Getting Started II: *Review of Sample Statistics and Optimization*

- ***Sample Statistics***
- ***Standardized/Normalized Variables***
- ***Optimization*: FOCs and SOCs**

Sample Statistics

1. We will make extensive use of Sample Statistics in this course, so it'll be useful to review those concepts (which you should have previously seen in your statistics course)… and to introduce the notation that we'll be using over the course of the semester.

2. You have a dataset consisting of n observations of two variables (x, y): $\{(x_i, y_i)\}\ i = 1, 2, \ldots n$. So, for example, you might have randomly selected fifty individuals from a population and observed their heights and weights. In that case, the *i*'s would track the individuals, and the *x*'s and *y*'s might reflect their heights and weights, respectively, so that x_i would be the height of person i and y_i would be his or her weight.

3. The ***sample mean*** (*average*):
   a. $\overline{x} = \frac{1}{n}\sum x_i$ and $\overline{y} = \frac{1}{n}\sum y_i$. Note that $\sum x_i = n\overline{x}$.

4. Deviations from means:
   a. $dx_i = (x_i - \overline{x})$ and $dy_i = (y_i - \overline{y})$
   b. By construction, the total/sum of the deviations from the means for any variable will be zero: $\sum dx_i = \sum (x_i - \overline{x}) = \left(\sum x_i\right) - n\overline{x} = 0$ and $\sum dy_i = \sum (y_i - \overline{y}) = 0$.

5. The ***sample variance***:
   a. $S_{xx} = S_x^2 = \frac{1}{n-1}\sum (dx_i)^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$ and likewise for the y's.
   b. This is almost the average squared deviation from the mean (except we divide by n-1, not n… the reason for this will become clear when we consider unbiased estimation).
   c. Also: Since $\sum x_i = n\overline{x}$, $S_{xx} = \frac{1}{n-1}\sum x_i^2 - \frac{n}{n-1}\overline{x}^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1}$.

6. The ***sample standard deviation*:**
   a. $S_x = \sqrt{S_{xx}} = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum (dx_i)^2} = \sqrt{\frac{1}{n-1}\sum (x_i - \overline{x})^2}$, and likewise for the y's.
   b. This is the square root of the Sample Variance. Since the Sample Variance is sort of an average squared deviation from the mean, this is sort of an average deviation from the

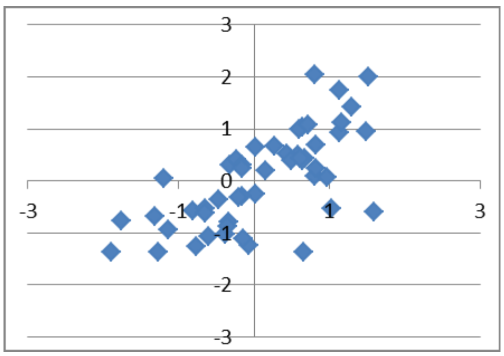
sample mean… but that's not quite right, of course. It is however a useful way to think of the sample standard deviation, sort of.

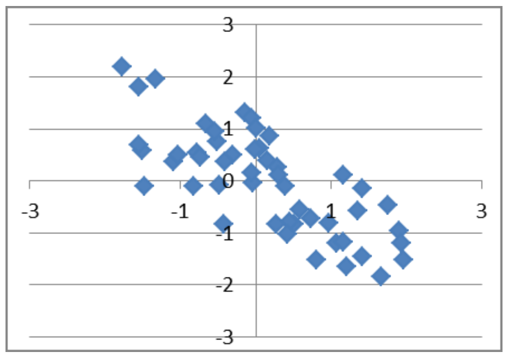
7. The ***sample covariance***:

   a. $\mathrm{cov}(x,y) = S_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}$, since $\sum x_i = n\bar{x}$ and $\sum y_i = n\bar{y}$.

   b. Again, almost the average product of the deviations from the means (except we again divide by n-1, not n… and yes, this is also related to unbiased estimation).

