Simple Linear Regression Notes

Session 2 Supplement

The Simple Linear Regression Model (SLRM) is used to estimate a population conditional expectation $E(Y \mid X)$ using a sample $\{X_i, Y_i\}_{i=1}^n$ from that population. Reasons for doing so include prediction, hypothesis testing, and quantifying causal effects. The SLRM will in many ways be inadequate for these purposes, but it is a good place to start.

The R code in these notes use the following libraries:

```
library(tidyverse) # For data manipulation
library(patchwork) # For composing graphics
library(latex2exp) # For annotating graphs with mathematics
```

The Simple Linear Regression Model

The SLRM assumes that the conditional expectation takes the form:

$$E(Y \mid X) = \beta_0 + \beta_1 X \,.$$

If we define $\epsilon = Y - \beta_0 - \beta_1 X$, then we can write

$$Y = \beta_0 + \beta_1 X + \epsilon \,.$$

Furthermore, we have

$$E(\epsilon \mid X) = E(Y - \beta_0 - \beta_1 X) = E(Y \mid X) - \beta_0 - \beta_1 X = 0 \,.$$

That is,

$$Y = \beta_0 + \beta_1 X + \epsilon \,,\ E(\epsilon \mid X) = 0 \,.$$

If your sample is a representative i.i.d. sample from the population (we assume this is the case), then it satisfies

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \,,\ E(\epsilon_i \mid X_i) = 0 \,.$$

The "independent part" of the iid assumptions allows us extend this further to

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \,,\ E(\epsilon_i \mid X_1, X_2, \ldots, X_n) = 0 \,.$$

The objective is to use the sample to estimate β_0 and β_1, which gives us the estimated conditional expectation

$$\widehat{E(Y \mid X)} = \hat{\beta}_0 + \hat{\beta}_1 X \,.$$

This is often written as $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Terminology:

- Y_i ~ "Regressand", "Dependent Variable", "Outcome Variable".
- X_i ~ "Regressor", "Independent Variable", "Predictor", "Feature".
- ϵ_i ~ "Noise" or "Error" term.
- β_1 is the slope coefficient or simply "coefficient" on X_i.
- β_0 is the (y-) intercept term or "constant" term.

In Machine Learning, β_0 is called the "bias". We will **not** use that terminology here.

For any potential estimator $\hat{\beta}_0$ and $\hat{\beta}_1$, define

- Fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_0 X_i$, $i = 1, 2, \dots, n$.
- Residuals: $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_0 X_i$, $i = 1, 2, \dots, n$.

The population conditional expectation $E(Y \mid X) = \beta_0 + \beta_1 X$ is also called the population regression function. The estimated conditional expectation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

is also called the sample regression function.

Example Suppose our data is as follows

```
df <- read_csv("data\\ols01.csv", show_col_types=F) # from tidyverse
glimpse(round(df,2))  # glimpse the data, rounded to 2 dec. places.
```

```
Rows: 10
Columns: 2
$ X <dbl> 2.51, 5.17, 1.73, 3.42, 4.03, 4.58, 8.19, 6.59, 8.72, 6.06
$ Y <dbl> 7.64, 10.67, 3.11, 1.85, 11.78, 10.58, 15.45, 9.63, 13.74, 11.80
```

This data is displayed as a scatterplot in Figure 1, with a potential fitted line, which would be our estimate of $E(Y \mid X)$. The residuals are marked out as dashed lines, with the (negative) residual for the fourth observation labelled. The hollow circles on the line are the points $(X_i, \hat{Y}_i)$.

How do we determine the fitted line, i.e., how do we choose $\hat{\beta}_0$ and $\hat{\beta}_1$?

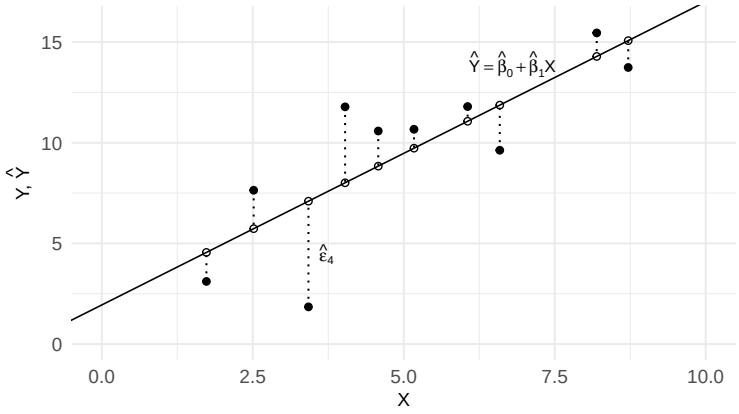


Figure 1

Estimating the Simple Linear Regression Model

The usual approach for estimating β_0 and β_1 is a method called Ordinary Least Squares, which proceeds by minimizing the sum of squared residuals. We begin our discussion, however, with an alternative approach (one that does not use calculus!).

