LINEAR REGRESSION SUMMARY

Model (Linear regression). $\mu_{Y|X=x} = \alpha_1 + \beta_1 x$ or $\boxed{y = \alpha_1 + \beta_1 x + \epsilon}$. For most regression analysis we require $\epsilon \sim N(0, \sigma_\epsilon^2)$.

Verify that the linear model is reasonable by looking at a plot of your data: `> plot(y~x)`.

A. Sample statistics

Sample: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Many of the statistics are calculated by R: `> model<-lm(y~x)` and `> summary(model)` will be useful.

$$\overline{x} = \frac{1}{n}\sum x_i \qquad \overline{y} = \frac{1}{n}\sum y_i$$

$$\hat{\alpha}_1 = \overline{y} - \hat{\beta}_1\overline{x} \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$S_{XX} = \sum (x_i - \overline{x})^2 \qquad S_{XY} = \sum (x_i - \overline{x})(y_i - \overline{y})$$

$$\hat{y}_i = \hat{\alpha}_1 + \hat{\beta}_1 x_i \qquad i^{\text{th}} \text{ residual: } e_i = y_i - \hat{y}_i$$

$$SST = S_{YY} = \sum (y_i - \overline{y})^2 \qquad SSE = \sum \left[y_i - (\hat{\alpha}_1 + \hat{\beta}_1 x_i)\right]^2 = \sum e_i^2$$

$$SSR = SST - SSE = \sum (\hat{y}_i - \overline{y})^2 \qquad s_\epsilon^2 = \frac{SSE}{n-2} \text{ (note: } s_\epsilon \text{ is residual standard error)}$$

$$\text{Coefficient of determination } r^2 = \frac{SSR}{SST} \qquad \text{Sample correlation } r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \pm\sqrt{r^2}$$

B. Test statistics and confidence intervals

All assume $\epsilon \sim N(0, \sigma_\epsilon^2)$; you should check on this assumption before proceeding using the plots of residuals vs fitted values and normal Q-Q: `> plot(model)`.

B.1. **Test and interval concerning β_1.** Hypothesis test $H_0 : \beta_1 = c$. Test stat: $t = \dfrac{\hat{\beta}_1 - c}{\frac{s_\epsilon}{\sqrt{S_{XX}}}}$ has a t dist with $n-2$ df. R tests $H_0 : \beta_1 = 0$ against $H_0 : \beta_1 \neq 0$ by default. $100(1-\alpha)\%$ confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{1-\alpha/2, n-2}\frac{s_\epsilon}{\sqrt{S_{XX}}}$$

B.2. **CI and PI for the regression line.** $100(1-\alpha)\%$ confidence interval for $\mu_{Y|X=x}$:

$$(\hat{\alpha}_1 + \hat{\beta}_1 x) \pm (t_{1-\alpha/2, n-2})(s_\epsilon)\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{XX}}}$$

$100(1-\alpha)\%$ prediction interval for Y given $X = x$:

$$(\hat{\alpha}_1 + \hat{\beta}_1 x) \pm (t_{1-\alpha/2, n-2})(s_\epsilon)\sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{XX}}}$$

Date: November 20, 2019.