Teresa Liu

CSCI 467 PS2

1) a) β_0 : If all other inputs are zero, then the baseline amount of funding that companies obtain on the crowdsourcing website will be \$964,800.

β_1 : For companies with the same # of employess hired (x_2), both in/not in IT (x_3), same age (x_4) and same founders history (x_5), amount of funding is predicted to be \$700200 higher for every unit increase in average annual founder salary

β_2 : For companies with the same avg founders' salary (x_1), bot in or not in IT (x_3), same age (x_4) and same founders' failure history (x_5), the amount of funding is predicted to be \$317,500 higher for every additional employee that startup hires.

β_3 : For companies with the same avg founders' salary (x_1), same # of employees (x_2), same age (x_4), and same founders' failure history, the amount of funding is predicted to decrease by \$200,200 if that company's field is information technology.

β_4 : For companies with the same avg founders salary (x_1), same # of employees (x_2), both in IT or not (x_3), and same founders' failure history (x_5), the predicted amount of funding is expected to increase by \$15,300 for every additional unit of age of the company.

β_5 : For companies with the same avg founders salary (x_1), same # of employees (x_2), both in or not in IT (x_3), and same age (x_4), the amounting of funding is expected to increase by \$17,100 if the founders had previous failures.

b) $n = 26$, $\alpha = 0.02$ $\quad t = \frac{\beta_i - 0}{SE(\beta_i)} = \frac{17.1 - 0}{2.3} = \boxed{7.434}$ deg of freedom: $26 - 5 - 1 = 20$

Reject if $t_{obs} > t_{n-2,\alpha/2}$ $\quad t_{20,0.01} = 2.528 < 7.434$ so we reject the null H_0 meaning that there is a relationship between output amount of funding and if founders' had previous failures (x_5)

c) $\beta_4 \pm (t^*_{20} \times SE(\beta_4))$ $\quad t^*_{20} = 3.552$ $\quad v = 0.001$

$\beta_4 \pm (3.552)(45.3) \Rightarrow 15.3 \pm 160.9056$

$\boxed{[-145.6056, 176.2056]}$ 99.8% conf interval for β_4

1c cont) For companies with the same average salary of founders (X_1), same # of employees hired (X_2), both in / not in IT (X_3), and same founder's failure history (X_5) we are 99.8% sure that the amount of funding will be between \$145,605.60 lower to \$176,205.60 higher for a company with 1 more unit of age.

d) Reg SS = 18147.5 RSS = 17136.5 $\alpha = 0.05$

$$F = \frac{(TSS - RSS)/p}{RSS/n-p-1} \sim F_{p,n-p-1} \qquad n=26 \quad p=5$$

$$TSS = Reg\,SS + RSS$$

$$F = \frac{((18147.5 + 17136.5) - 17136.5)/5}{17136.5/(26-5-1)}$$

$$= \frac{3629.5}{856.825} = \boxed{4.236} \qquad F_{5,20,0.05} = 2.7109$$

Reject H_0 if $F > F_{5,20,0.05}$

Since $4.236 > 2.7109$, we reject H_0 that states that none of the predictors are significant predictors of the output variable. Since we reject the null, this means that there is at least one predictor that is a significant predictor of the output variable.

2) a) $RSS = \sum (y_i - \hat{y}_i)^2$ $\quad y = \beta_1 X + \epsilon$

$\hat{y}_i = \hat{\beta}_1 x_i$

$$\boxed{RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2}$$

b) $$\frac{d}{d\beta_1} RSS = \frac{d}{d\beta_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2$$

$$0 = \sum_{i=1}^{n} \frac{d}{d\beta_1} (y_i - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_1 x_i) \cdot (-x_i)$$

$$0 = -2 \sum_{i=1}^{n} (x_i y_i - \hat{\beta}_1 x_i^2) = \sum_{i=1}^{n} x_i y_i - \sum_{j=1}^{n} \hat{\beta}_1 x_j^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \sum_{j=1}^{n} x_j^2$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2}}$$

c) With X_{new} as X_{n+1}: $\quad y = \hat{\beta}_1 X$

$$\boxed{\hat{y}_{new} = \frac{\sum_{i=1}^{n+1} x_i y_i}{\sum_{j=1}^{n+1} x_j^2} \cdot X_{new}}$$

d) We minimize RSS by taking into consideration every x_i and y_i from 1 to n to calculate the coef β_1. This is essentially similar to KNN Regression where $k = n$ since we look at all n nearest neighbors. Part C is a special case because we need to use the predicted y_{new} and X_{new} calculate the y_{new} point, so $k = n+1$ in this case. The similarity between (X_{new}, X_i) in this case would be $\frac{1}{X_{new} - X_i}$ since distances and similarity are inversely proportional.

3) $F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$ $\qquad R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

$$F = \frac{TSS - RSS}{p} \cdot \frac{n-p-1}{RSS} = \frac{n-p-1}{p} \cdot \frac{TSS - RSS}{RSS} \; *$$

$$* \; \frac{TSS - RSS}{RSS} = \frac{TSS \cdot R^2}{RSS} \quad \text{since} \quad R^2 \cdot TSS = TSS - RSS$$

$$= R^2 \cdot \frac{1}{1-R^2} \quad \text{since} \quad 1 - R^2 = \frac{RSS}{TSS} \Rightarrow \frac{1}{1-R^2} = \frac{TSS}{RSS}$$

so $\boxed{F_1 = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2}}$

4) Null $H_0 = \beta_1 = \beta_2 = \ldots = \beta_7 = 0$ $\quad \alpha = 0.05$

$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$ (2)

deg of freedom $= n - p - 1 = \overset{n}{20} - \overset{p}{7} - 1 = 12$

$F_{7,12,0.05} = 2.9134 \Rightarrow$ Reject H_0 if $F > \underset{\substack{\downarrow \\ 2.9134}}{F_{7,12,0.05}}$

$$F = \frac{12}{7} \frac{R^2}{1-R^2} = 2.9134$$

$$12R^2 = 2.9134(7 - 7R^2)$$

$$12R^2 = 20.3938 - 20.3938R^2$$

$$32.3938R^2 = 20.3938$$

$$R^2 = 0.6296$$

so threshold $\boxed{t = 0.6296}$

5) $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \ldots X_p^{\beta_p} \times \epsilon$

How the β_i's can be calibrated using linear regression:

$$\log(Y) = \log(\beta_0 X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} \ldots X_p^{\beta_p} \times \epsilon)$$
$$= \log(\beta_0) + \log(X_1^{\beta_1}) + \ldots + \log(X_p^{\beta_p}) + \log(\epsilon)$$
$$= \log(\beta_0) + \beta_1 \log(X_1) + \ldots + \beta_p \log(X_p) + \log(\epsilon)$$

Calibrate β_i's by taking the log of the input X vector and the log of the training output data Y to transform the multiplicative regression to linear regression. Calculate the β_i's from the linear regression model as normal.