   c. Some intuition/examples: In the following examples, $\bar{x} = 0$ and $\bar{y} = 0$. On the left, most of the data are in quadrants I and III, where $(x_i - \bar{x})(y_i - \bar{y}) > 0$, and so when you sum those products, as you do in calculating S_{xy}, you get a positive sample covariance. Most of the action on the right is in quadrants II and IV where $(x_i - \bar{x})(y_i - \bar{y}) < 0$, and so those products sum to a negative number, and we have a negative covariance.





   d. A few properties:

      i. The covariance of x with itself is the variance of x:

      $$\mathrm{cov}(x,x) = \frac{1}{n-1}\sum(x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = S_{xx}$$

      ii. The covariance of a sum is the sum of the variances plus twice the covariance:

      $$\mathrm{var}(x+y) = \frac{1}{n-1}\sum[(x_i + y_i) - (\bar{x} + \bar{y})]^2$$

      $$= \frac{1}{n-1}\sum\left[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2\right] = S_{xx} + 2S_{xy} + S_{yy}$$

      1. If $S_{xy} = 0$, then $\mathrm{var}(x+y) = S_{xx} + S_{yy} = \mathrm{var}(x) + \mathrm{var}(y)$

iii. The covariance of linear transformations of the x's and y's:

$$\operatorname{cov}(a+bx, c+dy) = \frac{1}{n-1}\sum\left[\left(a+bx_i\right)-\left(a+b\bar{x}\right)\right]\left[\left(c+dx_i\right)-\left(c+d\bar{x}\right)\right]$$

$$= \frac{1}{n-1}\sum\left[b\left(x_i-\bar{x}\right)\right]\left[d\left(y_i-\bar{y}\right)\right] = bdS_{xy} = bd\operatorname{cov}(x,y)$$

iv. The covariance of x with sums of variables:

$$\operatorname{cov}(x, y+z) = \frac{1}{n-1}\sum\left(x_i-\bar{x}\right)\left[\left(y_i+z_i\right)-\left(\bar{y}+\bar{z}\right)\right]$$

$$= \frac{1}{n-1}\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)+\frac{1}{n-1}\sum\left(x_i-\bar{x}\right)\left(z_i-\bar{z}\right) = S_{xy}+S_{xz}$$

$= \operatorname{cov}(x,y)+\operatorname{cov}(x,z)\ldots$ the sum of the covariances of x with each other variable.

v. And finally, since $\sum\bar{x}(y_i-\bar{y}) = \bar{x}\sum(y_i-\bar{y}) = \bar{x}\sum y_i - n\overline{xy} = n\overline{xy} - n\overline{xy} = 0$, we can drop either $\bar{x}$ or $\bar{y}$ (but not both!) from the equation for the sample covariance. So:

$$S_{xy} = \frac{1}{n-1}\sum x_i(y_i-\bar{y}) \text{ and } S_{xy} = \frac{1}{n-1}\sum(x_i-\bar{x})y_i .$$

vi. These formulas will be useful later in the semester.

8. The ***sample correlation*:**

a. $\rho_{xy} = \dfrac{S_{xy}}{S_x S_y}$, the ratio of the sample covariance to the product of the sample standard deviations.

b. It may not be obvious, but by construction, $\left|\rho_{xy}\right| \le 1, or\ -1 \le \rho_{xy} \le 1$.[1]

c. If $S_{xy} = 0$, the sample covariance is 0 and the sample correlation is also 0. And if the sample covariance is negative (positive), then so is the sample correlation (since sample standard deviations are always positive, so long as they are well defined and not zero).

d. If $\left|\rho_{xy}\right|$ is close to 1 then the relationship between x and y will look quite linear (with a positive slope if $\rho_{xy} \sim 1$, and a negative slope if $\rho_{xy} \sim -1$.

