Session 3

Multiple Linear Regression

Anthony Tay

This Version: 02 Sep 2025

Session 3

- Intro to **Multiple Linear Regression (MLR)** and **OLS estimation** of the MLR model
  - Solving omitted variable problem
  - Flexibility in functional specification

Recap OLS and SLR

Population with $E(Y \mid X) = \beta_0 + \beta_1 X$

Define $\epsilon = Y - \beta_0 - \beta_1 X$

We can write down

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

$$E(\epsilon \mid X) = 0 \Longrightarrow E(\epsilon) = 0 \;\text{ and }\; Cov(\epsilon, X) = E(X\epsilon) = 0 \,,$$

$$\beta_0 = E(Y) - \beta_1 E(X) \;\text{ and }\; \beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

Recap OLS and SLR

Representative iid sample $\{X_i, Y_i\}_{i=1}^n$ iid sample from population

Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$E(\epsilon_i \mid X_1, \ldots, X_n) = 0$$
$$E(\epsilon_i \epsilon_j \mid X_1, \ldots, X_n) = 0 \text{ for } i, j = 1, \ldots, n, \ i \neq j$$

Sometimes also

$$E(\epsilon_i^2 \mid X_1, \ldots, X_n) = \sigma^2$$

Recap OLS and SLR

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

- $\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X}$
- $\hat{\beta}_1^{ols} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Recap OLS and SLR

Different ways of writing $\hat{\beta}_1^{ols}$, including:

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

- Unbiased: $E(\hat{\beta}_1^{ols}) = \beta_1$ and consistent: $\hat{\beta}_1^{ols} \xrightarrow{p} \beta_1$
- Also: $\hat{\beta}_1^{ols}$ is a "linear estimator"

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

Recap OLS and SLR

FOC can be written as

$$\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = 0 \quad \text{and} \quad \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} X_i = 0 \quad \text{where} \quad \hat{\epsilon}_i^{ols} = Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i$$

- $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = 0 \Rightarrow \overline{\hat{\epsilon}^{ols}} = 0$
- $\overline{\hat{\epsilon}^{ols}} = 0$ and $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} X_i = 0 \implies \textit{Sample Cov.}(X_i, \hat{\epsilon}_i^{ols}) = 0$
- $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} X_i = 0 \equiv \hat{\epsilon}_i^{ols}$ and X_i are "**orthogonal**"

* I will use both $\hat{\epsilon}_i^{ols}$ and $\hat{\epsilon}_{i,ols}$ to denote OLS residuals

Recap OLS and SLR

- The fitted values $\hat{Y}_i^{ols} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i$ and the residuals $\hat{\epsilon}_i^{ols}$ are orthogonal

$$\sum_{i=1}^{n} \hat{Y}_i^{ols} \hat{\epsilon}_i^{ols} = \sum_{i=1}^{n} (\hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i) \hat{\epsilon}_i^{ols} = \hat{\beta}_0^{ols} \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} + \hat{\beta}_1^{ols} \sum_{i=1}^{n} X_i \hat{\epsilon}_i^{ols} = 0$$

- Sample covariance of $\hat{Y}_i^{ols}$ and $\hat{\epsilon}_i^{ols}$ is zero
- Simple regression of Y_i on X_i, with intercept, breaks Y_i into two uncorrelated parts

$$Y_i = \hat{Y}_i^{ols} + \hat{\epsilon}_i^{ols}$$

$\hat{Y}_i^{ols}$ perfectly correlated with X_i, $\hat{\epsilon}_i^{ols}$ perfectly uncorrelated with X_i

Recap OLS and SLR

- Variance Decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_{i,ols}-\overline{\widehat{Y}}_{ols})^2}_{FSS} + \underbrace{\sum_{i=1}^{n}\hat{\epsilon}_{i,ols}^2}_{RSS}. \quad (1)$$

- Measure of goodness-of-fit

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

Motivating MLR

Two issues with simple linear regression:

Omitted variables:

- Suppose X and Z affect Y

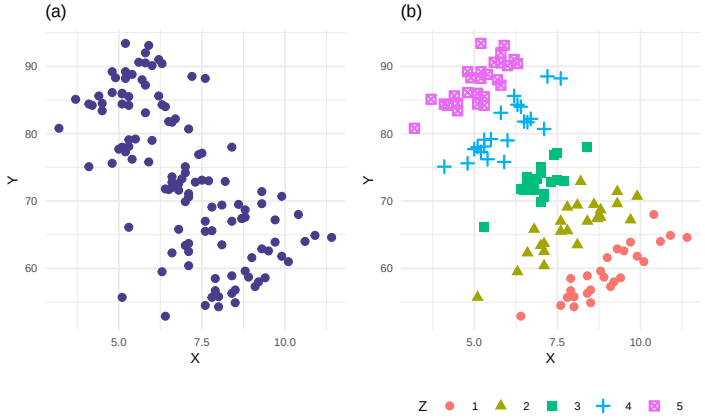
$$E(Y \mid X) = \alpha_0 + \alpha_1 X \text{ vs } E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

$$\alpha_1 = \beta_1 + \beta_2 \frac{Cov(X, Z)}{Var(Z)}$$

- OLS estimation of regression of Y on X gives you unbiased estimates of $E(Y \mid X)$
- Unless $Cov(X, Z) = 0$ or $\beta_2 = 0$, we have $\alpha_1 \neq \beta_1$

Motivating MLR

Data in `multireg_eg.csv`



Motivating MLR

Example illustrated by previous figure:

- $Y \sim$ final exam scores in a certain course;
- $Z \sim$ background preparedness of the students (1) very poor, to (5) excellent
- $X \sim$ study hours per week

Studying reduces exam score?

