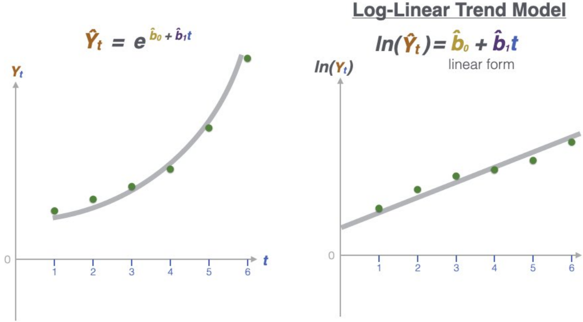
Fundamentals of Econometrics

Lecture 5: Multiple Regression Analysis: Further Issues



Section 1

Data Scaling

Rescaling Data

- Recalled in L2 we showed that units of measurement do not affect the relationship between y and x.
  - Instead, the scaling of the coefficients is affected.

How does scaling impact our statistical tests?

Question: Does changing units of measurement affect statistical tests?

- Answer: No.
  - Both the estimated coefficients and the standard errors are scale proportionally.

Rescaling Data

Let us consider the following models for the San Joaquin House price data:

```r
library(foreign) # used to read in STATA files
sanjoaquin <- read.dta("../../data/San_Joaquin.dta")
p1 <- lm(saleprice ~ ., data = sanjoaquin)
p2 <- lm(saleprice/1000 ~ ., data = sanjoaquin) # Price in '000s

# Extract t-statistics manually
t_stats1 <- coef(summary(p1))[, "t value"]
t_stats2 <- coef(summary(p2))[, "t value"]

stargazer(p1, p2, header = FALSE,
          se = list(t_stats1, t_stats2), # replace the se with t-stats
          notes = "T-stats are reported in parentheses.")
```

Table 1

	Dependent variable:	
	saleprice	saleprice/1000
	(1)	(2)
gdistance	−1.610	−0.002
	(−9.100)	(−9.100)
wdistance	0.665	0.001
	(4.190)	(4.190)
cdistance	2.440	0.002
	(2.630)	(2.630)
bathrooms	2,460.000***	2.460*
	(1.400)	(1.400)
bedrooms	−5,855.000***	−5.860
	(−5.030)	(−5.030)
sqftbuilding	72.800*	0.073
	(37.600)	(37.600)
sqftlot	0.506	0.001
	(10.400)	(10.400)
age	−506.000***	−0.506
	(−13.500)	(−13.500)
Constant	46,319.000***	46.300***
	(10.800)	(10.800)
Observations	2,661	2,661
R^2	0.678	0.678
Adjusted R^2	0.677	0.677

Standardized Coefficients

- Sometimes, our variables are measured on different scales, making direct interpretation of coefficients challenging.
- Standardizing variables allows us to compare the relative importance of predictors within the same model.

Examples:

- If we are working with indices, their units may not have an intuitive interpretation. Standardization ensures comparability.
- When analyzing standardized test scores, the scoring scale is often arbitrary (e.g., SAT vs. ACT). Standardization allows us to interpret results in terms of standard deviations rather than raw score differences.

Standardized Coefficients

Why Use Standardized Coefficients?

- Using standardized coefficients enables us to answer questions such as:
  - *Which predictor has the strongest association with the dependent variable?*
  - *How does a one standard deviation increase in an independent variable affect the dependent variable?*

Standardized Coefficients

Original OLS:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

Subtracting all variables by their means:

$$y - \bar{y} = \hat{\beta}_1 (x_{i1} - \bar{x}_1) + \hat{\beta}_2 (x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k (x_{ik} - \bar{x}_k) + \hat{u}_i$$

Dividing by the standard deviation:

$$\left(y - \bar{y}\right) / \hat{\sigma}_y = (\hat{\sigma}_1 / \hat{\sigma}_y) \hat{\beta}_1 \left({}^{x_{i1} - \bar{x}_1}/_{\hat{\sigma}_1} \right) + (\hat{\sigma}_2 / \hat{\sigma}_y) \hat{\beta}_2 \left({}^{x_{i2} - \bar{x}_2}/_{\hat{\sigma}_2} \right) + \ldots + (\hat{\sigma}_k / \hat{\sigma}_y$$

