Assignment 6 - Solutions

Question 1 - Roadways

Regression methods were used to analyze the data from a study investigating the relationship between roadway surface temperature (x) and pavement deflection (y). Summary quantities were $n = 15$, $\sum y_i = 11.75$, $\sum y_i^2 = 7.86$, $\sum x_i = 1348$, $\sum x_i^2 = 123,324.6$ and $\sum x_i y_i = 983.67$.

In [1]:
```
n=15;
y1i=11.75;
y1i2=7.86;
x1i=1348;
x1i2=123324.6;
xy1i=983.67;
n,y1i,y1i2,x1i,x1i2,xy1i
```

Out[1]: (15, 11.75, 7.86, 1348, 123324.6, 983.67)

(a) Calculate the least squares estimates of the slope and intercept. Graph the regression line. Estimate σ^2.

Substituting the summary quantities into the equations given in the condensed lecture notes we get:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum y_i x_i - \frac{(\sum y_i)(\sum x_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \\ &= \frac{983.67 - \frac{(11.75)(1378)}{15}}{123324.6 - \frac{(1378)^2}{15}}\end{aligned}$$

In [2]:
```
beta1=(xy1i-(y1i*x1i/n))/(x1i2-(x1i^2)/n)
```

Out[2]: -0.03308255760720271

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n} \\ &= \frac{11.75}{15} + 0.03308 \frac{1378}{15}\end{aligned}$$

In [3]:
```
beta0=y1i/n-beta1*x1i/n
```

Out[3]: 3.756352510300617

To calculate the estimate of the variance, first the Sum of Squares of the Errors needs to be calculated. Expanding the equation in the condensed lecture notes gives:

$$\begin{aligned}SS_E &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum (y_i^2 - 2\hat{\beta}_0 y_i + \hat{\beta}_0^2 - 2\hat{\beta}_1 x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2) \\ &= \sum y_i - 2\hat{\beta}_0 \sum y_i + n\hat{\beta}_0^2 - 2\hat{\beta}_1 \sum x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum x_i + \hat{\beta}_1^2 \sum x_i^2\end{aligned}$$

In [4]:
```
SSE=y1i2-2*beta0*y1i-2*beta1*xy1i+beta0^2*n+2*beta0*beta1*x1i+beta1^2*x1
i2
```

Out[4]: -3.7348225545551657

Now substituting this into the σ^2 equation

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2} = \frac{-3.7348}{15-2}$$

In [5]:
```
MSE=SSE/(n-2)
```

Out[5]: -0.28729404265808967

(b) Use the equation of the fitted line to predict what pavement deflection would be observed when the surface temperature is 75 deg F.

To find the fitted value, the value 75 is substituted for x in the linear equation

$$y(75) = \hat{\beta}_0 + \hat{\beta}_1 \times 75$$

In [6]:
```
beta0+beta1*75
```

Out[6]: 1.2751606897604137

So there is a pavement deflection of 1.2752.

(c) What is the mean pavement deflection when the surface temperature is 95 deg F?

Again substituting 95 into the linear equation

In [7]:
```
beta0+beta1*95
```

Out[7]: 0.6135095376163595

(d) What change in mean pavement deflection would be expected for a 1 deg F change in surface temperature?

With each 1 deg F change in surface temperature the pavement deflection would change by the slope of the linear model (β_1). So the change in mean pavement deflection that would be expected from a 1 deg F change in surface temperature is -0.033083.

Question 2. House Selling Prices

An article in *Technometrics* by S.C. Narula and J.F Wellington ["Prediction, Linear Regression, and a Minimum Sum of Relative Errors" (1977, Vol. 19)] presents data on the selling price and annual taxes for 24 houses. The data is stored in (*11-6.csv*).

(a) Assuming that a simple linear regression model is appropriate, obtain the least squares fit relating selling price to taxes paid. What is the estimate of σ^2?

Julia can be used to read in the data and then a model fitted using glm as shown below.