Background preparedness of students is a confounding factor

Motivating MLR

```
mdl1_sum <- lm(Y~X, data=df) %>% summary()
mdl1_sum %>% coef()
cat("R-squared:", round(mdl1_sum$r.squared,3))
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 102.86406  3.2199095 31.94626 6.373574e-60
X            -4.23213  0.4474755 -9.45779 3.857133e-16
R-squared: 0.431
```

```
mdl2_sum <- lm(Y~X+Z, data=df) %>% summary()
mdl2_sum %>% coef()
cat("R-squared:", round(mdl2_sum$r.squared,3))
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 21.188337  2.0816539 10.17861 8.212834e-18
X            3.114465  0.2054364 15.16024 2.219609e-29
Z           10.109717  0.2379845 42.48057 5.955647e-73
R-squared: 0.965
```

Motivating MLR

Example using data in earnings2019.csv

```
dat2_lm1_sum <- summary(lm(log(earn)~height, data=dat2))
dat2_lm1_sum %>% coefficients %>% round(4)
cat("R-squared:", round(dat2_lm1_sum$r.squared, 3))
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2382     0.1536  8.0617        0
height        0.0284     0.0023 12.4766        0
R-squared: 0.031
```

```
dat2_lm2_sum <- summary(lm(log(earn)~height+male, data=dat2))
dat2_lm2_sum %>% coefficients %>% round(4)
cat("R-squared:", round(dat2_lm2_sum$r.squared, 3))
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8140     0.2003  9.0568        0
height        0.0191     0.0031  6.1896        0
male          0.1109     0.0248  4.4668        0
R-squared: 0.034
```

Motivating MLR

```
dat2_lm3_sum <- summary(lm(log(earn)~height+age, data=dat2))
dat2_lm3_sum %>% coefficients %>% round(4)
cat("R-squared:", round(dat2_lm3_sum$r.squared, 3))
```

