Ch2

- Input variables, predictors, independent variables, features, variables
- Output variable, response, dependent variable
- Reducible error, irreducible error
- Training data, testing data
- Supervised learning, unsupervised learning
- Overfitting
- Bias-variance trade-off

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

$\hat{\beta}$ $se(\hat{\beta})$ $\hat{\sigma}$ $n-p$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5} + \varepsilon_i, i = 1, \ldots, n$$

Assume independent, $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$

P-value for the hypothesis testing:

$H_0: \beta_{Area} = 0$

$H_1: \beta_{Area} \neq 0$

$$RSS = \hat{\varepsilon}^T \hat{\varepsilon}$$

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

$$R^2 = 1 - \frac{RSS}{TSS}, TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Adjust R-square: $adj.\ R^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$

```
##
## Call:
## glm(formula = test ~ pregnant + glucose + bmi + diabetes + age,
##     family = "binomial", data = pima.rm.na)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## pregnant     0.083953   0.055031   1.526 0.127117
## glucose      0.036458   0.004978   7.324 2.41e-13 ***
## bmi          0.078139   0.020605   3.792 0.000149 ***
## diabetes     1.150913   0.424242   2.713 0.006670 **
## age          0.034360   0.017810   1.929 0.053692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
##
## Number of Fisher Scoring iterations: 5
```

$\hat{\beta}$

P-value for the hypothesis testing:

Ex: p-value =0.054

$$H_0: \beta_{age} = 0$$
$$H_1: \beta_{age} \neq 0$$

$$\log \frac{\hat{p}_i}{1-\hat{p}_i} = -9.992 + 0.084 \cdot (pregnamt) + 0.036 \cdot (glucose) + 0.078 \cdot (bmi) + 1.151 \cdot (diabetes) + 0.034 \cdot (age)$$