The Law of Iterated Expectations tells us that the condition $E(\epsilon \mid X) = 0$ implies

$$E(\epsilon) = 0 \;\text{ and }\; Cov(\epsilon, X) = 0\,. \tag{1}$$

Furthermore, since $Cov(\epsilon, X) = E(\epsilon X) - E(\epsilon)E(X)$, we can write (1) as

$$E(\epsilon) = 0 \;\text{ and }\; E(\epsilon X) = 0\,. \tag{2}$$

The "Method of Moments" approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to mimic these two population conditions in sample, i.e., we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to satisfy the sample moment conditions:

$$\begin{aligned} \overline{\hat{\epsilon}^{mm}} &= \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^{mm} = 0 \\ \overline{\hat{\epsilon}^{mm} X} &= \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^{mm} X_i = 0 \end{aligned} \tag{3}$$

where

$$\hat{\epsilon}_i^{mm} = Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_i\,. \tag{4}$$

We use the mm superscript to highlight the fact that the estimates of β_0 and β_1 are obtained using the Method of Moments approach. Using a different approach may (or may not) result in different estimators and different residuals.

Substituting (4) into the equations in (3), and dropping the $1/n$, we have

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm}X_i) &= 0 \\ \sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm}X_i)X_i &= 0\end{aligned} \tag{5}$$

Distributing the summation in the first equation of (5) and dividing by n gives

$$\hat{\beta}_0^{mm} = \overline{Y} - \hat{\beta}_1^{mm}\overline{X}\,. \tag{6}$$

Substituting (6) into the second equation gives

$$\begin{aligned}&\sum_{i=1}^{n}(Y_i - (\overline{Y} - \hat{\beta}_1^{mm}\overline{X}) - \hat{\beta}_1^{mm}X_i)X_i = 0 \\ \implies \quad &\sum_{i=1}^{n}(Y_i - \overline{Y})X_i - \hat{\beta}_1^{mm}\sum_{i=1}^{n}(X_i - \overline{X})X_i = 0\,.\end{aligned}$$

Solving for $\hat{\beta}_1^{mm}$ gives

$$\hat{\beta}_1^{mm} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})X_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{sample cov}(Y_i, X_i)}{\text{sample var}(X_i)}. \tag{7}$$

The estimated regression function is then

$$\hat{Y} = \hat{\beta}_0^{mm} + \hat{\beta}_1^{mm}X$$

where $\hat{\beta}_0^{mm}$ and $\hat{\beta}_1^{mm}$ given by (6) and (7) respectively. For our data set, we have

```
beta1mm <- cov(df$Y, df$X)/var(df$X)
beta0mm <- mean(df$Y) - beta1mm * mean(df$X)
cat(paste0("Estimated Model is Yhat = ", round(beta0mm,3), " + ", round(beta1mm,3), "X"))
```

```
Estimated Model is Yhat = 1.943 + 1.506X
```

It can be shown that if $E(Y \mid X) = \beta_0 + \beta_1 X$, which implies $E(\epsilon \mid X) = 0$, then[1]

$$\beta_0 = E(Y) - \beta_1 E(X) \quad \text{and} \quad \beta_1 = \frac{Cov(X,Y)}{Var(X)}\,. \tag{8}$$

Since our estimators are obtained from sample equations (3) that mimic the population moments (2), also implied by $E(\epsilon \mid X)$, it is not surprising that the MM estimators also mimic the equations in (8).

[1]We have $E(Y) = E(E(Y|X)) = E(\beta_0 + \beta_1 X) = \beta_0 + \beta_1 E(X)$ which gives the first equality. We also have

$$E(YX) = E(E(YX \mid X)) = E(XE(Y \mid X)) = E(X(\beta_0 + \beta_1 X)) = \beta_0 E(X) + \beta_1 E(X^2)$$

Substituting in $\beta_0 = E(Y) - \beta_1 E(X)$ into this equation and solving gives the second equality.

