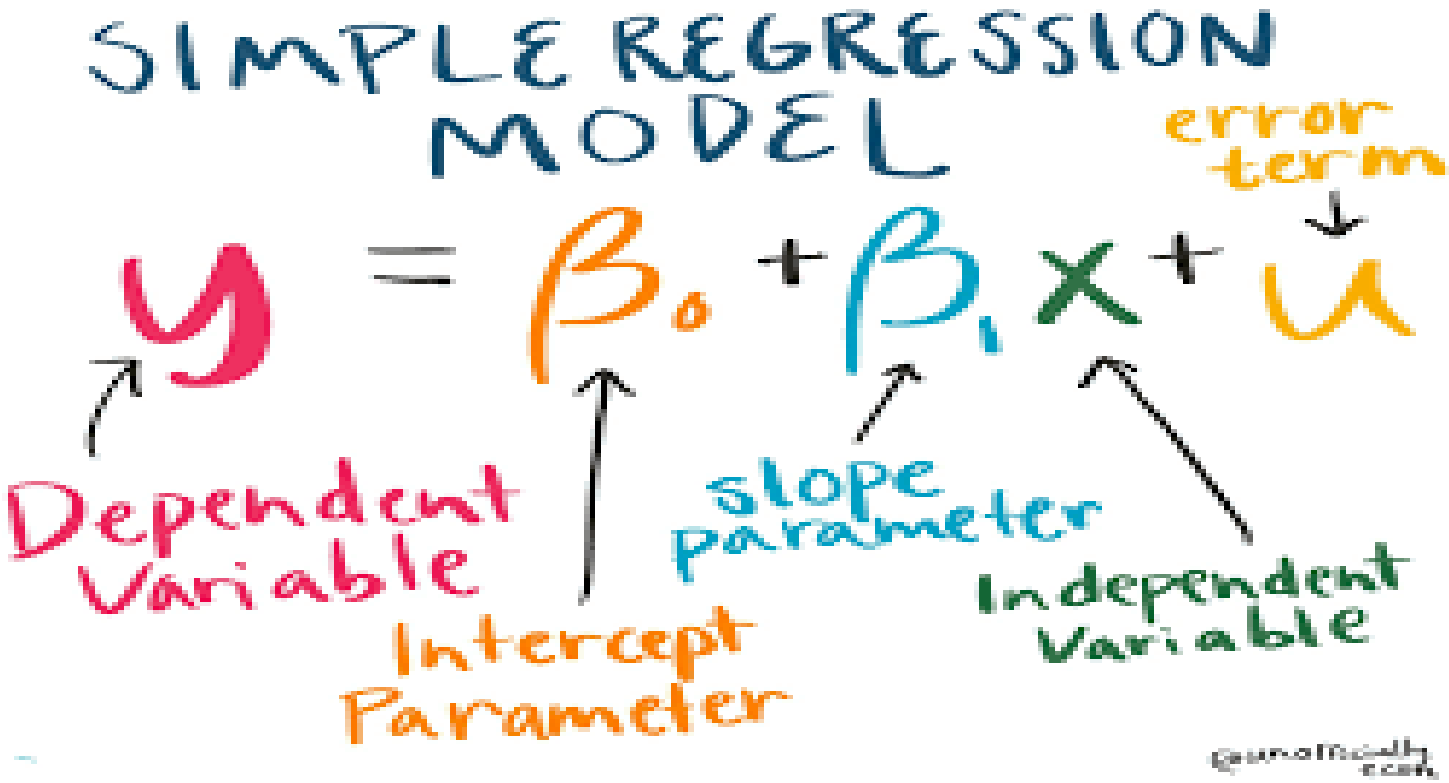
Fundamentals of Econometrics

Lecture 2: Simple Linear Regression Model



Section 1

The Basics

The Regression Problem

The premise: We have two variables, x and y, and we want to study **how y varies with changes in x.**

Regression analysis tries to answer the question: **How can we explain the average behavior of the dependent variable using that of the independent variable(s).**

$$y = \beta_0 + \beta_1 x_i + u$$

y: : **dependent** (or explained) variable, regressand

x: : **independent** (or explanatory) variable(s), regressor

parameters: : β_0 (**intercept**) and β_1 (**slope**)

u: : error, disturbance, innovation

Note: This is sometimes referred to as the **data generating process** (DGP) because we assume that the observable data follows this

y	x
Dependent Variable	Independent Variable
Explained Variable	Explaining (Explanatory) Variable
Endogenous Variable	Exogenous Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
LHS Variable	RHS Variable
Regressand	Regressor

We say we **regress** y on x.

The Regression Problem

The regression model studies how y varies with changes in x:

$$\frac{\delta y}{\delta x} = \beta_1 \quad \text{as long as} \quad \frac{\delta u}{\delta x} = 0$$

$\Downarrow$ $\Downarrow$

By how much does the dependent variable change if the independent variable is increased by one unit?

Interpretation of β_1 is only correct if all other things (in u) remain equal (or fixed) when the independent variable is increased by one unit.

The simple regression is rarely applicable in practice but offers a good starting point for understanding the more general multiple regression model.

Examples of SLR

Example: Does a new fertlizer increase soybean yield?

$$yeild_i = \beta_0 + \beta_1 fertilizer_i + u_i$$

We are interested in the causal effect of the amount of fertilizer used on the yield of soybeans.

Omitted Variables?

u contains all relevant factors which are unobserved by the researcher.

- What else is in u?

Examples of SLR

Example: Wages and education

$$wage_i = \beta_0 + \beta_1 education_i + u$$

- wage is dollars per hour
- education is years of schooling
- What else is in u?

Examples of SLR

Example: Corn yield and time

$$yeild_t = \beta_0 + \beta_1 t + u_t, \quad t = 1, \ldots, T$$

where t captures the effect of time on yield. T is the total number of time periods.

- What else is in u?

Note: Here, we are examining how corn yields change as a function of time (a linear trend). The implication is we have time series data

Key Condition for Causal Interpretation

- The key to identifying the causal effect of x on y is that we must restrict how u and x are related to each other.
- How do we restrict the dependence between u and x?

$$E(u|x) = E(u), \forall x$$

i.e. u is mean independent of x.

This is the assumption that allows you to interpret the result as "causal effect".

