w_{i'k}\epsilon_j\epsilon_k\right] \\ &= w_i^\intercal w_{i'} + 2\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} w_{ij}w_{i'k}\mathbb{E}\left[\epsilon_j\epsilon_k\right] \\ &= 0, \end{aligned}$$

where the first term disappears by since the two vectors are orthonormal, and the second term disappears because of independence of the errors.
Thus, we have that

$$R^\intercal R = \sigma^2\left(W^\intercal\epsilon\right)^\intercal W^\intercal W\left(W^\intercal\epsilon\right) = \sigma^2\left(W^\intercal\epsilon\right)^\intercal\left(W^\intercal\epsilon\right) \sim \sigma^2\chi^2_{n-p}. \tag{6}$$

Finally, we have that

$$\mathbb{E}\left[R^\intercal R\right] = \sigma^2\left(n-p\right) \Rightarrow \mathbb{E}\left[\frac{\sum_{i=1}^{n}\left(y - X\hat{\beta}\right)^2}{n-p}\right] = \sigma^2.$$

Our consistent estimator is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y - X\hat{\beta}\right)^2}{n-p}. \tag{7}$$

Applying Equation 7, we obtain $\boxed{\hat{\sigma} = 0.3345267301243203.}$

Hypothesis Testing

We can rewrite y as $y = X\beta + \sigma\epsilon$, where each element of ϵ is drawn from $\mathcal{N}\left(0,1\right)$. Substituting, we have that

$$\begin{aligned} \hat{\beta} &= \left(X^\intercal X\right)^{-1} X^\intercal\left(X\beta + \sigma\epsilon\right) \\ &= \beta + \sigma\left(X^\intercal X\right)^{-1} X^\intercal\epsilon. \end{aligned} \tag{8}$$

Thus, $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 \left(X^\intercal X\right)^{-1}_{jj}\right)$.
This gives us that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left(X^\intercal X\right)^{-1}_{jj}}} \sim \mathcal{N}(0,1).$$

From Equations 6 and 7,

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}. \tag{9}$$

$\hat{\beta}$ and $\hat{\sigma}^2$ are independent by Basu's theorem: $\hat{\sigma}^2$ is an ancillary statistic that does not depend on the model parameters, β. Thus, we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left(X^\intercal X\right)^{-1}_{jj}}} \Bigg/ \sqrt{\frac{(n-p)\frac{\hat{\sigma}^2}{\sigma^2}}{n-p}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \left(X^\intercal X\right)^{-1}_{jj}}} \sim t_{n-p}. \tag{10}$$

That is, we have t distribution with $n-p$ degrees of freedom. The denominator of Equation 10 gives the second column of Table 1.
For each β_j, our null hypothesis is $H_0 : \beta_j = 0$. Thus, our t-test statistic is obtain from substituting β_j into Equation 10,

$$\hat{t}_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \left(X^\intercal X\right)^{-1}_{jj}}},$$

which gives us the third column of Table 1.
The fourth column is the probability of obtaining evidence that contradicts the null hypothesis at least as much. Let $F^{-1}_{t_{n-p}}$ be the inverse cumulative distribution function. The p-value is

$$\mathbb{P}\left(|T_{n-p}| \geq \left|\hat{t}_j\right| \mid \hat{t}_j\right) = 2\left(1 - F^{-1}_{n-p}\left(\left|\hat{t}_j\right|\right)\right).$$

These calculations are carried out in `chredlin_explore.ipynb`.

(d) What assumptions are valid for:

i. An unbiased estimate of β_j, $j = 1, \ldots, 5$.

Solution: From Equation 8, we have that

$$\mathbb{E}\left[\hat{\beta}\right] = \beta + \left(X^\intercal X\right)^{-1} X^\intercal \mathbb{E}\left[\epsilon\right] \tag{11}$$

since expectation is a linear operator. In our previous calcuations, we assumed that the ϵ_i were independent and normally distributed.
It's sufficient, however, that $\boxed{\mathbb{E}\left[\epsilon\right] = \mathbf{0}.}$ Then, we'll have

$$\text{bias}\left(\hat{\beta}\right) = \mathbb{E}\left[\hat{\beta}\right] - \beta = \beta - \beta = 0.$$

ii. An accurate estimate of the standard error of $\hat{\beta}_j$, $j = 1, \ldots, 5$.