In [27]:
```
using DataFrames, Distributions, GLM, PyPlot
prices = readtable("11-6.csv")
model = glm(@formula(SalePrice_1000~Taxes_local_school_county_1000),pric
es,Normal(),IdentityLink())
```

Out[27]:
```
DataFrames.DataFrameRegressionModel{GLM.GeneralizedLinearModel{GLM.GlmRes
p{Array{Float64,1},Distributions.Normal{Float64},GLM.IdentityLink},GLM.De
nsePredChol{Float64,Base.LinAlg.Cholesky{Float64,Array{Float64,2}}}},Arra
y{Float64,2}}

Formula: SalePrice_1000 ~ 1 + Taxes_local_school_county_1000

Coefficients:
                                 Estimate Std.Error z value Pr(>|z|)
(Intercept)                       13.3202   2.57172 5.17948    <1e-6
Taxes_local_school_county_1000   3.32437  0.390276   8.518   <1e-16
```

In [9]:
```
modelcoeff=coef(model)
```

Out[9]:
```
2-element Array{Float64,1}:
 13.3202
  3.32437
```

So the least squares model for this data is $\text{Sale Price} = 13.3202 + 3.32437\ \text{Taxes}$.To obtain the estimate of σ^2 the following code is used:

In [10]:
```
sum((prices[:SalePrice_1000].-modelcoeff[1].-modelcoeff[2].*prices[:Taxe
s_local_school_county_1000]).^2)/(size(prices,1)-2)
```

Out[10]:
```
8.767752951991381
```

Therefore $\hat{\sigma}^2 = 8.76775295199138.$

(b) Find the mean selling price given that the taxes paid are $x = 6.00.$

Substituting the value $x = 6.00$ into the equation obtained above the following result is obtained

In [11]:
```
modelcoeff[1]+modelcoeff[2]*6.00
```

Out[11]:
```
33.26640668148228
```

So the mean selling price given that the taxes paid are $x = 6.00$ is $ 33266.42

(c) Calculate the fitted value of y corresponding to $x = 5.8980.$ Find the corresponding residual.

Again to calculate the fitted values the value $x = 5.8980$ is substituted into the equation for the model.

In [12]:
```
fitval=modelcoeff[1]+modelcoeff[2]*5.8980
```

Out[12]:
```
32.9273208156939
```

To calculate the residual, the observed value at this x value is required. The observed value is 30.9. The residual is then the difference between the observed value and the fitted value

In [13]:
```
30.9-fitval
```

Out[13]:
```
-2.027320815693905
```

(d) Calculate the fitted $\hat{y}_i$ for each value of x_i used to fit the model. Then construct a graph of $\hat{y}_i$ versus the corresponding observed value y_i and comment on what this plot would look like if the relationship between y and x was a deterministic (no random error) straight line. Does the plot actually obtained indicate that taxes paid is an effective regressor variable in predicting selling price?

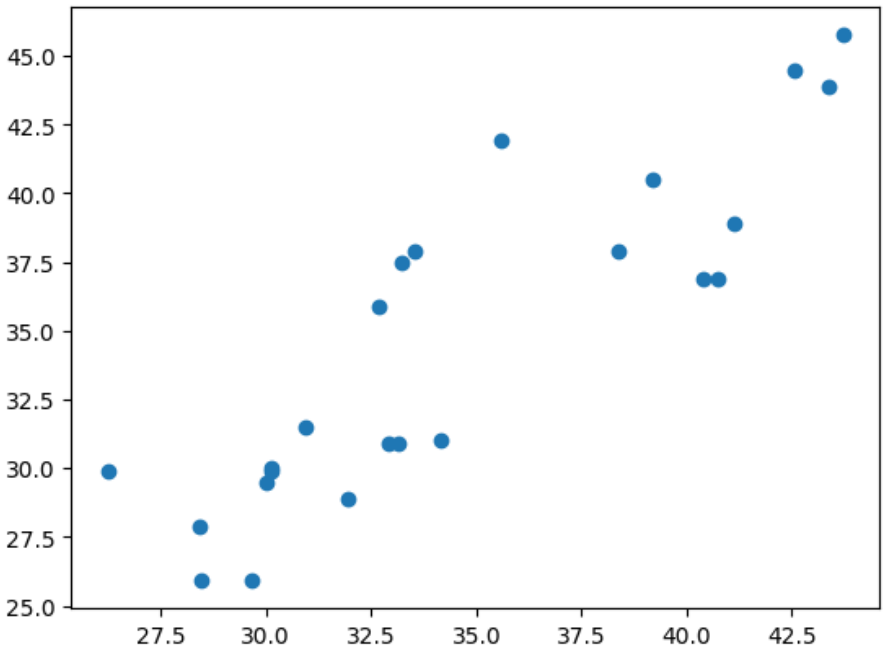
First the values of $\hat{y}_i$ need to be calculated, which can be done with the following Julia code.

```
In [14]: yhat=modelcoeff[1].+modelcoeff[2].*prices[:Taxes_local_school_county_1000]
```

```
Out[14]: 24-element DataArrays.DataArray{Float64,1}:
          29.6681
          30.0112
          28.4225
          28.4703
          30.1405
          26.2553
          32.9273
          31.9496
          32.6953
          30.9403
          34.168
          33.1308
          30.1083
          40.7343
          35.5832
          39.1974
          43.3672
          33.2312
          38.3933
          42.5584
          33.5427
          41.1142
          40.3806
          43.7103
```

