SLR, MR, Dummy variable, Quadratic term, Model selection, Logistic regression.

SLR $y = \alpha + \beta x + \varepsilon$ (y Quant, x Quant), $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

- estimates: $a = \bar{y} - b\bar{x}$, $b = r\frac{S_y}{S_x}$, $S = \sqrt{MSE}$
- interpretation: y-int (a), slope (b)
- prediction: $\hat{y} = a + bx$ → residual: $y_i - \hat{y}_i$
- correlation: $-1 \le r \le 1$
- coefficient of determination: $R^2 = (r)^2 = SSR/SST$
- inference: C.I., t-test, ANOVA [For SLR, t-test ⇔ ANOVA]
  1. $b \pm t_{n-2, \alpha/2} \cdot S.E._b$
  2. $t_{obs} = \frac{b}{S.E._b}$
  3.

Source	df	SS	MS	F	p-val
Regression	1	SSR	MSR	MSR/MSE	$P(X \ge F)$
Error	n-2	SSE	MSE		
Total	n-1	SST			

where $X \sim F_{1, n-2}$

- Understanding of SS (Page 74)
- CI v.s. PI for response — interpretations: average response (CI) / new individual (PI)
- Diagnostics via residual analysis [Check $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$]
  - Histogram of residual
  - NPP
  - Residual vs fit
  - Residual vs Order/x (Page 79)

Multiple regression $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$, $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

y Quant, x can be categorical

- estimates: $a \to \alpha$, $b_i \to \beta_i$, $\sqrt{MSE} \to \sigma$.
- interpretation: similar to SLR
- prediction

Source	df	...
Reg	p	
Error	n-p-1	...
Total	n-1	

- inference: ANOVA → any good predictor?
  t-test for β_i → Is x_i good pred AFTER all other pred variables are accounted for?

- oversaturated model
- multicollinearity
- C.I. for β_i } both bad.
- R^2 vs. R^2_{adj} (pros & cons)
- Order of testing.

※ Categorical variables in regression

- Dummy variable $X = \begin{cases} 0 \\ 1 \end{cases}$
- Baseline group / model
- Interpretation of coefficients in the full model (α, β_i)
- Interpretation of estimates (a, b_i)
- Prediction
- Interaction (between quantitative terms and categorical terms)
- Type of lines the model allows (parallel? Line or curve?)
- How to test intercept/slope for ... is sig diff than ....

※ Quadratic regression $y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$

- When to use it? ∪ ∩ /
- Interpretation: only for b_2 [sig diff from zero? sign?]
- Prediction
- Comparison to SLR.

※ Selecting the best model

- Backwards elimination, forward selection, best subsets regression
- Goal: Simple model that does a decent job of prediction.
- R^2_{adj}, C_p, AIC/BIC
- Data cleaning: - Unusual observations { "R" - large standardized residual; "X" - large leverage
  - Outlier, influential outlier

※ Logistic regression $P(Y=1) = \frac{e^{\alpha+\beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\alpha+\beta_1 x_1 + \cdots + \beta_p x_p}}$

Y: Binary
we predict a probability

- Interpretation: a - Don't care, b_i - (Sign and) log odds ratio interpretation
- Properties of $f(x) = \frac{e^x}{1+e^x}$ and $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ $\left[\ln\left(\frac{x}{1-x}\right)\right]$
- Equivalent formulation: - $P(Y=1) = \frac{1}{1 + e^{-(\alpha+\beta_1 x_1 + \cdots + \beta_p x_p)}}$
  - $\text{logit}(P(Y=1)) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$
- Prediction ($\hat{p}$) • Solve for x when $\hat{p} = 0.5$.