Another approach, called "Ordinary Least Squares" (OLS) is to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of squared residuals

$$SSR = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

That is,

$$\text{OLS}: \quad \hat{\beta}_0^{ols}, \hat{\beta}_1^{ols} = \underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \,. \tag{9}$$

It can be shown that SSR is a convex function in $\hat{\beta}_0$ and $\hat{\beta}_1$ (we omit details of the argument here). Basic optimization theory then tells us that we can minimize SSR by choosing $\hat{\beta}_0^{ols}$ and $\hat{\beta}_1^{ols}$ that solves the first-order conditions[2]

$$\begin{aligned} (1) \quad & \left.\frac{\partial SSR}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}} &= -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i) = 0 \\ (2) \quad & \left.\frac{\partial SSR}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}} &= -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i) X_i = 0 \end{aligned} \tag{10}$$

Dropping the inconsequential -2 from the equations in (10), we find that the OLS first-order conditions are exactly the same equations as the sample moment conditions (3) that we solved using the Method of Moments approach. That is, the OLS estimators are exactly the same as the MM estimators.

$$\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols} \overline{X}$$

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y}) X_i}{\sum_{i=1}^{n} (X_i - \overline{X}) X_i} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Moving forward, we will refer to these estimators as OLS estimators rather than Method-of-Moments estimators. Likewise, the residuals and fitted values will be called the OLS residuals and OLS fitted values:

- OLS fitted values: $\hat{Y}_i^{ols} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i$, $i = 1, \ldots, n$
- OLS residuals: $\hat{\epsilon}_i^{ols} = Y_i - \hat{Y}_i^{ols} = Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i$, $i = 1, \ldots, n$

We will discuss OLS estimator standard errors and other associated statistics in the next class. The Method-of-Moments approach will be extended when we discuss estimation using instrumental variables (Session 8).

[2] If you are having trouble seeing how the differentiation was done, you may find it helpful to write the SSR out in full before differentiating, i.e.,

$$SSR = (Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1)^2 + \cdots + (Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n)^2 \,.$$

You may find this expression easier to work with. After partially differentiating with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$, collapse the expression back using the summation notation.

Unbiasedness and Consistency of OLS estimators

The OLS estimators turn out to be unbiased estimators. We focus on $\hat{\beta}_1^{ols}$ and show unbiasedness of $\hat{\beta}_0^{ols}$ later when we treat the general case. First rewrite $\hat{\beta}_1^{ols}$ as

$$\begin{aligned}\hat{\beta}_1^{ols} &= \frac{\sum_{i=1}^n (Y_i - \overline{Y})X_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\ &= \frac{\beta_0 \sum_{i=1}^n (X_i - \overline{X}) + \beta_1 \sum_{i=1}^n (X_i - \overline{X})X_i + \sum_{i=1}^n (X_i - \overline{X})\epsilon_i)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\ &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\sum_{i=1}^n (X_i - \overline{X})X_i}\end{aligned}$$

Then

$$E(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) = \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})E(\epsilon_i \mid X_1, \ldots, X_n)}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \beta_1$$

since $E(\epsilon_i \mid X_1, \ldots, X_n) = 0$. Furthermore, $E(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) = \beta_1$ implies $E(\hat{\beta}_1^{ols}) = \beta_1$ so $\hat{\beta}_1^{ols}$ is unbiased.

Note that the key condition that gives us unbiasedness is $E(\epsilon_i \mid X_1, \ldots, X_n) = 0$. If this condition does not hold, then we will not have unbiasedness.

The OLS estimators are also consistent. Focusing again on β_1, we have

$$\hat{\beta}_1^{ols} = \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}\,.$$

The denominator of the second term converges in probability to the population variance of X_i. The numerator is the sample covariance of X_i and ϵ_i (the latter is not observed, but that doesn't matter). Since $X_i\epsilon_i$ is iid and $E(X\epsilon) = 0$ under our assumptions, the Law of Large Numbers guarantees that it converges in probability to zero. i.e.,

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_1^{ols} = \operatorname*{plim}_{n\to\infty} \beta_1 + \frac{\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} = \beta_1 + \frac{0}{var(X)} = \beta_1\,.$$

Alternatively, we can note that

$$\hat{\beta}_1^{ols} = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\text{sample cov}(X_i, Y_i)}{\text{sample var}(X_i)} \xrightarrow{p} \frac{Cov(Y, X)}{Var(X)} = \beta_1$$

where the last equality holds if

$$E(Y \mid X) = \beta_0 + \beta_1 X\,. \tag{11}$$

The fact that OLS estimators satisfy equations (10) means that the sample covariance of the OLS residuals and the regressors will always be zero, i.e.,

$$\text{sample cov.}(\hat{\epsilon}_i^{ols}, X_i) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i^{ols} - \bar{\hat{\epsilon}})(X_i - \overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^{ols} X_i = 0$$