SLR Assumption #1

Zero Conditional Mean Assumption

- Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (a normalization) gives the Zero Conditional mean assumption:

$$\boxed{E(u_i|x) = 0}, \ \forall x$$

- $E(u|x) = 0$ implies that:

$$\begin{aligned} E(y|x) &= E(\beta_0 + \beta_1 x + u|x) & &= E(\beta_0 + \beta_1 x + E(u|x)) \\ &= \beta_0 + \beta_1 x + \cancelto{0}{E(u|x)} & &= \beta_0 + \beta_1 x \end{aligned}$$

- **The population regression function is a linear function of x.**
- Implies that the explanatory variable must not contain information about the mean of the unobserved factors that affect the dependent.
- Holds regardless of whether x is fixed or stochastic (in repeated samples).

SLR Assumption #1

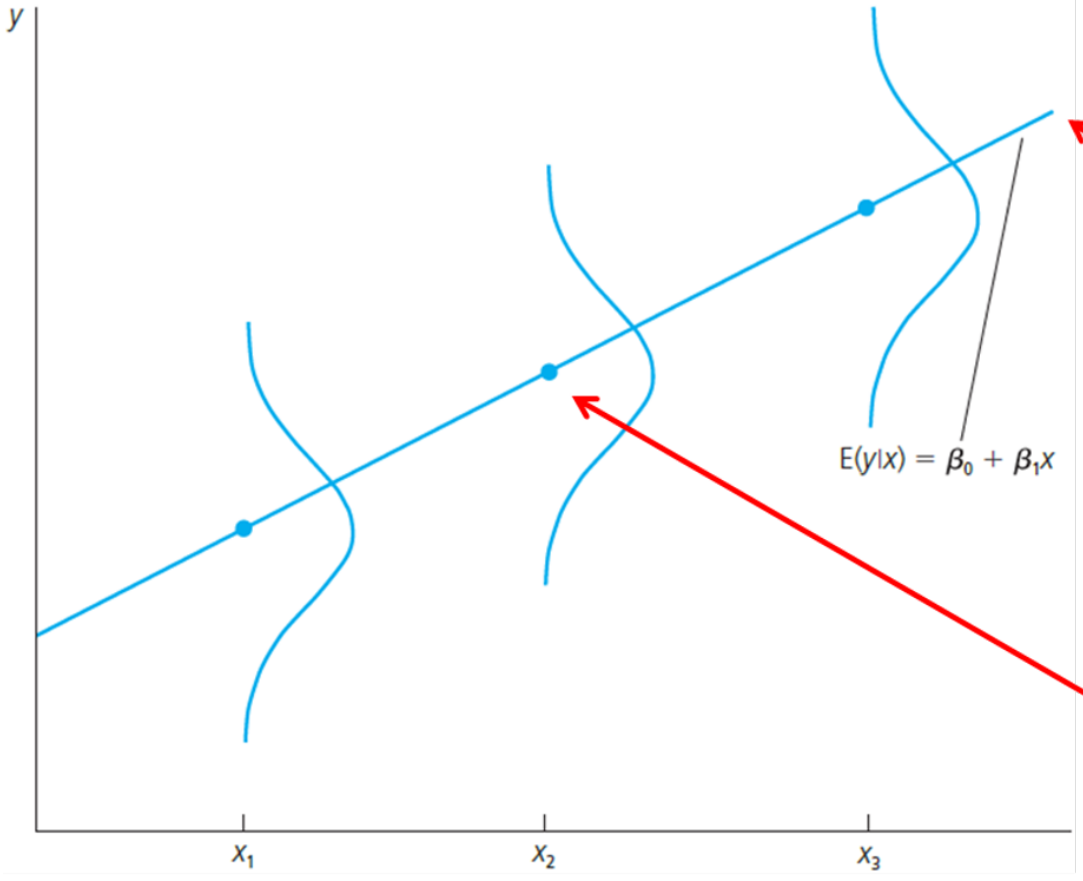
Example: wage equation

$$wage_i = \beta_0 + \beta_1 educ_i + u$$

- u includes factors such as ability, motivation, intelligence, etc.

The conditional mean independence assumption is unlikely to hold because individuals with more education will also be more intelligent on average

The SLR



Population regression function

For individuals with $x = x_2$, the average value of y is $\beta_0 + \beta_1 x_2$

Meaning of "Linear" Regression

Linearity in Variables vs. Linearity in Parameters

- Models that are nonlinear in variables but linear in parameters

$$y_i = \beta_0 + \beta_1 x_i^2 + u_i$$

$$y_i = \beta_0 + \beta_1 \frac{1}{x_i^2} + u_i$$

- Models that are nonlinear in parameters but linear in variables

$$y_i = \beta_0 + \beta_1^2 x_i + u_i$$

$$y_i = \beta_0 + \sqrt{\beta_1} x_i + u_i$$

- Models that are nonlinear in both variables and parameters

$$y_i = \beta_0 x_i^{\beta_1} + u_i$$

$$log(y_i) = \beta_0 + \frac{1}{\beta_1 x_i} + u_i$$

- For estimation purposes, what matters is linearity in parameters.

Section 2

Deriving the OLS Estimators

Deriving the OLS Estimator

- One way of deriving the OLS estimator is to **minimize the sum of squared residuals**.
- The regression residuals, $\hat{u}_i$, are the difference between the observed (actual) and fitted (predicted) values of the dependent variable:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

