Homework Three

Chiayu Tu (Louis Tu)

2022-10-24

Question One

For the prostate data, fit a model with lpsa as the response and the other variables as predictors.

a

Suppose a new patient with the following values arrives:

lcavol lweight age lbph svi lcp 1.44692 3.62301 65.00000 0.30010 0.00000 -0.79851

gleason pgg45 7.00000 15.00000

Predict the lpsa for this patient along with an appropriate 95% CI.

```
data(prostate)
model_p <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate)
model_p_1 <- data.frame(lcavol  = 1.45000,
                        lweight = 3.62301,
                        age     = 65.00000,
                        lbph    = 0.30010,
                        svi     = 0.00000,
                        lcp     = -0.79851,
                        gleason = 7.00000,
                        pgg45   = 15.00000)
predict(model_p, model_p_1, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 2.390861 0.9664447 3.815277
```

```
print("We can observe that the lpsafor this patient along with an appropriate 95% is (0.9664447, 3.8152
```

```
## [1] "We can observe that the lpsafor this patient along with an appropriate 95% is (0.9664447, 3.815
```

b

Repeat the last question for a patient with the same values except that he is age 20. Explain why the CI is wider.

```
model_p_2 <- data.frame(lcavol  = 1.45000,
                        lweight = 3.62301,
                        age     = 20.00000,
                        lbph    = 0.30010,
                        svi     = 0.00000,
                        lcp     = -0.79851,
                        gleason = 7.00000,
                        pgg45   = 15.00000)
predict(model_p, model_p_2, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.274534 1.540492 5.008575
```

c

For the model of the previous question, remove all the predictors that are not significant at the 5% level. Now recompute the predictions of the previous question. Are the CIs wider or narrower? Which predictions would you prefer? Explain.

```
#summary(model_p)
model_p_3 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate)
model_p_4 <- data.frame(lcavol  = 1.45000,
                        lweight = 3.59801,
                        svi     = 0.00000)
predict(model_p_3, model_p_4, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 2.361519 0.9272333 3.795805
```

Question Two

Using the teengamb data, fit a model with gamble as the response and the other variables as predictors.

a

Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI.

```
#view(teengamb)
model_t <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
model_t_1 <- data.frame(sex    = 0,
                        status = mean(teengamb$status),
                        income = mean(teengamb$income),
                        verbal = mean(teengamb$verbal))
predict(model_t, model_t_1, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
```

b

Repeat the prediction for a male with maximal values (for this data) of status, income and verbal score. Which CI is wider and why is this result expected?

```
model_t_2 <- data.frame(sex    = 0,
                        status = max(teengamb$status),
                        income = max(teengamb$income),
                        verbal = max(teengamb$verbal))
predict(model_t, model_t_2, interval = "prediction")
```

```
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
```

c

Fit a model with sqrt(gamble) as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

```
model_t_3 <- lm(sqrt(gamble) ~ sex + status + income + verbal, data = teengamb)
predict(model_t_3, model_t_1, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 4.049523 3.180676 4.918371
```

d

Repeat the prediction for the model in (c) for a female with status=20, income=1, verbal = 10. Comment on the credibility of the result.

```
model_t_4 <- data.frame(sex    = 1,
                        status = 20,
                        income = 1,
                        verbal = 10)
predict(model_t_3, model_t_4, interval = "confidence")
```

```
##        fit       lwr      upr
## 1 -2.08648 -4.445937 0.272978
```

