AMS 380.01: Problem Set 4

Due on 03/09

Wei Zhu

Harris Temuri

Problem 1

Is there a simple linear relationship between income and happiness? The data-set 'income.data.csv' tabulates these two variables from a random sample of 498 people. Please write up the entire R code necessary to answer the following questions.

Solution

(a) Find the least squares regression line.

```
# 1.1 Finding Least squares regression line
lin_fit <- lm(happiness ~ income, data=income_data)
summary(lin_fit)
```

$$Happiness = (0.71383) * Income + 0.20427$$

(b) Plot the points and the regression line in the same figure.

```
# 1.2 Plotting Linear Fit
ggplot(data = income_data, aes(x=income, y=happiness)) + geom_point() + stat_smooth
    (method = lm)
```

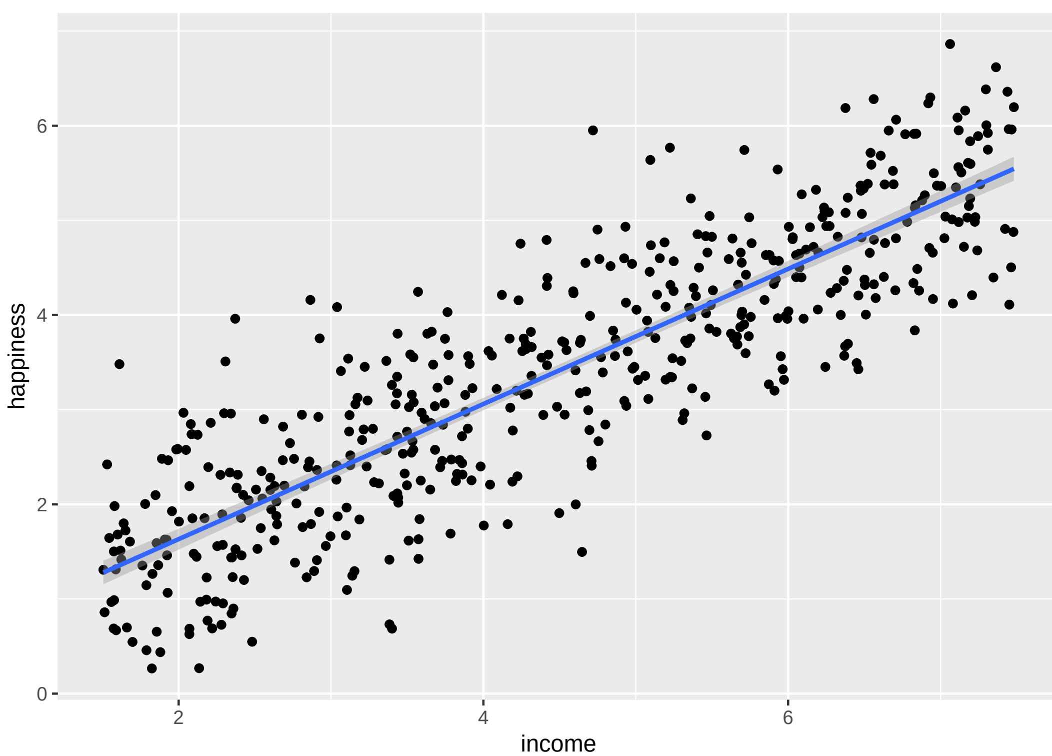

Figure 1: Linear fit Income vs Happiness

Problem 1 continued on next page. . .

(c) Test at $\alpha = 0.05$ whether there is a significant linear relationship between these two variables. What assumptions are necessary? Please test these assumptions.

```
# 1.3 Testing linear relationship along with assumptions
# Normality Test of residuals
shapiro.test(lin_fit$residuals)
# Significance test (look at p-value of variable)
summary(lin_fit)
```

The p-value of the normality test of the residuals is 0.4237 which is above 0.05 so do not reject the null hypothesis.

The p-value of the coefficient of income in the linear regression is ¡¡0.05 so the variable is significant.

(d) Compute the sample correlation coefficient between the two variables and test whether the corresponding population correlation is zero or not at $\alpha = 0.05$.

```
# 1.4 Correlation coefficient test
cor.test(income_data$happiness, income_data$income)
```

The p-value of the Pearson test is ¡¡0.05 so reject null hypothesis. There is significance between the variables happiness and income. The sample correlation coefficient between the two variables is 0.8656337.