1. Minimizing the sum of squared residuals:

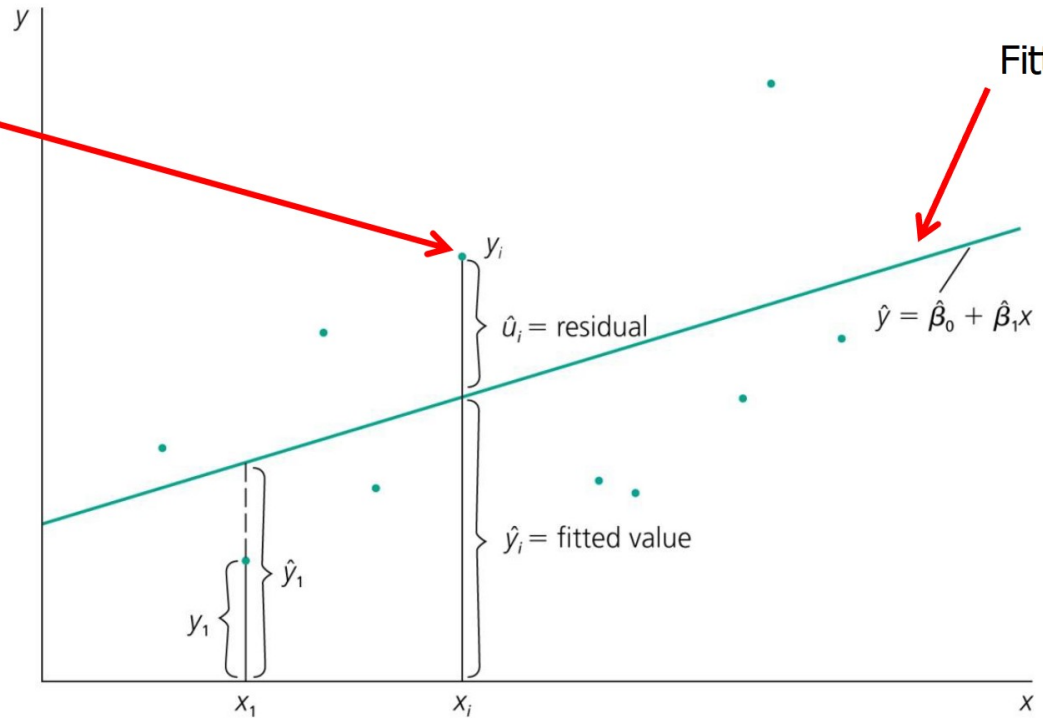
$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

2. OLS Estimators:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}; \quad \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

For example, the i-th data point (x_i, y_i)

Fitted regression line



Deriving the OLS Estimator

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that $\hat{u}_i^2$ is "small".

$$\min_{\{\beta_0,\beta_1\}} \sum_{i=1}^{n} Q^2 = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Differentiating with respect to β_0 and β_1 and setting the derivatives equal to zero gives the OLS estimators:

$$\frac{\partial Q^2}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \tag{1}$$

$$\frac{\partial Q^2}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \tag{2}$$

Basic properties: summation operator (Digression)

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

$$\sum_{i=1}^{n} \alpha = n\alpha$$

$$\sum_{i=1}^{n} \alpha x_i = \alpha x_1 + \alpha x_2 + \ldots + \alpha x_n = \alpha \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} \alpha(x_i + y_i) = \alpha \left[\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \right]$$

Deriving the OLS Estimator

Using these properties, we can rewrite (1) and (2) as:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \boxed{\sum_{i=1}^{n} \hat{u}_i = 0} \tag{3}$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = \boxed{\sum_{i=1}^{n} \hat{u}_i x_i = 0} \tag{4}$$

(3) is particularly important because, as long as there is an intercept in the model, the sum of the residuals is always zero.

(3) and (4) are the so called "Least Squares Normal Equations" or just "Normal Equations". We now have two equations and two unknowns, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Deriving the OLS Estimator, β_0

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6}$$

Note that (6) naturally implies that:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Therefore, the estimated regression line always passes through the point $(\bar{x}, \bar{y})$, the sample means!

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0} \tag{7}$$

$$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{var(x_i)} = \sum_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x}\cancelto{0}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$= \sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i$$

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{8}$$

Deriving the OLS Estimator

Useful Properties

$$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}_{\text{covariance(x,y)}} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - \bar{y}\cancelto{0}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i$$

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i \implies \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}} \tag{9}$$

Deriving the OLS Estimator, β_1

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Substitute $\hat{\beta}_0$ from (6)

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\underbrace{\boxed{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}}_{cov(x,y)} = \beta_1 \underbrace{\boxed{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}}_{var(x)}$$

Deriving the OLS Estimator, β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{10}$$

$$\hat{\beta}_1 = \frac{cov(x, y)}{\hat{\sigma}_x^2} \tag{11}$$

Section 3

Estimating the OLS Coefficients

Example: Hours studied and exam score

Below is a sample dataset of 10 observations on the relationship between the number of hours studied and the exam score.

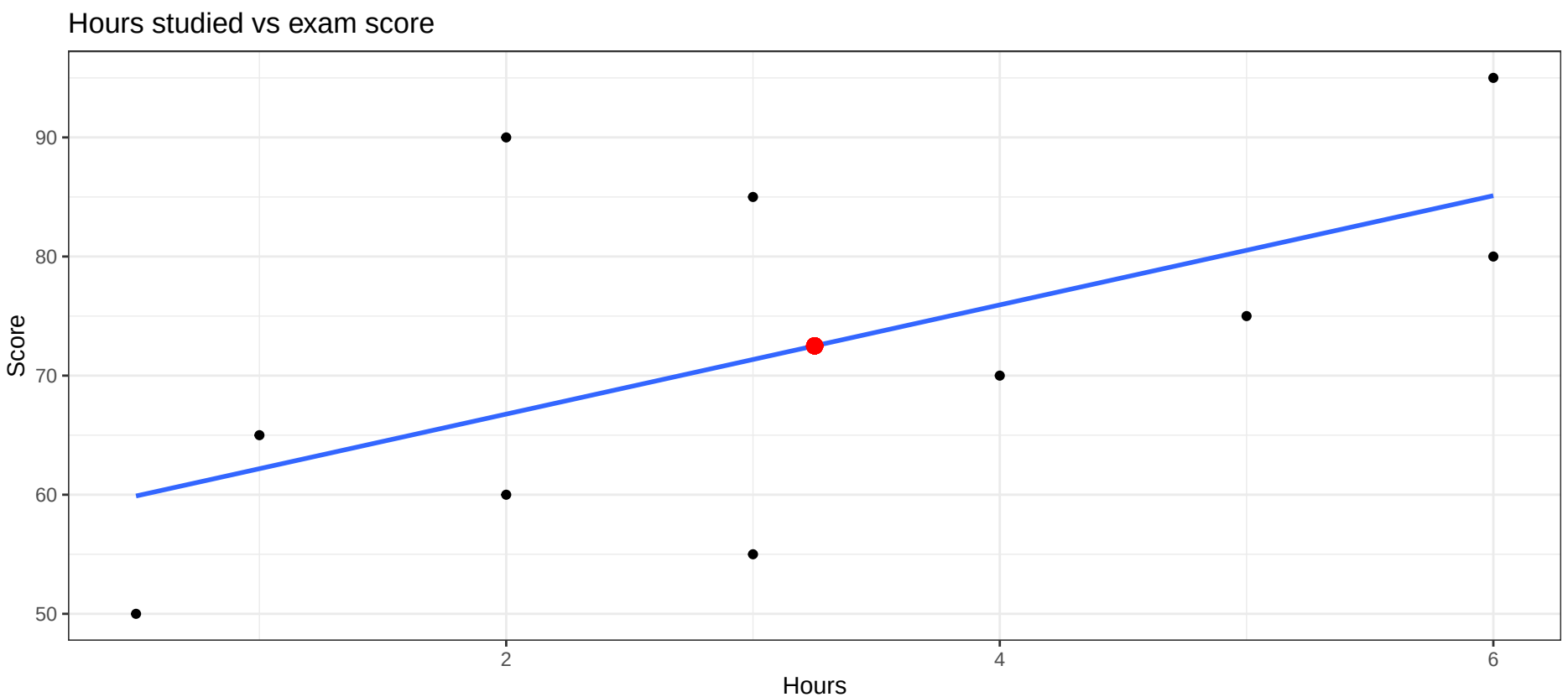
$$score_i = \beta_0 + \beta_1 hours_i + u_i$$

```
hours <- c(0.5,3,2,1,4,5,6,3,2,6)
score <- c(50, 55, 60, 65, 70, 75, 80, 85, 90, 95)
df <- data.frame(hours, score)
df
```

```
##    hours score
## 1    0.5    50
## 2    3.0    55
## 3    2.0    60
## 4    1.0    65
## 5    4.0    70
## 6    5.0    75
## 7    6.0    80
## 8    3.0    85
## 9    2.0    90
## 10   6.0    95
```

