HW-Section4

Rob McCulloch

July 24, 2020

1 Homework for Section 4

1.1 Housing Plug-in Prediction

(a)

Get the midcity.csv data from the webpage.

Regress price on size, number of bathrooms, and number of bedrooms.

Using plug-in prediction, what is your 95% predictive interval for the price of a house which has size = 1.6 (thousands of square feet), nbed = 3, and nbath =2?

(b)

Using plug-in prediction, what is your 95% predictive interval for the price of a house which has size = 2.6, nbed = 3, and nbath =2?

(c)

Now regress price on size alone. What is your plug-in prediction interval for price given size=1.6?

(d)

What is your plug-in prediction interval for price given size=2.6?

(e)

How do you answers to (a) and (b) compare to your answers to (c) and (d) ?

Solution

(a)

$\hat{\sigma} = 20.36$.

-5.641 + 35.643*1.6 + 10.463 + 13.546*2 = 109.8598

interval is 109.9 +/- 40.72

(b)

Using plug-in prediction, what is your 95% predictive interval for the price of a house which has size = 2.6, nbed = 3, and nbath =2?

-5.641 + 35.643*2.6 + 10.463 + 13.546*2 = 145.5028

interval is 145.5 +/- 40.72.

Note that our model assumes we should use the same $\pm$ in each case!

(c)

a=-10.1, b = 70.2, $\hat{\sigma}$=22.48.

-10.1+70.2*1.6 = 102.22

interval is 102.22 +/- 45.

(d)

-10.1+70.2*2.6 = 172.42

172.42 +/- 45.

(e)

In simple linear regression on size, when size increases by 1, typically, the number of bedrooms and bathrooms will increase as well and the the coefficient 70.2 takes this into account. In the multiple regression, the size coefficient of 35.6 give the effect of a change is size *with the number of bathrooms and bedrooms fixed.*

1.2 Zagat Plug-in Prediction

The data for this question is in the file zagat.csv .

The data is from the Zagat restaurant guide.
There are 114 observations and each observation corresponds to a restaurant.

There are 4 variables:
price: the price of a typical meal
food: the zagat rating for the quality of food.
service: the zagat rating for the quality of service.
decor: the zagat rating for the quality of the decor.

We want to see how the price of a meal relates the quality characteristics
of the restaurant experience as measured by the variables food, service, and decor.

(a)

Plot price vs. each of the three x's. Does it seem like our y (price) is related to the x's (food, service, and decor) ?

(b)

Suppose a restaurant has food = 18, decor=16, and service=14.

Run the regression of price on food, decor, and service and give the 95% plug-in predictive interval for the price of a meal.

(c)

What is the interpretation of the coefficient estimate for the explanatory variable food in the multiple regression from part (b) ?

(d)

Suppose you were to regress price on the one variable food in a simple linear regression?
What would be the interpretation of the slope?

Plot food vs. service. Is there a relationship? Does it make sense?

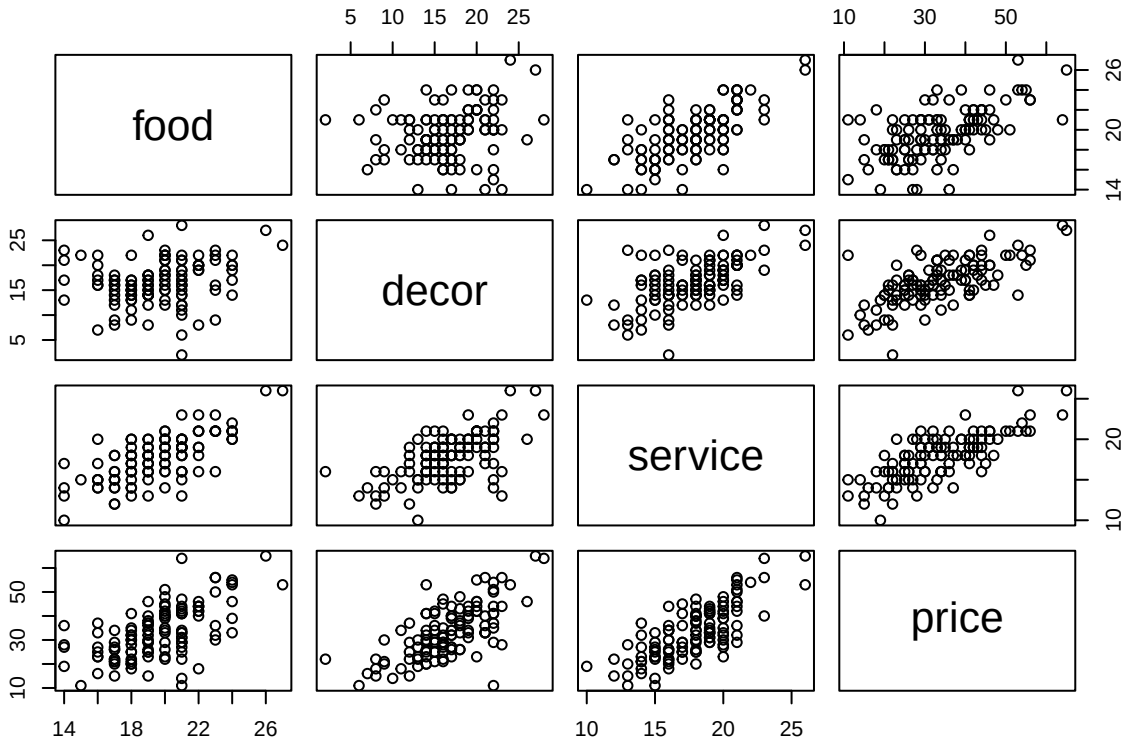
What is your prediction for how the estimated coefficient for the variable food in the regression of price on food will compare to the estimated coefficient for food in the regression of price on food, service, and decor?

