In these notes, you will find material for:

- model with categorical predictors
- multiple linear regression models
- logistic regressions

MODEL WITH CATEGORICAL PREDICTOR(S)

$$y = \alpha + \beta x + \text{error}$$

x : categorical variables $\quad x = \begin{cases} 0 \\ 1 \end{cases}$

[x takes values 0, 1]

when $x = 0$, $\quad y = \alpha$

when $x = 1$, $\quad y = \alpha + \beta$

$$(y \mid x = 1) - (y \mid x = 0) = \hat{\beta}$$

$\hat{\beta}$:- represents the average change in y when x switches from 0 to 1

MODEL WITH CATEGORICAL PREDICTOR(S)

Example : $\text{Wage} = \alpha + \hat{\beta} \times \text{Female}$

Run the model on the dataset 'wage1'

$$\text{Wage} = 7.10 - 2.51 \times \text{Female}$$

$$\hat{\beta} = -2.51$$

interpretation : the average wage for women is \$2.51 lower than that for men.

(check the p-value)

MULTIPLE LINEAR MODEL

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \text{error}$$

$\hat{\beta}_1$: the association b^n x_1 and y holding x_2 constant

$\hat{\beta}_2$: the association b^n x_2 and y holding x_1 constant

MULTIPLE LINEAR MODEL: EXAMPLE

$$Wage = \alpha + \hat{\beta}_1 \times educ + \hat{\beta}_2 \times exper$$

educ : education (in years)

exper : years of experience

The estimated model is :

$$Wage = -3.39 + 0.64 \times educ + 0.07 \times exper$$

interpretation ($\hat{\beta}_1$) : when you add one more year of education, the average wage goes up by 64 cents

interpretation ($\hat{\beta}_2$) : one additional year of experience translates into 7 additional cents in wages.

MODEL WITH CATEGORICAL PREDICTORS (Revisited)

$$y = \alpha + \hat{\beta}_1 \times cat_1 + \hat{\beta}_2 \times cat_2 + \hat{\beta}_3 \times cat_3$$

Let's imagine a predictor x with four categories $cat_1, cat_2, cat_3, cat_0$.

The estimated model will expand the predictor x into three new predictors.

If there are T categories, you will have $T-1$ predictors.

cat_0 will be the "base category".

MODEL WITH CATEGORICAL PREDICTORS (REVISITED)

Example: $GDP = \alpha + \hat{\beta}_1 \times North + \hat{\beta}_2 \times South + \beta_3 \times West$

there are four regions: N, W, S, E (east is the base)

$GDP = 100 + 0.1 \times North + 0.2 \times South + 0.3 \times West$

interpretation($\hat{\beta}_1$): the Northern region of this country has a GDP 0.1 units more than that for the eastern region.

GOODNESS OF FIT

R^2	adjusted R^2
$R^2 = 1 - \frac{SSR}{SST}$	$R^2_{adj} = 1 - \left[\frac{SSR}{SST} \times \frac{n-1}{n-k-1}\right]$
SSR : total sum of squared for residuals	n : number of observations
SST : total sum of squared	k : number of predictors.
	SSR and SST can be computed using ANOVA

GOODNESS OF FIT AND MODEL SELECTION

1. R^2 increases when you add more predictors
2. Adjusted R^2 penalizes addition of predictors
3. Pick the model with the highest adjusted R^2

Model Conditions : Multiple Regression Model

Just like the one variable model, the following conditions should be checked:

- residuals are nearly normal
  - * plot the residuals and check
- residual variance is constant
  - * plot residuals versus each predictor
- linear relationship b^n the outcome and each predictor [* plot Y vs X]

Story so far : we have dealt with numeric outcome variables.

$$Y = \alpha + \beta X$$

Y: Numeric → OLS (linear model); Categorical ?

X: Numeric; Categorical

LOGISTIC REGRESSIONS

- One of the ways in which you can model categorical outcome variable is logistic regressions
- Logistic regressions use a function of odds of an event as the outcome variable
- Outcome variable in this case are binary (yes/no, default/not, pass/fail, etc)

Odds

for any event E,

p = probabilty of success

odds ratio $= \dfrac{p}{1-p}$

recall that our outcome variable is binary.

logit function

$$\text{logit}(p) = \log(\text{odds ratio})$$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

where $0 \le p \le 1$

OUTCOME VARIABLE

BINOMIAL DISTRIBUTION

Logistic Regression Model

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \times x$$

predictor

log of odds ratio

$$\frac{p}{1-p} = e^{\alpha+\beta x}$$

$$\Rightarrow \quad p_i = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

probability of success

Logistic Regression Model : Interpretation

$\hat{\beta}$: for one unit increase in x, by how much will the log odds ratio change.

$\hat{\alpha}$: the log odds ratio when $x = 0$

Logistic Regression Model : Example

$$\text{Callback} = \alpha + \beta \times \text{Black}$$

$$\text{Callback} = -2.34 - 0.44 \times \text{Black}$$

interpretation ($\hat{\beta}$) : when the race changes from black to white, the log of odds of getting a callback falls by -0.44 units.

This is not very insightful. We can retrieve the predicted probabilities of callback for each group.

Predicted Probabilities : example

Model : $\log\left(\frac{p}{1-p}\right) = -2.34 - 0.44 \times black$

Probability of callback for black CV:

$\log\left(\frac{p}{1-p}\right) = -2.34 - 0.44 \times 1 \Rightarrow \frac{p}{1-p} = \exp(-2.78)$

$\Rightarrow \boxed{\hat{p}^{BLACK} = 0.06}$

Probability of callback for white CV:

$\log\left(\frac{p}{1-p}\right) = -2.34 - 0.44 \times 0 \Rightarrow \left(p/1-p\right) = \exp(-2.34)$

$\Rightarrow \boxed{\hat{p}^{WHITE} = 0.1}$