Visualizing the data

```r
# Scatter plot of hours studied and exam score
ggplot(df, aes(x = hours, y = score)) +
# Add points
  geom_point() +
# Add regression line
  geom_smooth(method = "lm", se = FALSE) +
  # Add mean of x and y to plot
  geom_point(aes(x = mean(hours), y = mean(score)), color = "red", size = 3) +
# Add title and axis labels
  labs(title = "Hours studied vs exam score", x = "Hours", y = "Score") +
# Change theme
  theme_bw()
```



Estimating the OLS coefficients

```r
# Calculate the means of hours and score
xbar <- mean(df$hours)
ybar <- mean(df$score)
# Calculate the variance of hours
varx <- sum((df$hours - xbar)^2)
# Calculate the covariance of hours and score
cov_xy <- sum((df$hours - xbar) * (df$score - ybar))
# Calculate the OLS estimator for beta_1
beta_1_hat <- cov_xy / varx
# Calculate the OLS estimator for beta_0
beta_0_hat <- ybar - beta_1_hat * xbar

# Print the OLS estimators
cat("The OLS estimator for beta_0 is: ", beta_0_hat,
    "\n beta_1 is: ", beta_1_hat)
```

```
## The OLS estimator for beta_0 is:  57.59928
##  beta_1 is:  4.584838
```

{.plain

Using the `lm()` function in R, we can estimate the OLS coefficients.

```
model1 <- lm(score ~ hours, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = score ~ hours, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.354  -6.561  -5.316   8.123  23.231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   57.599      8.220   7.007 0.000112 ***
## hours          4.585      2.195   2.089 0.070158 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.92 on 8 degrees of freedom
```

Estimating the OLS coefficients

We hypothesize the following DGP:

$$score_i = \beta_0 + \beta_1 hours_i + u_i$$

The fitted regression line is:

$$\hat{score}_i = 57.599 + 4.585 hours_i$$

- If the number of hours studied increases by one unit, the exam score is expected (predicted) to increase by 4.585 points.
- Every extra hour of study increases the exam score by approximately 4.585 points on average.

Computing the OLS residuals (manually)

Using the OLS estimators, we can calculate the residuals.

```
# Calculate the residuals
# y - yhat = u
df$resids.man <- df$score - (beta_0_hat + beta_1_hat * df$hours)
df
```

```
##    hours score resids.man
## 1    0.5    50  -9.891697
## 2    3.0    55 -16.353791
## 3    2.0    60  -6.768953
## 4    1.0    65   2.815884
## 5    4.0    70  -5.938628
## 6    5.0    75  -5.523466
## 7    6.0    80  -5.108303
## 8    3.0    85  13.646209
## 9    2.0    90  23.231047
## 10   6.0    95   9.891697
```

What is the sum of the residuals?

Computing the OLS residuals

Using the `residuals()` function in `R`, we can calculate the residuals.

```
# Calculate the residuals
df$resids.lm <- residuals(model1)
# Print the residuals
df
```

```
##    hours score resids.man  resids.lm
## 1    0.5    50  -9.891697  -9.891697
## 2    3.0    55 -16.353791 -16.353791
## 3    2.0    60  -6.768953  -6.768953
## 4    1.0    65   2.815884   2.815884
## 5    4.0    70  -5.938628  -5.938628
## 6    5.0    75  -5.523466  -5.523466
## 7    6.0    80  -5.108303  -5.108303
## 8    3.0    85  13.646209  13.646209
## 9    2.0    90  23.231047  23.231047
## 10   6.0    95   9.891697   9.891697
```

Computing the OLS residuals

Are all the residuals the same **pairwise**?

```
# Check if the residuals are the same
all.equal(df$resids.lm,df$resids.man)
```

```
## [1] TRUE
```