Using these values the following scatter plot can be obtained

```
In [15]: PyPlot.scatter(yhat,prices[:SalePrice_1000])
```




```
Out[15]: PyObject <matplotlib.collections.PathCollection object at 0x7f7aeffbd438>
```

Given that there is little variation around the diagonal line that would be expected, it can be said that taxes paid is an effective regressor of selling price.

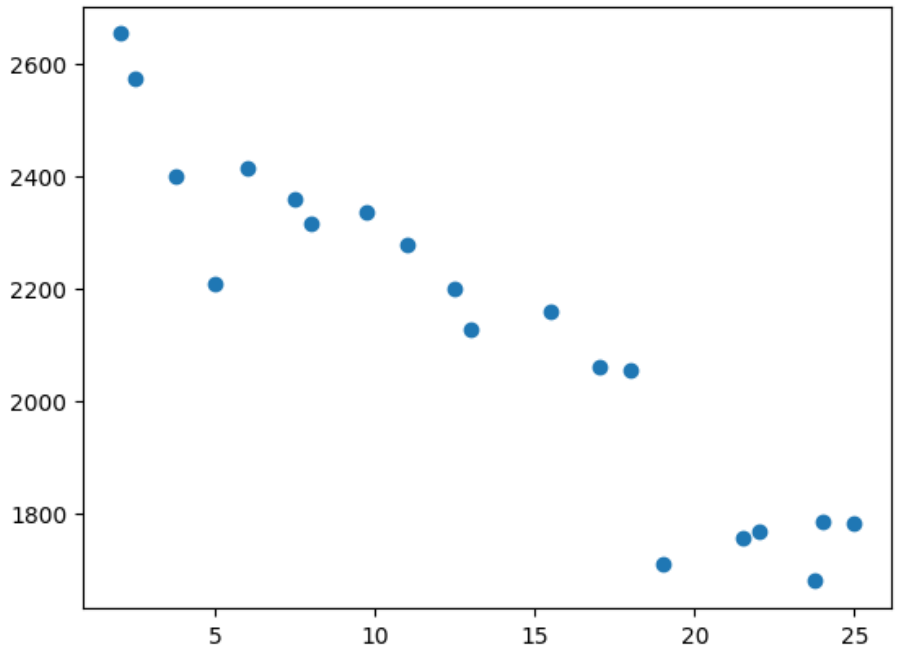
Question 3. Rocket Motor

A rocket motor is manufactured by bonding together two types of propellants, an igniter and a sustainer. The shear strength of the bond y is thought to be a linear function of the age of the propellant x when the motor is cast. The data is stored in (*11-13.csv*).

(a) Draw a scatter diagram of the data. Does the straight-line regression model seem to be plausible?

To draw the scatter plot of the data the following Julia code is used:

In [16]:
```
rocket=readtable("11-13.csv")
# Correct unusual value (typo)
rocket[:Strength_y_psi_][rocket[:ObservationNumber].==11]=rocket[:Streng
th_y_psi_][rocket[:ObservationNumber].==11]/10
PyPlot.scatter(rocket[:Age_x_weeks_],rocket[:Strength_y_psi_])
```



Out[16]:
```
PyObject <matplotlib.collections.PathCollection object at 0x7f7aec2f76a0>
```

As seen in the figure above, there is a negative linear trend in the data so a linear model would be appropriate.

(b) Find the least squares estimates of the slope and intercept in the simple linear regression model. Find an estimate of σ^2.