Solution: From Equation 8, we can estimate the standard error exactly if σ^2 is known. For $\hat{\beta}_j$, we get $\sigma\sqrt{(X^\intercal X)^{-1}_{jj}}$.

When σ^2 is unknown, but our errors are still independent and normally distributed, we apply Equation 10. Since $\hat{\beta}_j$ has Student's t-distribution, we can estimate the standard error for $\hat{\beta}_j$ with $\sqrt{\hat{\sigma}^2\,(X^\intercal X)^{-1}_{jj}}$.

If our errors are not normally distributed, our estimate is only accurate if the number of observations is large, and our errors have a distribution that converges to a normal distribution.

iii. Accurate coverage probabilities for $100\,(1-\alpha)\,\%$ confidence intervals of the form

$$\hat{\beta}_j \pm \sqrt{\hat{\sigma}^2_j} z_{1-\alpha/2}, \tag{12}$$

where $z_{1-\alpha/2}$ represents the $(1-\alpha/2)$ quantile of an $\mathcal{N}\,(0,1)$ random variable, and $\hat{\sigma}^2_j = \hat{\sigma}^2\,(X^\intercal X)^{-1}_{jj}$.

Solution: Firstly, the assumptions from the previous part must hold for $\hat{\sigma}^2_j$ to be meaningful.

From Equation 10, $\hat{\sigma}^2_j$ has Student's t-distribution, so the normal approximation for the the confidence interval (Equation 12) only holds when n is large.

iv. Accurate coverage probabilities for $100\,(1-\alpha)\,\%$ confidence intervals of the form

$$\hat{\beta}_j \pm \sqrt{\hat{\sigma}^2_j} t_{n-p}\,(1-\alpha/2)\,, \tag{13}$$

where $p = \dim(\beta)$ and $t_{n-p}\,(1-\alpha/2)$ represents the $(1-\alpha/2)$ quantile of standard Student's t random variable with $n-p$ degrees of freedom.

Solution: Equation 10 shows that this is exactly the correct distribution when the ϵ_i are independent and identically distributed as normal random variables with mean zero.

It may still prove to be an accurate confidence interval if the errors have distributions that are well-approximated by the normal distribution and the number of observations is large.

v. An accurate prediction for an *observed* outcome at x_0.

Solution: Suppose we were to observe $(x_0, y_0 = x_0^\intercal\beta + \epsilon_0)$. Let our prediction be $\hat{y}_0 = x_0^\intercal\hat{\beta}$. If the conditions in Part 1(d)i are satisfied, the error has mean zero, and our estimate for $\hat{\beta}$ is unbiased, so

$$\mathbb{E}\,[y_0] = x_0^\intercal\beta = \mathbb{E}\,[\hat{y}_0]\,.$$

We want to compare our prediction with $\hat{y}_0$ with some hypothetical observed response y_0. We'll call our prediction accurate within $\delta > 0$ if

$$\hat{y}_0 - \delta \le y_0 \le \hat{y}_0 + \delta.$$

We want the probability of this event to be high, so we'll say accurate within δ at confidence level $1-\alpha$ if

$$\mathbb{P}\,(\hat{y}_0 - \delta \le y_0 \le \hat{y}_0 + \delta) = \mathbb{P}\,(-\delta \le y_0 - \hat{y}_0 \le \delta) \ge 1-\alpha.$$

Our prediction is accurate if for small α, we have small δ.
If we assume normality, we can calculate the minimum δ for a specific α, which we'll denote δ_α.
Since $\hat{\beta}$ satisfies $(X^\intercal X)\,\hat{\beta} = X^\intercal y$, we have that the intercept estimate is

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{X}_{:,j}. \tag{14}$$

Consider trying to predict $\hat{y} = x^\intercal \hat{\beta}$ for some x. We have that

$$\hat{y} = \bar{y} + \sum_{j=1}^{p} \left(x_i - \bar{X}_{:,j}\right) \hat{\beta}_j, \tag{15}$$

so the variance of the prediction increases with values far from data.
Let $\bar{X}$ be the vector of column-wise means of X. Since $\bar{\epsilon}$ is an ancillary statistic, this can also be written as

$$\hat{y} \mid x \sim \mathcal{N}\left(x^\intercal \beta, \sigma^2 \left(\frac{1}{n} + \left(x - \bar{X}\right)^\intercal (X^\intercal X)^{-1} \left(x - \bar{X}\right)\right)\right). \tag{16}$$