Computing the fitted values

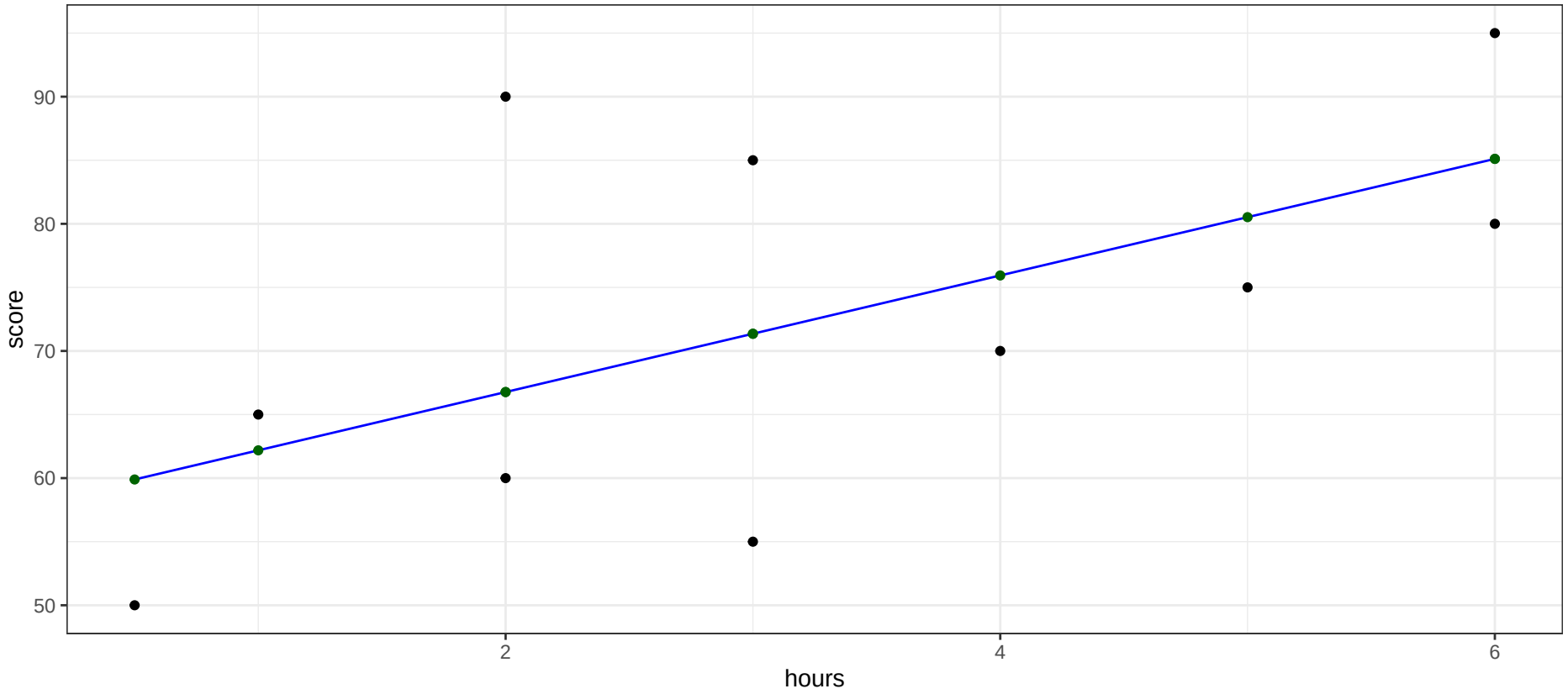
We might be interested in knowing the fitted values implied by the OLS estimators.

```
# Calculate the fitted values (y-hat)
df$fit.lm <- fitted(model1)
df$fit.man <- beta_0_hat + beta_1_hat * df$hours
df
```

```
##    hours score resids.man  resids.lm   fit.lm  fit.man
## 1    0.5    50  -9.891697  -9.891697 59.89170 59.89170
## 2    3.0    55 -16.353791 -16.353791 71.35379 71.35379
## 3    2.0    60  -6.768953  -6.768953 66.76895 66.76895
## 4    1.0    65   2.815884   2.815884 62.18412 62.18412
## 5    4.0    70  -5.938628  -5.938628 75.93863 75.93863
## 6    5.0    75  -5.523466  -5.523466 80.52347 80.52347
## 7    6.0    80  -5.108303  -5.108303 85.10830 85.10830
## 8    3.0    85  13.646209  13.646209 71.35379 71.35379
## 9    2.0    90  23.231047  23.231047 66.76895 66.76895
```

Plotting the fitted vs actual values

```r
# Scatter plot of hours studied and exam score
ggplot(df, aes(x = hours, y = score)) +
# Add points (use a diff point type for each obs)
  geom_point() +
# Add regression line
geom_line(aes(x = hours, y = fit.lm), color = "blue") +
# Show points of fitted values
  geom_point(aes(x = hours, y = fit.lm),color = "darkgreen") +
  theme_bw()
```



Section 4

Goodness of Fit

Coefficient of Determination, R^2

- How well does our explanatory variable explain the dependent variable?
- More generally: How well does the regression line fit the data?

One approach is to measure the share of observed variation in y that can be explained by the variation in x.

- Total Sum of Squares (TSS) $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Explained Sum of Squares (ESS) $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Residual Sum of Squares (SSR) $= \sum_{i=1}^{n} \hat{u}_i^2$

$$SST = ESS + SSR$$

Coefficient of Determination, R^2

- Hard to interpret the value of any of the sum of squares because they depend on the scale of the data. Much easier to interpret a ratio which would remove the scaling effect.

$$
\begin{aligned}
SST &= ESS + SSR \\
1 &= \frac{ESS}{SST} + \frac{SSR}{SST} \\
1 &= R^2 + \frac{SSR}{SST} \\
R^2 &= 1 - \frac{SSR}{SST}
\end{aligned}
$$

Properties of R^2:

- $0 \leq R^2 \leq 1$ (once there is an intercept in the model)
- Measures statistical correlation, not causation

Coefficient of Determination, R^2

CEO Salary and return on equity:

$$\widehat{salary}_i = 963.191 + 18.501 roe_i$$
$$n = 209, \quad R^2 = 0.0132$$

The regression explains only 1.3% of the total variation in salaries.

Voting outcomes and campaign expenditures

$$\widehat{voteA} = 26.81 + 0.464 shareA$$
$$n = 173, \quad R^2 = 0.856$$

The percentage of total campaign expenditures accounted for by Candidate A ($shareA$) explains 85.6% of the total variation in election outcomes for candidate A.