In [17]:
```
rockmodel= glm(@formula(Strength_y_psi_~Age_x_weeks_),rocket,Normal(),Id
entityLink())
```

Out[17]:
```
DataFrames.DataFrameRegressionModel{GLM.GeneralizedLinearModel{GLM.GlmRes
p{Array{Float64,1},Distributions.Normal{Float64},GLM.IdentityLink},GLM.De
nsePredChol{Float64,Base.LinAlg.Cholesky{Float64,Array{Float64,2}}}},Arra
y{Float64,2}}

Formula: Strength_y_psi_ ~ 1 + Age_x_weeks_

Coefficients:
              Estimate Std.Error  z value Pr(>|z|)
(Intercept)     2623.3   45.3275  57.8743   <1e-99
Age_x_weeks_  -36.9501   2.96555 -12.4598   <1e-34
```

So the model that is obtained is $\text{Strength} = 2623.3 - 36.9501\ \text{Age}$.To estimate σ^2 the following code is used:

In [18]:
```
rockcoeff= coef(rockmodel)
rockyhat= rockcoeff[1].+rockcoeff[2].*rocket[:Age_x_weeks_]
sum((rocket[:Strength_y_psi_].-rockyhat).^2)./(size(rocket,1)-2)
```

Out[18]:
```
9802.84515529937
```

So $\hat{\sigma}^2 = 9802.84515529937.$

(c) Estimate the mean shear strength of a motor made from propellant that is 20 weeks old.

Substituting 20 weeks into the model given above

```
In [19]: rockcoeff[1]+rockcoeff[2]*20
```

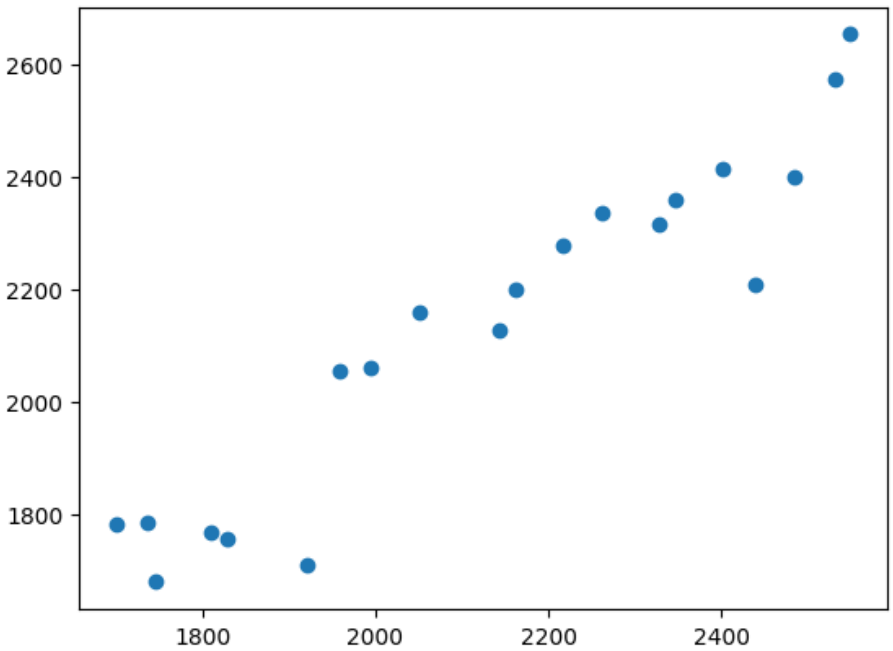
```
Out[19]: 1884.293558850658
```

So the mean sheet strength would be 1884.293558850658 psi.

(d) Obtain the fitted values $\hat{y}_i$ that correspond to each observed value y_i. Plot $\hat{y}_i$ versus y_i and comment on what this plot would look like if the linear relationship between shear strength and age were perfectly deterministic (no error). Does this plot indicate that age is a reasonable choice of regressor variable in this model?

Earlier the $\hat{y}_i$ have been calculated so only the scatter plot code is needed:

```
In [20]: PyPlot.scatter(rockyhat,rocket[:Strength_y_psi_])
```




```
Out[20]: PyObject <matplotlib.collections.PathCollection object at 0x7f7aec2a9f98>
```

The deviation of the points along the diagonal is evenly spread with no other patter, so the linear model appears to be a reasonable choice here.

