Regression Analysis (I) Quiz 1 solution

Po-Wei Chen

October 28, 2022

1. (a) A simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad 3\%$$

where:

(1) Y_i: the value of the response variable in the ith trial.

(2) β_0 and β_1: parameters to be estimated.

(3) X_i: the value of the predictor variable in the ith trial.

(4) ϵ_i: a random error term with mean $E(\epsilon_i) = 0$ and variance $\sigma^2(\epsilon_i) = \sigma^2$. 1%

(5) ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma(\epsilon_i, \epsilon_j) = 0$ for all $i, j; i \neq j$) $i = 1, \cdots, n$. 1%

(b) The parameters β_0 and β_1, in simple regression model are called regression coefficients.

(1) The parameter β_0 is the Y intercept of the regression line. β_1, is the slope of the regression line. 2%

(2) β_1 indicates the change in the mean of the probability distribution of Y per unit increase in X. 2%

(3) When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$. When the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a separate term in the regression model. 1%

5% (c) To set up interval estimates and make tests, however, we need to make an assumption about the form of the distribution of the error terms ϵ_i: they are normally distributed.

2. $\because Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, for all $i = 1, ..., n$

$\Rightarrow E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i) = \beta_0 + \beta_1 X_i + 0 = \beta_0 + \beta_1 X_i$

and $\sigma^2(Y_i) = \sigma^2(\beta_0 + \beta_1 X_i + \epsilon_i) = 0 + 0 + \sigma^2(\epsilon_i) = \sigma^2$, for all $i = 1, ..., n$

Since a linear combination of normally distributed random variables is also normally distributed.

$\Rightarrow Y_i \overset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, for all $i = 1, ..., n$

$\Rightarrow f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$, for all $i = 1, ..., n$

3% $\Rightarrow L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2\right]$$

3% $\Rightarrow \log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

$$\Rightarrow \begin{cases} \frac{\partial \log L}{\partial \beta_0} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-1) \\ \frac{\partial \log L}{\partial \beta_1} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) \\ \frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 \end{cases}$$

Let the partial derivatives equal to zero, replacing β_0, β_1 and σ^2 by the estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$.

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\ \sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\ \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} &= \hat{\sigma}^2 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \quad \text{3\%} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{3\%} \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \quad \text{3\%} \end{cases}$$

3. (a) $\because \sum_{i=1}^{n}(X_i-\bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0$

5%

$$\Rightarrow \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n}(X_i-\bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i-\bar{X}) = \sum_{i=1}^{n}(X_i-\bar{X})Y_i$$

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})Y_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \sum_{i=1}^{n} k_i Y_i, \text{ where } k_i = \frac{X_i-\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Hence b_1 is a linear combination of the observation Y_i.

5% (b) For normal error regression model, we know that $Y_i \overset{ind}{\sim} \mathcal{N}(\beta_0+\beta_1 X_i, \sigma^2)$, for all $i = 1, ..., n$.

Since b_1 is a linear combination of the observation Y_i and a linear combination of normally distributed random variables is also normally distributed.

Hence the sampling distribution of b_1 is normal.

(c) $\because b_1 = \sum_{i=1}^{n} k_i Y_i$, where $k_i = \frac{X_i-\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

1% $$\Rightarrow \sum_{i=1}^{n} k_i = \sum_{i=1}^{n} \frac{(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n} X_i - n\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = 0$$

$$\because \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X}) = \sum_{i=1}^{n}(X_i-\bar{X})X_i - \bar{X}\sum_{i=1}^{n}(X_i-\bar{X}) = \sum_{i=1}^{n}(X_i-\bar{X})X_i$$

1% $$\Rightarrow \sum_{i=1}^{n} k_i X_i = \frac{\sum_{i=1}^{n}(X_i-\bar{X})X_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})X_i}{\sum_{i=1}^{n}(X_i-\bar{X})X_i} = 1$$

1% $$\Rightarrow E(b_1) = E\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} k_i E(Y_i) = \sum_{i=1}^{n} k_i(\beta_0+\beta_1 X_i) = \beta_0\sum_{i=1}^{n} k_i + \beta_1\sum_{i=1}^{n} k_i X_i = \beta_1$$

1% $$\because \sum_{i=1}^{n} k_i^2 = \sum_{i=1}^{n}\left(\frac{(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(\sum_{i=1}^{n}(X_i-\bar{X})^2)^2} = \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

1% $$\Rightarrow \sigma^2(b_1) = \sigma^2\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} k_i^2\sigma^2(Y_i) = \sum_{i=1}^{n} k_i^2\sigma^2 = \sigma^2\sum_{i=1}^{n} k_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