Caution: A high R-squared does not necessarily mean that the regression has a causal interpretation!

```
##
## Call:
## lm(formula = score ~ hours, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.354  -6.561  -5.316   8.123  23.231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   57.599      8.220   7.007 0.000112 ***
## hours          4.585      2.195   2.089 0.070158 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 12.92 on 8 degrees of freedom
## Multiple R-squared:  0.3529, Adjusted R-squared:  0.272
## F-statistic: 4.363 on 1 and 8 DF,  p-value: 0.07016
```

Coefficient of Determination, R^2

```r
# Calculate the total sum of squares
TSS <- sum((df$score - mean(df$score))^2)
# Calculate the explained sum of squares
ESS <- sum((predict(model1) - mean(df$score))^2)
# Calculate the residual sum of squares
SSR <- sum(residuals(model1)^2)
# Calculate R-squared
R2 <- 1 - SSR / TSS
R2
```

```
## [1] 0.3528936
```

Coefficient of Determination, R^2

Key Limitations:

- R^2 always increases when more explanatory variables are added to the model, even if they are irrelevant.
- R^2 does not indicate whether a regression model is adequate. You can have a low R^2 for a good model, or a high R^2 for a model that does not fit the data.

Are Low R-squared values bad?

No! If your R^2 is low but your estimates are statistically significant and make sense in the context of the problem (economic significance), you can still draw useful conclusions from your model.

Unit of Measurement

How does changing the units of measurement affect the OLS estimators?

Example: Regress CEO salary on return on equity

Original model:

- Salary in **thousands of dollars** and roe in percent

$$\widehat{salary}_i = 963.191 + 18.501 roe_i$$

Transformed model:

- Express salary (dependent variable) in **dollars** and roe in percent

$$\widehat{salary}_i = 963,191 + 18,501 roe_i$$

Unit of Measurement

```r
# library(wooldridge) # Load datasets
# Store original model
model.ceo <- lm(salary ~ roe, data = ceosal1)

# Store transformed model
model.ceo2 <- lm(salary * 1000 ~ roe, data = ceosal1)

# compare the coefficients
data.frame(Original = coef(model.ceo), Dep.Transformed = coef(model.ceo2))
```

```
##                Original Dep.Transformed
## (Intercept) 963.19134       963191.34
## roe          18.50119        18501.19
```

General Takeaways?

Unit of Measurement

What happens if we were to express the independent variable in decimal form instead of percentages?

New model: Express salary (dependent variable) in **thousand of dollars** and roe in **percent/100**

```
# Store transformed model
model.ceo3 <- lm(salary ~ I(roe/100), data = ceosal1)
coef(model.ceo3)
```

```
## (Intercept)  I(roe/100)
##    963.1913   1850.1186
```

Unit of Measurement

What happens to R^2 in both cases?

```
#Extract R-squared from the original model
R2_original <- summary(model.ceo)$r.squared
#Extract R-squared from dep. variable transformed model
R2_dep <- summary(model.ceo2)$r.squared
#Extract R-squared from indep. variable transformed model
R2_ind <- summary(model.ceo3)$r.squared
# Store in a data frame
R2 <- data.frame(Original = R2_original, Dependent = R2_dep,
                 Independent = R2_ind)
R2
```

```
##     Original  Dependent Independent
## 1 0.01318862 0.01318862  0.01318862
```

Functional Form

What happens if I think there are increasing returns to years of education. This might imply that:

$$wage_i = \exp \beta_0 + \beta_1 educ_i + u_i$$

If we want to estimate with OLS then the model still needs to be **linear in the parameters**. We could convert this to a linear model by taking the natural log of both sides:

$$\log (wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

This will change the interpretation of the coefficients.

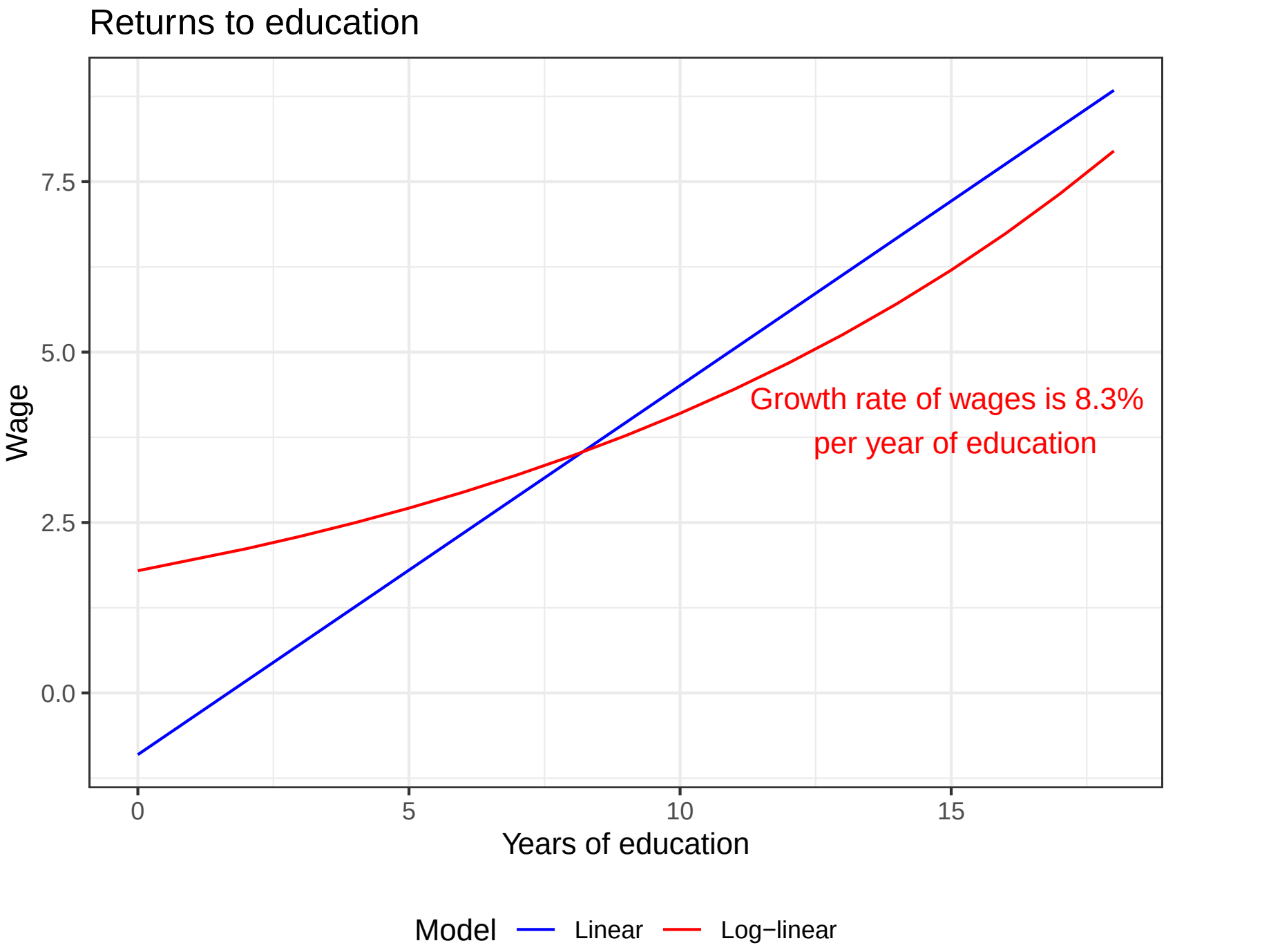
$$\beta_1 = \frac{\Delta \log (wage)}{\Delta educ} = \frac{1}{wage} \cdot \frac{\Delta wage}{\Delta educ} = \frac{\frac{\Delta wage}{wage}}{\Delta educ}$$