Question 4. Regression without the Intercept Term

Suppose that we wish to fit a regression model for which the true regression line passes through the point (0,0). The appropriate model is $y = \beta x + \epsilon$. Assume that we have n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

(a) Find the least squares estimate of β.

First define L

$$L = \sum (y_i - \beta x_i)^2$$

Differentiate to find critical point

$$\frac{\partial L}{\partial \beta} = -2 \sum (y_i - \beta x_i) x_i = 0$$

Rearrange to find β

$$\begin{aligned} 0 &= -2 \sum y_i x_i + 2\beta \sum x_i^2 \\ \beta \sum x_i^2 &= \sum y_i x_i \\ \beta &= \frac{\sum y_i x_i}{\sum x_i^2} \end{aligned}$$

(b) Fit the model $y = \beta x + \epsilon$ to the chloride concentration roadway area data stored in (*11-22.csv*). Plot the fitted model on a scatter diagram of the data and comment on the appropriateness of the model.

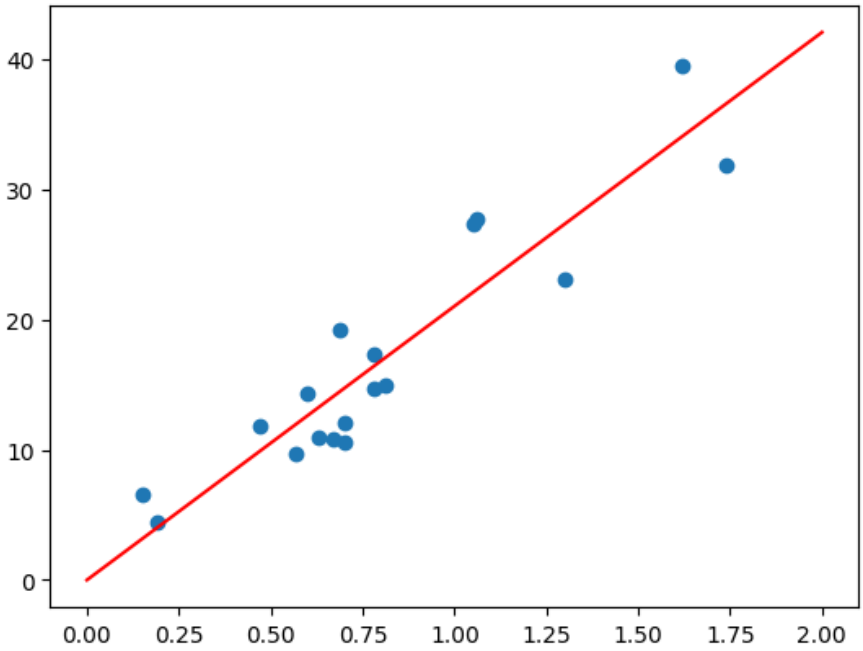
First read in the data and then using the equation above calculate β

In [21]:
```
chloride = readtable("11-22.csv")
βCh=sum(chloride[:ChlorideConcentration_y].*chloride[:RoadwayArea_x])/su
m(chloride[:RoadwayArea_x].^2)
```

Out[21]:
```
21.031460567201325
```

Now using this to plot the model on top of the scatter plot of the data

In [22]:
```
PyPlot.scatter(chloride[:RoadwayArea_x],chloride[:ChlorideConcentration_
y])
PyPlot.plot([0,2],[0;2]*βCh,"r")
```



Out[22]:
```
1-element Array{PyCall.PyObject,1}:
 PyObject <matplotlib.lines.Line2D object at 0x7f7aec2f7ac8>
```

Looking at the plot of the data, it follows the model well with small deviation away from the line. Given that it is expected that Chloride Concentration would be zero when the Roadway Area is zero this model is appropriate.

Question 5. Body Mass Index

The World Health Organization defines obesity in adults as having a body mass index (BMI) higher than 30. In a study of 250 men at Bingham Young University, 23 are by this definition obese. How good is waist (size in inches) as a predictor of obesity? A logistic regression model was fit to the data:

$$\log\left(\frac{p}{1-p}\right) = -41.828 + 0.9864 \text{ waist}$$

where p is the probability of being classified as obese.

(a) Does the probability of being classified as obese increase or decrease as a function of waist size? Explain.