(d) $\because \frac{b_1-\beta_1}{s(b_1)} \sim t_{(n-2)}$ 2%

$$\Rightarrow P\left(t_{\alpha/2;n-2} \le \frac{b_1-\beta_1}{s(b_1)} \le t_{1-\alpha/2;n-2}\right) = 1-\alpha \quad 1\%$$

$\because t_{\alpha/2} = -t_{1-\alpha/2}$ 1%

$$\Rightarrow P\left(b_1 - t_{1-\alpha/2;n-2}\times s(b_1) \le \beta_1 \le b_1 + t_{1-\alpha/2;n-2}\times s(b_1)\right) = 1-\alpha$$

Hence the $(1-\alpha)\%$ confidence interval for β_1 is $b_1 \pm t_{1-\alpha/2;n-2}\times s(b_1)$. 1%

4. (a) For the simple linear regression case, the full model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad 3\%$$

The error sum of squares for the full model is

$$SSE(F) = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE \qquad 2\%$$

(b) Consider $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, the model when H_0 holds is called the reduced model:

$$Y_i = \beta_0 + \epsilon_i \qquad 3\%$$

The error sum of squares for the reduced model is

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \bar{Y})^2 = SSTO \qquad 2\%$$

5% (c) The actual test statistic is a function of $SSE(R) - SSE(F)$,

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \Big/ \left(\frac{SSE(F)}{df_F} \right)$$

which follows the F distribution when H_0 holds.

5% (d) For testing whether or not $\beta_1 = 0$, we therefore have:

$$SSE(R) = SSTO, \quad SSE(F) = SSE, \quad df_R = n-1, \quad df_F = n-2,$$

so that we obtain

$$F^* = \left(\frac{SSTO - SSE}{(n-1) - (n-2)} \right) \Big/ \left(\frac{SSE}{n-2} \right) = \left(\frac{SSR}{1} \right) \Big/ \left(\frac{SSE}{n-2} \right) = \frac{MSR}{MSE}$$

which is identical to the analysis of variance test statistic.

5. (a) This is not necessarily correct. In the Toluca Company example, we saw that the coefficient of determination was high ($R^2 = 0.82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to pennit management to schedule workers effectively. 5%

5% (b) This is not necessarily correct. Figure 1(a) shows a scatter plot where the coefficient of determiination is high ($R^2 = 0.69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.

5%

(c) This is not necessarily correct. Figure 1(b) shows a scatter plot where the coefficient of determination between X and Y is $R^2 = 0.02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.

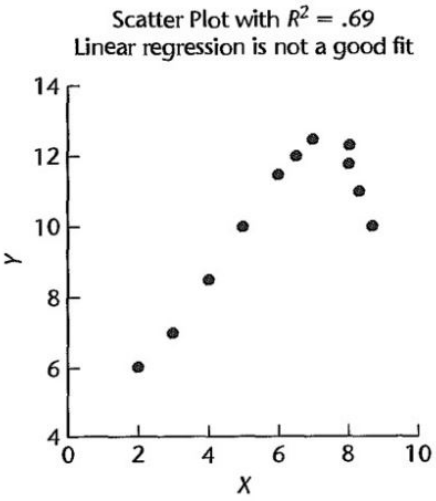


Figure 1(a)

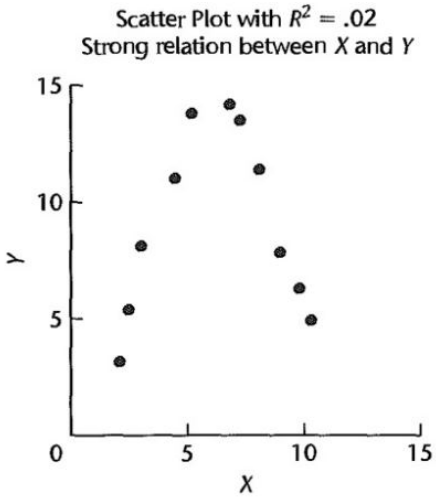


Figure 1(b)