(why does the second equality hold?) This has some important implications. First note that

$$Y_i = \hat{Y}_i^{ols} + \hat{\epsilon}_i^{ols}\,,\; i = 1, \ldots, n\,.$$

Since $\hat{Y}_i^{ols} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i$, it is perfectly correlated with X_i. Since X_i is uncorrelated with $\hat{\epsilon}_i^{ols}$, $\hat{Y}_i^{ols}$ is uncorrelated with $\hat{\epsilon}_i^{ols}$. Here "uncorrelated" means that the sample covariance (and hence the sample correlation) is zero. In other words, we have "decomposed" Y_i into two perfectly uncorrelated parts $\hat{Y}_i^{ols}$ and $\hat{\epsilon}_i^{ols}$. One of these parts ($\hat{Y}_i^{ols}$) is perfectly correlated with the regressor X_i, the other part ($\hat{\epsilon}_i^{ols}$) perfectly *uncorrelated* with X_i.

Since $\hat{Y}_i^{ols}$ and $\hat{\epsilon}_i^{ols}$ are uncorrelated, the sample variance of the sum of the two is the sum of their sample variances. That is,

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i^{ols} - \overline{\hat{Y}})^2 + \sum_{i=1}^{n}\hat{\epsilon}_{i,ols}^2\,. \tag{12}$$

We can skip the division by n since that cancels out. Equation (12) is usually stated as "Sum of Squared Total = Sum of Squared Explained + Sum of Squared Residuals", or

$$SST = SSE + SSR\,.$$

Furthermore, we can use this to define a measure of goodness-of-fit:

$$R^2 = 1 - \frac{SSR}{SST} \tag{13}$$

which of course lies between zero and one. If $R^2 = 1$, then it must be that $SSR = 0$, which means that the data points all lie on a straight line, and you have a perfect fit. If $R^2 = 0$, then it must be that $SSR = SST$, i.e.,

$$\frac{\sum_{i=1}^{n}\hat{\epsilon}_{i,ols}^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{ols})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1\,.$$

This will be the case when $\hat{Y}_i^{ols} = \overline{Y}$ for all i, which happens when $\hat{\beta}_1^{ols} = 0$. All other intermediate levels of fit result in $0 < R^2 < 1$.

Of course, we can also write $R^2 = SSE/SST$, which shows that the R^2 shows the proportion of the variation in Y_i that is accounted for (some say "explained") by the regressor X_i. It is conventional, however, to define the R^2 as in (13). It can also be shown that the R^2 is the square of the sample correlation between Y_i and $\hat{Y}_i^{ols}$, which is where it gets its name.

We have assumed regression with an intercept term throughout. In regressions without an intercept term, (12) doesn't hold necessarily, and the R^2, as defined in (13), can fall below zero.

Alternative Specifications

Although the SLRM assumes (11), there is actually considerable flexibility. The SLRM only assumes linearity-in-parameters, not linearity-in-variables. Possible specifications include:

- $E(\ln Y \mid X) = \beta_0 + \beta_1 \ln X$
- $E(\ln Y \mid X) = \beta_0 + \beta_1 X$
- $E(Y \mid X) = \beta_0 + \beta_1 \ln X$
- $E(Y \mid X) = \beta_0 + \beta_1 X$
- $E(Y \mid X) = \beta_0 + \beta_1 X^2$

and many more. Which specification is appropriate depends on a number of factors, including which fit the data best, and which conform best with prior economic considerations. For instance, if Y is earnings and X is educ, then

$$earn = \beta_0 + \beta_1 educ + \epsilon \tag{14}$$

says that, holding all factors in ϵ fixed, an additional 1 year of education is associated with a increase in earnings of β dollars, and *that this is true at all levels of educ*, which seems unlikely to be true. On the other hand, if you assume that

$$\ln earn = \beta_0 + \beta_1 educ + \epsilon\,, \tag{15}$$

then a plus one year difference in *educ* is associated with a $100\beta_1$ percent increase in wages, that that this percentage difference is the same regardless of education levels. This seems more plausible (or is it?)

