# Assignment 2: Selected solutions

October 13, 2021

- Problem 1:
  Relate $Y$ to $X_2$: $\hat{Y} = 42.1087 + 0.4239X_2$, with residuals $e_{Y\cdot X_2}$
  Relate $X_1$ to $X_2$: $\hat{X}_1 = 34.3196 + 0.6075X_2$, with residuals $e_{X_1\cdot X_2}$
  Relate $e_{Y\cdot X_2}$ to $e_{X_1\cdot X_2}$ : $\hat{e}_{Y\cdot X_2} = 0.78035e_{X_1\cdot X_2}$

- Problem 2: see R file

  (a) see R file.

  (b) In all three models, the null hypothesis $\beta_0 = 0$ is rejected according the output in R.

  (c) $P_2$ is better, for a larger $R^2$.

  (a) $P_3$ is the best, for the largest adjusted $R^2$. In this case, $\hat{F} = 80.71282$ with prediction interval $(71.79724, 89.6284)$

- Problem 3:

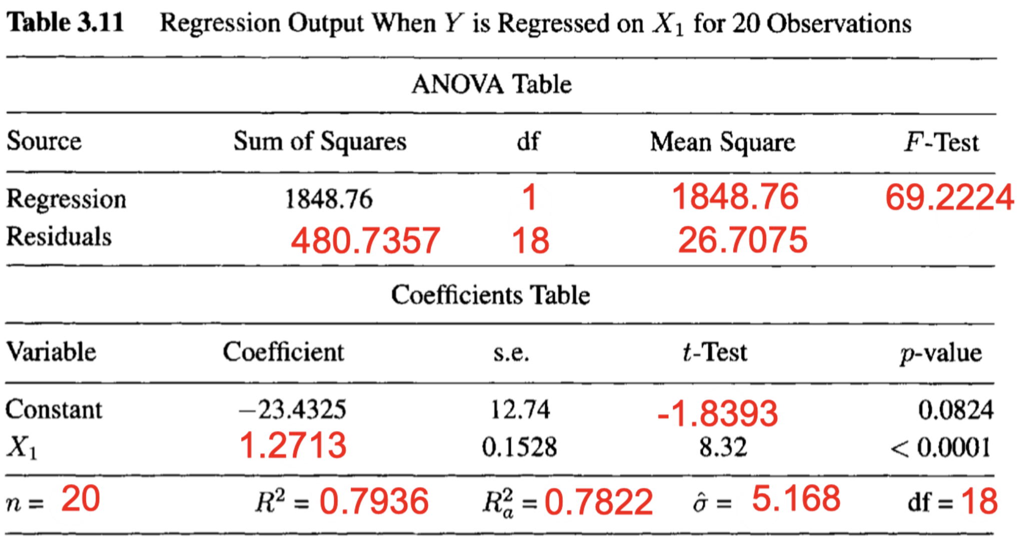

**Table 3.11** Regression Output When $Y$ is Regressed on $X_1$ for 20 Observations

ANOVA Table

| Source | Sum of Squares | df | Mean Square | $F$-Test |
|---|---|---|---|---|
| Regression | 1848.76 | 1 | 1848.76 | 69.2224 |
| Residuals | 480.7357 | 18 | 26.7075 | |

Coefficients Table

| Variable | Coefficient | s.e. | $t$-Test | $p$-value |
|---|---|---|---|---|
| Constant | −23.4325 | 12.74 | -1.8393 | 0.0824 |
| $X_1$ | 1.2713 | 0.1528 | 8.32 | < 0.0001 |
| $n =$ 20 | $R^2 =$ 0.7936 | $R^2_a =$ 0.7822 | $\hat{\sigma} =$ 5.168 | df = 18 |

- Problem 4: see R file. For both model (a) and model (b), we do not reject the null hypothesis according to the F-statistic.

- Problem 5:

(a) $F = 22.98 > F_{(4,88;0.05)} = 2.475277$. Note the $F_{(4,88;0.05)}$ is the upper tail probability of F distribution with $\alpha = 0.05$, which can be checked via R using command

```
qf(0.05,4,88, lower.tail =FALSE) or qf(0.95,4,88, lower.tail =TRUE)
```

(b) Conside the hypothesis testing $H_0 : \beta_3 = 0$ v.s. $H_1 : \beta_3 > 0$. Since $t_3 = 2.16 > t_{0.05,88} = 1.662354$, hence the null hypothesis is rejected and there is a positive linear relationship between salary and experience.

(c)&(d) $\widehat{Salary} = 3526.4 + 722.5 Gender + 90.02 Education + 1.2690 Experience + 23.406 Months = 5692.92$

(e) $\widehat{Salary} = 3526.4 + 722.5 Gender + 90.02 Education + 1.2690 Experience + 23.406 Months = 4970.42$

- Problem 6: Let $X_1 = Gender, X_2 = Education, X_3 = Experience, X_4 = Months$. The hypothesis testing becomes

$$H_0 : Y = \beta_0 + \beta_2 X_2 + \varepsilon \quad \text{v.s.} \quad H_1 : Y = \beta_0 + \beta_2 X_1 + \beta_2 X_2 + \beta_2 X_3 + \beta_2 X_4 + \varepsilon$$

Then

$$F = \frac{[SSE(RM) - SSE(FM)]/(4+1-2)}{SSE(FM)/(93-4-1)} = 20.459 > F_{(3,88;0.05)} = 2.708186$$

The null hypothesis is rejected and hence at least one of $Gender, Experience, Months$ would be statistically significant in predicting $Salary$, we shall prefer the full model.

- Problem 7: see R file. According to the residual versus fitted value plot, the linearity assumption is violated. Also, potential-residual plot implies that the assumption of equally reliability of each observation is kind of violated. Assumptions about the constant variance of error terms is also violated, according to standardized residual against the predictor variable.

- Problem 8: Point 1 and 2 are high-leverage points/outliers in X-space (not influential points as their residuals are small). Point 3 is high-leverage/outlier in both X and Y space/influential point. Point 4 is an outlier in Y-space.

- Problem 9:

(a) scatter plot of response against each predictor variable / rotating plot / scatter plot of the standardized residual against each predictor variable / scatter plot of the standardized residual versus the fitted values

(b) index plot of residuals (meaningful when index is in natural order like time)

(c) scatter plot of standardized residual against fitted value

(d) scatter plot of standardized residual against each predictor variable

(e) normal q-q plot

(f) index plot of Cook's distance / potential-residual plot

- Problem 10: see R file. Omitted.

- Problem 11: You could use potential-residual plot/index plot of Cook's distance to identify the unusual points. The point $(0, 8.11)$ is a high-leverage point and $(5, 11.00)$ is an outlier (both influential).