Multiple Regression
Homework Problems
Homework Solutions

Rob McCulloch

2.1. Problem: 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?

(*see the next two pages for notes on running multiple regression in Excel and R*).

Remember, if you do this in excel with the standard add-in (/tools/Data Analysis/Regression) you will have to have all the columns for the three "x" variables (size, nbed, and nbath) beside each other. I found it easiest to just create 4 columns (size, nbed, nbath, price) all beside each other. So, the first three columns are my x's and the fourth column is my y. You may want to divide price and size by 1000 to make the numbers easier to look at.

Then, when you get into excel and it asks you for the x's you give it the range of first value of the first x and the last value of the last x. Another useful thing to remember for running the regression in excel is that if you have variable labels in the top row (you should) then you can use these labels in your regression output by (i) including the labels in the variable ranges (eg. y is d1:d129 instead of d2:d129) and then clicking the "Labels" box right below "Input X range".

In R try:

```
#if midcity.csv is in your working director, check getwd()
md = read.csv("midcity.csv",header=TRUE)
# or just
md = read.csv("http://www.rob-mcculloch.org/data/midcity.csv",header=TRUE)

# change the units of Price and Size
md$Price = md$Price/1000
md$Size = md$SqFt/1000

#get the regression
lmsbb = lm(Price~Size+Bedrooms+Bathrooms,md)
#print the summary of the regression
print(summary(lmsbb))
```

(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 + 10.46*3 + 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 + 10.46*3 + 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, s_e=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 bedroosm fixed*.

2.2. Problem: 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?

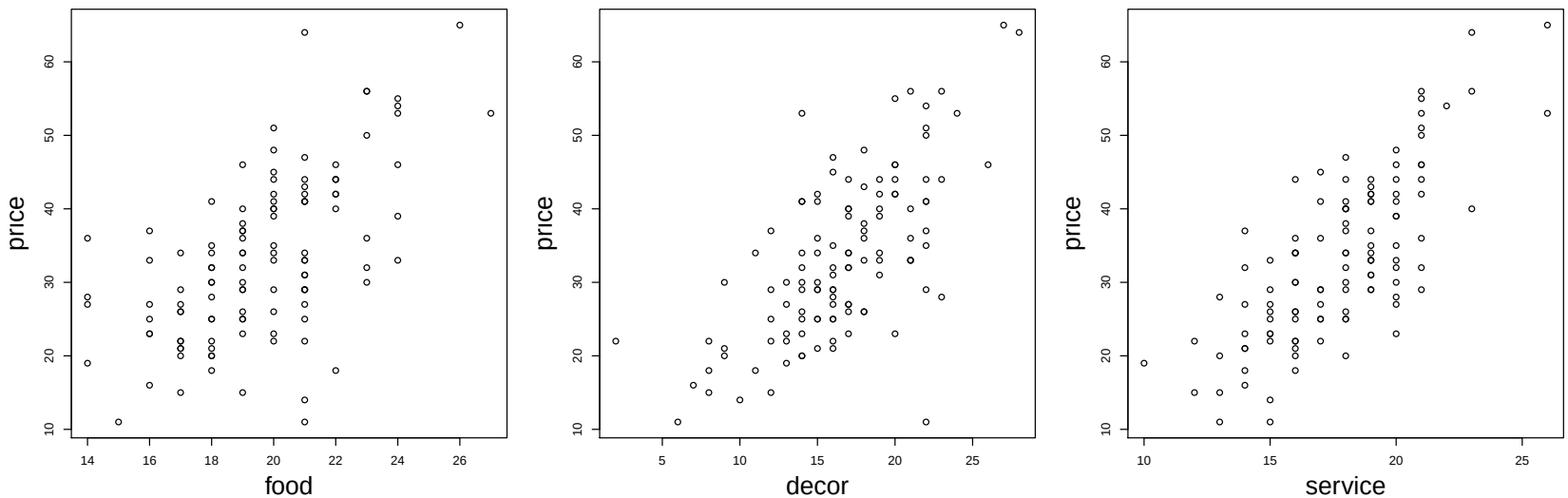
What is your prediction for how the estimated coefficient for the variable food in the regression of price on food will compare to the estimated coefficient for food in the regression of price on food, service, and decor?