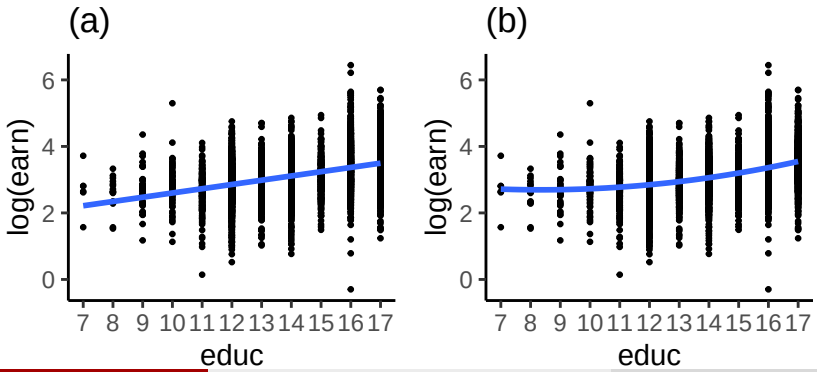
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9072     0.1557  5.8253        0
height        0.0286     0.0023 12.7034        0
age           0.0075     0.0008  9.8971        0
R-squared: 0.049
```

- When will including confounding factor change $\hat{\beta}_1^{ols}$?
- Will s.e.$(\hat{\beta}_1^{ols})$ go up or go down? Depends on what?
- Will R^2 always go up?

Recap of SLR

Insufficient flexibility in functional form

(a) $\ln earn = beta_0 + \beta_1 educ + \epsilon$ vs (b) $\ln earn = \beta_0 + \beta_1 educ + \beta_2 educ^2 + \epsilon$



Multiple Linear Regression

We start with the two-regressor multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i \, , \, i = 1, \dots , n$$

and consider estimation of the model by OLS

$$\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols} = \underset{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i)^2$$

Extend to more regressors later

Multiple Linear Regression

Main result: if in population,

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

or equivalently

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon\,,\ E(\epsilon \mid X, Z) = 0$$

and you have a representative iid sample $\{X_i, Y_i, Z_i\}_{i=1}^n$ from the population, then

$\hat{\beta}_0^{ols}$, $\hat{\beta}_1^{ols}$ and $\hat{\beta}_2^{ols}$ are unbiased and consistent estimators for β_0, β_1 and β_2

We consider also standard errors, hypotheses testing, etc.

OLS Estimation of the Multiple Linear Regression Model

Sample $\{Y_i, X_i, Z_i\}_{i=1}^n$

For any estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ (whether or not obtained by OLS), define

- **Fitted values**: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i$
- **Residuals**: $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i$

$$\begin{aligned}\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols} &= \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} SSR \\ &= \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i)^2\end{aligned}$$

OLS Estimation of the Multiple Linear Regression Model

OLS estimators can be found by solving FOC:

$$\left.\frac{\partial SSR}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0^{ols},\hat{\beta}_1^{ols},\hat{\beta}_2^{ols}} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols}X_i - \hat{\beta}_2^{ols}Z_i) = 0$$

$$\left.\frac{\partial SSR}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0^{ols},\hat{\beta}_1^{ols},\hat{\beta}_2^{ols}} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols}X_i - \hat{\beta}_2^{ols}Z_i)X_i = 0$$

$$\left.\frac{\partial SSR}{\partial \hat{\beta}_2}\right|_{\hat{\beta}_0^{ols},\hat{\beta}_1^{ols},\hat{\beta}_2^{ols}} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols}X_i - \hat{\beta}_2^{ols}Z_i)Z_i = 0$$

Can also write FOC as

$$\sum_{i=1}^{n}\hat{\epsilon}_i^{ols} = 0\,, \quad \sum_{i=1}^{n}\hat{\epsilon}_i^{ols}X_i = 0\,, \quad \text{and} \quad \sum_{i=1}^{n}\hat{\epsilon}_i^{ols}Z_i = 0$$

OLS Estimation of the Multiple Linear Regression Model

Can solve FOC directly for $\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}$

We take an alternative indirect approach

- more illustrative approach
- highlight how including Z controls for its confounding effects

Focus on $\hat{\beta}_1^{ols}$

- can get the solution for $\hat{\beta}_2^{ols}$ by switching X_i with Z_i in the steps shown
- can obtain $\hat{\beta}_0^{ols}$ from $\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X} - \hat{\beta}_2^{ols}\overline{Z}$

OLS Estimation of the Multiple Linear Regression Model

First recall (again!!)

- OLS estimation of SLR $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ breaks Y_i into two parts

$$Y_i = \hat{Y}_i^{ols} + \hat{\epsilon}_i^{ols}$$

  where $\hat{Y}_i^{ols} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i$ is perfectly correlated with X_i and $\hat{\epsilon}_i^{ols}$ is perfectly uncorrelated with X_i

- residuals $\hat{\epsilon}_i^{ols}$ is that part of Y_i that is uncorrelated with X_i
- residuals $\hat{\epsilon}_i^{ols}$ have sample mean equal to zero