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + error$$

where z_y is the z-score of y, z_i is the z-score of x_i, and $\hat{b}_i$ is the standardized coefficient.

$$\hat{b}_i = (\hat{\sigma}_i / \hat{\sigma}_y) \hat{\beta}_i$$

Interpretation: If x_i increases by one standard deviation, then y will increase by $\hat{b}_i$ standard deviations.

Effect of Pollution on House Prices

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u$$

```r
lm(price ~ nox + crime + rooms + dist + stratio, data = hprice2) |> summary()
```

```
##
## Call:
## lm(formula = price ~ nox + crime + rooms + dist + stratio, data = hprice2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13914  -3201   -662   2110  38064
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20871.1     5054.6    4.13  4.3e-05 ***
## nox          -2706.4      354.1   -7.64  1.1e-13 ***
## crime         -153.6       32.9   -4.66  4.0e-06 ***
## rooms         6735.5      393.6   17.11  < 2e-16 ***
## dist         -1026.8      188.1   -5.46  7.6e-08 ***
## stratio      -1149.2      127.4   -9.02  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5590 on 500 degrees of freedom
## Multiple R-squared:  0.636,  Adjusted R-squared:  0.632
## F-statistic:  174 on 5 and 500 DF,  p-value: <2e-16
```

$$z_{price} = b_0 + b_1 z_{nox} + b_2 z_{crime} + b_3 z_{rooms} + b_4 z_{dist} + b_5 z_{stratio} + error$$

```r
# Drop the intercept
lm(scale(price) ~ -1 + scale(nox) + scale(crime) + scale(rooms) +
     scale(dist) + scale(stratio), data = hprice2) |> summary()
```

```
##
## Call:
## lm(formula = scale(price) ~ -1 + scale(nox) + scale(crime) +
##     scale(rooms) + scale(dist) + scale(stratio), data = hprice2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.511 -0.348 -0.072  0.229  4.133
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## scale(nox)      -0.3404     0.0445   -7.65  1.0e-13 ***
## scale(crime)    -0.1433     0.0307   -4.67  3.9e-06 ***
## scale(rooms)     0.5139     0.0300   17.13  < 2e-16 ***
## scale(dist)     -0.2348     0.0430   -5.46  7.3e-08 ***
## scale(stratio)  -0.2703     0.0299   -9.03  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.606 on 501 degrees of freedom
## Multiple R-squared:  0.636,  Adjusted R-squared:  0.632
## F-statistic:  175 on 5 and 501 DF,  p-value: <2e-16
```

Section 2

Functional Forms

Choosing a Functional Form

Problems: - Often times, economic theory provides guidance on variables to include, but not the function. - Several alternative functions exist.

How should we choose a form?

- Understand the relationship between x and y.
- Depends on the problem that is being modeled.
- Look at existing literature.
  - For instance, wage equations are often log-linear models.
  - Doesn't mean existing literature is correct…
- Use Taylor series approximation for continuous independent variables.

Commonly Used Forms

Name	Functional Form	Marginal Effect	Elasticity
Linear	$y = \beta_0 + \beta_1 x + u$	β_1	$\beta_1 x/y$
linear-log	$y = \beta_0 + \beta_1 \ln(x) + u$	β_1/x	β_1/y
log-linear	$log(y) = \beta_0 + \beta_1 x + u$	$\beta_1 y$	$\beta_1 x$
Reciprocal	$y = \beta_0 + \beta_1(1/x) + u$	$-\beta_1/x^2$	$-\beta_1/xy$
Quadratic	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	$\beta_1 + 2\beta_2 x$	$(\beta_1{+}2\beta_2 x)/(x/y)$
Interaction	$y = \beta_0 + \beta_1 x + \beta_2 xz + u$	$\beta_1 + \beta_3 z$	$(\beta_1 + \beta_3 z)/(x/y)$
log-reciprocal	$log(y) = \beta_0 + \beta_1(1/x) + u$	$\beta_1 y/x^2$	$-\beta_1/x$
log-quadratic	$log(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	$y(\beta_1 + 2\beta_2 x)$	$x(\beta_1 + 2\beta_2 x)$
log-log	$log(y) = \beta_0 + \beta_1 \log(x) + u$	$\beta_1 y/x$	β_1

