STAT40790 – Predictive Analytics 1 Project

Brian Buckley 14203480

1. Introduction

The objective of this study was to determine whether the population density of an area is a good indication of the crime rate.

Data were collected on the population density (number of people per unit area) and the crime rate (per 100,000 people) for 6 cities.

Given the question asked, we therefore take the population density as the explanatory variable (x) and the Robbery rate as the output (predicted) variable (y). The regression method of least squares was used in this analysis together with ANOVA analysis.

2. Data

The data is shown in table 1 below.

Population density (number of people per unit area	Robbery rate (per 100,000 people)
59	209
45	180
75	195
72	186
89	200
70	204

S_{XX} = 1119.33, S_{XY} = 304.67, SS_E = 522.41, $\bar{X}$ = 68.33, $\bar{Y}$ = 195.67

Table 1: Population versus crime rate for 6 cities

3. Analysis

Figure 1 is a scatter plot of the data and a linear regression fit with 95% CI to the data using R. The plot visually suggests the fitted positive linear relationship between population density and robbery rate could be tenuous given the spread of the confidence interval and the large residuals.

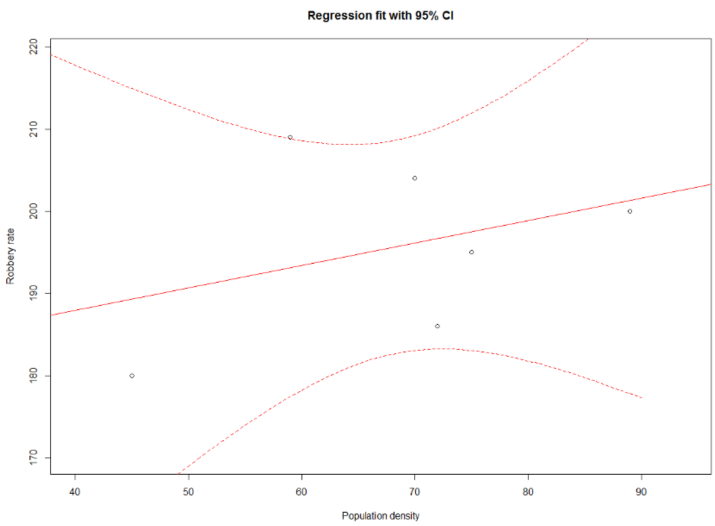

Figure 1: Scatter plot of population density against robbery rate with a linear model fitter using R

For a linear model to hold for the data we must assume the expected value of the sum of the residuals is zero and the sum of the observed values (Y_i) equals the sum of the fitted values ($\hat{Y}_i$). Both hold for this data set so we conclude a linear model is appropriate in this instance notwithstanding the concerns raised above.

The least squares estimates for intercept and slope are shown in the analysis results below in table 2 ($\hat{\beta}_0$ = 177.084, $\hat{\beta}_1$ = 0.272).

The t-test and formal F-test suggest that crime rate is not related to population density. A search for further evidence resulted in corroborative evidence from the research community (e.g. see reference 1).

We carried out both a t-test hypothesis and ANOVA with F-test hypothesis. The t-test result (0.796) is much less than t-critical (2.7) so we fail to reject the null hypothesis that the slope is zero. The ANOVA SS_R is smaller than SS_E so the error term is more significant. Also the F-test for the data (0.644) is much less than F-critical read from the tables (7.709) so again we fail to reject the null hypothesis that the slope is probably zero. We conclude that this is not a good model as the error component is greater than the regression component.

This model estimates the robbery rate is 200 per 100,000 people when the population density is 85 people per unit area.

4. Analysis Results

Least squares estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$	$\hat{\beta}_0$ = **177.084**, $\hat{\beta}_1$ = **0.272**
Mean Squared Error	MS_E = **130.603**
95% CI for $\hat{\beta}_0$	(**112.82**, **241.35**)
95% CI for $\hat{\beta}_1$	(**-0.65**, **1.19**)
H_0: $\hat{\beta}_1 = 0$ vs H_1: $\hat{\beta}_1 \neq 0$, $\alpha = 0.05$	$T = 0.796 < t_{crit} = 2.7$ therefore **fail to reject H_0**
E(Y) when X* = 85	Robbery rate ~ **200** per 100,000 people
95% CI for E(Y) when X* = 85	(**180.32**, **220.08**)
95% PI for Y* when X* = 85	(**163.5**, **236.9**)
ANOVA Analysis	MS_R = **82.87** , MS_E = **128.66**
Formal F-test	$F = 0.644 < F_{1,4}(95\%) = 7.709$ therefore **fail to reject H_0**

Table 2: Full analysis results

5. References

[1] Nolan, Establishing the statistical relationship between population size and UCR crime rate: Its impact and implications, Journal of Criminal Justice 32 (2004) 547 - 555

Appendix 1

Mathematical Workings

(a) Linear model assumptions

1) $\sum_{i=1}^{n} \epsilon_i = 0$ $\qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \; i=1 \ldots n$

x_i	Y_i	$\hat{Y}_i$	$(Y_i - \hat{Y}_i) = \epsilon_i$
59	209	193	16
45	180	189	−9
75	195	198	−3
72	186	197	−11
89	200	201	−1
70	204	196	8
410	1174	1174	$\emptyset = \sum_{i=1}^{n} \epsilon_i$