Run the simple linear regression of price on food and see if you are right!!!

Why are the coeficients different in the two regressions?

Solution (a)

```
zagd = read.csv("http://www.rob-mcculloch.org/data/zagat.csv")
pairs(zagd)
```



As we might expect, it looks like price is strongly related to each characteristic.

(b)

```
lmz = lm(price~.,zagd)
summary(lmz)
```

```
##
## Call:
## lm(formula = price ~ ., data = zagd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0442  -4.0530   0.2109   4.6547  13.0864
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30.6640     4.7872  -6.405 3.82e-09 ***
## food          1.3795     0.3533   3.904 0.000163 ***
## decor         1.1043     0.1761   6.272 7.18e-09 ***
## service       1.0480     0.3811   2.750 0.006969 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.298 on 110 degrees of freedom
## Multiple R-squared:  0.6875, Adjusted R-squared:  0.6789
```

```
## F-statistic: 80.66 on 3 and 110 DF,  p-value: < 2.2e-16
```

-30.664 + 1.38*18 + 1.1*16 + 1.05*14 = 26.476
2(6.3) =12.6

so we get 26.476 +/- 12.6.

(c)

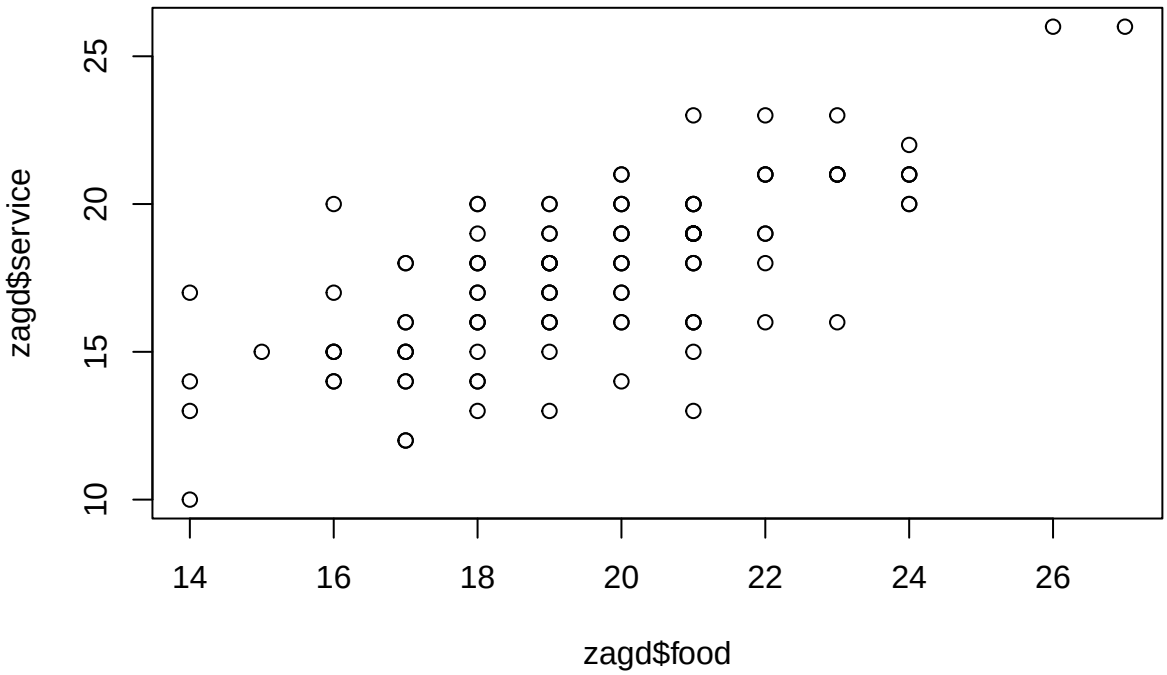
If you hold service and decor fixed and increase food by 1 then price goes up (on average) by 1.38.

