Regression Checklist

For Simple Linear Regression

Checklist[1]

(1) Look at raw data scatterplot (is it linear?)

(2) State the population model and identify its components

- $y = \beta_0 + \beta_1 x + \epsilon_i$
  - y: response
  - β_0: y-intercept (value of y when $x = 0$)
  - β_1: slope
  - ϵ_i: residual (error) term

(3) Use regression analysis output (provided) to obtain the sample regression equation

- Use equation for estimations
  - $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$ with $\hat{\beta_0}$ and $\hat{\beta_1}$ as numerical values found on the provided regression output
- Calculate residuals
  - $e = y - \hat{y_i} = observed - estimated$

(4) Interpret slope and intercept **in context**. If one does not make logical sense, state the reason(s)

- Slope: the change in y (either increase or decrease based on the value of the slope) due to a one-unit increase in x
- Intercept: the value of y when $x = 0$. May not make logical sense in context, especially when $x = 0$ is not a value in the observed dataset

(5) Hypothesis tests for slope (and intercept if appropriate)

- Results for hypothesis tests on provided regression analysis output
- State hypotheses, test statistic, *pvalue*, results, and conclusion (same basic steps as learned starting in Module 8)
  - Slope: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
  - Intercept (only if appropriate – see step 4): $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$

(6) Correlation (r) and the Coefficient of Determination (R^2)

- R^2 is listed as `Multiple R-square` on the output; R^2 (convert to a percent) is the percent of the variation in the estimated response that can be explained by the model
  - Want $R^2 \geq 60\%$
- r is not on the output but if $R^2 = (r)^2$, then $r = \pm\sqrt{R^2}$ and the sign is the same as the slope (if the slope is negative, r is negative; if the slope is positive, r is positive)
  - $|r| \geq 0.8$: strong
  - $0.6 \leq |r| < 0.8$: moderate
  - $0.4 \leq |r| < 0.6$: fair
  - $|r| < 0.4$: weak

[1] Disclaimer: Some results may vary. In other words, no two of my checklists to date have been identical but all contain the same basic procedures :-)

(7) List assumptions and check them (make brief but specific mentions of how they are/are not met)

   (1) $E(\epsilon_i) = 0$ (mean of residuals ≈ 0); histogram of residuals should be centered at 0 (largest bar right at around 0 on the x-axis)

   (2) $V(\epsilon_i) = \sigma_\epsilon^2$ (variance of residuals is constant); plot of `Residiuals vs. Predicted` shows no pattern in the plot

   (3) $Cov(\epsilon_i, \epsilon_j) = 0$ (independence of residuals); no check for this, assume it is met

   (4) $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ (normality of residuals); histogram of residuals should be approximately normal/symmetric **OR** most points on the QQplot are along the $y = x$ line

(8) Overall assessment of the model using specific references to numbers 5, 6, and 7 from this checklist. If all are "good", then you have a good (decent) model.[2]

[2] For a detailed example that follows the basics of this checklist, see `Final Exam Review` on the class website.