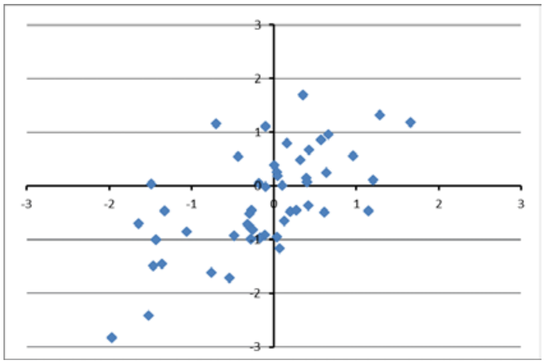
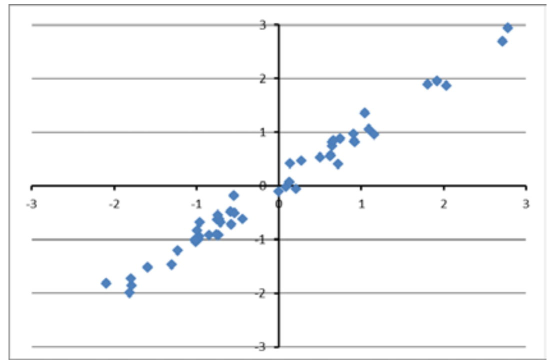
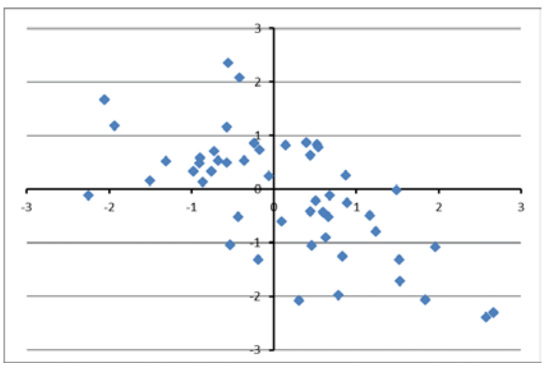
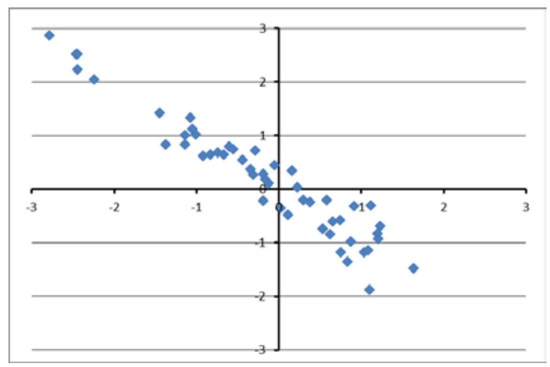
i. If there is in fact an exact linear relationship between the *x's* and *y's* (so that $y_i = \beta_0 + \beta_1 x_i$, where β_0 is the *intercept* and β_1 is the *slope*)… then the sample correlation between the x's and the y's is +1 if $\beta_1 > 0$, –1 if $\beta_1 < 0$, and 0 if $\beta_1 = 0$.

e. And as $\left|\rho_{xy}\right|$ gets closer to 0, the relationship between *x* and *y* looks less and less linear.

f. So**: Correlation captures the extent to which x and y are moving together *in a linear fashion*.**

[1] This follows from the Cauchy–Schwarz inequality.

g. Here are some examples:[2]

	
$\rho_{xy} = .635$	$\rho_{xy} = .991$
	
$\rho_{xy} = -.652$	$\rho_{xy} = -.959$

Standardized/Normalized Variables

9. For reasons that will later become clear, it is sometimes useful to *standardize*, or *normalize*, variables. We do this with a particular *linear transformation*… by first subtracting the variable's mean from each observation, and then dividing each new value by the variable's standard deviation: $z_i = \dfrac{x_i - \overline{x}}{S_x}$.

[2] Sampling 50 times from a bivariate Standard Normal distribution.

10. ***Means and variances***: The result is a transformed variable, z, with mean 0 and variance 1:

    a. Sample Mean of the z_i's: $\bar{z} = \dfrac{\bar{x}-\bar{x}}{S_x} = 0$

    b. Sample Variance of the z_i's: $S_{zz} = \dfrac{1}{n-1}\sum (z_i - \bar{z})^2 = \dfrac{1}{n-1}\sum z_i^2$ since $\bar{z} = 0$, and so

    $$S_{zz} = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{S_x}\right)^2 = \frac{1}{S_x^2}\frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{S_{xx}} S_{xx} = 1$$