- If $\overline{X} = 0$, then $\hat{\beta}_1^{ols}$ takes the form $\hat{\beta}_1^{ols} = \dfrac{\sum_{i=1}^n (X_i - \overline{X}) Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \dfrac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$

OLS Estimation of the Multiple Linear Regression Model

Solve for $\hat{\beta}_1^{ols}$ using the following "auxiliary" regressions:

- Regress X_i on Z_i, and collect residuals $r_{i,x|z}$

$$r_{i,x|z} = X_i - \hat{\delta}_0^{ols} - \hat{\delta}_1^{ols} Z_i \ , \ i = 1, 2, ..., n$$

  - $r_{i,x|z}$ contains movements in X_i that are perfectly uncorrelated with Z_i
  - $r_{i,x|z}$ has sample mean zero

- Regress Y_i on Z_i, and collect residuals $r_{i,y|z}$

$$r_{i,y|z} = Y_i - \hat{\alpha}_0^{ols} - \hat{\alpha}_1^{ols} Z_i \ , \ i = 1, 2, ..., n$$

  - $r_{i,y|z}$ contains movements in Y_i that are perfectly uncorrelated with Z_i
  - $r_{i,y|z}$ has sample mean zero

OLS Estimation of the Multiple Linear Regression Model

We are going to show that

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^{n} r_{i,x|z}^2}$$

Consider regression of $r_{i,y|z}$ on $r_{i,x|z}$ using OLS

$$r_{i,y|z} = \gamma_0 + \gamma_1 r_{i,x|z} + u_i$$

we have

$$\hat{\gamma}_1^{ols} = \frac{\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^{n} r_{i,x|z}^2}$$

We will show $\hat{\beta}_1^{ols} = \hat{\gamma}_1^{ols}$

OLS Estimation of the Multiple Linear Regression Model

Consider numerator

$$\begin{aligned}\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z} &= \sum_{i=1}^{n} r_{i,x|z}(Y_i - \hat{\alpha}_0^{ols} - \hat{\alpha}_1^{ols} Z_i) \\ &= \sum_{i=1}^{n} r_{i,x|z} Y_i - \hat{\alpha}_0^{ols} \sum_{i=1}^{n} r_{i,x|z} - \hat{\alpha}_1^{ols} \sum_{i=1}^{n} Z_i r_{i,x|z} \\ &= \sum_{i=1}^{n} r_{i,x|z} Y_i\end{aligned}$$

OLS Estimation of the Multiple Linear Regression Model

If we had solved the MLR FOC for $\hat{\beta}_0^{ols}$, $\hat{\beta}_1^{ols}$ and $\hat{\beta}_2^{ols}$, we would have

$$Y_i = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i + \hat{\beta}_2^{ols} Z_i + \hat{\epsilon}_i^{ols}$$

Substitute this into $\sum_{i=1}^{n} r_{i,x|z} Y_i$ gives

$$\begin{aligned}\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z} = \sum_{i=1}^{n} r_{i,x|z} Y_i &= \sum_{i=1}^{n} r_{i,x|z} (\hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i + \hat{\beta}_2^{ols} Z_i + \hat{\epsilon}_i^{ols}) \\ &= \hat{\beta}_1^{ols} \sum_{i=1}^{n} r_{i,x|z}^2 + \sum_{i=1}^{n} r_{i,x|z} \hat{\epsilon}_i^{ols}\end{aligned}$$

OLS Estimation of the Multiple Linear Regression Model

Now we use MLR FOC. We have

$$\sum_{i=1}^{n} r_{i,x|z}\hat{\epsilon}_i^{ols} = \sum_{i=1}^{n} \hat{\epsilon}_i^{ols}(X_i - \hat{\delta}_0^{ols} - \hat{\delta}_1^{ols} Z_i) = \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} X_i - \hat{\delta}_0 \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} - \hat{\delta}_1 \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} Z_i = 0$$

Therefore $\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z} = \hat{\beta}_1^{ols} \sum_{i=1}^{n} r_{i,x|z}^2$ which gives

$$\hat{\gamma}_1^{ols} = \frac{\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^{n} r_{i,x|z}^2} = \hat{\beta}_1^{ols}$$

Likewise, we have $\hat{\beta}_2^{ols} = \dfrac{\sum_{i=1}^{n} r_{i,z|x} r_{i,y|x}}{\sum_{i=1}^{n} r_{i,z|x}^2}$

OLS Estimation of the Multiple Linear Regression Model

Derivation of $\hat{\beta}_1^{ols}$ shows how confounding factors are 'controlled' in a multiple regression analysis

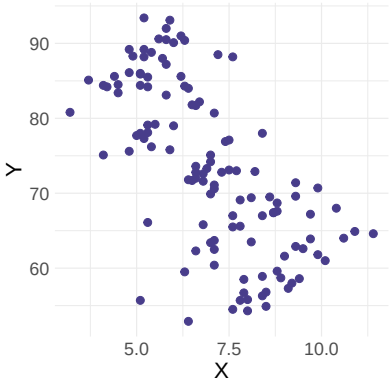
- Suppose we want to measure how Y_i is affected by X_i
- Suppose Z_i is an important determinant of Y_i that is correlated with X_i
- Omission of Z_i distorts measurement of the causal influence of X_i on Y_i
- Multiple regression estimates β_1 (coefficient on X_i) by
  - stripping out all variation in Y_i and X_i that are correlated with Z_i
  - measuring the correlation between the remaining variation in Y_i and X_i

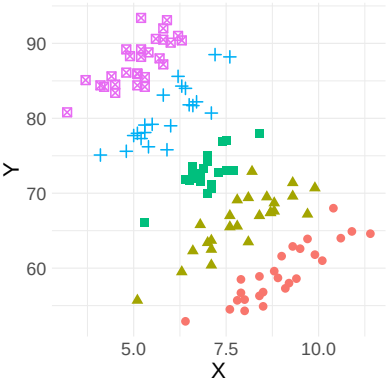
OLS Estimation of the Multiple Linear Regression Model

Example using data in **multireg_eg.csv**