(d)

If food goes up by one price goes up by the slope (on average).

Here is the plot of food vs. service, definitely related!!

```
plot(zagd$food,zagd$service)
```



Better restaurants will tend to have better food, decor, and service, and higher prices, so all 4 variables will "move together' '.

1.3 Housing and Zagat, Intervals and Tests

(a)

In the regression of house price on size, nbath, and nbed, what is the 95% confidence interval for the true slope for size?

Is it big or small?

(b)

In the zagat regression, give the 95% confidence interval for the true slope for the variable food.

Is it big or small?

(c)

In the zagat regression, test the null hypothesis that the true slope for the variable food is 0 (at level .05).

(d)

In the zagat regression test the null hypothesis that the true coefficient for food is equal to 1.

(e)

In the zagat regression, test the null hypothesis that the slope for service = 1.

In the zagat regression, test the null hypothesis that the slope for decor = 1.

What would be a simple way to summarize the relationship bewtween price and food, service, and decor that might be approximately correct?

Solution

(a)

35.643 +/- 21.3

It is big.

(b)

1.38 +/- .7

Again, pretty big.

(c)

t from output is 3.9, p-value is .000163, clear reject.

(d)

t=(1.3795-1)/.3533 = 1.074158 fail to reject.

(e)

In both cases you fail to reject.

Since all the coefficients are not too differeent from 1, you might just relate price to the sum of food , decor, and service. You could make a new variable with is the sum and do a simple linear regression of price on the sum.

I tried it (let fds denote the sum of food, decor, and service) and got price = -28.5 + 1.148 fds +/- 2(6.26)

($\hat{\sigma} = 6.26$)

which is just as good a $\pm$ as the multiple regression!!

1.4 Fits, Resids, and R^2 in Zagat

Get the fitted values and residuals for the zagat regression of price on food, decor, and service.

In excel, there is a box you can check to get the fits and resids.

In R:

```
zd = read.csv("http://www.rob-mcculloch.org/data/zagat.csv",header=T)
lmz = lm(price~.,zd)
print(names(lmz))
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
e = lmz$residuals #the residuals
yhat = lmz$fitted.values #the fitted values
```

(a)

Verify the first fitted value and resid "by hand'.\ That is, compute $\hat{y} = -30.6640 + 1.3795 * 18 + 1.1043 * 22 + 1.0480 * 17$ and make sure it is the first fitted value.

Compute $y_1 - \hat{y}$ =41-$\hat{y}$ and make sure it is the first residual.

(b)

Plot the residuals vs. food.

What is the correlation between the residuals and food?

(c)

Plot the residuals vs. the fitted values.

What is the correlation between the residuals and the fitted values?

(d)

Plot y=price vs yhat=the fitted values.

What is the correlation between y and yhat?

How does the square of this correlation compare to R^2?

Solution

(a)

```
yhat1 = -30.6640 + 1.3795*18 + 1.1043*22+ 1.0480*17
e1 = zd$price[1]-yhat1
print(yhat1)
```

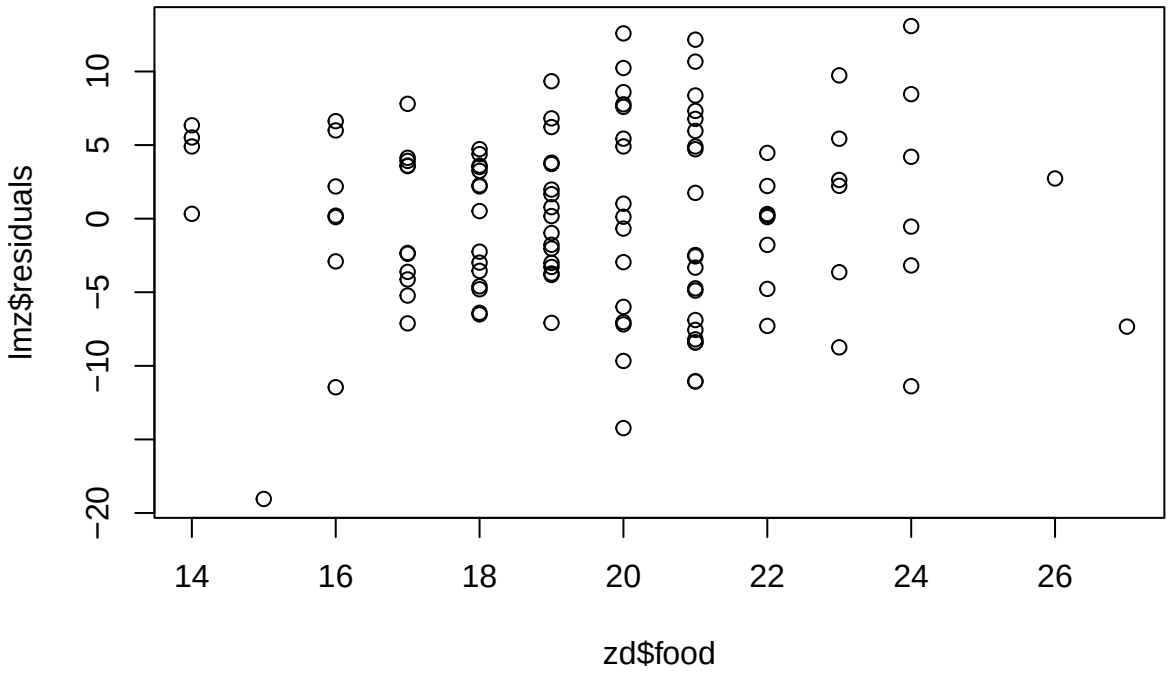
```
## [1] 36.2776
```

```
print(e1)
```

```
## [1] 4.7224
```