Example Consider the data in `earnings2019.csv`

```
dat <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)  # read_csv from tidyverse
head(dat, 3)  # show first three rows
```

```
# A tibble: 3 x 11
    age height  educ feduc meduc tenure  wexp race   male  earn totalwork
  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <chr> <dbl> <dbl>     <dbl>
1    59     67    12     3     3      5    30 White     0 36.3       1652
2    43     63    10     4     3      7    13 White     1  6.46      1548
3    28     74    12     2     3      6     9 White     1 13.1       2460
```

The table was imported as a `tidyverse tibble` object (a kind of dataframe). The base `R` function `head()` is used to show the first three rows. Another way to view the data is to use the `tidyverse` library's `glimpse()` function.

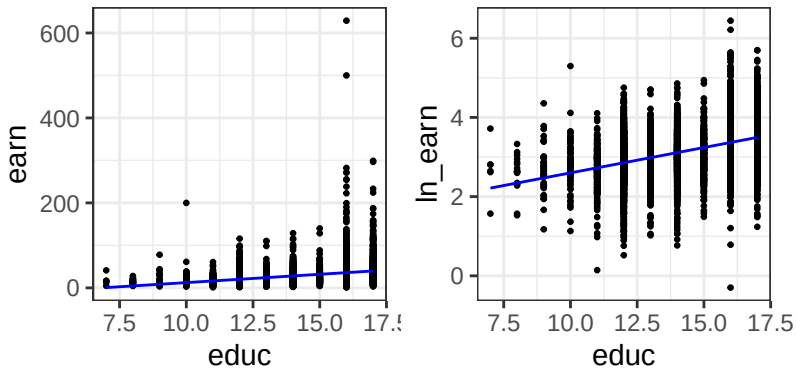
```
glimpse(dat)
```

```
Rows: 4,946
Columns: 11
$ age       <dbl> 59, 43, 28, 66, 63, 42, 64, 38, 37, 44, 38, 27, 26, 24, 55, ~
$ height    <dbl> 67, 63, 74, 66, 61, 70, 75, 73, 75, 64, 65, 73, 75, 67, 65, ~
$ educ      <dbl> 12, 10, 12, 16, 10, 12, 17, 17, 17, 11, 12, 10, 12, 12, 14, ~
$ feduc     <dbl> 3, 4, 2, 4, 2, 4, 2, 8, 8, 4, 3, 3, 4, 4, 7, 4, 4, 3, 4, 3, ~
$ meduc     <dbl> 3, 3, 3, 4, 2, 5, 4, 7, 7, 3, 3, 3, 2, 4, 6, 6, 4, 6, 2, 3, ~
$ tenure    <dbl> 5, 7, 6, 3, 24, 11, 28, 1, 2, 5, 2, 4, 6, 4, 1, 2, 2, 6, 13,~
$ wexp      <dbl> 30, 13, 9, 46, 38, 19, 4, 6, 14, 15, 2, 2, 1, 2, 9, 4, 19, 6~
$ race      <chr> "White", "White", "White", "White", "White", "White", "White~
$ male      <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, ~
$ earn      <dbl> 36.319613, 6.459948, 13.100000, 45.098039, 12.890625, 19.230~
$ totalwork <dbl> 1652, 1548, 2460, 2040, 3328, 2600, 3750, 2340, 3084, 2080, ~
```

There are altogether 4946 obs. You can also use `summary(dat)` to get a summary of each variable in the data set. This is left for you to explore.

We fit both specifications (14) and (15) and plot the fitted lines over the scatterplots. In the code below, we estimate the regressions using the `lm()` function from base R:

```
mdl1 <- lm(earn ~ educ, data=dat) # earn ~ educ - 1 for regression without intercept term
dat1 <- tibble(earn = dat$earn,
               educ = dat$educ,
               yhat = fitted(mdl1),
               ehat = residuals(mdl1))
mdl2 <- lm(log(earn) ~ educ, data=dat)
dat2 <- tibble(ln_earn = log(dat$earn),
               educ = dat$educ,
               yhat = fitted(mdl2),
               ehat = residuals(mdl2))
p1 <- ggplot(data=dat1) + geom_point(aes(x=educ, y=earn), size=0.5) +
      geom_line(aes(x=educ, y=yhat), color="blue") + theme_bw() + theme(aspect.ratio=1)
p2 <- ggplot(data=dat2) + geom_point(aes(x=educ, y=ln_earn), size=0.5) +
      geom_line(aes(x=educ, y=yhat), color="blue") + theme_bw() + theme(aspect.ratio=1)
p1 | p2
```



Specification (15) seems to work better than specification (14). The fitted model under specification (15) is

```
cat("ln_earn_hat = ", round(coef(mdl2)[1],3), " + ", round(coef(mdl2)[2], 3), "educ")
```

```
ln_earn_hat =  1.32  +  0.128 educ
```

Each one-year increase in *educ* is associated with a 12.8 percent increase in hourly earnings. This estimate seems plausible, and appears to fit the data well, except perhaps for lower levels of *educ*.