Functional Form

Example: returns of education to wage

```
wage.lin <- lm(wage ~ educ, data = wage1)
wage.log <- lm(log(wage) ~ educ, data = wage1)

# Plot the model fits
ggplot(wage1) +
  geom_line(aes(x = educ, y = fitted(wage.lin), color = "blue")) +
  # Add the log-linear model, must exponentiate the fitted values
  geom_line(aes(x = educ, y = exp(fitted(wage.log)), color = "red")) +
  labs(title = "Returns to education", x = "Educ", y = "Wage") +
  # Add a legend
  scale_color_manual(name = "Model", values = c("blue", "red"),
                     labels = c("Linear", "Log-linear")) +
  # Add an arrow and text to quadratic line
  annotate("text",x = 15, y = 4,
           label = "Growth rate of wages is 8.3% \n per year of educ.",
           color = "red") +
  theme_bw() +
  theme(legend.position = "bottom")
```


Returns to education
Wage
7.5
5.0
2.5
0.0
0
5
10
15
Years of education
Growth rate of wages is 8.3%
per year of education
Model
Linear
Log−linear

Example: CEO salary and firm sales

What about a model where logs appear on both sides?

$$log(salary) = \beta_0 + \beta_1 log(sales) + u$$

- Again, this changes the interpretation of the coefficients.

$$\beta_1 = \frac{\Delta \log(salary)}{\Delta \log(sales)} = \frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}}$$

- A 1% increase in sales is associated with a β_1% increase in salary.
- This is a measure of the **elasticity** of salary with respect to sales, whereas the semi-log form assumes a **semi-elasticity**.

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales)$$

```
##
## Call:
## lm(formula = log(salary) ~ log(sales), data = ceosal1)
##
## Coefficients:
## (Intercept)   log(sales)
##      4.8220       0.2567
```

As sales increase by 1%, CEO salary is expected to increase by 0.257%.

Functional Form

The functional forms can be summarized as follows:

- level-level: $wage_i = \beta_0 + \beta_1 educ_i + u_i$
- log-level (semi-log): $\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$
- level-log (semi-log): $wage_i = \beta_0 + \beta_1 \log(educ_i) + u_i$
- log-log: $\log(wage_i) = \beta_0 + \beta_1 \log(educ_i) + u_i$

You should be able to interpret the coefficients in each of these models.

Properties of OLS Estimators

The estimated regression coefficients are random variables because they are calculated from a random sample.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Our data is random and depends on particular sample that has been drawn.

This raises the question of what the estimators will estimate on average and how large will their variability be in repeated samples.

$$E(\hat{\beta}_1) \stackrel{?}{=} \beta_1, \; E(\hat{\beta}_0) \stackrel{?}{=} \beta_0 \quad var(\hat{\beta}_1) \stackrel{?}{=} \sigma^2_{\hat{\beta}_1}, \quad var(\hat{\beta}_0) =?, \; var(\hat{\beta}_1) =?$$

Digression

Show that β_1 is a random variable since it depends on the sample drawn.

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(\textcircled{x_i} - \bar{x})(\textcircled{y_i} - \bar{y})}{\sum_{i=1}^{n}(\textcircled{x_i} - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(\beta_0 + \beta_1 x_i + u_i) \\
&= \beta_0 \frac{\cancelto{0}{\sum_{i=1}^{n}(x_i - \bar{x})}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}u_i \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(\textcircled{x_i} - \bar{x})}{\sum_{i=1}^{n}(\textcircled{x_i} - \bar{x})^2}\textcircled{u_i}
\end{aligned}
$$

Similar logic holds for $\hat{\beta}_0$

OLS Assumptions

1. **SLR.1 Linearity in parameters:** The model is linear in the parameters reflecting the assumption that the true relationship between the dependent and independent variables (in the population) is indeed linear.
2. **SLR.2 Random sampling:** The data (each x, y pair) are a random sample from the population of interest. Each data point therefore follows the population equation...
3. **SLR.3 Sample variation in the independent variable:** The independent variable has some variation in the sample: $\implies \sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$.
4. **SLR.4 Zero conditional mean:** The value of the explanatory variable must not contain information about the mean of the unobserved factors that might affect y: $\implies E(u_i|x_i) = 0$.

OLS Assumptions

Unbiasedness of OLS Estimators:

SLR.1 - SLR.4 $\implies$ $E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1$

Interpretation: If we were to draw many samples from the population and estimate the regression line in each sample, the average of the estimated slopes would be the true population slope.

- The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw. Some could be well-off from their true values too.
- However, on average, they will be equal to the values that characterize the true relationship between y and x in the population.