Question Three

Use the teengamb data with gamble as the response. We focus on the effect of sex on the response and so we include this predictor in all models. There are eight possible models that include all, some, or none of the other three predictors. Fit all these models and report on the coefficient and significance of sex in each case. Comment on the stability of the effect.

```
model_T <- lm(gamble ~ sex, data = teengamb)
summary(model_T)
```

```
##
## Call:
## lm(formula = gamble ~ sex, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.775 -18.325  -3.766   6.334 126.225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.775      5.498   5.415 2.28e-06 ***
## sex          -25.909      8.648  -2.996  0.00444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.09 on 45 degrees of freedom
## Multiple R-squared:  0.1663, Adjusted R-squared:  0.1478
## F-statistic: 8.977 on 1 and 45 DF,  p-value: 0.004437
```

```
print(model_T$coefficients['sex'])
```

```
##       sex
## -25.90921
```

```
model_T_1 <- lm(gamble ~ sex + status, data = teengamb)
summary(model_T_1)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.873 -15.755  -3.007  10.924 111.586
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.2233    15.1347   3.979 0.000255 ***
## sex         -35.7094     9.4899  -3.763 0.000493 ***
## status       -0.5855     0.2727  -2.147 0.037321 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.99 on 44 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2111
## F-statistic: 7.154 on 2 and 44 DF,  p-value: 0.002042
```

```
print(model_T_1$coefficients['sex'])
```

```
##       sex
## -35.70937
```

```r
model_T_2 <- lm(gamble ~ sex + income, data = teengamb)
summary(model_T_2)
```

```
##
## Call:
## lm(formula = gamble ~ sex + income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.757 -11.649   0.844   8.659 100.243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.041      6.394   0.632  0.53070
## sex          -21.634      6.809  -3.177  0.00272 **
## income         5.172      0.951   5.438 2.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.75 on 44 degrees of freedom
## Multiple R-squared:  0.5014, Adjusted R-squared:  0.4787
## F-statistic: 22.12 on 2 and 44 DF,  p-value: 2.243e-07
```

```r
print(model_T_2$coefficients['sex'])
```

```
##       sex
## -21.63439
```

```r
model_T_3 <- lm(gamble ~ sex + verbal, data = teengamb)
summary(model_T_3)
```

```
##
## Call:
## lm(formula = gamble ~ sex + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.034 -14.702  -3.700   5.595 113.450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   60.662     16.237   3.736 0.000535 ***
## sex          -27.722      8.417  -3.294 0.001957 **
## verbal        -4.528      2.249  -2.013 0.050209 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.15 on 44 degrees of freedom
## Multiple R-squared:  0.2366, Adjusted R-squared:  0.2019
## F-statistic:  6.82 on 2 and 44 DF,  p-value: 0.002631
```

```r
print(model_T_3$coefficients['sex'])
```

```
##       sex
## -27.72208
```

```r
model_T_4 <- lm(gamble ~ sex + status + income, data = teengamb)
summary(model_T_4)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.682 -12.169  -0.268   9.161  97.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.0315    15.8676   0.821  0.41603
## sex         -24.3393     8.1274  -2.995  0.00454 **
## status       -0.1496     0.2413  -0.620  0.53856
## income        4.9280     1.0352   4.760 2.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.92 on 43 degrees of freedom
## Multiple R-squared:  0.5058, Adjusted R-squared:  0.4713
## F-statistic: 14.67 on 3 and 43 DF,  p-value: 1.014e-06
```

```r
print(model_T_4$coefficients['sex'])
```

```
##       sex
## -24.33934
```

```r
model_T_5 <- lm(gamble ~ sex + status + verbal, data = teengamb)
summary(model_T_5)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.108 -15.290  -3.998   8.401 108.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.2216    17.5619   3.942 0.000293 ***
## sex         -33.7520     9.6839  -3.485 0.001144 **
## status       -0.4039     0.3267  -1.237 0.222971
## verbal       -2.7037     2.6784  -1.009 0.318410
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.99 on 43 degrees of freedom
## Multiple R-squared:  0.2629, Adjusted R-squared:  0.2114
## F-statistic: 5.111 on 3 and 43 DF,  p-value: 0.004106
```

```
print(model_T_5$coefficients['sex'])
```

```
##       sex
## -33.75202
```

```
model_T_6 <- lm(gamble ~ sex + income + verbal, data = teengamb)
summary(model_T_6)
```

```
##
## Call:
## lm(formula = gamble ~ sex + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.639 -11.765  -1.594   9.305  93.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1390    14.7686   1.634   0.1095
## sex         -22.9602     6.7706  -3.391   0.0015 **
## income        4.8981     0.9551   5.128 6.64e-06 ***
## verbal       -2.7468     1.8253  -1.505   0.1397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.43 on 43 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.4933
## F-statistic: 15.93 on 3 and 43 DF,  p-value: 4.148e-07
```

```
print(model_T_6$coefficients['sex'])
```

```
##       sex
## -22.96022
```

```
model_T_7 <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
summary(model_T_7)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```
print(model_T_7$coefficients['sex'])
```

```
##       sex
## -22.11833
```