Logarithmic Functional Forms

- Convenient percentage/elasticity interpretation
- Slope coefficients of logged variables are invariant to rescalings
- Taking logs often eliminates/mitigates problems with outliers
- Taking logs often helps to secure normality and homoskedasticity
- Variables measured in units such as years should not be logged
- Variables measured in percentage points should also not be logged
- Logs must not be used if variables take on zero or negative values
  - Workaround: add a constant to the variable before taking logs, say $log(x+1)$ instead.
- It is hard to reverse the log-operation when constructing predictions

Logarithmic Functional Forms

$$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 rooms + u$$

- β_1 is the elasticity of price with respect to *nox* (pollution).
- β_2 is the change in log(price) when Δ rooms = 1.
    - $\beta_2 \cdot 100$ is the **approximate** percentage change in *price*.
    - $\beta_2 \cdot 100$ is sometimes called the **semi-elasticity**.

Using `hprice2`:

$$\widehat{\log(\text{price})} = \underset{(0.188)}{9.234} + \underset{(0.066)}{-0.718} \log(\text{nox}) + \underset{(0.019)}{0.306}\text{rooms}$$

- When *nox* increases by 1%, prices fall by 0.718%, holding only rooms constant.
- **Approximating:** When rooms increase by 1, prices increase by 30.592%, holding only *nox* constant.
    - Issue: As $log(price)$ becomes larger and larger, this approximation becomes less accurate.
- **Exact change:** $100 \cdot (\exp(0.306) - 1) = 35.787$

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2 + u$$

Keeping x_1 fixed:

$$\Delta \widehat{\log(y)} = \hat{\beta}_2 \Delta x_2$$

Taking exponentials:

$$\begin{aligned}
\exp(\Delta \widehat{\log(y)}) &= \exp(\hat{\beta}_2 \Delta x_2) \\
\frac{\Delta \hat{y}}{\hat{y}} + 1 &= \exp(\hat{\beta}_2 \Delta x_2) \\
\frac{\Delta \hat{y}}{\hat{y}} &= \exp(\hat{\beta}_2 \Delta x_2) - 1 \\
\% \Delta \hat{y} &= 100 \cdot \Big[\exp(\hat{\beta}_2 \Delta x_2) - 1 \Big]
\end{aligned}$$

$$\% \widehat{\Delta Price} = 100 \cdot [\exp(0.306) - 1] = 35.787$$

What happens if we decreased the number of rooms by 1 instead?

$$\% \widehat{\Delta Price} = 100 \cdot [\exp(-1 \cdot 0.306) - 1] = -26.355\%$$

Quadratic Functional Forms

- Often used in economics to capture increasing or decreasing marginal effects.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

- Although β_1 measures the change in y with respect to x, it makes no sense holding x^2 constant, while changing x.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