(b)

```
plot(zd$food,lmz$residuals)
```

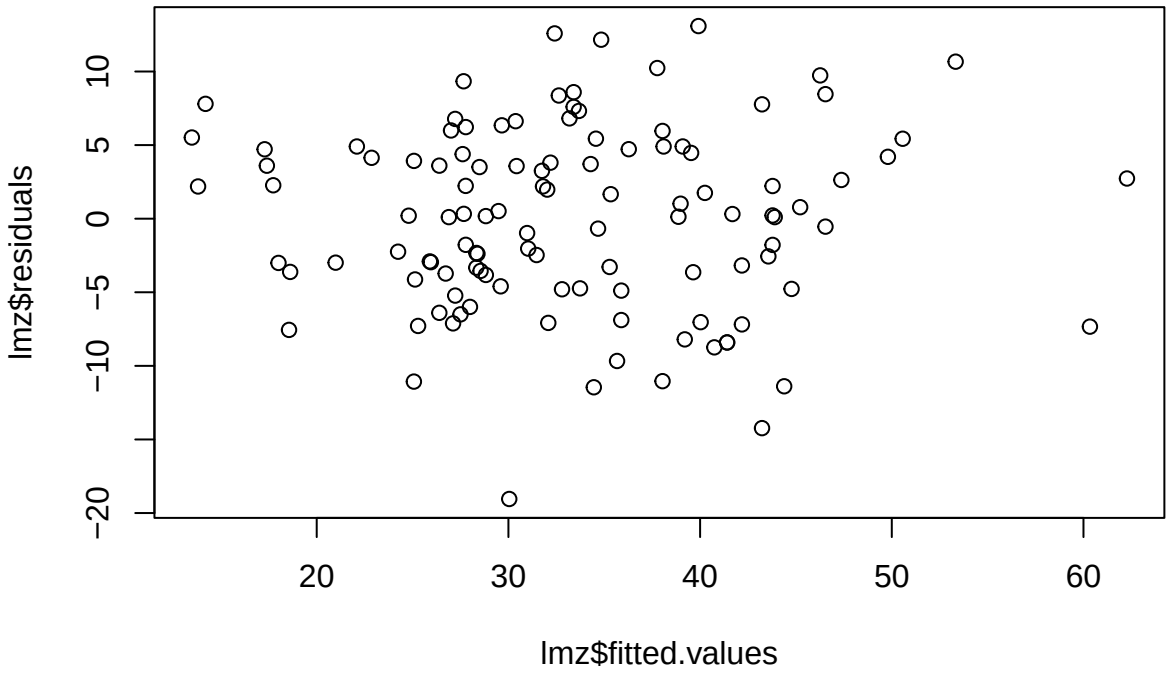



```
print(cor(zd$food,lmz$residuals))
```

```
## [1] -4.817066e-17
```