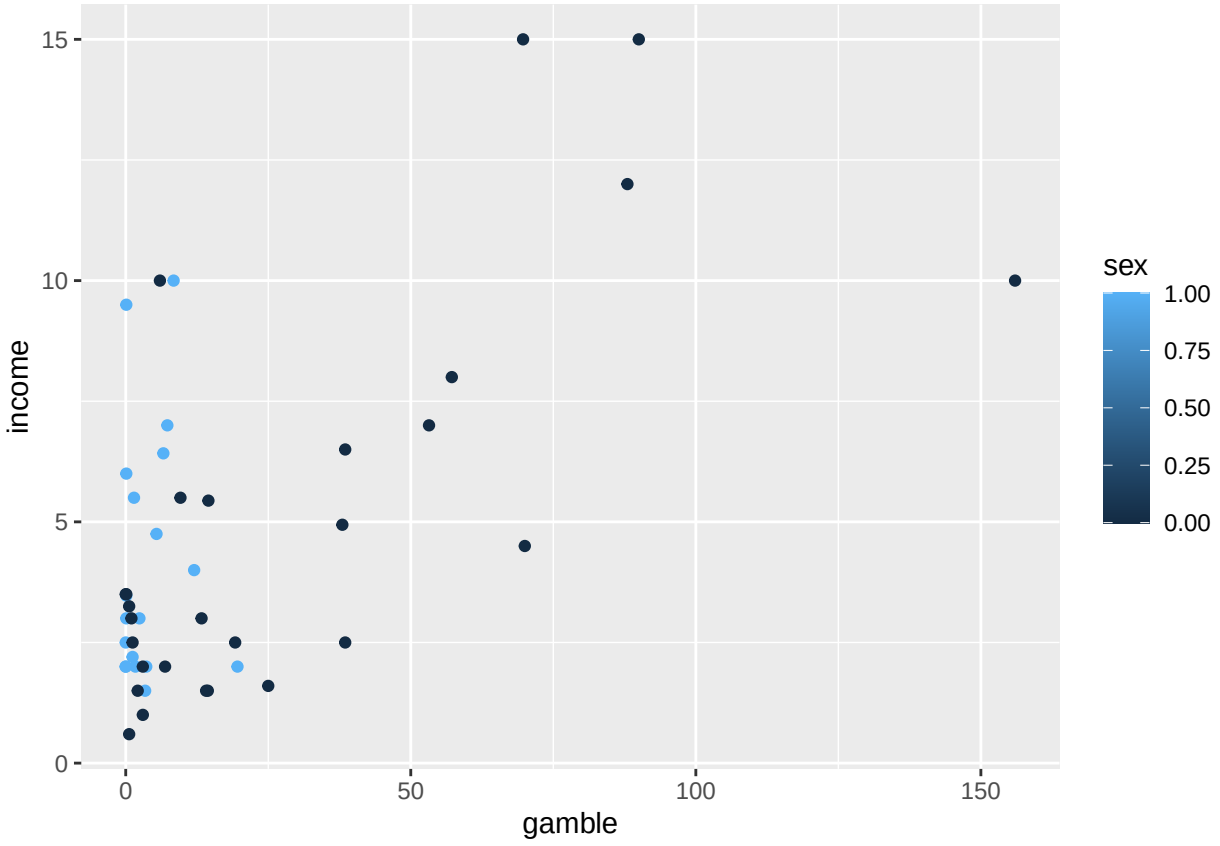
Question Four

Use the teengamb data for this question.

a

Make a plot of gamble on income using a different plotting symbol depending on the sex.