11. ***Covariances and correlations***: While sample covariances will typically be impacted by standardization, sample correlations will not. Let's use * to indicate normalized /standardized, so: $x_i^* = \dfrac{x_i - \bar{x}}{S_x}$ and $y_i^* = \dfrac{y_i - \bar{y}}{S_y}$. Then it's easy to show that:

    a. Sample Covariances: $S_{x^*y^*} = \dfrac{1}{n-1}\sum x_i^* y_i^* = \dfrac{1}{S_x S_y}\dfrac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) = \dfrac{S_{xy}}{S_x S_y} = \rho_{xy}$.

        i. Note that the sample covariance of two standardized variables is also their sample correlation.

    b. Sample Correlations: $\rho_{x^*y^*} = \dfrac{S_{x^*y^*}}{S_{x^*}S_{y^*}} = S_{x^*y^*}$.since $S_{x^*} = S_{y^*} = 1$, and so:

    $$\rho_{x^*y^*} = S_{x^*y^*} = \frac{S_{xy}}{S_x S_y} = \rho_{xy}.$$

12. The correlation result will be especially useful later, so to repeat:

    Standardization will typically affect sample means, variances and covariances of variables… but it does not impact sample correlations.

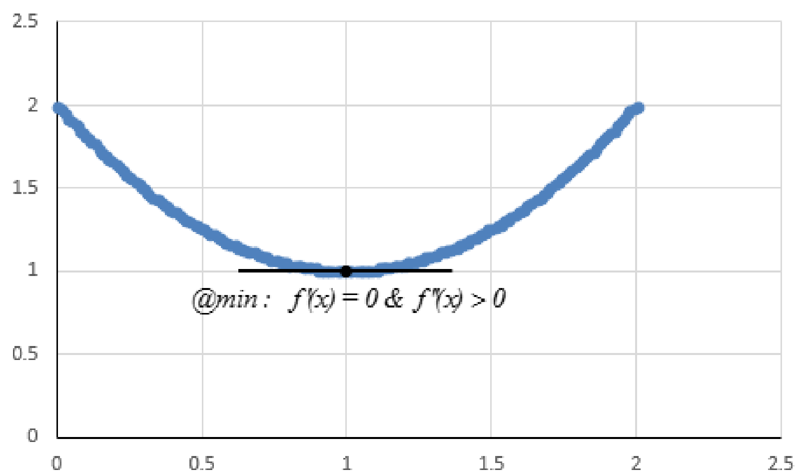
Optimization: **FOCs and SOCs**

13. For most of the semester we'll be focusing on using *Ordinary Least Squares (OLS)* to estimate unknown parameter values. When running OLS, we are solving an optimization problem: What coefficients minimize the sum of the squared differences between predicted and actual values?

14. This is a minimization problem. We'll call these differences between predicted and actual values *residuals*… and the sum of the squared residuals *SSR*s, for, well, *Sum of Squared Residuals*. Since we are trying to minimize SSRs, we call SSR the *Objective Function* (which is to be minimized).

15. And if time permits, we might also look at a second approach to estimation called *Maximum Likelihood Estimation (MLE)*. When running MLE models, the objective is to find the coefficient values that maximize the value of the associated *likelihood* function. This is a maximization problem.

16. There are many ways to solve optimization problems. Probably the most common approach is to use what are called:
    a. ***First Order Conditions (FOCs)*** to identify solution candidates, and
    b. ***Second Order Conditions (SOCs)*** to establish that the candidates do in fact minimize or maximize the objective function.[3]
17. You'll see below that I will distinguish between *local* and *global* optimums. To explain, I focus on the case of minimization:
    a. *local* minimum (in a neighborhood of x*): x* is a local minimum if the value of the function at x* is no greater that the value of the function in a small neighborhood around x*.
    b. *global* minimum (everywhere): And x* provides a global minimum if the value of the function at x* is no greater that all other values of the function.
18. ***A Picture***: The following Figure shows FOCs and SOCs in action, and considers a minimization problem.[4]
    a. In this Figure, and moving x left to right, the function $f(x)$ is decreasing as x increases towards 1, reaches a minimum value when x=1 and increases as x moves to higher values.
    b. Notice also that to the left of x=1, the derivative (slope) of the function is negative, and to the right of x=1 it is positive.
    c. And most importantly, when x=1, the derivative is 0 (the function flattens out for a brief moment)… and that only happens at x=1.