You can get a fuller report on the regression by using the `summary()` function:

```
summary(mdl2)
```

```
Call:
lm(formula = log(earn) ~ educ, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6653 -0.3722  0.0037  0.3695  3.0761

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.319989   0.057541   22.94   <2e-16 ***
educ        0.127996   0.003979   32.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5936 on 4944 degrees of freedom
Multiple R-squared:  0.1731,    Adjusted R-squared:  0.1729
F-statistic:  1035 on 1 and 4944 DF,  p-value: < 2.2e-16
```

The output contains quite a few statistics that we will discuss later. If you want to skip the details on the residuals and report only the main estimates and R^2, you can do the following:

```
summary(mdl2)$call
cat("\n")
summary(mdl2)$coefficients %>% round(4)
cat("\n")
cat("R-squared:", round(summary(mdl2)$r.squared, 4))
```

```
lm(formula = log(earn) ~ educ, data = dat)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.320     0.0575 22.9399        0
educ           0.128     0.0040 32.1700        0

R-squared: 0.1731
```

Causal Interpretations and Extensions

The fact that

$$\beta_1 = \frac{Cov(X,Y)}{Var(X))} \text{ and } \hat{\beta}_1^{ols} = \frac{\text{sample cov}(X_i, Y_i)}{\text{sample var}(X_i)}$$

shows you that what you have estimated is correlation, and of course, correlation does not imply causality. We have merely estimated a predictive relationship. We will discuss in class several situations where correlation is a misleading indicator of causality, but as a quick example, suppose there are two variables X and Z that have direct causal effects on Y, and that all other variation in Y not due to X and Z is pure "noise". In particular, suppose

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u\,.$$

Since u is pure random noise, we have $E(u \mid X) = 0$. If we subsume $\beta_2 Z$ and u into a composite term ϵ, then we have

$$Y = \beta_0 + \beta_1 X + \epsilon \text{ where } \epsilon = \beta_2 Z + u\,.$$

If Z and X happen to be perfectly uncorrelated, then $E(Z \mid X) = 0$, and we have

$$E(\epsilon \mid X) = E(\beta_2 Z + u \mid X) = \beta_2 E(Z \mid X) + E(u \mid X) = 0\,.$$

In this case, we still able to obtain an unbiased estimate of β_1 from a regression of Y on X. However, if Z and X are correlated, then $E(Z \mid X) \neq 0$, and $E(\epsilon \mid X) \neq 0$, and we would not get unbiased estimates of β_1.

Note that in this example, it is still possible for the conditional expectation $E(Y \mid X)$ to be linear even if Z and X are correlated. That is, we could still have

$$E(Y \mid X) = \alpha_0 + \alpha_1 X\,,$$

but in this case α_1 will not be equal to β_1, and a simple linear regression of Y_i on X_i will give you unbiased estimates of α_1, not β_1.

In order for our estimates to reflect a causal effect of X on Y, we will either have to sample our data in such a way such that the error term in the regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ is uncorrelated with the regressor (such as in randomized controlled trials), or we have to control for all other factors affecting Y that are correlated with X. The latter can be done by estimating multiple linear regression models such as

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i\,.$$

Coming sessions will discuss the details of this approach. Note that even if X and Z were not correlated, so that a simple linear regression of Y on X gives you an unbiased estimate of the causal effect of X on Y, it is often still useful to include Z into the regression, because doing so will often reduce the standard error on your estimates.

There are some situations where multiple linear regression is not feasible, or cannot solve the underlying problem causing biased estimation of causal relationships. In these cases we have to turn to other techniques. We will discuss some of these situations in class.