As the slope is greater than zero, the log odds of being classified obese would increase as waist size increased. As such the probability p would increase as the ratio of p to $1-p$ would need to get larger.

(b) What is the estimated probability of being classified as obese for a man with a waist size of 36 inches?

First rearrange the equation given so that the probability is given

$$\begin{aligned}
\log\left(\frac{p}{1-p}\right) &= -41.828 + 0.9864 \text{ waist} \\
\frac{p}{1-p} &= \exp(-41.828 + 0.9864 \text{ waist}) \\
p &= (1-p)\exp(-41.828 + 0.9864 \text{ waist}) \\
p(1+\exp(-41.828 + 0.9864 \text{ waist})) &= \exp(-41.828 + 0.9864 \text{ waist}) \\
p &= \frac{\exp(-41.828 + 0.9864 \text{ waist})}{1+\exp(-41.828 + 0.9864 \text{ waist})} \\
&= \frac{1}{1+\exp(41.828 - 0.9864 \text{ waist})}
\end{aligned}$$

Now substitute the waist size of 36 inches into this equation

$$p = \frac{1}{1+\exp(41.828 - 0.9864 \times 36)}$$

```
In [23]: 1/(1+exp(41.828-0.9864*36))
```

```
Out[23]: 0.0018010190366111484
```

(c) What is the estimated probability of being classified as obese for a man with a waist size of 42 inches?

Again substituting into the equation for probability, a man with a waist size of 42 inches has a probability of being classified as obese of

```
In [24]: 1/(1+exp(41.828-0.9864*42))
```

```
Out[24]: 0.40150456364997983
```

(d) What is the estimated probability of being classified as obese for a man with a waist size of 48 inches?

Again substituting into the equation for probability, a man with a waist size of 48 inches has a probability of being classified as obese of

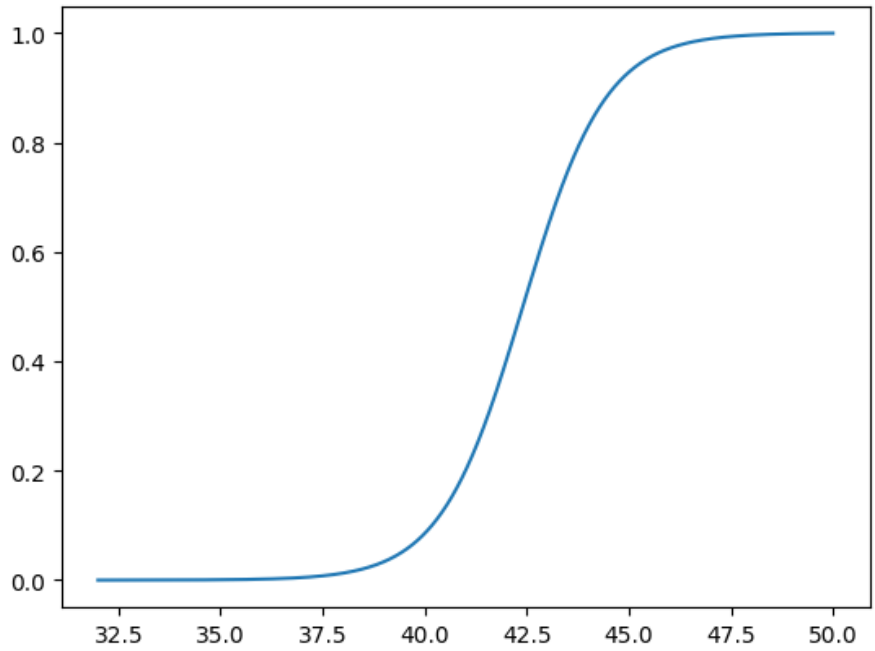
```
In [25]: 1/(1+exp(41.828-0.9864*48))
```

```
Out[25]: 0.9960069544322595
```

(e) Make a plot of the estimated probability of being classified as obese as a function of waist size

To create the plot, first create a vector of waist sizes (size 10^6) and then plot the function worked out above for the probability.

In [26]:
```
waist=linspace(32,50,10^6)
PyPlot.plot(waist,1./(1.+exp.(41.828.-0.9864.*waist)))
```



Out[26]:
```
1-element Array{PyCall.PyObject,1}:
 PyObject <matplotlib.lines.Line2D object at 0x7f7ae1c330b8>
```