2) $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

(a) Linear model assumptions:

$$\epsilon_i \sim N(\phi, \sigma^2) \text{ i.i.d }, \text{ and}$$

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

(b) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$; $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$\therefore$ $\hat{\beta}_1 = \frac{304.67}{1119.33} = \boxed{0.272}$

$\hat{\beta}_0 = 195.67 - (0.272)(68.33) = \boxed{177.084}$

(c) $MS_E = \frac{SS_E}{n-2} = \frac{522.41}{4} = \boxed{130.603}$

95% CI for $\hat{\beta}_0$

$$\hat{\beta}_0 \pm t_{1-\frac{\alpha}{2}, n-2} \sqrt{MS_E\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)}$$

$$\pm t_{0.975, 4} \sqrt{130.603\left(\frac{1}{6} + \frac{68.33^2}{1119.33}\right)}$$

$= 177.084 \pm 64.266$

$= \boxed{(112.818, 241.35)}$

95% CI for $\hat{\beta}_1$:

$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}, n-2} \sqrt{\frac{MS_E}{S_{xx}}}$$

$$= 0.272 \pm 2.7 \sqrt{\frac{130.603}{1119.33}}$$

$$= 0.272 \pm 0.92$$

$$= \boxed{(-0.648, 1.192)}$$

Hypothesis test for $H_0: \beta_1 = 0$

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$, $\alpha = 0.05$

NCSS, 9 $t_{crit} = t_{0.975, 4} = 2.7$

$$T_{data} = \frac{\hat{\beta}_1 - m}{\sqrt{\frac{MS_E}{S_{xx}}}} = \frac{0.272 - 0}{\sqrt{\frac{130.603}{1119.33}}} = 0.796$$

compare T_{data} to t_{crit}:

$$0.796 < 2.7$$

$\therefore$ Fail to reject $H_0: \beta_1 = 0$. therefore there is evidence to suggest slope = 0. and therefore no evidence to suggest crime rate is related to population density.

(d) ANOVA

Source of Variation	SS	df	MS
Regression	82.87	1	82.87
Error	514.63	4	128.66
Total_c	597.5	5	

Source of Variation	SS	df	MS
Regression	82.87	1	82.87
Error	514.63	4	128.66
Correction	229720.50	1	
Total_u	230318	6	

Data comes from:

$$SS_R = \hat{\beta}_1 S_{xy} = (0.272)(304.67) = 82.87$$

$$SS_{TOU} = \sum Y_i^2 = 230318$$

$$\text{Correction factor} = n\bar{Y}^2 = 229720.5$$

$$SS_{TO} = SS_{TOU} - n\bar{Y}^2 = 230318 - 229720.5 = 597.5$$

$$SS_E = SS_{TO} - SS_R = 597.5 - 82.87 = 514.63$$

Formal F-test:

$$F = \frac{MS_R}{MS_E} = \frac{82.87}{128.66} = 0.644$$

From table 12b:

$$F_{1,4}(95\%) = 7.709$$

$F < F_{1,4}(95\%)$ therefore we accept H_0 that β_1 is not significant.

(e) Find estimated robbery rate when the population density is 85.

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* \quad , \text{ here } x^* = 85$$

$$\hat{y}^* = 177.084 + (0.272)(85) = \boxed{200.2}$$

95% CI for $E(y^*)$

$$\hat{y}^* \pm t_{0.975,4} \sqrt{MS_E\left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}\right)}$$

$\alpha = 0.05$, $1 - \frac{\alpha}{2} = 0.975$

$n = 6$

$n - 2 = 4$

$$\hat{y}^* = 200.2 \pm (2.7)\sqrt{130.6\left(\frac{1}{6} + \frac{(85-68.33)^2}{1119.33}\right)}$$

$$= (180.32, \ 220.08)$$

95% CI for y^*

$$= \hat{y}^* \pm t_{0.975,4} \sqrt{MS_E\left(1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}\right)}$$

$$= 200.2 \pm (2.7)\sqrt{130.6\left(1 + \frac{1}{6} + \frac{(85-68.33)^2}{1119.33}\right)}$$

$$= (163.5, \ 236.9)$$

Appendix 2

R code for Figure 1

```
# STAT40790 Predictive Analytics I

# Project

# Brian Buckley

# 1. Plot the data
x <- c(59,45,75,72,89,70)                                  # population density
y <- c(209,180,195,186,200,204)                            # robbery rate
plot(x, y, xlim=c(min(x)-5, max(x)+5), ylim=c(min(y)-10, max(y)+10),
    main='Regression fit with 95% CI', xlab='Population density', ylab='Robbery rate')

# 2. Construct a linear predictor model
linearModel<-lm(y ~ x)
abline(linearModel, col="red")
newx<-seq(20,90)
prd<-predict(linearModel,newdata=data.frame(x=newx),interval = c("confidence"),
          level = 0.95,type="response")
lines(newx,prd[,2],col="red",lty=2)
lines(newx,prd[,3],col="red",lty=2)

# 3. Perform ANOVA test
mod1.anova<-aov(y ~ x)
summary(mod1.anova)

#              Df Sum Sq Mean Sq F value Pr(>F)
# x             1   82.9   82.93   0.635   0.47
# Residuals     4  522.4  130.60
```