```
df <- read_csv(
   ".\\data\\multireg_eg.csv",
   col_types = c("n","n","n"))
head(df, 8) # first 8 obs.
```

```
# A tibble: 8 x 3
      Z     X     Y
  <dbl> <dbl> <dbl>
1     1  10.1  61
2     2   7.1  63.7
3     1   9.2  58
4     5   6    90.1
5     2   9.7  67.2
6     4   7.6  88.2
7     2   8.1  63.5
8     3   7.1  70.6
```





OLS Estimation of the Multiple Linear Regression Model

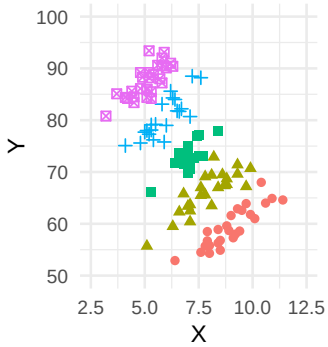
```
mdl1 <- lm(Y~X, data=df)
coef(summary(mdl1)); cat("R-squared:", summary(mdl1)$r.squared,"\n\n")
```

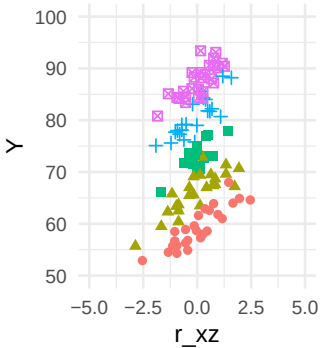
```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 102.86406  3.2199095 31.94626 6.373574e-60
X            -4.23213  0.4474755 -9.45779 3.857133e-16
R-squared: 0.4311877
```

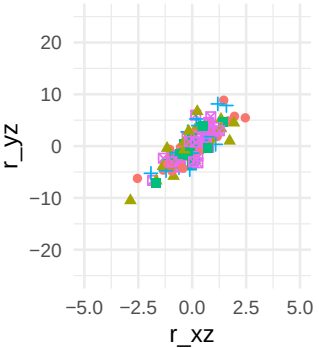
```
mdl2 <- lm(Y~X+Z, data=df)
coef(summary(mdl2)); cat("R-squared:", summary(mdl2)$r.squared,"\n\n")
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 21.188337  2.0816539 10.17861 8.212834e-18
X            3.114465  0.2054364 15.16024 2.219609e-29
Z           10.109717  0.2379845 42.48057 5.955647e-73
R-squared: 0.9653668
```

OLS Estimation of the Multiple Linear Regression Model







OLS Estimation of the Multiple Linear Regression Model

When is OLS infeasible?

To compute $\hat{\beta}_1^{ols} = \dfrac{\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^{n} r_{i,x|z}^2}$, we must have

- $\sum_{i=1}^{n} r_{i,x|z}^2 \neq 0$, i.e., X_i and Z_i cannot be *perfectly* correlated

To run auxiliary regression of X_i on Z_i in the first place

- There must be variation in Z_i, i.e., Z_i cannot be equal to some constant value c for all i

Similarly, to obtain $\hat{\beta}_2^{ols}$, we require variation in X_i, no perfect correlation between X_i and Z_i

OLS Estimation of the Multiple Linear Regression Model

We summarize these requirements by saying that OLS requires:

$$c_1 + c_2 X_i + c_3 Z_i = 0 \text{ for all } i = 1, 2, ..., n \text{ iff } (c_1, c_2, c_3) = (0, 0, 0)$$

Examples of when this condition does not hold:

- If $X_i = c$ for all i, $(-c) + (1)X_i + (0)Z_i = -c + c + 0 = 0$
- If $Z_i = c$ for all i, $(-c) + (0)X_i + (1)Z_i = -c + 0 + c = 0$
- If X_i and Z_i are perfectly correlated, i.e., $X_i = \gamma_0 + \gamma_1 Z_i$ for all i, then

$$\gamma_0 + (-1)X_i + \gamma_1 Z_i = 0$$

In any of these cases, we cannot regress Y_i on X_i and Z_i

Properties of OLS Estimators (MLR)

Many of the algebraic properties carry over from the simple linear regression model

- FOC can be written as

$$\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = 0, \quad \sum_{i=1}^{n} X_i \hat{\epsilon}_i^{ols} = 0 \quad \text{and} \quad \sum_{i=1}^{n} Z_i \hat{\epsilon}_i^{ols} = 0$$

- Fitted values $\hat{Y}_i^{ols}$ and residuals $\hat{\epsilon}_i^{ols}$ are also uncorrelated
- The fact that $\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X} - \hat{\beta}_2^{ols}\overline{Z}$ means that the point $(\overline{X}, \overline{Y}, \overline{Z})$ lies on the sample regression function

Properties of OLS Estimators (MLR)

- $\overline{Y} = \overline{\hat{Y}^{ols}}$
- TSS = FSS + RSS equality continues to hold in the multiple regression case

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

- We can use this to define the goodness-of-fit measure:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Properties of OLS Estimators (MLR)

- Let R_1^2 be R^2 from $Y_i = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i + \hat{\epsilon}_i^{ols}$
- Let R_2^2 be R^2 from $Y_i = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i + \hat{\beta}_2^{ols} Z_i + \hat{\epsilon}_i^{ols}$
- We have $R_2^2 \geq R_1^2$
  - R^2 will never decrease as we add more variables to the regression
  - OLS minimizes RSS, and therefore maximizes R^2
- "Adjusted R^2" is sometimes used instead:

$$\text{Adj.-}R^2 = 1 - \frac{RSS/(n-K)}{TSS/(n-1)} = 1 - \frac{RSS}{TSS}\frac{n-1}{n-K}$$

where K is no. of slope coefficients plus intercept ($K = 3$ for the 2-regressor case)

Properties of OLS Estimators (MLR)

When will including Z **not change** $\hat{\beta}_1$?

In the auxiliary regression X on Z (i.e., $\hat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i$)

$$\textit{Sample Cov.}(X_i, Z_i) = 0 \;\Rightarrow\; \hat{\delta}_1 = 0 \;\Rightarrow\; \hat{\delta}_0 = \overline{X} \;\Rightarrow\; r_{i,x|z} = X_i - \overline{X}$$

Then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} r_{i,x|z} Y_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) Y_i}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

This is, of course, just the OLS estimator for the coefficient on X_i in the *simple* linear regression of Y on X

Properties of OLS Estimators (MLR)

In our derivation, we obtained the following

$$\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z} = \sum_{i=1}^{n} r_{i,x|z}(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 Z_i) = \sum_{i=1}^{n} r_{i,x|z} Y_i.$$

This implies $\hat{\beta}_1$ can also be obtained as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^{n} r_{i,x|z}^2} = \frac{\sum_{i=1}^{n} r_{i,x|z} Y_i}{\sum_{i=1}^{n} r_{i,x|z}^2}$$

- you can also get the OLS estimator $\hat{\beta}_1$ by regressing Y_i on $r_{i,x|z}$ without first stripping out the covariance between Y_i and Z_i.
- But this does not fully reflect what happens in a regression of Y_i on X_i and Z_i

Properties of OLS Estimators (MLR)

- $\hat{\beta}_1$ is a linear estimator, i.e.,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} r_{i,x|z} Y_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = \sum_{i=1}^{n} \left(\frac{r_{i,x|z}}{\sum_{i=1}^{n} r_{i,x|z}^2} \right) Y_i = \sum_{i=1}^{n} w_i Y_i$$