(e) Report the coefficient of determination – does this statistic indicate a good linear model fit? (Note: Recall that for simple linear regression, the coefficient of determination is simply the squared sample Pearson correlation coefficient.)

```
# 1.5 Coefficient of determination
cor(income_data$happiness, income_data$income)^2
```

The value of the coefficient of determination is 0.7493218.

Problem 2

Scientists wish to analyze the effect of fertilizer type on crop yield. The dataset 'crop.data.csv' tabulates crop yields from 3 different fertilizers.

Solution

(a) Please draw side-by-side box plots to visually compare the yields from the three fertilizers.

```
# 1.1 Box Plots of fertilizer yield
ggplot(data = crop_data, aes(x = fertilizer, y = yield)) + geom_boxplot()
```

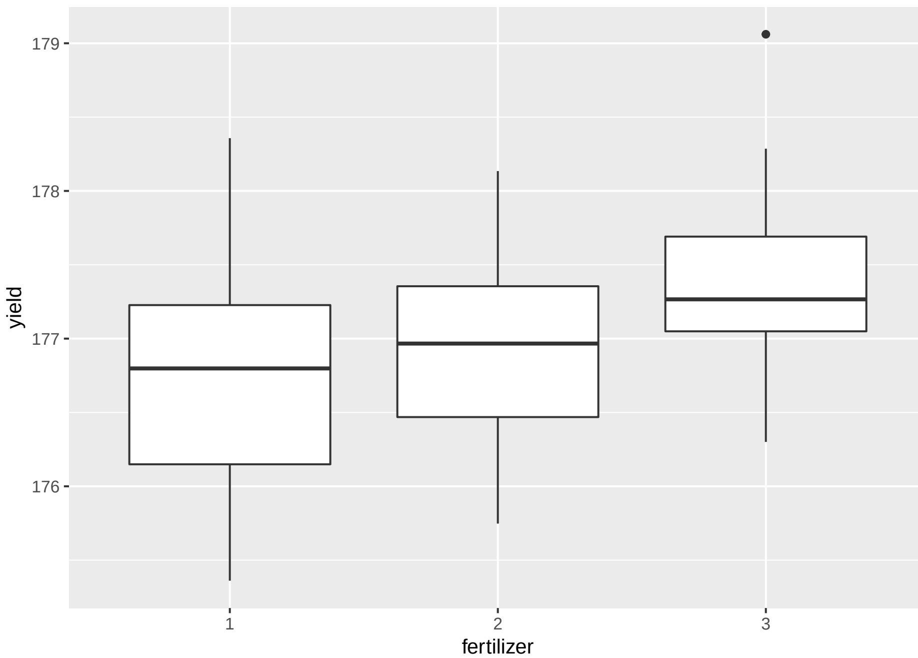

Figure 2: Box Plot of Three Fertilizers

(b) Test at $\alpha = 0.05$ whether the three fertilizers are equally effective. What assumptions are necessary? Please test these assumptions.

```
# 1.2 Test for fertilizer being equally effective
res_aov <- aov(yield ~ fertilizer, data = crop_data)
summary(res_aov)
# Testing Assumptions
# Normality Test of residuals
shapiro.test(res_aov$residuals)
# Homogeneity of Variance test
bartlett.test(yield ~ factor(fertilizer), data = crop_data)
```

(c) At the familywise error rate of $\alpha = 0.05$, please perform pairwise comparison of the three fertilizers using the Tukey HSD test.

```
# 1.4 Tukey Test
TukeyHSD(res_aov)
```

Group 1 and 3 is not significantly different

Group 1 and 2 is significantly different

Group 3 and 2 is significantly different

(d) Please compare fertilizers 2 and 3 using the usual pooled-variance t-test at the significance level $\alpha = 0.05$. What assumptions are necessary? Please test these assumptions.

```
# 1.5 Comparing fertilizers 2 and 3
t.test(crop_data$yield[crop_data$fertilizer == 2], crop_data$yield[crop_data$
    fertilizer == 3], var.equal = T)
# Testing Assumptions
# Normality of Groups
shapiro.test(crop_data$yield[crop_data$fertilizer == 2])
shapiro.test(crop_data$yield[crop_data$fertilizer == 3])
# Var test
var.test(crop_data$yield[crop_data$fertilizer == 2], crop_data$yield[crop_data$
    fertilizer == 3])
```

Group 2 and 3 are significantly different because p ¡ 0.05 and we reject the null hypothesis.