Solutions Pattern Recognition 2015
Linear Models for Regression and Classification

1 Linear Regression

(a)

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 31 & 25 & 27 & 23 & 32 & 22 & 29 \end{bmatrix} \begin{bmatrix} 1 & 31 \\ 1 & 25 \\ 1 & 27 \\ 1 & 23 \\ 1 & 32 \\ 1 & 22 \\ 1 & 29 \end{bmatrix} = \begin{bmatrix} 7 & 189 \\ 189 & 5193 \end{bmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{t} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 31 & 25 & 27 & 23 & 32 & 22 & 29 \end{bmatrix} \begin{bmatrix} 80 \\ 105 \\ 120 \\ 105 \\ 70 \\ 120 \\ 100 \end{bmatrix} = \begin{bmatrix} 700 \\ 18540 \end{bmatrix}$$

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{t} = \frac{1}{630} \begin{bmatrix} 5193 & -189 \\ -189 & 7 \end{bmatrix} \begin{bmatrix} 700 \\ 18540 \end{bmatrix} = \frac{1}{630} \begin{bmatrix} 131040 \\ -2520 \end{bmatrix} = \begin{bmatrix} 208 \\ -4 \end{bmatrix}$$

So the fitted model is

$$y(x) = 208 - 4x$$

(b) $w_0 = 208$: this is the expected productivity at a temperature of 0 degrees. This doesn't make any sense: the model is only supposed to hold for temperatures between 20 and 35 degrees.

$w_1 = -4$: this is the change in expected productivity when the temperature increases with one degree.

(c)

$$y(x = 20) = 208 - 4 \times 20 = 128.$$

(d) Some bookkeeping:

n	x_n	t_n	y_n	$t_n - y_n$	$(t_n - y_n)^2$	$t_n - \bar{t}$	$(t_n - \bar{t})^2$
1	31	80	84	−4	16	−20	400
2	25	105	108	−3	9	5	25
3	27	120	100	20	400	20	400
4	23	105	116	−11	121	5	25
5	32	70	80	−10	100	−30	900
6	22	120	120	0	0	20	400
7	29	100	92	8	64	0	0
$\sum$	189	700	700	0	710	0	2150

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{710}{2150} \approx 0.67.$$

2 Linear Models for Classification

(a) `whtvict` and `stranger` ($\alpha = 0.05$) In addition: `aggcirc` and `multstab` ($\alpha = 0.1$)

(b) The fitted probability is $-0.18679 - 0.08692 = -0.27371$. Negative probabilities are not possible according to the axioms of probability. This highlights a shortcoming of the linear probability model.

(c) 0.35639

(d) It appears that black defendants have a lower probability of getting the death penalty, but the coefficient of `blkdef` is not significantly different from zero (p-value: 0.43) at any conventional significance level. On the other hand, if you kill a white person, you have a higher probability of getting the death penalty, and the coefficient of `whtvict` is significant (p-value: 0.013) at $\alpha = 0.05$. One could argue that this is also a form of racial discrimination.

(e) The fitted probability is

$$\hat{p}(t = 1|\boldsymbol{x}) = (1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^{-1} = \left(1 + e^{3.5675 + 0.5308}\right)^{-1} = 0.0166$$

(f) The fitted response function is given by

$$\hat{p}(t = 1|\boldsymbol{x}) = (1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^{-1}.$$

Applying the chain rule twice, and noting that $\frac{d\, e^z}{d\, z} = e^z$, we get

$$\frac{\partial \hat{p}(t = 1|\boldsymbol{x})}{\partial x_i} = -(1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^{-2} \times e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}} \times -w_i = w_i \times \frac{e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}}}{(1 + e^{-\boldsymbol{w}_{\text{ML}}^\top \boldsymbol{x}})^2}$$

Hence we see that the marginal effect of an increase in x_i depends on the value of x_i and also on the value of the other variables. However, the quantity

$$\frac{e^{-\boldsymbol{w}_{\text{ML}}^{\top}\boldsymbol{x}}}{(1+e^{-\boldsymbol{w}_{\text{ML}}^{\top}\boldsymbol{x}})^2}$$

is always positive, so the sign of the influence of an increase in x_i can be read from the sign of w_i. **Note**: in fact,

$$f(z) = \frac{e^z}{(1+e^z)^2}$$

is the probability density function of the standard logistic distribution, and the cumulative distribution function of the standard logistic distribution is given by

$$F(z) = \frac{e^z}{1+e^z},$$

which is the logistic response function (activation function or transfer function in neural network terminology).

3 Logistic Regression

Note: $\exp(x) \equiv e^x$. I use both notations interchangeably.

(a) Not surprising at all. Explanatory variable x denotes the additional time taken by public transport. The more additional time, the higher the probability that a person will take the car. This is exactly what the positive coefficient says.