then we have the approximation

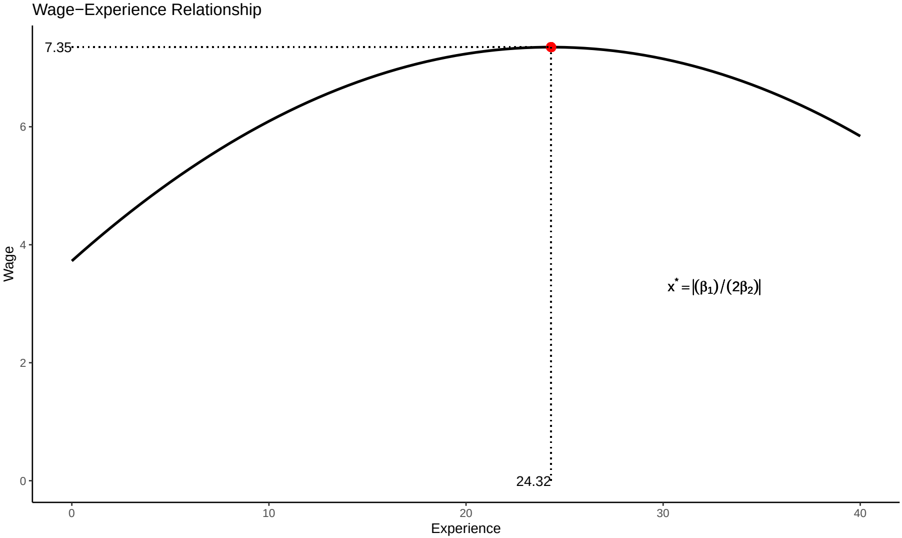
$$\Delta\hat{y} = (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x \implies {}^{\Delta\hat{y}}/_{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$$

- If $x = 0$, β_1 captures the approximate slope in going from $x = 0$ to $x = 1$. After that, we must account for the second term, $2\beta_2 x$.

Can you compute the turning point?

Wage-Experience Relationship (`wage1`):

$wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$



Models with Interactions

- There may be situations where two independent variables jointly affect y

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Example:

- Total degree days and total precipitation affect crop yields.
- Are yields higher when high temperatures are offset by more precipitation?
- What is the marginal effect of more heating degree days?
- Expected sign of β_3?

$$yield = \beta_0 + \beta_1 HDD + \beta_2 PCPN + \beta_3 HDD \times PCPN$$

House Price Example

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + bthrms + u$$

The partial effect of *bdrms* on *price* (holding all other variables constant) is

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft \tag{1}$$

- If $\beta_3 > 0$: marginal effect of an additional bedroom is larger for larger houses.
    - In other words, there is an interaction effect between square footage and the number of bedrooms.
- To get at the effect of another bedroom on price, we must evaluate (1) at different levels of square footage– mean, lower or upper quartile, median, etc.
- Interpreting the original variables might get tricky when interactions are included.

House Price Example

We could reparameterize the model to make interpretation easier:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$
$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$

where μ_1 and μ_2 are the means of x_1 and x_2, respectively.

- Now, δ_2 is the partial effect of x_2 on y at the mean value of x_1.
- With some algebra, we can show that $\delta_2 = \beta_2 + \beta_3 \mu_1$.

Advantages:

- Easy interpretation of all parameters.
- S.e.s for partial effects at the means are readily available.
- If necessary, interactions may be centered at other interesting values.

Average Partial Effects (APE)

- In models with quadratics, interactions, and other nonlinear functional forms, the partial effect depend on the values of one or more explanatory variables
- Average partial effect (APE) is a summary measure to describe the relationship between dependent variable and each explanatory variable
- After computing the partial effect and plugging in the estimated parameters, average the partial effects for each unit across the sample

