Coursework 1: STAT 570

Philip Pham

October 2, 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

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

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

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

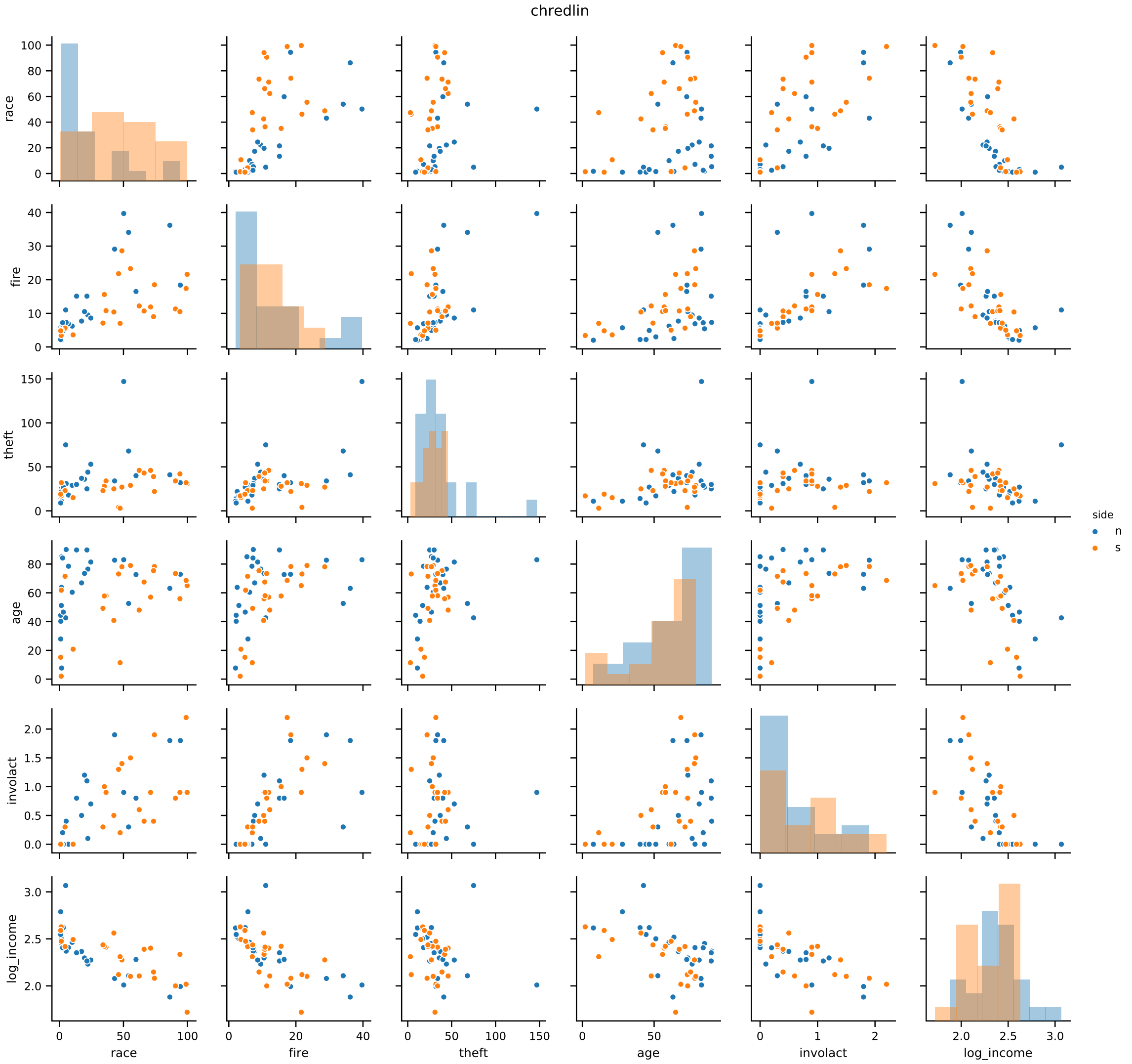

Figure 1: The empirical univariate and joint distributions for the `chredlin` dataset.