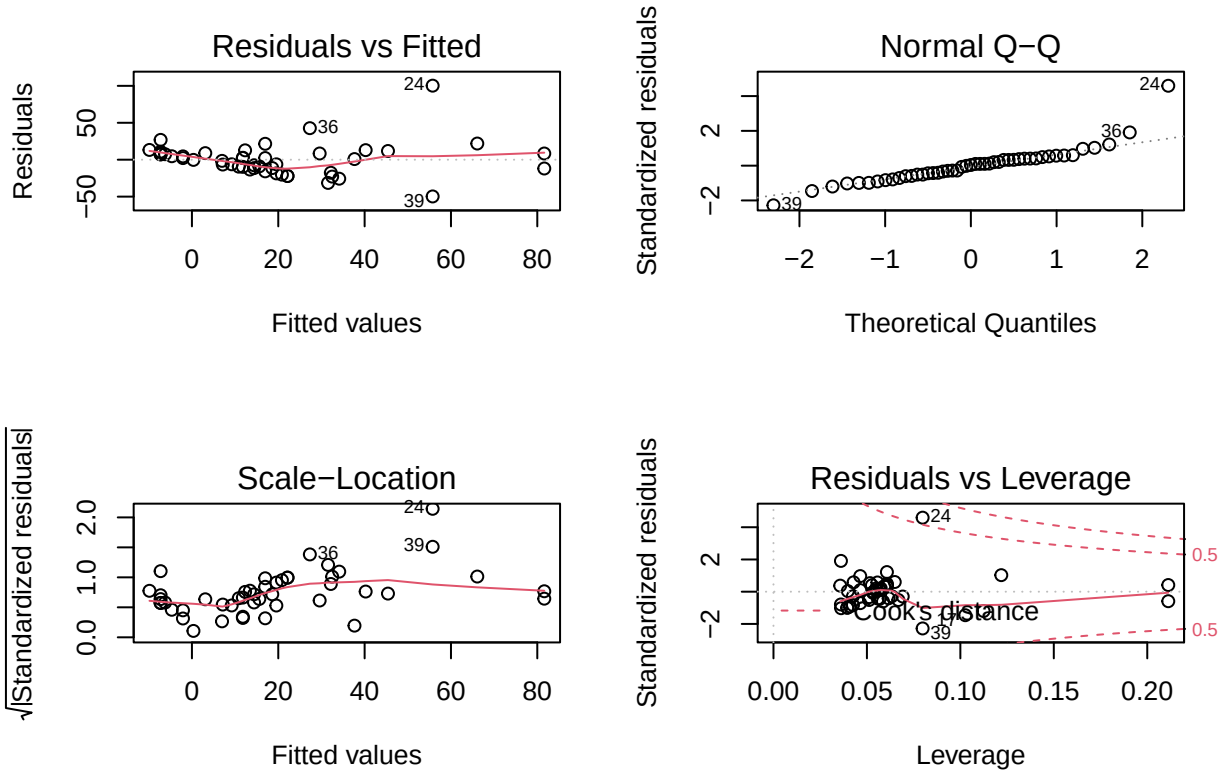
```
ggplot(teengamb, aes(gamble, income, color = sex)) +
  geom_point()
```



b

Fit a regression model with gamble as the response and income and sex as predictors. Display the regression fit on the plot.

```
model_i <- lm(gamble ~ income + sex, data = teengamb)
par(mfrow=c(2,2))
plot(model_i)
```



Question Five

Thirty-nine MBA students were asked about happiness and how this related to their income and social life. The data are found in happy.

a

Fit a regression model with happy as the response and the other four variables as predictors. Give an interpretation for the meaning of the love coefficient.

```
#view(happy)
model_h <- lm(happy ~ money + sex + love + work, data = happy)
summary(model_h)
```

```
##
## Call:
## lm(formula = happy ~ money + sex + love + work, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7186 -0.5779 -0.1172  0.6340  2.0651
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -0.072081   0.852543  -0.085   0.9331
## money        0.009578   0.005213   1.837   0.0749 .
## sex         -0.149008   0.418525  -0.356   0.7240
## love         1.919279   0.295451   6.496 1.97e-07 ***
## work         0.476079   0.199389   2.388   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 34 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.6761
## F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

b

The love predictor takes three possible values but mostly takes the value 2 or 3. Create a new predictor called clove which takes the value zero if love is 2 or less. Use this new predictor to replace love in the regression model and interpret the meaning of the corresponding coefficient. Do the results differ much from the previous model?

```
happy$clove <- ifelse(happy$love <= 2, 0, 1)
model_h_1 <- lm(happy ~ money + sex + love + clove + work, data = happy)
summary(model_h_1)
```