The weights have the following properties:

$$\sum_{i=1}^{n} w_i = 0 \,, \quad \sum_{i=1}^{n} w_i Z_i = \frac{\sum_{i=1}^{n} r_{i,x|z} Z_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = 0 \,,$$

$$\sum_{i=1}^{n} w_i X_i = \frac{\sum_{i=1}^{n} r_{i,x|z} X_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = 1 \,, \quad \sum_{i=1}^{n} w_i^2 = \frac{\sum_{i=1}^{n} r_{i,x|z}^2}{(\sum_{i=1}^{n} r_{i,x|z}^2)^2} = \frac{1}{\sum_{i=1}^{n} r_{i,x|z}^2}$$

Properties of OLS Estimators (MLR)

$E(\epsilon \mid \mathsf{x}, \mathsf{z}) = 0$ in population, representative i.i.d. sample from the population, implies

(A1) $E(\epsilon_i \mid \mathsf{x}, \mathsf{z}) = 0$ for all $i = 1, ..., n$

(A2) $E(\epsilon_i \epsilon_j \mid \mathsf{x}, \mathsf{z}) = 0$ for all $i \neq j,\ i, j = 1, ..., n$ where x denotes $X_1, X_2, ..., X_n$, and z to denotes $Z_1, Z_2, ..., Z_n$

We will also assume homoskedastic errors

(A3) $E(\epsilon_i^2 \mid \mathsf{x}, \mathsf{z}) = \sigma^2$ for all $i = 1, ..., n$

Properties of OLS Estimators (MLR)

- $\hat{\beta}_1$ is unbiased

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i = \sum_{i=1}^{n} w_i(\beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i) = \beta_1 + \sum_{i=1}^{n} w_i \epsilon_i$$

Taking conditional expectations, noting that w_i comprise only x and z,

$$E(\hat{\beta}_1 \mid \mathsf{x}, \mathsf{z}) = \beta_1 + \sum_{i=1}^{n} w_i E(\epsilon_i \mid \mathsf{x}, \mathsf{z}) = \beta_1$$

It follows that the unconditional mean is $E(\hat{\beta}_1) = \beta_1$

- $\hat{\beta}_1$ also consistent (since $Cov(w_i, \epsilon_i) = 0$)

Properties of OLS Estimators (MLR)

- Conditional variance of $\hat{\beta}_1$

$$
\begin{aligned}
Var(\hat{\beta}_1 \mid \mathsf{x}, \mathsf{z}) &= Var\left(\beta_1 + \sum_{i=1}^{n} w_i \epsilon_i \,\middle|\, \mathsf{x}, \mathsf{z} \right) \\
&= \sum_{i=1}^{n} w_i^2 \, Var(\epsilon_i \mid \mathsf{x}, \mathsf{z}) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} r_{i,x|z}^2}
\end{aligned}
$$

Properties of OLS Estimators (MLR)

Since the R^2 from the regression of X_i on Z_i is $R^2_{x|z} = 1 - \dfrac{\sum_{i=1}^n r^2_{i,x|z}}{\sum_{i=1}^n (X_i - \overline{X})^2}$

We have

$$Var(\hat{\beta}_1 \mid \mathsf{x}, \mathsf{z}) = \frac{\sigma^2}{(1 - R^2_{x|z}) \sum_{i=1}^n (X_i - \overline{X})^2} \,.$$

Properties of OLS Estimators (MLR)

Tradeoffs? Suppose

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon\,,\ E(\epsilon \mid X, Z) = 0\,,\ X, Z \text{ correlated}\,,\ Var(\epsilon \mid X, Z) = \sigma^2$$

In the regression $Y = \beta_0 + \beta_1 X + u$ (nb: $u = \beta_2 Z + \epsilon,\ Var(u \mid X) = \sigma_u^2$)

$$\hat{\beta}_1 \text{ biased and inconsistent}\,,\ Var(\hat{\beta}_1 \mid \mathrm{x}) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

In the regression $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$

$$\hat{\beta}_1 \text{ unbiased, consistent}\,,\ Var(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z}) = \frac{\sigma^2}{(1 - R^2_{x|z})\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Properties of OLS Estimators (MLR)

To compute a numerical estimate for $Var(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z})$, have to estimate σ^2

An unbiased estimator for σ^2 in the two-regressor case is

$$\widehat{\sigma^2} = \frac{1}{n-3} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Therefore

$$\widehat{Var}(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z}) = \frac{\widehat{\sigma^2}}{(1 - R^2_{x|z}) \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Take square root for standard error

Hypothesis Testing

To test: $\beta_k = r_k$ in population, use the t-statistic as in SLR:

- If the noise terms ϵ are conditionally normally distributed, then

$$t = \frac{\hat{\beta}_k - r_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}} \sim \text{t}(n-K)$$

  - $K = 3$ in the two-regressor case with intercept
- If we do not assume normality of the noise terms, then (as long as CLT applies)

$$t = \frac{\hat{\beta}_k - r_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}} \overset{a}{\sim} \text{Normal}(0,1)$$

Hypothesis Testing

You can also test linear combinations of the parameters

E.g. Suppose regression is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

To test, say, $H_0 : a_1\beta_1 + a_2\beta_2 = r_1$ vs $H_A : a_1\beta_1 + a_2\beta_2 \neq r_1$

$$t = \frac{a_1\hat{\beta}_1 + a_2\hat{\beta}_2 - r_1}{\sqrt{\widehat{Var}(a_1\hat{\beta}_1 + a_2\hat{\beta}_2)}}$$

Hypothesis Testing

(Useful-to-know trick) Reparameterize regression equation to generate t-stat automatically

To test $H_0 : \beta_1 + \beta_2 = 1$, reparameterize the regression equation as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X - X + X - \beta_2 X + \beta_2 Z + \epsilon$$

$$Y - X = \beta_0 + (\beta_1 + \beta_2 - 1)X + \beta_2 (Z - X) + \epsilon$$

By regressing $Y - X$ on X and $Z - X$ (and an intercept), you can test $H_0 : \beta_1 + \beta_2 = 1$ by testing if the coefficient on X is equal to zero

Hypothesis Testing

Using data in the `earnings2019.xlsx`, we estimate the regression

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon\,.$$

To test $H_0 : \beta_2 = \beta_3$ (equivalent to $H_0 : \beta_2 - \beta_3 = 0$), re-parameterize regression equation as

$$\begin{aligned}\ln earn &= \beta_0 + \beta_1 educ + (\beta_2 - \beta_3) black + \beta_3 (female + black) + \beta_4 black.female + \epsilon \\ &= \pi_0 + \pi_1 educ + \pi_2 black + \pi_3 (female + black) + \pi_4 black.female + \epsilon\end{aligned}$$

Test $\pi_2 = 0$

Hypothesis Testing