(c)

```
plot(lmz$fitted.values,lmz$residuals)
```

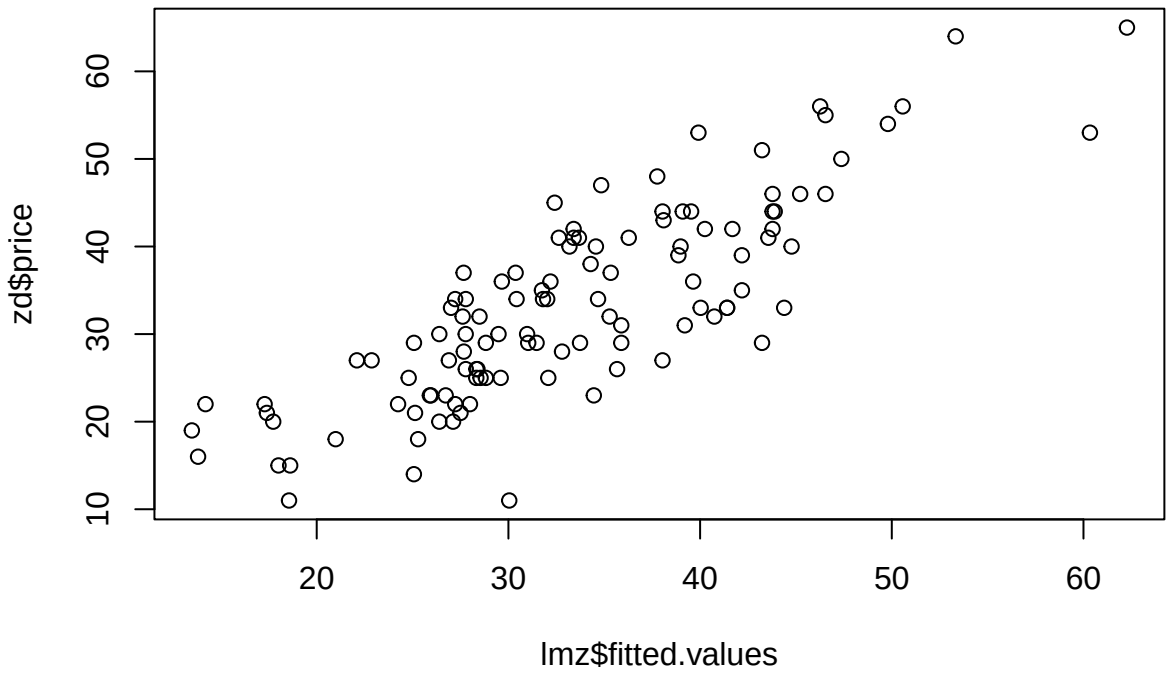



```
cor(lmz$fitted.values,lmz$residuals)
```

```
## [1] 8.961454e-17
```

(d)

```
plot(lmz$fitted.values,zd$price)
```



```
Rmultiple = cor(lmz$fitted.values,zd$price)
Rsquared = Rmultiple^2
cat("multiple R and R-squared:",Rmultiple,Rsquared,"\n")
```

```
## multiple R and R-squared: 0.8291379 0.6874697
```

check:

```
summary(lmz)
```

```
##
## Call:
## lm(formula = price ~ ., data = zd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0442  -4.0530   0.2109   4.6547  13.0864
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30.6640     4.7872  -6.405 3.82e-09 ***
## food          1.3795     0.3533   3.904 0.000163 ***
## decor         1.1043     0.1761   6.272 7.18e-09 ***
## service       1.0480     0.3811   2.750 0.006969 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.298 on 110 degrees of freedom
## Multiple R-squared:  0.6875, Adjusted R-squared:  0.6789
## F-statistic: 80.66 on 3 and 110 DF,  p-value: < 2.2e-16
```

1.5 Gender and the Beer Data

Get the beer data:

```
bd = read.csv("http://www.rob-mcculloch.org/data/nbeer.csv")
head(bd)
```

```
##   nbeer weight height age gender
## 1    12    192     72  26      0
## 2    12    160     66  27      0
## 3     5    155     65  25      0
## 4     5    120     66  28      0
## 5     7    150     67  28      0
## 6    13    175     71  31      0
```

```
dim(bd)
```

```
## [1] 50  5
```

```
table(bd$gender)
```

```
##
##  0  1
## 41  9
```

Each observation corresponds to an MBA Students.
`nbeer` is the number of beers they can drink without getting drunk.
`weight`, `height`, and `age` are self-explanatory.

In the beer data (nbeer.csv),
how does nbeers relate to gender?

Note that the variable `gender` in the data is already coded as a binary dummy, 0=male, 1=female.

(a)

Plot nbeer vs gender.

(b)

Regress nbeer on the gender dummy:

$$nbeer = \beta_0 + \beta_1 gender + \epsilon$$

Interpret the estimate, confidence interval, and p-value corresponding to β_1.

(c)

Regress nbeer on weight and gender:

$$nbeer = \beta_0 + \beta_1 gender + \beta_2 weight + \epsilon.$$

Interpret the estimate, confidence interval, and p-value corresponding to β_1.