```
##
## Call:
## lm(formula = happy ~ money + sex + love + clove + work, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7014 -0.6051 -0.1178  0.6303  2.0127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.320774   1.526796   0.210   0.8349
## money        0.009261   0.005380   1.721   0.0946 .
## sex         -0.180499   0.436029  -0.414   0.6816
## love         1.716097   0.716586   2.395   0.0225 *
## clove        0.288093   0.923082   0.312   0.7569
## work         0.474189   0.202180   2.345   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 33 degrees of freedom
## Multiple R-squared:  0.7111, Adjusted R-squared:  0.6673
## F-statistic: 16.24 on 5 and 33 DF,  p-value: 4.425e-08
```

c

Fit a model with only clove as a predictor and interpret the coefficient. How do the results compare to the previous outcome.

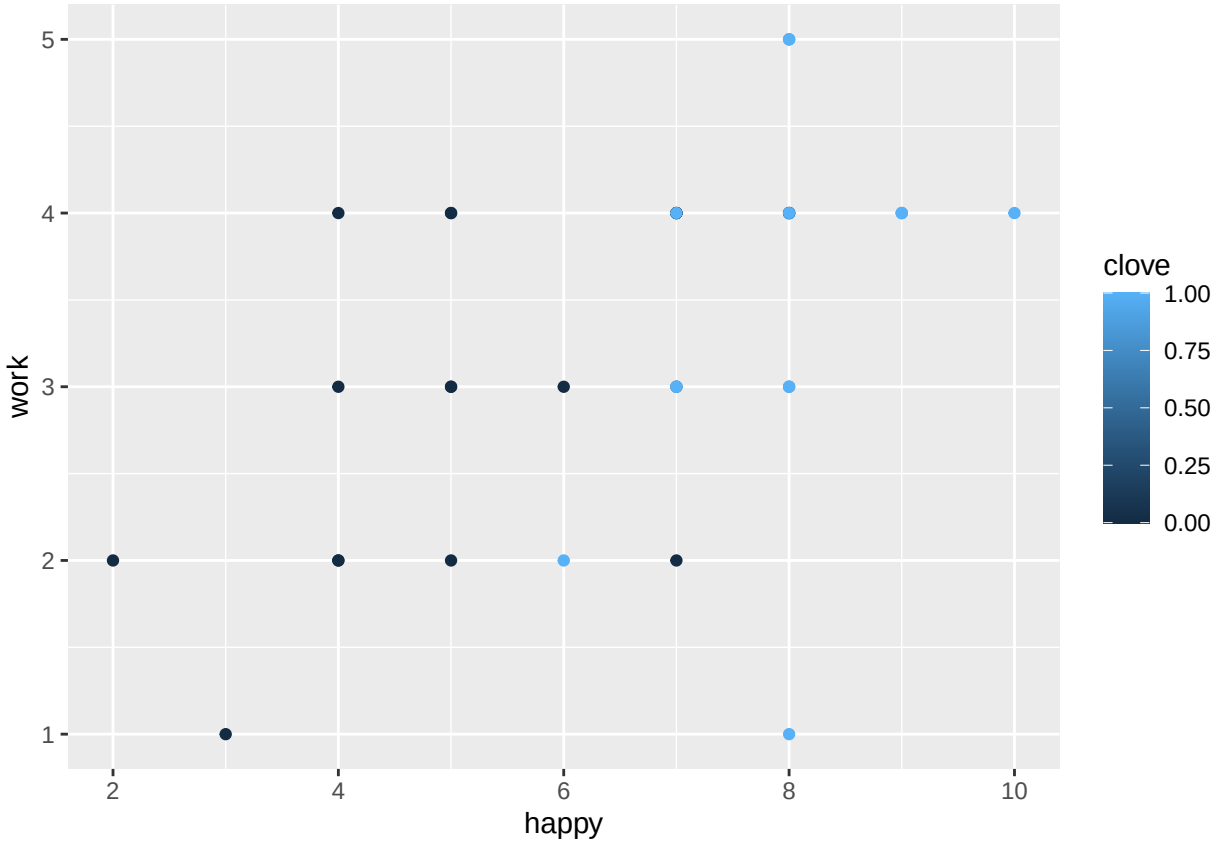
```
model_h_2 <- lm(happy ~ clove, data = happy)
summary(model_h_2)
```

```
##
## Call:
## lm(formula = happy ~ clove, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2778 -0.6389  0.0000  0.8611  2.7222
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.2778     0.2992  17.641  < 2e-16 ***
## clove         2.7222     0.4077   6.677 7.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.269 on 37 degrees of freedom
## Multiple R-squared:  0.5465, Adjusted R-squared:  0.5342
## F-statistic: 44.58 on 1 and 37 DF,  p-value: 7.667e-08
```

d

Make a plot of happy on work, distinguishing the value clove by using a plotting symbol. Use jittering to distinguish overplotted points.

```
ggplot(happy, aes(happy, work, color = clove)) +
  geom_point()
```


work
happy
clove
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
2
4
6
8
10