Using the same method as in deriving Equation 10, if we replace β with $\hat{\beta}$ and σ^2 with $\hat{\sigma}^2$, we have

$$\frac{\hat{y} - x^\intercal \hat{\beta}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \left(x - \bar{X}\right)^\intercal (X^\intercal X)^{-1} \left(x - \bar{X}\right)\right)}} \sim t_{n-p}. \tag{17}$$

Noting that $y_0 \sim \mathcal{N}\left(x_0^\intercal \beta, \sigma^2\right)$, we can apply Equation 17 to (x_0, y_0), which gives us

$$\boxed{\delta_\alpha = t_{n-p}\left(1 - \alpha/2\right) \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \left(x_0 - \bar{X}\right)^\intercal (X^\intercal X)^{-1} \left(x_0 - \bar{X}\right)\right)}.}$$

Thus, our predictions will always have standard error of at least σ, but they will be more accurate when x_0 is close to $\bar{X}$.

(e) Summarize the relationship between y, and x_1, x_2, x_3, x_4, x_5, fitting any other models that you see fit to.

Solution: The relationship between y and the covariates was described in Parts 1a and 1b.
Particularly, we see `income` does not explain much about `involact` due to multicollinearity: it is correlated with `race` and `fire`.
Removing `log_income` from the model gives us the model parameters in Table 2. The residual standard error for this model was 0.3335165792871865, which is actually ever so slightly smaller than the model that includes income.
I tried adding an indicator for `side` but it suffers from the same issue as `income`: its effect is already explained by the other covariates.

	estimate	std_error	t-statistic	p-value
(intercept)	-0.243118	0.145054	-1.676054	0.101158
race	0.008104	0.001886	4.296913	0.000100
fire	0.036646	0.007916	4.629173	0.000035
theft	-0.009592	0.002690	-3.565847	0.000921
age	0.007210	0.002408	2.994369	0.004595

Table 2: The result of fitting a model without considering income.

2. Consider the following distributions:

Poisson:

$$p\left(y \mid \mu\right) = \frac{\exp\left(-\mu\right)\mu^{y}}{y!}, \tag{18}$$

for $y = 0, 1, 2, \ldots$.

Gamma:

$$p\left(y \mid \alpha, \beta\right) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y) \tag{19}$$

for $y > 0$ and with α known.

Inverse Gaussian:

$$p\left(y \mid \mu, \delta\right) = \left(\frac{\delta}{2\pi y^{3}}\right)^{1/2} \exp\left[\frac{-\delta\left(y-\mu\right)^{2}}{2\mu^{2} y}\right]$$

for $y > 0$ and δ known.

A distribution is said to be a member of the one parameter exponential family of distributions if it can be written as

$$p\left(y \mid \eta_1, \eta_2\right) = h(y) \exp\left[\eta_1 y + \eta_2 T_2(y) - A\left(\eta_1, \eta_2\right)\right], \tag{20}$$

where η_2 is known.

(a) Show that each of the above distributions is a member of the exponential family and identify η_1, η_2, $T_2(y)$, $A\left(\eta_1, \eta_2\right)$, and $h(y)$.

Solution: For each distribution, we can do some algebra.

Poisson: We can rewrite Equation 18 as Equation 20, where

$$\begin{aligned} \eta_1 &= \log \mu \\ \eta_2 &= 0 \\ T_2(y) &= 0 \\ A\left(\eta_1, \eta_2\right) &= \exp(\eta_1) \\ h(y) &= \frac{1}{y!}. \end{aligned}$$

Gamma: We can rewrite Equation 19 as Equation 20, where

$$\begin{aligned}
\eta_1 &= -\beta \\
\eta_2 &= \alpha - 1 \\
T_2(y) &= \log(y) \\
A(\eta_1, \eta_2) &= -(\eta_2 + 1)\log(-\eta_1) + \log\Gamma(\eta_2 + 1) \\
h(y) &= 1.
\end{aligned}$$

