Tutorial Notes 5 of MATH3424

1 Summary of course material

1.1 The Standard Regression Assumptions

- Assumptions about the form of the model (linearity assumption):

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \varepsilon$$

- Assumptions about the error: The errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are assumed to be independently (independent-errors assumption) and identically distributed normal (normality assumption) random variables with mean zero and a common variance σ^2 (constant variance assumption).

- Assumptions about the predictors:

  1. The predictor variables $X_1, \cdots, X_p$ are nonrandom.
  2. The values $x_{1j}, \cdots, x_{nj}; j = 1, \cdots, p$ are measured without error.
  3. The predictor variables $X_1, \cdots, X_p$ are assumed to be linearly independent of each other.

- Assumptions about the observations: All observations are equally influential on least squares results

1.2 Various Types of Residuals

- $\hat{y} = \underbrace{X(X^\top X)^{-1}X^\top}_{P} y$, elements of projection matrix P is $p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum (x_i - \bar{x})^2}$

- Leverage $p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$

- Studentized residual $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}$ (also referred as standardized residual).

1.3 Checking Linearity and Normality Assumptions

- Normal probability plot of the standardized residuals

- Scatter plots of the standardized residual against each of the predictor variables

- Scatter plot of the standardized residual versus the fitted values

1.4 Leverage, Influence and Outliers

- Points with standardized residuals larger than 2 or 3 standard deviations away from the mean (zero) are called outliers (in response variable).
- The i-th data point is influential if

$$p_{ii} \geq 2\left(\frac{\sum_{i=1}^{n} p_{ii}}{n}\right) = \frac{2(p+1)}{n}$$

  which also measures the outlyingness of the predictor variables.
- Cook's Distance, Welsch and Kuh Measure, Potential-Residual Plot

2 Questions

The table below shows the regression output from a simple linear regression model relating a response variable Y to one predictor variable X. The model is fitted based on $n = 26$ observations.

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-statistic
Regression	(a)	1	(b)	(c)
Residuals	100	(d)	(e)	

Coefficients Table

Variable	Coefficient	s.e.	t-statistic	p-value
Constant	-4	2	-2	0.0569
X	5	(f)	(g)	<0.0001

Other Statistics

$n = 26$	$R^2 =$ (h)	$R^2_{adjusted} =$ (i)	Root MSE $\hat{\sigma} =$ (j)	$\text{var}(Y) = 8$