(d)

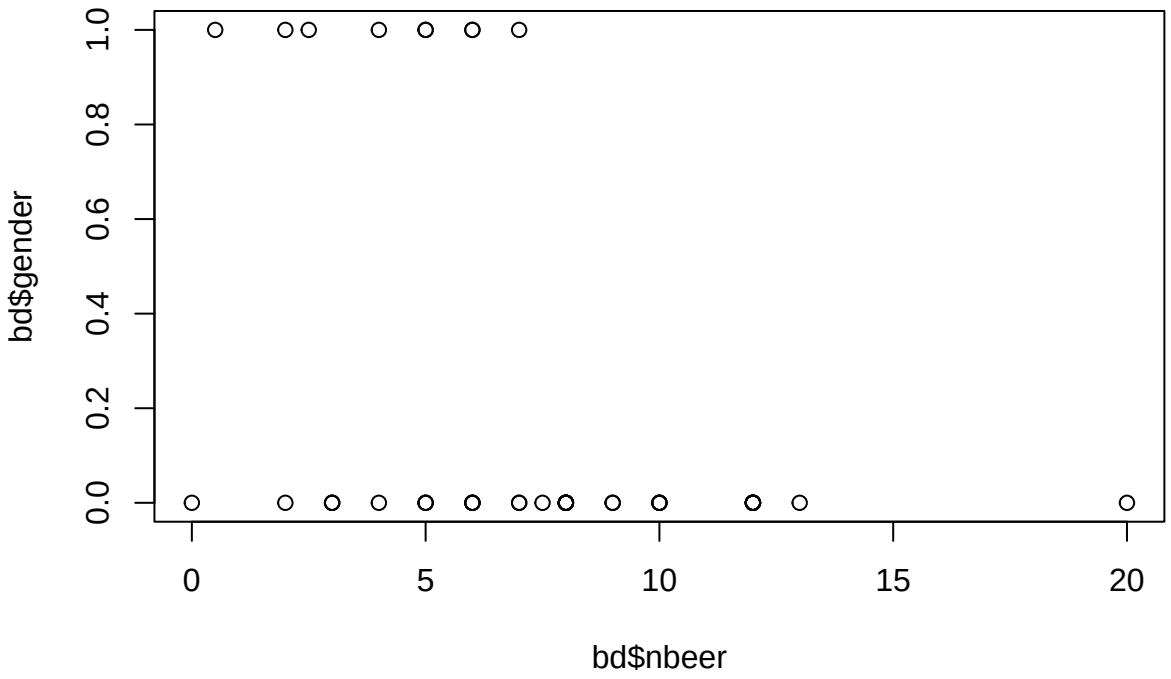
Is gender related to number of beers?

Discuss in light of your results in (a), (b), and (c).

Solution

(a)

```
plot(bd$nbeer,bd$gender)
```



Clearly, the men claim to be able to drink more.

With this few observations, this is not a bad plot, but with more observations this is not a good way to plot a binary variable vs a numeric variable.

(b)

```
lmg = lm(nbeer~gender,bd)
summary(lmg)
```

```
##
## Call:
## lm(formula = nbeer ~ gender, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1585 -2.1585 -0.1585  1.8415 11.8415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.1585     0.5463  14.935  < 2e-16 ***
## gender       -3.9363     1.2876  -3.057  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.498 on 48 degrees of freedom
## Multiple R-squared:  0.163,  Adjusted R-squared:  0.1455
## F-statistic: 9.346 on 1 and 48 DF,  p-value: 0.003646
```

On average, the women can drink -3.9 more (4 fewer) beers than the men.

The confidence interval for the "gender effect'' is -3.9 $\pm$ 2*1.3 = -3.9 $\pm$ 2.6.
While there is a lot of uncertainty even if the gender effect was as small as -1.3, that is still more than a beer.

The t stat is -3.057 and the p-value is .00365, so we have a clear reject of $\beta_1 = 0$.

(c)

```
lmgw = lm(nbeer~gender+weight,bd)
summary(lmgw)
```

```
##
## Call:
## lm(formula = nbeer ~ gender + weight, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7410 -2.0751 -0.0431  1.6605  5.4103
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.83031    3.01315  -2.599   0.0125 *
## gender       0.52841    1.32046   0.400   0.6908
## weight       0.09748    0.01818   5.362 2.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.784 on 47 degrees of freedom
## Multiple R-squared:  0.4807, Adjusted R-squared:  0.4586
## F-statistic: 21.75 on 2 and 47 DF,  p-value: 2.052e-07
```

The estimated coefficent for gender is now positive but the confidence interval is so big (.5 $\pm$ 2.6) that we don't take it seriously.

The t-tstat (.4) and p-value (.69) indicate we should fail to reject the hypothesis that the coefficient for gender is 0.