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)



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

(b)

-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.

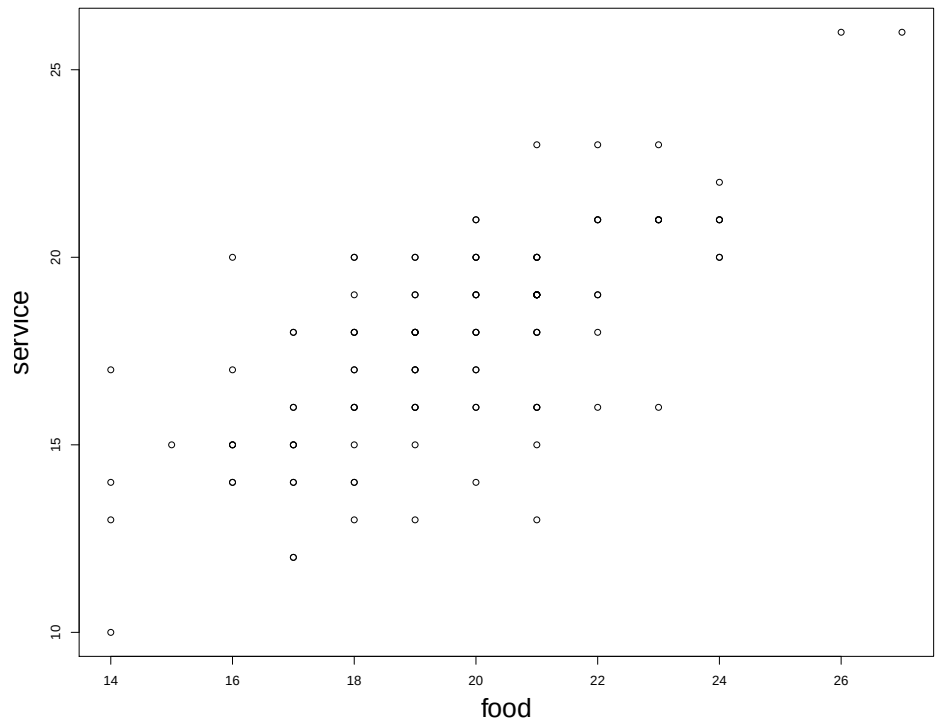
(c)

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!!



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

I would expect that the coefficient for food in the regression of price on food alone would be bigger than the coefficient for food in the multiple regression of price on food, decor, and service.

I ran the regression of price on food and the coefficient for food is 2.94!!!

If all you know is food went up, service and decor probably went up as well to you expect a bigger price increase than if food went up with service and decor held fixed.

3.1. Problem: 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)

($s_e = 6.26$)
which is just as good a $\pm$ as the multiple regression!!

4.1. Problem: 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("zagat.csv",header=T)
lmz = lm(price~.,zd)
print(names(lmz))
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.