Inverse Gaussian: We can rewrite Equation 2 as Equation 20, where

$$\begin{aligned}
\eta_1 &= -\frac{\delta}{2\mu^2} \\
\eta_2 &= -\frac{\delta}{2} \\
T(y) &= \frac{1}{y} \\
A(\eta_1, \eta_2) &= -2\sqrt{\eta_1\eta_2} - \frac{1}{2}\log(-2\eta_2) \\
h(y) &= \frac{1}{\sqrt{2\pi y^3}}.
\end{aligned}$$

(b) Identify $\mathbb{E}[Y \mid \theta]$ and $\mathrm{Var}(Y \mid \theta)$.

Solution: We can derive a general formula for computing the mean and variance from Equation 20.
The log-likelihood function is

$$l(\eta_1, \eta_2) = \log h(y) + \eta_1 y + \eta_2 T_2(y) - A(\eta_1, \eta_2). \tag{21}$$

If η_2 is known, the score function is

$$S(\eta_1, \eta_2) = \frac{\partial l(\eta_1, \eta_2)}{\partial \eta_1} = y - \frac{\partial A(\eta_1, \eta_2)}{\partial \eta_1}. \tag{22}$$

The expectation of the score is 0, so

$$\boxed{\mathbb{E}[y \mid \eta_1, \eta_2] = \frac{\partial A(\eta_1, \eta_2)}{\partial \eta_1}.} \tag{23}$$

The variance of the score is Fisher information, so

$$\begin{aligned}
\mathcal{I}(\eta_1, \eta_2) &= \mathrm{Var}(S(\eta_1, \eta_2)) \\
&= \mathbb{E}\left[\left(y - \frac{\partial A(\eta_1, \eta_2)}{\partial \eta_1}\right)^2 \mid \eta_1, \eta_2\right] \\
&= \mathrm{Var}(y \mid \eta_1, \eta_2)
\end{aligned} \tag{24}$$

by Equation 23 and using that the mean of the score function is 0.
An alternative definition of the Fisher information is the expected value of the observed information:

$$\mathcal{I}(\eta_1, \eta_2) = -\frac{\partial^2 l(\eta_1, \eta_2)}{\partial \eta_1^2} = \frac{\partial^2 A(\eta_1, \eta_2)}{\partial \eta_1^2}. \tag{25}$$

Combining Equations 24 and 25, we obtain

$$\boxed{\text{Var}\left(y \mid \eta_1, \eta_2\right) = \frac{\partial^2 A\left(\eta_1, \eta_2\right)}{\partial \eta_1^2}.} \tag{26}$$

We can now apply Equations 23 and 26 to our the results from Part 2a.
Poisson:

$$\begin{aligned}\mathbb{E}\left[y \mid \eta_1, \eta_2\right] &= \exp\left(\eta_1\right) = \mu \\ \text{Var}\left(y \mid \eta_1, \eta_2\right) &= \exp\left(\eta_1\right) = \mu.\end{aligned}$$

Gamma:

$$\begin{aligned}\mathbb{E}\left[y \mid \eta_1, \eta_2\right] &= -\frac{\eta_2 + 1}{\eta_1} = \frac{\alpha}{\beta} \\ \text{Var}\left(y \mid \eta_1, \eta_2\right) &= \frac{\eta_2 + 1}{\eta_1^2} = \frac{\alpha}{\beta^2}.\end{aligned}$$

Inverse Gaussian:

$$\begin{aligned}\mathbb{E}\left[y \mid \eta_1, \eta_2\right] &= \sqrt{\frac{\eta_2}{\eta_1}} = \mu \\ \text{Var}\left(y \mid \eta_1, \eta_2\right) &= \frac{1}{2}\sqrt{\frac{\eta_2}{\eta_1^3}} = \frac{\mu^3}{\delta}.\end{aligned}$$

(c) The canonical link function is such that $g(\mu) = \eta_1$. Determine the canonical link for each distribution.

Solution: We can use the results from Part 2a.

Poisson: $g(\mu) = \exp(\mu)$.

Gamma: $g(\mu) = -\frac{\alpha}{\mu} \propto \mu^{-1}$, where α is known.