Given the confidence interval, "fail to reject'' is about right, there is a lot of uncertainty and no evidence for a gender effect.

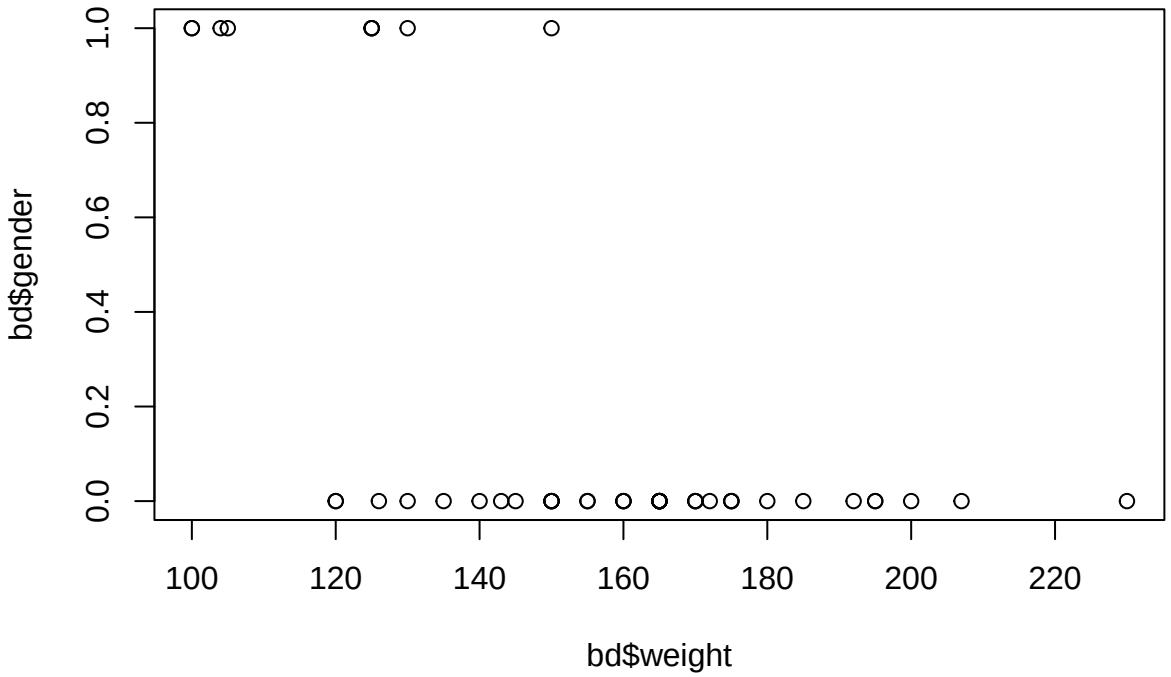
(d)

It does seem to be the case the men claim to drink more.

But, holding weight constant (after we *control* for weight), the is no clear evidence for a gender effect. If we compare a man and women *with the same weight* we do not have evidence that the man can drink more.

Below is a plot of gender vs weight.
The men are bigger, so they can drink more.

```
plot(bd$weight,bd$gender)
```



1.6 Mid City Dummies

In the notes we used dummies for neighborhoods 1 and 2 to capture the neighborhood.

Let's just try to repeat the analysis but use dummies for neighborhoods 2 and 3.

First of all, create the three dummies, one for each neighborhood.

In Excel you can use the if function.
For example, if Nbhd is in b2:b129 then you can copy the formula `{=if(b2=1,1,0)}` down a column to create the dummy for neighborhood 1.

In R we can create the dummies in lots of ways. Here is one approach.

```
### read in the housing data:
hd = read.csv("http://www.rob-mcculloch.org/data/midcity.csv")
## collect the variables the way you want them in a new data.frame
hdf = data.frame(price=hd$Price/1000, size=hd$SqFt/1000)

##make dummies
dn1 = ifelse(hd$Nbhd==1,1,0) #dum for Neighborhood 1
dn2 = ifelse(hd$Nbhd==2,1,0)
dn3 = ifelse(hd$Nbhd==3,1,0)

## add the dummies to the data frame hdf
hdf$dn1=dn1
hdf$dn2=dn2
hdf$dn3=dn3

## have a look
head(hdf)
```

```
##   price size dn1 dn2 dn3
## 1 114.3 1.79   0   1   0
## 2 114.2 2.03   0   1   0
## 3 114.8 1.74   0   1   0
## 4  94.7 1.98   0   1   0
## 5 119.8 2.13   0   1   0
## 6 114.6 1.78   1   0   0
```

(a)

Regress Price on SqFt and dn1 and dn2 and make sure you get the same thing as in the notes.

(b)

Using the regression in (a) plot SqFt vs. the fitted values.

