Indian Institute of Technology Bombay

SI 422: Regression Analysis

Formula sheet for Problem Set 1

January 22, 2020

Data: We observe a paired random sample $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ of size n on the response variable y and the predictor variable x.

Simple linear regression model and assumptions: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; \forall i = 1, 2, \ldots, n$ where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ is random error corresponding to the i-th observation and β_0, β_1 are the unknown intercept and slope parameter of the model. Additionally, we assume that the predictor variable x is non-stochastic.

1. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}_i) = \sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}$,
   $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x}_i)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$
2. t_ν is a t-distributed random variable or t-distribution with degrees of freedom (DF) ν.
3. $t'_{\nu,\delta}$ is a non-central t-distribution with DF ν and non-centrality parameter δ.
4. $(t_\nu)_\alpha$ is the upper α-point or $100(1-\alpha)$-th percentile of the t_ν distribution.
5. $F_{\nu,\eta}$ is a F-distributed random variable or F-distribution with DF ν, η.
6. $F'_{\nu,\eta,\lambda}$ is a non-central F-distribution with DF ν, η and non-centrality parameter λ.
7. $(F_{\nu,\eta})_\alpha$ is the upper α-point or $100(1-\alpha)$-th percentile of the $F_{\nu,\eta}$ distribution.
8. Let X be a random variable and $f(\cdot)$ be a measurable function. Then $f(X)_{\text{observed}}$ is the observed value of $f(X)$.
9. LSE of the slope parameter β_1 and the intercept parameter β_0 are respectively $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.
10. Fitted simple linear regression model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
11. Fitted values are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, 2, \ldots, n$.
12. Residuals due to fit are $e_i = y_i - \hat{y}_i$ for $i = 1, 2, \ldots, n$.
13. Total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ measures total variation of y in the observed data.

14. Sum of squares due to regression fit $SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ measures the variation of y explained by the underlying regression model.

15. Error sum of squares or residual sum of squares $SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (\hat{y}_i - y_i)^2$ measures the variation of y which cannot be explained by the underlying regression model.

16. Fundamental ANOVA identity for regression model is $TSS = SS_R + SSE$.

17. Coefficient of determination $R^2 = \frac{SS_R}{TSS} = 1 - \frac{SSE}{TSS}$ is the proportion of variation of y explained by the underlying regression model. Note that $0 \le R^2 \le 1$.

18. ANOVA table

Source of variation	SS	DF	MS	F stat	p-value
Regression	SS_R	1	$MS_R = SS_R/1$	$F_0 = \frac{MS_R}{MSE}$	$P(F_{1,n-2} > (F_0)_{\text{observed}})$
Error	SSE	n-2	$MSE = SSE/(n-2)$		
Total	TSS	n-1			

19. To test $H_0 : \beta_0 = \beta_{00}$ against $H_1 : \beta_0 \neq \beta_{00}$, we use the test statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

which follows t_{n-2} and t'_{n-2,δ_0} with $\delta_0 = \frac{\beta_0 - \beta_{00}}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$ under H_0 and H_1 respectively. We reject H_0 at $100\alpha\%$ level of significance if

$$|T_0|_{\text{observed}} > (t_{n-2})_{\alpha/2} \quad \text{or equivalently}$$
$$P(|t_{n-2}| > |T_0|_{\text{observed}}) < \alpha.$$

20. To test $H_0 : \beta_1 = \beta_{10}$ against $H_1 : \beta_1 \neq \beta_{10}$, we use the test statistic

$$T_1 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

which follows t_{n-2} and t'_{n-2,δ_1} with $\delta_1 = \frac{\beta_1 - \beta_{10}}{\sigma\sqrt{\frac{1}{S_{xx}}}}$ under H_0 and H_1 respectively. We reject H_0 at $100\alpha\%$ level of significance if

$$|T_1|_{\text{observed}} > (t_{n-2})_{\alpha/2} \quad \text{or equivalently}$$
$$P(|t_{n-2}| > |T_1|_{\text{observed}}) < \alpha.$$

21. A $100(1-\alpha)\%$ confidence interval of β_0 is

$$\left(\hat{\beta}_0 - (t_{n-2})_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \hat{\beta}_0 + (t_{n-2})_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right).$$

22. A $100(1-\alpha)\%$ confidence interval of β_1 is

$$\left(\hat{\beta}_1 - (t_{n-2})_{\alpha/2}\sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + (t_{n-2})_{\alpha/2}\sqrt{\frac{MSE}{S_{xx}}}\right).$$

23. Mean of Y when $x = x_0$ is $E(Y|x_0) = \beta_0 + \beta_1 x_0$ and its unbiased estimator is $\widehat{E(Y|x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Moreover,

$$\widehat{E(Y|x_0)} \sim \mathcal{N}\left(E(Y|x_0), \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right).$$

Thus a $100(1-\alpha)\%$ confidence interval of $E(Y|x_0)$ is

$$\left(\widehat{E(Y|x_0)} - (t_{n-2})_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}, \widehat{E(Y|x_0)} + (t_{n-2})_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right).$$

24. Let y_0 be the observed value of Y when $x = x_0$. The fitted value of Y at $x = x_0$ is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Moreover, we have

$$y_0 - \hat{y}_0 \sim \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right).$$

Thus a $100(1-\alpha)\%$ prediction interval of y_0 is

$$\left(\hat{y}_0 - (t_{n-2})_{\alpha/2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}, \hat{y}_0 + (t_{n-2})_{\alpha/2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right).$$

25. Consider an equidistant partition of $(x_{\min}, x_{\max})$ of length k (say). Suppose it is $x_{new} = ((x_{new})_1, (x_{new})_2, \ldots, (x_{new})_k)$. Then a $100(1-\alpha)\%$ confidence band is

$$\left(E(\widehat{Y|(x_{new}})_j) - (t_{n-2})_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{((x_{new})_j - \bar{x})^2}{S_{xx}}\right)},\right.$$
$$\left. E(\widehat{Y|(x_{new}}_j)) + (t_{n-2})_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{((x_{new})_j) - \bar{x})^2}{S_{xx}}\right)}\right), \quad j = 1, 2, \ldots, k.$$

26. A $100(1-\alpha)\%$ prediction band is

$$\left(E(\widehat{Y|(x_{new}})_j) - (t_{n-2})_{\alpha/2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{((x_{new})_j - \bar{x})^2}{S_{xx}}\right)},\right.$$
$$\left. E(\widehat{Y|(x_{new}}_j)) + (t_{n-2})_{\alpha/2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{((x_{new})_j) - \bar{x})^2}{S_{xx}}\right)}\right), \quad j = 1, 2, \ldots, k.$$