Wage Example

```
lm(lwage ~ educ + exper + tenure + tenure*exper, data = wage1) |> summary()
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + tenure + tenure * exper,
##     data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0211 -0.2710 -0.0186  0.2697  1.3959
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.26731    0.10244    2.61   0.0093 **
## educ          0.08718    0.00728   11.97  < 2e-16 ***
## exper         0.00730    0.00184    3.97  8.1e-05 ***
## tenure        0.05394    0.00779    6.93  1.3e-11 ***
## exper:tenure -0.00102    0.00023   -4.45  1.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Section 3

Goodness of Fit and Selection of Regressors

Goodness of Fit

General Comments:

- A high R-squared does not imply that there is a causal interpretation
- A low R-squared does not preclude precise estimation of partial effects

R^2**:**

- What is R-squared supposed to measure?

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)} \text{ is an estimate for } \underbrace{\boxed{1 - \frac{\sigma_u}{\sigma_y^2}}}_{\text{Pop. } R^2}$$

Adjusted R^2

- Penalizes the R^2 for each additional variable we add to the model

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

- Adding an additional variable decreases both the SSR and n-k-1
- Adding a variable increases $\bar{R}^2$ if |t-stat| > 1
- Adding a group of variables increases $\bar{R}^2$ if the F-stat > 1
- Can have a negative $\bar{R}^2$

Relationship between R^2 and $\bar{R}^2$

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right)$$

Comparing Non-nested Models

- Recall that models are non-nested if neither model is a special case of the other

$$y = \beta_0 + \beta_1 x_1 + u; \qquad R^2 = 0.06, \bar{R}^2 = 0.030 \tag{2}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u; \qquad R^2 = 0.14, \bar{R}^2 = 0.090 \tag{3}$$

- Which model is “better”?

Comapring with different dependent variables

- R^2 or $\bar{R}^2$ must not be used to compare models which differ in their definition of the dependent variable