Even if we are interested only in estimating predictive relationships, it is often still necessary to use the multiple linear regression framework. For instance, in the earnings example we may want to model the conditional expectation $E(\ln earn \mid educ)$ as

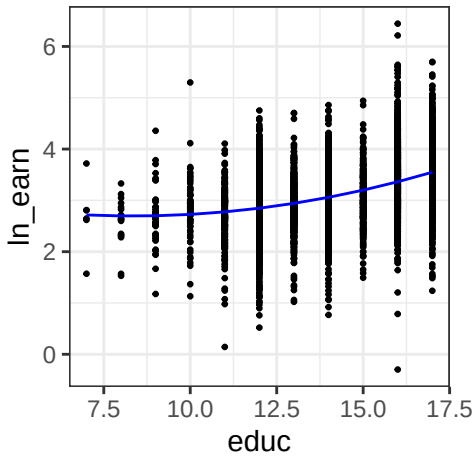
$$E(\ln earn \mid educ) = \beta_0 + \beta_1 educ + \beta_2 educ^2 .$$

In this case we would estimate the multiple linear regression model

$$\ln earn_i = \beta_0 + \beta_1 educ_i + \beta_2 educ_i^2 + \epsilon_i$$

which requires using the multiple linear regression framework. This gives us the following fit:

```
mdl3 <- lm(log(earn) ~ educ + I(educ^2), data=dat)
dat3 <- tibble(ln_earn = log(dat$earn),
               educ = dat$educ,
               yhat = fitted(mdl3),
               ehat = residuals(mdl3))
p3 <- ggplot(data=dat3) + geom_point(aes(x=educ, y=ln_earn), size=0.5) +
      geom_line(aes(x=educ, y=yhat), color="blue") + theme_bw() + theme(aspect.ratio=1)
p3
```



This should give better predictions at lower levels of `educ`. In this example, we would expect that the prediction errors will be quite large, since `educ` only explains a small proportion of the variation in `ln earn`. You can try yet more flexible functional forms, though in this case it seems unlikely you will be able to improve on the fit by doing so. Gains are more likely to come from including yet other predictors, perhaps

$$\ln earn_i = \beta_0 + \beta_1 educ_i + \beta_2 educ_i^2 + \beta_3 male_i + \beta_4 age_i + \epsilon_i .$$

More details on predictive regressions in Session 7.

Appendix: A Bit of Optimization Theory

We will only need to deal with the simplest cases. Suppose we have a function $f(x)$, and want to find x^* such that $f(x^*)$ is maximized or minimized. We can use the following results:

- If $f''(x) > 0$ for all x (function is convex), then $f'(x^*) = 0 \Rightarrow x^*$ minimizes $f(x)$
- If $f''(x) < 0$ for all x (function is concave), then $f'(x^*) = 0 \Rightarrow x^*$ maximizes $f(x)$

See Figure 2(a) and (b). These conditions are usually stated as "first order conditions" and "second order conditions".

E.g., Find the minimum of $f(x) = e^{2(x-1)} - 1$. Since $f'(x) = 2e^{2(x-1)}$ and $f''(x) = 4e^{2(x-1)}$:

FOC: $f'(x^*) = 2e^{2(x^*-1)} = 0$, so $x^* = 1$ is a candidate minimum point.

SOC: Since $f''(x) = 4e^{2(x-1)} > 0$ for all x, $x^* = 1$ is a global minimum point.

Note that the conditions stated above are sufficient, not necessary. See Figure 2(c) for an example where the function is neither fully concave or fully convex, yet it has a global maximum, a global minimum, and a local maximum.

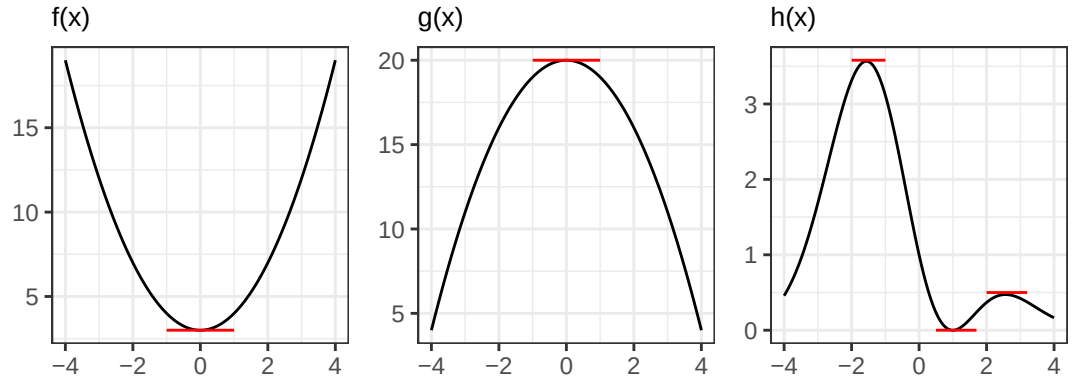


Figure 2: Three optimization examples.