6. (a) $\because \sum_{i=1}^{10} X_i = 10$ and $\sum_{i=1}^{10} X_i^2 = 20$

$$\Rightarrow \sum_{i=1}^{10}(X_i - \bar{X})^2 = \sum_{i=1}^{10} X_i^2 - 10\bar{X}^2 = 20 - 10 = 10$$

$$\because \sum_{i=1}^{10} X_i = 10, \sum_{i=1}^{10} Y_i = 142, \text{ and } \sum_{i=1}^{10} X_i Y_i = 182$$

$$\Rightarrow \sum_{i=1}^{10}(X_i - \bar{X})(Y_i - \bar{Y}_i) = \sum_{i=1}^{10} X_i Y_i - 10\bar{X}\bar{Y} = 182 - 10 \times 14.2 = 40$$

$$\because b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

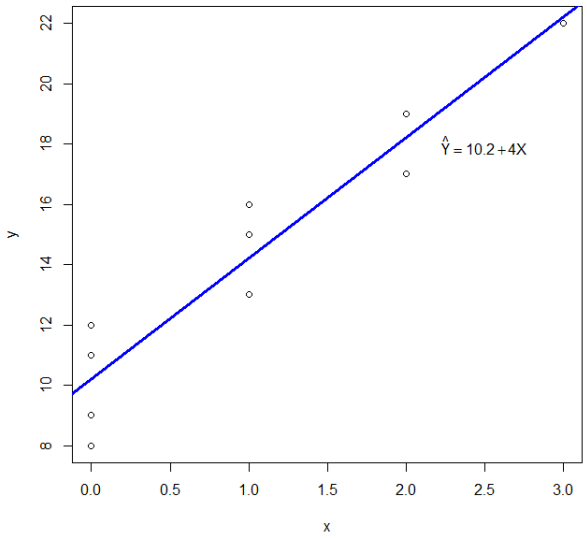
$$\Rightarrow b_1 = \frac{40}{10} = 4$$ 2%

$$\because b_0 = \bar{Y} - b_1\bar{X}$$

$$\Rightarrow b_0 = 14.2 - 4 \times 1 = 10.2$$ 2%

Hence the estimated regression function is $\hat{Y} = 10.2 + 4X$. 1%

5% (b)



5% (c) When $X = 1$.

$\Rightarrow \hat{Y} = 10.2 + 4 \times 1 = 14.2$

Hence we estimate that the expected number of broken ampules of $X = 1$ transfer is made is 14.2.

5% (d) We know that $\hat{Y} = 10.2 + 4X$.

Hence we estimate that the mean the number of broken ampules increases by 4 units for each additional unit time the carton was transferred.

7. (a) $\because \sum_{i=1}^{10}(Y_i - \bar{Y})^2 = \sum_{i=1}^{10} Y_i^2 - 10\bar{Y}^2 = 2194 - 10 \times 14.2^2 = 177.6$

$\Rightarrow SSE = SSTO - SSR = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - SSR = 177.6 - 160 = 17.6$

$\Rightarrow MSE = \frac{SSE}{n-2} = \frac{17.6}{10-2} = 2.2$

$\Rightarrow F = \frac{MSR}{MSE} = \frac{SSR/1}{MSE} = \frac{160}{2.2} \approx 72.7273$

Source of Variation	SS	df	MS	F
Regression	160	1	160	72.7273
Error	17.6	8	2.2	
Total	177.6	9		

(b) (1) $\begin{cases} H_0: & \beta_1 = 0 \\ H_a: & \beta_1 \neq 0 \end{cases}$

(2) Significance leval: $\alpha = 0.05$

(3) Test statistic: $F^* = \dfrac{MSR}{MSE}$

(4) Decision rule: $\begin{cases} \text{If } F^* \leq F(1-\alpha; 1, n-2), \text{ conclude } H_0 \\ \text{If } F^* > F(1-\alpha; 1, n-2), \text{ conclude } H_a \end{cases}$

(5) Decision:

$\because \alpha = 0.05$ and so $1 - \alpha = 0.95$

$\Rightarrow F(1-\alpha; 1, n-2) = F(0.95; 1, 8) = 5.3177$

$\because F^* \approx 72.7273 > 5.3177 = F(0.95; 1, 8)$

$\Rightarrow$ reject H_0 at $\alpha = 0.05$

(6) Conclusion:

Hence there is a linear association between the number of times a carton is transferred and the number of broken ampules.

(c) $\because SSR = 160$ and $SSTO = 177.6$

$\Rightarrow R^2 = \dfrac{SSR}{SSTO} = \dfrac{160}{177.6} \approx 0.9009$

$\Rightarrow r = \pm\sqrt{R^2} \approx \pm 0.9492$

$\because$ the slope of the fitted regression line is positive (i.e., $b_1 > 0$)

$\Rightarrow r \approx 0.9492$

$R^2 \approx 0.9009$, it means the variation in the number of broken ampules is reduced by 90.09 percent when the number of times a carton is transferred is considered.