(c)

Now regress Price on SqFt and dn2 and dn3.

Make sure you understand how the regression output in (a) relates to that of (b).

Solution

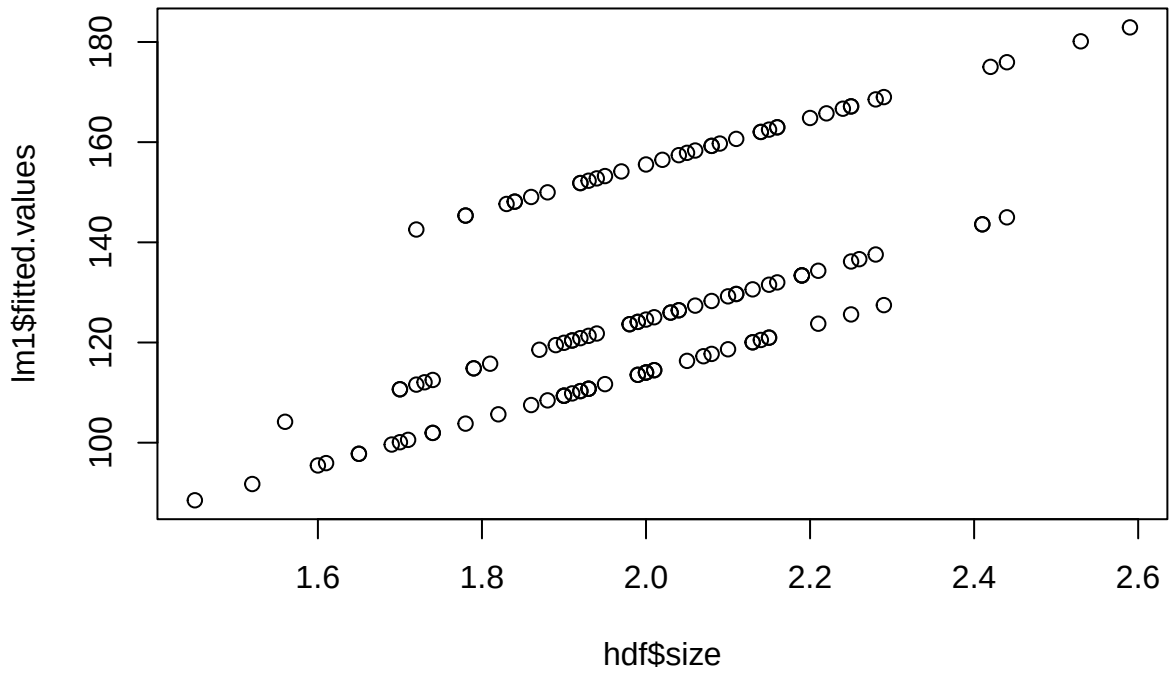
(a)

```
lm1 = lm(price~size+dn1+dn2,hdf)
summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ size + dn1 + dn2, data = hdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.107 -10.924  -0.305   9.643  38.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.776     14.248   4.406 2.25e-05 ***
## size          46.386      6.746   6.876 2.67e-10 ***
## dn1          -41.535      3.534 -11.754  < 2e-16 ***
## dn2          -30.967      3.369  -9.192 1.13e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

(b)

```
plot(hdf$size,lm1$fitted.values)
```



(c)

```
lm2 = lm(price~size+dn2+dn3,hdf)
summary(lm2)
```

```
##
## Call:
## lm(formula = price ~ size + dn2 + dn3, data = hdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.107 -10.924  -0.305   9.643  38.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.241     13.134   1.617  0.10835
## size          46.386      6.746   6.876 2.67e-10 ***
## dn2           10.569      3.301   3.202  0.00174 **
## dn3           41.535      3.534  11.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

Notice the $\hat{\sigma}$ is exactly the same (15.26) in both regressions.
In the first regression, our estimates tell us the neighborhood 1 is -41.535 different from 3.
In the second regression, our estimates tell us that neighborhood 3 is 41.535 different from 1.

Note that we can get the regression just by coding neighborhood up as a *factor* in R which is what R calls a categorical variable.

```
hdf$nf = as.factor(hd$Nbhd)
lm3 = lm(price ~ size + nf,hdf)
summary(lm3)
```

```
##
## Call:
## lm(formula = price ~ size + nf, data = hdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.107 -10.924  -0.305   9.643  38.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.241     13.134   1.617  0.10835
## size          46.386      6.746   6.876 2.67e-10 ***
## nf2           10.569      3.301   3.202  0.00174 **
## nf3           41.535      3.534  11.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 124 degrees of freedom
```

```
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

We get exactly the same results that we got by explicitly putting in the dummies for neighboroods 2 and 3.