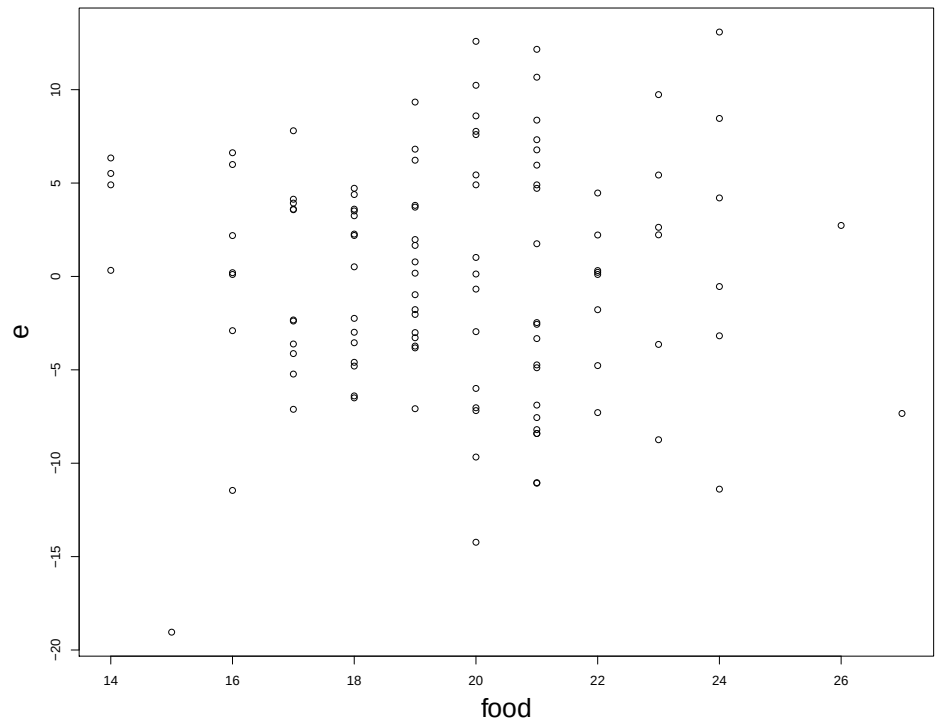
What is the correlation between the residuals and the fitted values?

(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

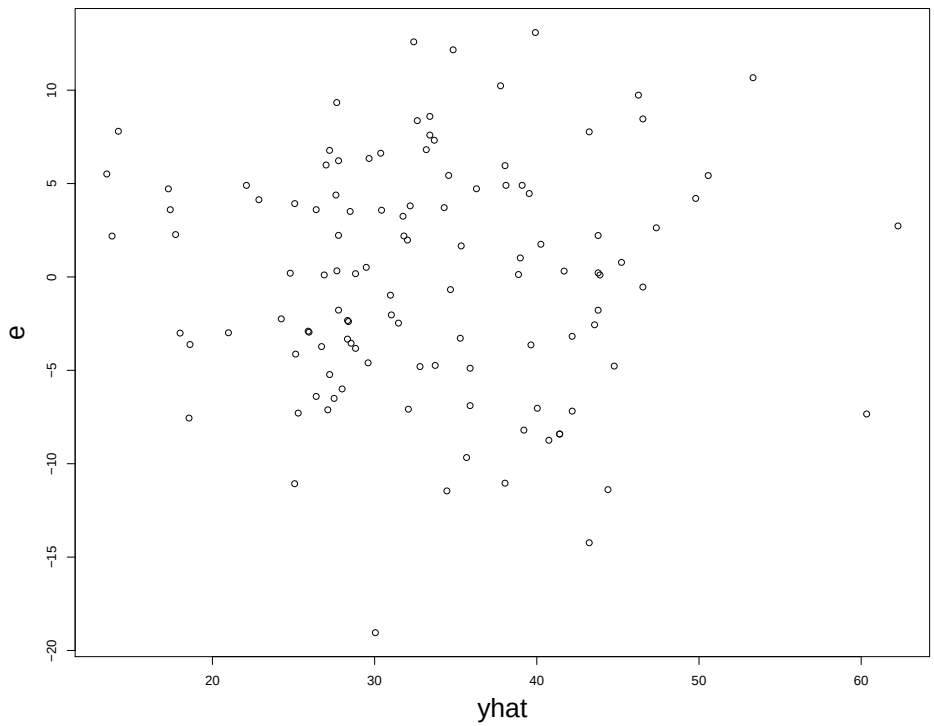
(b)

Residuals vs. food, correlation is 0.