Inverse Gaussian: $g(\mu) = -\frac{\delta}{2\mu^2} \propto \mu^{-2}$, where δ is known.

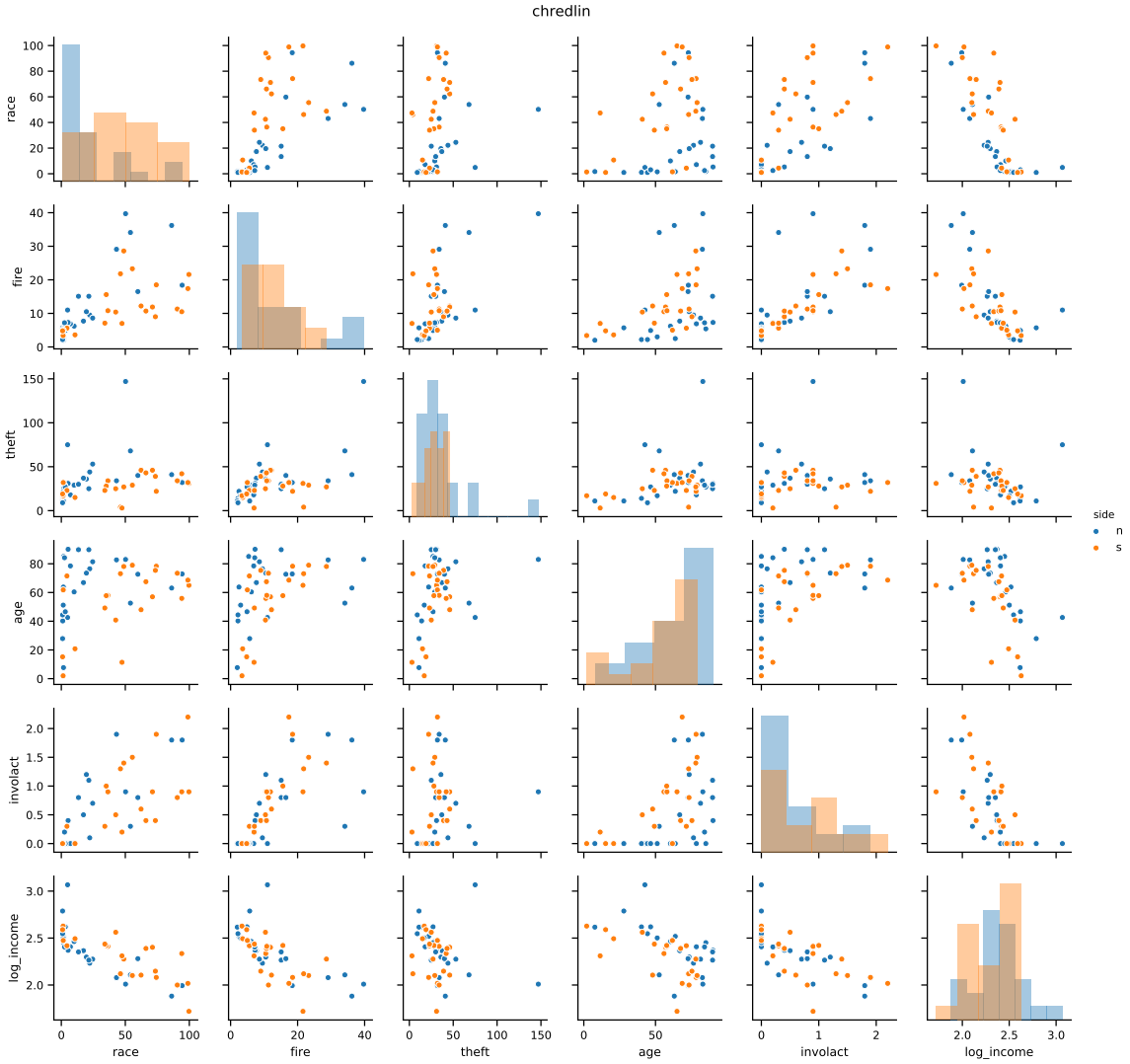

Figure 1: The empirical univariate and joint distributions for the `chredlin` dataset.