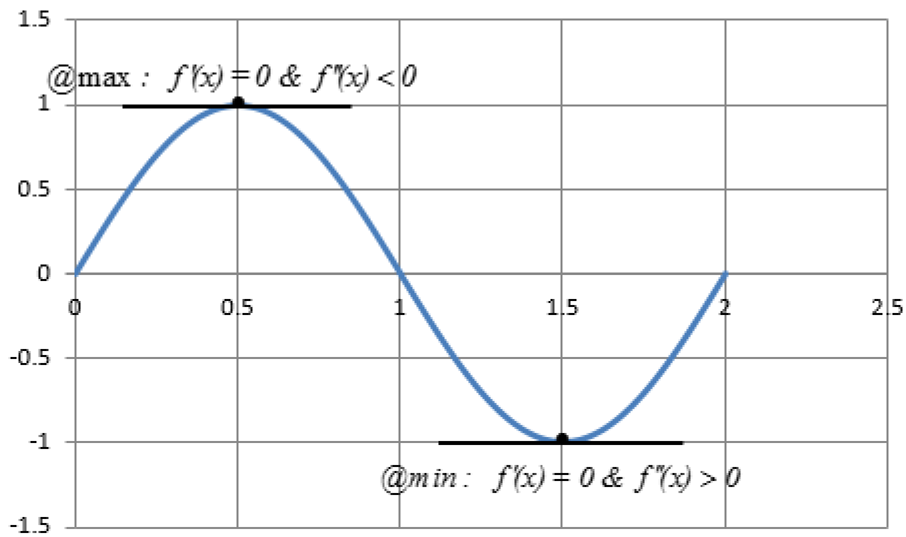
[3] We'll assume that the objective function is continuously differentiable.
[4] You should also have seen FOCs and SOCs in action in you Micro Theory course.

d. FOC: First Order Condition

i. Optimization candidates must have a zero first derivative: $f'(x^*) = 0$.

ii. If that is not the case, then small movements left or right of x* will lead to smaller or larger values of the objective function… which is to say that there are better candidates, and x* does not give us a maximum or a minimum value of the objective function.

e. As the following Figure illustrates, there may be multiple candidates for which the FOC is satisfied. If we are fortunate, we'll be able to choose between the candidates using a SOC:



f. **SOC: Second Order Condition**

i. ***Minimization***:

1. We have a *local* minimum at x* if the FOC is satisfied, so $f'(x^*) = 0$, and if the *function* is *concave up* (we used to say *convex*) at x*, so that $f''(x^*) > 0$.
2. This second order condition (involving the second derivative) assures us that in the neighborhood of x*, the objective function is declining to the left of x* and increasing to the right… which means that x* is a *local* minimum. In the Figure above, this happens at x* = 1.5.
3. If $f''(x) > 0$ for all x's then the function is strictly concave up and we have a *global* minimum at x*.

ii. ***Maximization***:

1. We have a *local* maximum at x* if the FOC is satisfied, so $f'(x^*) = 0$, and if the function is now *concave down* (we used to say *concave*) at x*, so that $f''(x^*) < 0$.

2. This second order condition assures us that we have a *local* maximum at x*, since the function is increasing to the left of x* and decreasing to the right. In the Figure above, this happens at x* = 0.5.
3. If $f''(x) < 0$ for all x's then the function is strictly concave down and we have a *global* maximum at x*.

19. ***Summary***:

a. **FOCs - Identify solution candidates:** Use FOCs to identify candidates for solving the optimization problem. So start by finding the x* values for which $f'(x^*) = 0$.

b. **SOCs - Check to see if have a min or max:** Sign the SOC for each identified candidate: What is the sign of $f''(x)$? If the second derivative at x* is negative (so $f''(x^*) < 0$), then we have a *local* maximum, and if it's positive (so $f''(x^*) > 0$) then we have a *local* minimum,

c. **Local v. global:** And if the SOC is always of the same sign ($f''(x)$ is always positive or always negative, for any x), then we have *global* maximums or minimums.