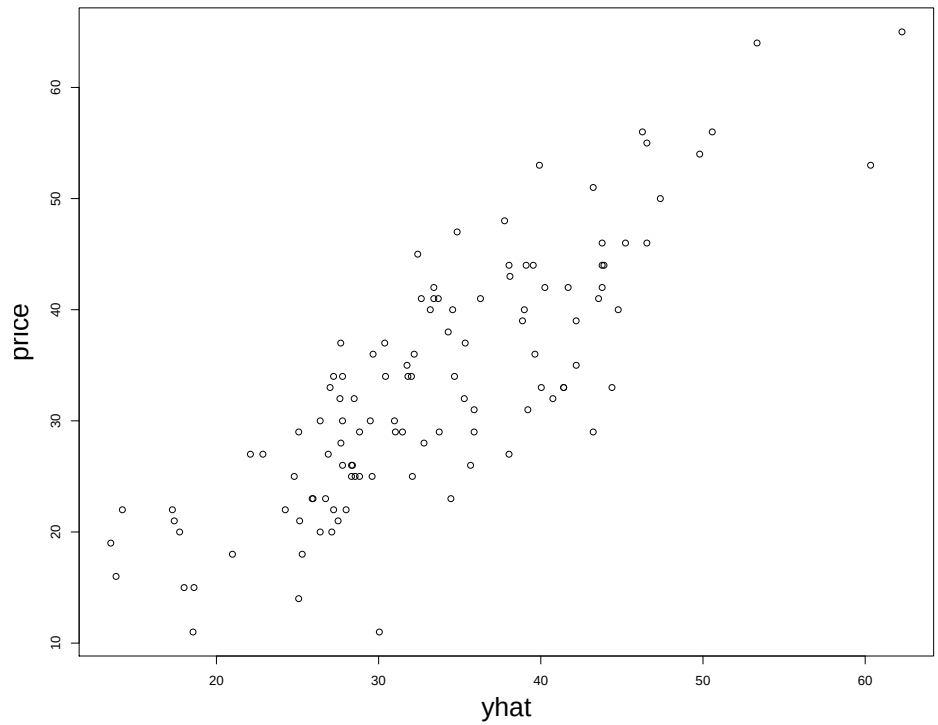
(c)

Residuals vs. fits, correlation is 0.



(c)

y=price vs. fits, correlation is .83.
$.83^2 = 0.6889 \sim$ R-squared.



5.1. Problem: Gender and the Beer Data

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:

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

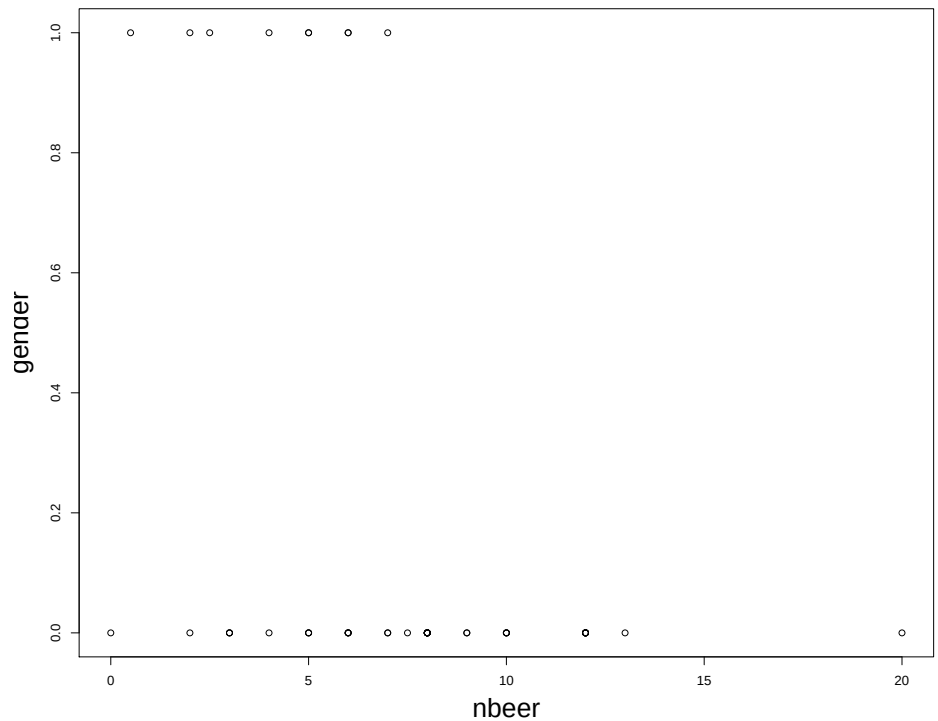
Interpret the estimate, confidence interval, and p-value corresponding to β_1.