$$\widehat{salary} = \beta_0 + \beta_1 sales + \beta_2 roe + u \tag{4}$$

$$\boxed{log\widehat{(salary)}} = \beta_0 + \beta_1 sales + \beta_2 roe + u \tag{5}$$

- **There is inherently less variation in log(salary) that needs to be explained than in salary.**

Model Selection

It is possible to compare the models in (4) and (5) if we were to convert the predictions of (5) back to levels.

1. Run the regression in (5) and obtain the predicted values $\widehat{log(salary)}$.
2. Take exponential to convert the predictions back to levels, $m = \exp(\widehat{log(salary)})$.
3. Regress $salary$ on m and compare the R^2 values. You can drop the intercept in this regression.

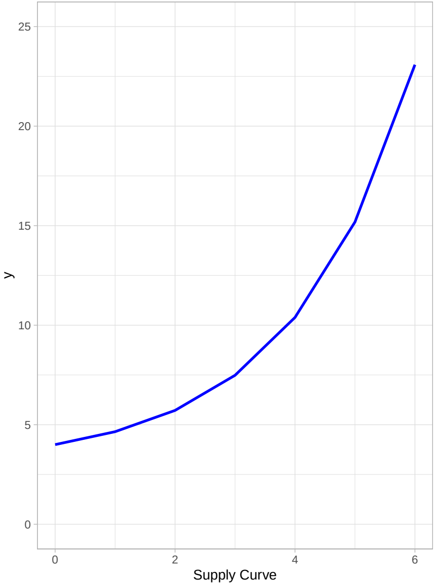
Controlling for too many factors

- In some cases, certain variables should not be held fixed.
  - In a regression of traffic fatalities on state beer taxes (and other factors) one should not directly control for beer consumption.
  - In a regression of family health expenditures on pesticide usage among farmers one should not control for doctor visits.
- Different regressions may serve different purposes.
  - In a regression of house prices on house characteristics, one would only include price assessments if the purpose of the regression is to study their validity; otherwise one would not include them.

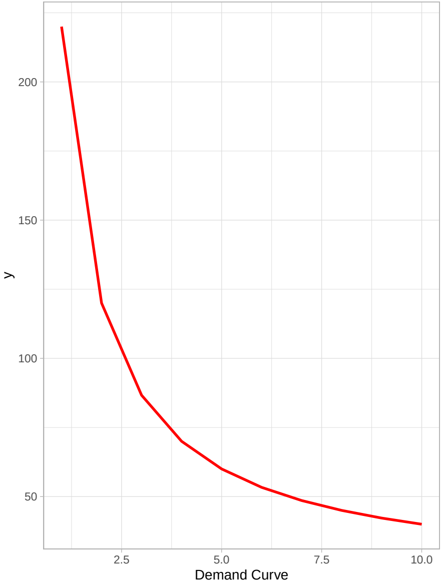
Section 4

Appendix: Functional Forms

Nonlinear Supply and Demand

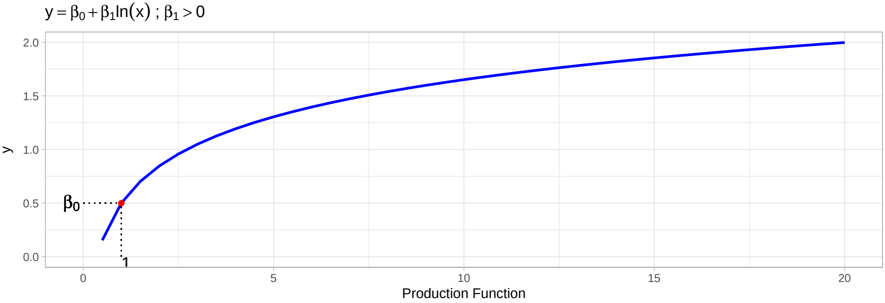


$\ln(y) = \beta_0 + \beta_1 x$

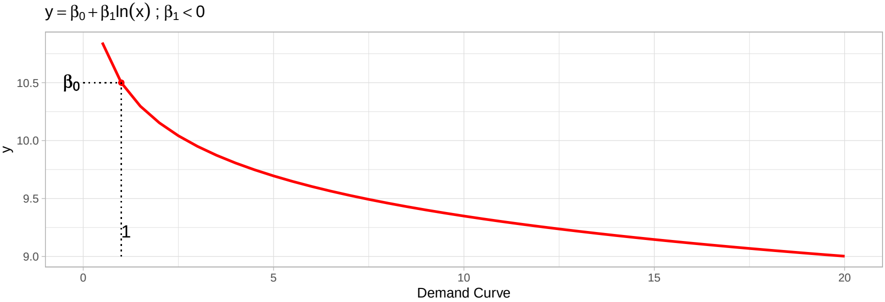


$y = \beta_0 + \beta_1 (1/x)$

Logarithmic Functions



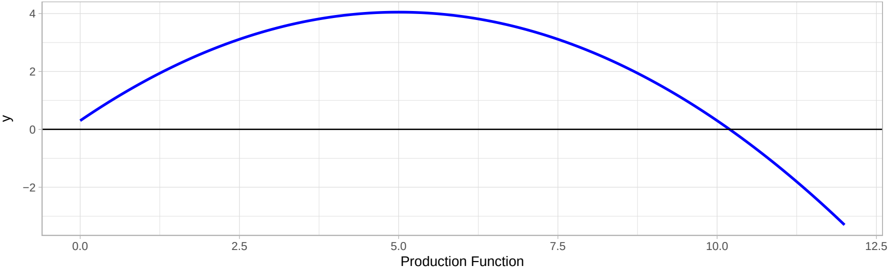
$y = \beta_0 + \beta_1 \ln(x)$; $\beta_1 > 0$



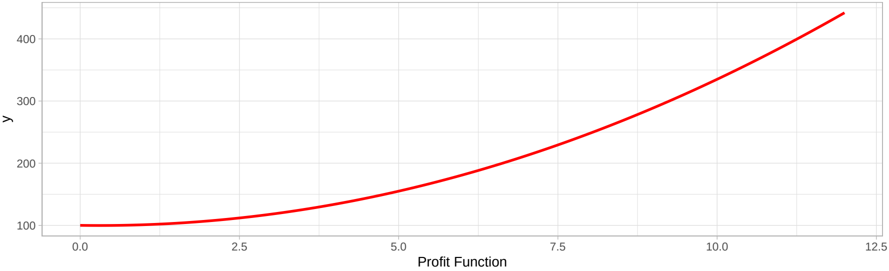
$y = \beta_0 + \beta_1 \ln(x)$; $\beta_1 < 0$

Quadratic Functions

$y = \beta_0 + \beta_1 x + \beta_2 x^2$; $\beta_2 < 0$, $\beta_1 > 0$



$y = \beta_0 + \beta_1 x + \beta_2 x^2$; $\beta_1 < 0$, $\beta_2 > 0$


y
Profit Function