(b) If traveling by car and public transport takes the same time ($x = 0$), then there is a preference for public transport, because

$$\frac{e^{-0.24}}{1+e^{-0.24}} \approx 0.44 < 0.5.$$

(c) Fill in $x = 30$:

$$\hat{p}(t = 1 \mid x = 30) = \frac{\exp(-0.24 + 0.053 \cdot 30)}{1 + \exp(-0.24 + 0.053 \cdot 30)} \approx 0.794$$

(d) The marginal effect of an increase in x is

$$\frac{\partial \hat{p}(t=1|x)}{\partial x} = 0.053 \times \frac{e^{0.24-0.053x}}{(1+e^{0.24-0.053x})^2}$$

For $x = 5$ this evaluates to 0.016, for $x = 30$ to 0.009. So an increase from 5 to 6 minutes time difference has a larger effect than an increase from 30 to 31 minutes time difference.

(e) If $-0.24 + 0.053x > 0$, predict that someone will take the car, otherwise predict public transport. Further simplification gives: if $x > 4.53$ then car, otherwise public transport. Since travel time is measured in whole minutes, an appropriate verbal description would be: *If, for a given person, travelling by public transport takes 5 minutes or more longer than travelling by car, predict that this person will take the car, otherwise predict that this person will take public transport.*

4 Linear Regression and Logistic Regression

Unfortunately this won't work because in the travel data and the death penalty data the target variable is binary, i.e. $t_n \in \{0, 1\}$. The transformed variable would be

$$z_n = \ln\left(\frac{t_n}{1 - t_n}\right)$$

If $t_n = 0$, the transformed variable has the value $\ln 0$ and if $t_n = 1$ it has the value $\ln \frac{1}{0}$. In either case the value is not defined. In the Google Flu example, the target was already a fraction between 0 and 1 (the fraction of all physician visits that was flu-related in a particular week) and therefore the transformation was possible.

5 The Multinomial Logit Model

(a) Recall that according to the multinomial logit model

$$p(t = k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{\sum_{j=0}^{K-1} \exp(\mathbf{w}_j^\top \mathbf{x})}$$

Therefore

$$\begin{aligned}\ln\left\{\frac{p(t = k|\mathbf{x})}{p(t = \ell|\mathbf{x})}\right\} &= \ln\left\{\frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{\exp(\mathbf{w}_\ell^\top \mathbf{x})}\right\} \\ &= \ln \exp(\mathbf{w}_k^\top \mathbf{x}) - \ln \exp(\mathbf{w}_\ell^\top \mathbf{x}) \\ &= \mathbf{w}_k^\top \mathbf{x} - \mathbf{w}_\ell^\top \mathbf{x} = (\mathbf{w}_k - \mathbf{w}_\ell)^\top \mathbf{x}\end{aligned}$$

(b) For $K = 2$ the multinomial logit model becomes:

$$p(t = 0|\mathbf{x}) = \frac{\exp(\mathbf{w}_0^\top \mathbf{x})}{\exp(\mathbf{w}_0^\top \mathbf{x}) + \exp(\mathbf{w}_1^\top \mathbf{x})} = \frac{1}{1 + \exp(\mathbf{w}_1^\top \mathbf{x})},$$

since $\mathbf{w}_0 \equiv \mathbf{0}$. Furthermore we have:

$$p(t = 1|\mathbf{x}) = \frac{\exp(\mathbf{w}_1^\top \mathbf{x})}{\exp(\mathbf{w}_0^\top \mathbf{x}) + \exp(\mathbf{w}_1^\top \mathbf{x})} = \frac{\exp(\mathbf{w}_1^\top \mathbf{x})}{1 + \exp(\mathbf{w}_1^\top \mathbf{x})}.$$

Summarizing: the multinomial logit model with $K = 2$ is the binary logistic regression model (rename $\mathbf{w}_1$ to $\mathbf{w}$).

(c) For class 2 we compute: $\exp(4.76 - 0.55 \times 16 + 0.43) = 0.03$. Likewise, for class 3 we compute $\exp(-26.01 + 1.63 \times 16 - 2.11) = 0.13$. Hence, we get

$$p(t = 1) = \frac{1}{1 + 0.03 + 0.13} = 0.86$$
$$p(t = 2) = \frac{0.03}{1 + 0.03 + 0.13} = 0.03$$
$$p(t = 3) = \frac{0.13}{1 + 0.03 + 0.13} = 0.11$$

(d) We can interpret $w_{3,1} = 1.63$ as follows. When *years of education* increases with one year, the log-odds of class *management job* versus class *administrative job* increase with 1.63.

Loosely speaking: the more years of education someone has, the relatively more likely it becomes that this person will have a management job as opposed to an administrative job.

That makes sense.