(d)
Is gender related to number of beers?
Discuss in light of your results in (a), (b), and (c).

Solution

Solution.
(a)

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)

```
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 **
```

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)

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.

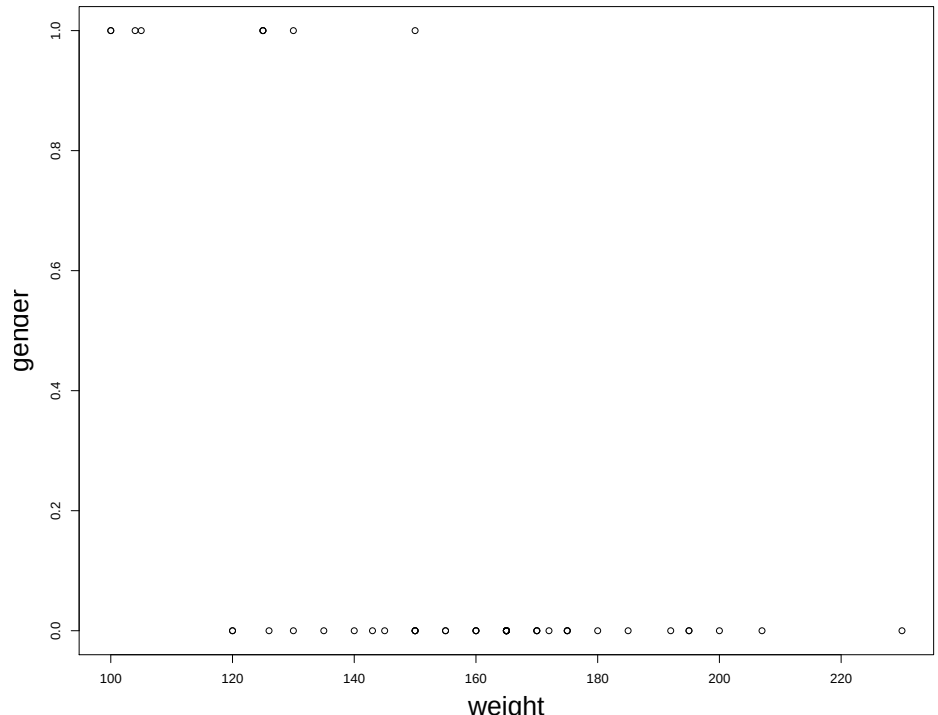
Given the confidence interval, "fail to reject" is about right, there is a lot of uncertainty and no evidence for a gender effect.

```
            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 ***
```

(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.



5.2. Problem: 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:

```
##read in data
md = read.csv("midcity.csv",header=TRUE)
##make dummies
dn1 = ifelse(md$Nbhd==1,1,0) #dum for Neighborhood 1
dn2 = ifelse(md$Nbhd==2,1,0)
dn3 = ifelse(md$Nbhd==3,1,0)
## add dummies to data frame
md$dn1=dn1
md$dn2=dn2
md$dn3=dn3
##change units of price and size
md$Price = md$Price/1000
md$size = md$SqFt/1000
##run regression and print summary
lmsn = lm(Price~size+dn1+dn2,md)
print(summary(lmsn))
##also try
md = read.csv("midcity.csv",header=TRUE)
md$Price = md$Price/1000
md$size = md$SqFt/1000
md$Nbhd = as.factor(md$Nbhd)
lmsn = lm(Price~size+Nbhd,md) #lm automatically makes dummies for a factor.
print(summary(lmsn))
```

(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),(b) in the notes.

```
            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 ***
```

(c)

```
            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 ***
```

In the regression with dummies for 1 and 2 we have:
N1: 63-42= 21.
N2: 63-31 = 32.
N3: 63.

In the regression with dummies for 2 and 3 we have:
N1: 21
N2: 21+11=32
N3: 21+42=63.