```
dat2$female <- 1 - dat2$male
mdl_earnings1 <- lm(log(earn) ~ educ + black*female, data=dat2)
mdl_earnings1 %>% summary %>% coefficients %>% round(4)
```

```
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)    1.5369     0.0571  26.9219   0.0000
educ           0.1279     0.0039  32.9577   0.0000
black         -0.2600     0.0270  -9.6468   0.0000
female        -0.2807     0.0196 -14.3238   0.0000
black:female   0.0873     0.0355   2.4601   0.0139
```

```
mdl_earnings2 <- lm(log(earn) ~ educ + black + I(female+black) + female:black, data=dat2)
mdl_earnings2 %>% summary %>% coefficients %>% round(4)
```

```
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)         1.5369     0.0571  26.9219   0.0000
educ                0.1279     0.0039  32.9577   0.0000
black               0.0206     0.0271   0.7607   0.4469
I(female + black)  -0.2807     0.0196 -14.3238   0.0000
black:female        0.0873     0.0355   2.4601   0.0139
```

Hypothesis Testing

The hypothesis is not rejected

In some cases, we may wish to test multiple hypotheses, e.g., in the two-variable regression, we may wish to test

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ vs } H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

One possibility would be to do individual t-tests for each of the two hypotheses, but we should be aware that two individual 5% tests is not equivalent to a joint 5% test

Hypothesis Testing

- Simulate 100 observations of three uncorrelated variables X, Y and Z
- Regress Y on X and Z, and collect the t-statistics on X and Z
- Repeat the experiment 1000 times (with different draws each time

```
set.seed(3)
nreps <- 1000
tx <- tz <- rep(NA,nreps)
n <- 100
for (i in 1:nreps){
  X <- rnorm(n, mean=0, sd=2)
  Z <- rnorm(n, mean=0, sd=2)
  Y <- rnorm(n, mean=0, sd=2)
  df_test <- data.frame(X,Y,Z)
  mdlsim <- lm(Y~X+Z, data=df_test)
  tx[i] <- coef(summary(mdlsim))[2,'t value']
  tz[i] <- coef(summary(mdlsim))[3,'t value']
}
rjt_x <- sum(tx<qt(0.025,n-3) | tx>qt(0.975,n-3))/nreps
rjt_z <- sum(tz<qt(0.025,n-3) | tz>qt(0.975,n-3))/nreps
rjt_x_or_z <- sum(tz<qt(0.025,197) | tz>qt(0.975,197) |
                  tx<qt(0.025,197) | tx>qt(0.975,197))/nreps
```

Hypothesis Testing

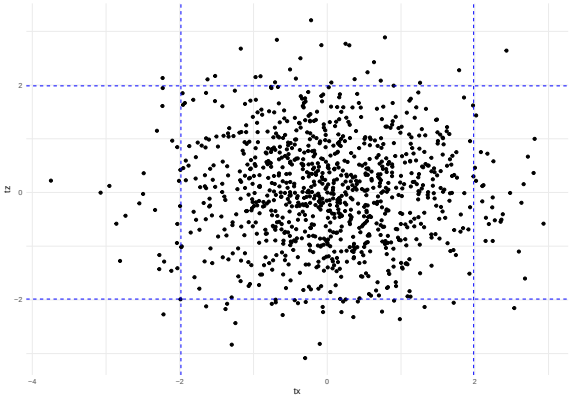
```
cat("Freq. of rejection of Beta_X = 0:", rjt_x, "\n")
cat("Freq. of rejection of Beta_Z = 0:", rjt_z, "\n")
cat("Freq. of rejection of Beta_X = 0 and Beta_Z = 0 using two t-tests:", rjt_x_or_z, "\n")
```