The same ideas applies to functions of two variables $f(x, y)$. Stationary points of such functions are points (x^*, y^*) such that $f'_x(x^*, y^*) = 0$ and $f'_y(x^*, y^*) = 0$. If f concave, then the stationary point is the maximum point. If f is convex, then the stationary point is the minimum point. The following conditions can be used to check for concavity / convexity of functions of two variables:

- If $v_1^2 \dfrac{\partial^2 f(x,y)}{\partial x^2} + 2v_1 v_2 \dfrac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \dfrac{\partial^2 f(x,y)}{\partial y^2} < 0$ for all v_1, v_2 not both zero

then $f(x, y)$ is concave

- If $v_1^2 \dfrac{\partial^2 f(x,y)}{\partial x^2} + 2v_1 v_2 \dfrac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \dfrac{\partial^2 f(x,y)}{\partial y^2} > 0$ for all v_1, v_2 not both zero

then $f(x,y)$ is convex.

A rough explanation is as follows. Let $x = x_0 + v_1 s$, $y = y_0 + v_2 s$, $v_1^2 + v_2^2 = 1$, and

$$z(s) = f(x(s), y(s))$$

Note that $z(0) = f(x_0, y_0)$, $dx/ds = v_1$, $dy/dx = v_2$. The "Directional Derivative" at (x_0, y_0) in direction $v = (v_1, v_2)$ is then

$$\frac{dz}{ds} = f'_x(x,y)\frac{dx}{ds} + f'_y(x,y)\frac{dy}{ds} = v_1 f'_x(x,y) + v_2 f'_y(x,y)$$

This is a "directional derivative" since a one-unit increase in s leads to a increase in x by v_1 and an increase in y by v_2, so (x,y) moves in the direction (v_1, v_2). We are asking what happens to the function when we move in that direction. For a point (x_0, y_0) to be a minimum or maximum point, the slope in all directions must be zero. This is guaranteed by

$$f'_x(x_0, y_0) = 0, f'_x(x_0, y_0) = 0\,.$$

The second directional derivative is

$$\begin{aligned}
\frac{d^2z}{ds^2} &= f'_x(x,y)\frac{dx}{ds} + f'_y(x,y)\frac{dy}{ds} \\
&= \left[f''_{xx}(x,y)\frac{dx}{ds} + f''_{xy}(x,y)\frac{dy}{ds}\right]\frac{dx}{ds} + f'_x(x,y)\frac{d^2x}{ds^2} + \\
&\qquad \left[f''_{yx}(x,y)\frac{dx}{ds} + f''_{yy}(x,y)\frac{dy}{ds}\right]\frac{dy}{ds} + f'_y(x,y)\frac{d^2y}{ds^2} \\
&= v_1^2 f''_{xx}(x,y) + 2v_1v_2 f''_{xy}(x,y) + v_2^2 f''_{yy}(x,y)
\end{aligned}$$

since $dx/ds = v_1$, $dy/ds = v_2$, and $d^2x/ds^2 = d^2y/ds^2 = 0$.

The function f is convex if its slope is always increasing in all directions, i.e., $z''(0) > 0$ for all v and for all (x_0, y_0). It is concave if its slope is always decreasing in all directions, i.e., $z''(0) < 0$ for all v and for all (x_0, y_0).

Example: Find minimum point of $f(x,y) = x^2 + xy + y^2$.

We have $f'_x(x,y) = 2x + y$ and $f'_y(x,y) = y + 2y$

Therefore

$$\text{FOC:} \quad \begin{aligned} f'_x(x^*, y^*) &= 2x^* + y^* = 0 \\ f'_y(x^*, y^*) &= 2y^* + x^* = 0 \end{aligned} \quad \Rightarrow \quad (x^*, y^*) = (0,0) \text{ stationary point}$$

SOC: We have $f''_{xx}(x,y) = 2$, $f''_{xy}(x,y) = f''_{yx}(x,y) = 1$ and $f''_{yy}(x,y) = 2$, therefore

$$\begin{aligned}
& v_1^2 f''_{xx}(x,y) + 2v_1v_2 f''_{xy}(x,y) + v_2^2 f''_{yy}(x,y) \\
&= 2(v_1^2 + v_1v_2 + v_2^2) \\
&= 2[(v_1 + 0.5v_2)^2 + 0.75v_2^2] > 0
\end{aligned}$$

for all v_1, v_2 not both equal to zero, i.e., the function $f(x,y)$ is convex.

Therefore $(x^*, y^*) = (0,0)$ is a minimum point of $f(x,y) = x^2 + xy + y^2$.