20. **An Example**: ***Estimate the unknown mean of a distribution***

a. You are interested in estimating μ, the mean of the distribution of some random variable Y, and decide to randomly sample n times from this distribution. Your dataset consists of n observations: $\{y_i\}$ $i = 1, 2, ... n$.

b. There are many many ways to estimate the unknown mean μ with the given sampled data. Here's one that perhaps you haven't previously encountered:

To estimate μ, find the number m that is *closest* to the observed sample.

c. So implement this estimator, you'll need to decide on how you'll be measuring *closeness*. There are lots of such metrics. Here's one, which plays a prominent role in *least squares* regression analysis:

Sum Squared Residuals (SSR): $SSR = \sum \left(y_i - m \right)^2$

The difference between the observed (sampled) value y_i and the estimate m, $y_i - m$, is called the *residual* (sometimes we refer to this as the difference between the actual and the estimate). To generate SSRs, you square the residuals and then add them up.

d. To measure *closeness*, you might be inclined to just add up the residuals. But then you'd allow positive and negative residuals to offset one another, which makes no sense. You avoid this by squaring the residuals first before summing them.

   i. You might ask: Why not just sum the absolute values of the residuals? That, of course makes lots of sense. However it turns out that that approach is not as analytically simple/straightforward, and so we turn to SSRs.

e. The minimization problem:

Find the *m* that minimizes $SSR = \sum (y_i - m)^2$

f. We have the following FOC and SOC for the minimization problem:

i. FOC: $\frac{dSSR}{dm} = \sum 2(y_i - m)(-1) = 0$ and so $\sum y_i = \sum m^* = nm^*$ and $m^* = \frac{1}{n}\sum y_i = \overline{y}$.

And so the only SSR minimization candidate satisfying the FOC is the ***sample mean***, $\overline{y}$.

ii. SOC: $\frac{d^2SSR}{dm^2} = \sum 2(-1)(-1) = 2n > 0$.

Since $\frac{d^2SSR}{dm^2} > 0$ for all m, SSR is concave up in m, and we have a *global* minimum at the m value that satisfies the FOC.

g. Since the SOC is always satisfied and since the FOC is satisfied at $m^* = \frac{1}{n}\sum y_i = \overline{y}$, we find that the value of m that minimizes $SSR = \sum (y_i - m)^2$ is sample mean of the y's.

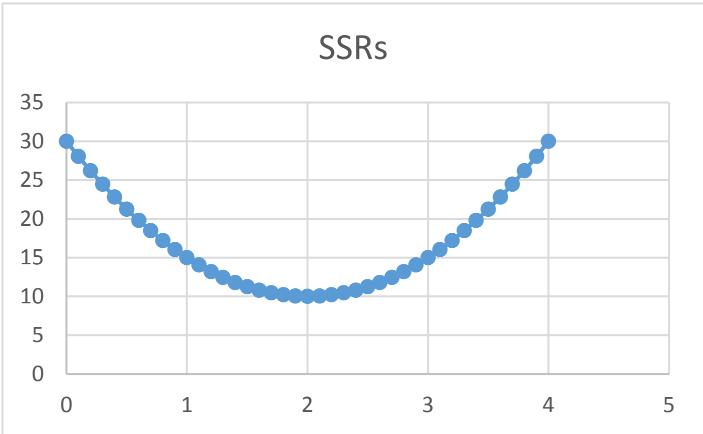
h. ***Who knew?*** And so the sample mean is a least squares estimator of the unknown population mean μ. We'll be returning to this example later in the semester.

21. ***… with five data points***

a. You have n=5 observations of the variable y, $\{0,1,2,3,4\}$. The sample mean for these observations is 2. For this sample,

$$SSR = \left((0-m)^2 + (1-m)^2 + (2-m)^2 + (3-m)^2 + (4-m)^2\right).$$

b. The following Figure shows the SSRs for different m values given these five datapoints.



Notice that SSRs are declining as m increases to 2, reach a minimum at m=2 and increase thereafter.

c. Not surprisingly, the *eyeball test* confirms what you saw above:

i. the FOC is satisfied at m*=2, and

ii. the SOC is also satisfied.