```
Freq. of rejection of Beta_X = 0: 0.058
Freq. of rejection of Beta_Z = 0: 0.054
Freq. of rejection of Beta_X = 0 and Beta_Z = 0 using two t-tests: 0.112
```

- Using a 5% t-test, reject $\beta_x = 0$ in about 6% of the time
- Likewise, the 5% t-test for $\beta_z = 0$ rejects the hypothesis 5.5% of the time, roughly five percent
- However, if we reject $\beta_x = 0$ **and** $\beta_z = 0$ if either t-tests rejects the corresponding hypothesis, then the frequency of rejection is much larger, roughly double.

Hypothesis Testing

We plot the t-stats below, indicating the critical values for the individual t-tests.



Hypothesis Testing

To jointly test multiple hypotheses, we can use the F-test

E.g. Suppose in the regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

we wish to jointly test the hypotheses

$$H_0 : \beta_1 = 1 \text{ and } \beta_2 = 0 \text{ vs } H_A : \beta_1 \neq 1 \text{ or } \beta_2 \neq 0 \text{ (or both)}$$

Run the regression twice:

- Unrestricted regression: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$
- Restricted regression: $Y = \beta_0 + X + \epsilon$

(Restricted OLS estimator for β_0 is $\widehat{\beta}_0^{rls} = (1/n) \sum_{i=1}^{n} (Y_i - X_i)$)

Hypothesis Testing

Calculate the RSS from both equations. The "unrestricted RSS" and "restricted RSS" are

$$RSS_{ur} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 \quad \text{and} \quad RSS_r = \sum_{i=1}^{n} \hat{\epsilon}_{i,rols}^2$$

respectively, where

- $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i$
- $\hat{\epsilon}_{i,rols} = Y_i - \hat{\beta}_0^{rols} - X_i$
- Note: $RSS_r \geq RSS_{ur}$

Hypothesis Testing

It can be shown that if the hypotheses in H_0 are true (and the noise terms are normally distributed), then

$$F = \frac{(RSS_r - RSS_{ur})/J}{RSS_{ur}/(n-K)} \sim F(J, n-K)$$

- J is the number of restrictions being tested (in our example, $J = 2$)
- K is the no. of coefficients in unrestricted regression (including intercept; in our example, $K = 3$)

Hypothesis Testing

Idea:

- if H_0 true in population, then imposing the restrictions will not increase RSS by much, F-statistic will be close to zero
- if one or more of the hypotheses in H_0 are false in population, then imposing them will cause the RSS to increase substantially, F-statistic will be large

If $F > \mathsf{F}_{1-\alpha}(J, n-K)$, reject H_0:

- $\mathsf{F}_{1-\alpha}(J, n-K)$ is $(1-\alpha)$-percentile of the $\mathsf{F}(J, n-K)$ distribution, α is typically 0.10, 0.05 or 0.01

Hypothesis Testing

- Since $R^2 = 1 - RSS/TSS$, we can write the F-statistic as

$$F = \frac{(R^2_{ur} - R^2_r)/J}{(1 - R^2_{ur})/(n - K)}$$

- If you cannot assume that the noise terms are conditionally normally distributed, then you will have to use an asymptotic approximation

$$JF \xrightarrow{d} \chi^2(J)$$

as $n \to \infty$, where J is the number of hypotheses being jointly tested

This is the "Chi-square Test"

Hypothesis Testing

Take the regression

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon$$

Test the hypothesis $H_0 : \beta_2 = \beta_3$ and $\beta_4 = 0$, vs the alternative that at least one of the two restrictions do not hold

The restricted regression is

$$\begin{aligned} \ln earn &= \beta_0 + \beta_1 educ + \beta_2 black + \beta_2 female + 0 black.female + \epsilon \\ &= \beta_0 + \beta_1 educ + \beta_2 (black + female) + \epsilon \end{aligned}$$

We carry out the F-test of our hypotheses below (next slide)

Hypothesis Testing

```
mdl_unres <- lm(log(earn) ~ educ + black*female, data=dat2)
mdl_res <- lm(log(earn) ~ educ + I(black + female), data=dat2)
RSS_unres <- sum(residuals(mdl_unres)^2)
RSS_res   <- sum(residuals(mdl_res)^2)
J = 2
n_minus_K <- mdl_unres$df.residual
Fstat = ((RSS_res - RSS_unres) / J) / (RSS_unres / n_minus_K)
pval = 1-pf(Fstat, J, n_minus_K)
cat("F-stat:", Fstat, "   pvalue:", pval)
```

```
F-stat: 4.546114    pvalue: 0.01065276
```

The joint hypothesis is rejected (at 0.05 significance level, but not 0.01 significance level, marginally)

Hypothesis Testing

We can use `linearHypothesis()` from the `car` package to carry out F tests

```
library(car)
mdl_unres <- lm(log(earn) ~ educ + black*female, data=dat2)
linearHypothesis(mdl_unres, c("black=female", "black:female=0"))
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   4943 1606.5
2   4941 1603.5  2    2.9507 4.5461 0.01065 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Testing

To get the chi-square version of the test:

```
linearHypothesis(mdl_unres, c("black=female", "black:female=0"), test="Chisq")
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
1   4943 1606.5
2   4941 1603.5  2    2.9507 9.0922    0.01061 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The chi-square statistic is just J times the F statistic

Hypothesis Testing

You can use the F-test to test a single hypothesis, e.g., to test $\beta_2 = \beta_3$:

```
Linear hypothesis test:
black - female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   4942 1603.7
2   4941 1603.5  1   0.18778 0.5786 0.4469
```

- F-stat for testing single hypothesis is square of t-stat for the same hypothesis
- The p-value is the same

Hypothesis Testing

```
mdl_unres %>% summary()
```

```
Call:
lm(formula = log(earn) ~ educ + black * female, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5992 -0.3448  0.0022  0.3515  2.8614

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.536856   0.057086  26.922   <2e-16 ***
educ          0.127858   0.003879  32.958   <2e-16 ***
black        -0.260050   0.026957  -9.647   <2e-16 ***
female       -0.280661   0.019594 -14.324   <2e-16 ***
black:female  0.087299   0.035486   2.460   0.0139 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5697 on 4941 degrees of freedom
Multiple R-squared:  0.2389,    Adjusted R-squared:  0.2383
F-statistic: 387.7 on 4 and 4941 DF,  p-value: < 2.2e-16
```

- `summary()` output of `lm` `object` contains a F-statistic
- Tests joint hypotheses that all coefficients in the regression (not including intercept) are zero, versus the alternative that at least one is not zero