Unbiasedness of OLS Estimators (Adopted from URFIE, Heiss)

```r
set.seed(1234567) # Set the random seed

# set sample size and number of simulations
n <- 1000; r <- 10000
# set true parameters: betas and sd of u
b0 <- 1; b1 <- 0.5; su <- 2
# initialize b0hat and b1hat to store results later:
b0hat <- numeric(r); b1hat <- numeric(r)

x <- rnorm(n,4,1)# Draw a sample of x, fixed over replications:

for(j in 1:r) {# repeat r times:
# Draw a sample of y:
  u <- rnorm(n,0,su) # u changes with each draw
  y <- b0 + b1*x + u
# estimate parameters by OLS and store them in the vectors
  bhat <- coefficients( lm(y~x) )
  b0hat[j] <- bhat["(Intercept)"]; b1hat[j] <- bhat["x"]
}
```

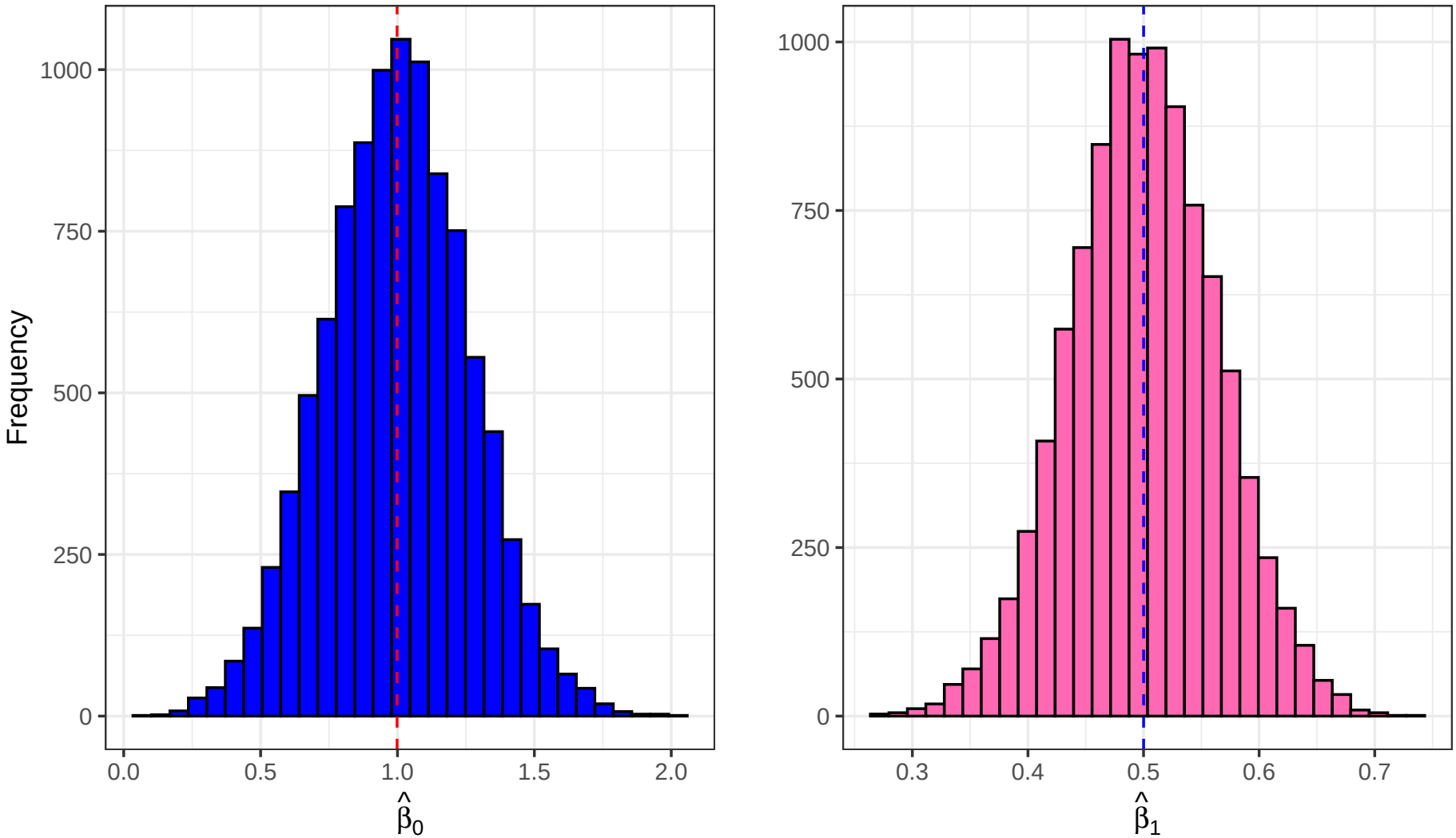
```r
#library(patchwork)

p1 <- ggplot() +
  geom_histogram(aes(x = b0hat), fill = "blue",
                 color = "black", bins = 30) +
  labs(title = NULL, x = expression(hat(beta)[0]), y = "Frequency") +
  geom_vline(xintercept = mean(b0hat), color = "red",
             linetype = "dashed") +
  theme_bw()

p2 <- ggplot() +
  geom_histogram(aes(x = b1hat), fill = "hotpink",
                 color = "black", bins = 30) +
  labs(title = NULL, x = expression(hat(beta)[1]), y = "") +
  geom_vline(xintercept = mean(b1hat), color = "blue",
             linetype = "dashed") +
  theme_bw()

(p1|p2) + plot_annotation(title = "Distribution of OLS estimators")
```

Distribution of OLS estimators



Proving Unbiasedness of OLS Estimators

Recall from earlier that:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} u_i$$

taking conditional expectations:

$$E(\hat{\beta}_1|x_i) = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cancelto{0}{E(u_i|x_i)}$$

$$E(\hat{\beta}_1|x_i) = \beta_1$$

Therefore, $\hat{\beta}_1$ is an unbiased estimator of the population slope coefficient, β_1.

In the case of $\hat{\beta}_0$:

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big(\beta_0 + \beta_1 x_i + u_i\Big) - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\cdot n\beta_0 + \frac{1}{n}\Big(\beta_1 - \hat{\beta}_1\Big)\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} u_i
\end{aligned}$$

taking conditional expectations:

$$E(\hat{\beta}_0|x_i) = \beta_0 + \boxed{\Big(\beta_1 - E(\hat{\beta}_1|x_i)\Big)}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} E\Big(u_i|x_i\Big)$$

$$E(\hat{\beta}_0|x_i) = \beta_0$$

Hence, $\hat{\beta}_0$ is an unbiased estimator of the population intercept term.

Variance of the OLS Estimators

- As we noted earlier, depending on the sample, the estimates will be nearer or farther away from the true population values.
- How far can we expect our estimates to be away from the true population values on average (= sampling variability)?
- Sampling variability is measured by the estimator's variances

$$var(\hat{\beta}_0), \quad var(\hat{\beta}_1)$$

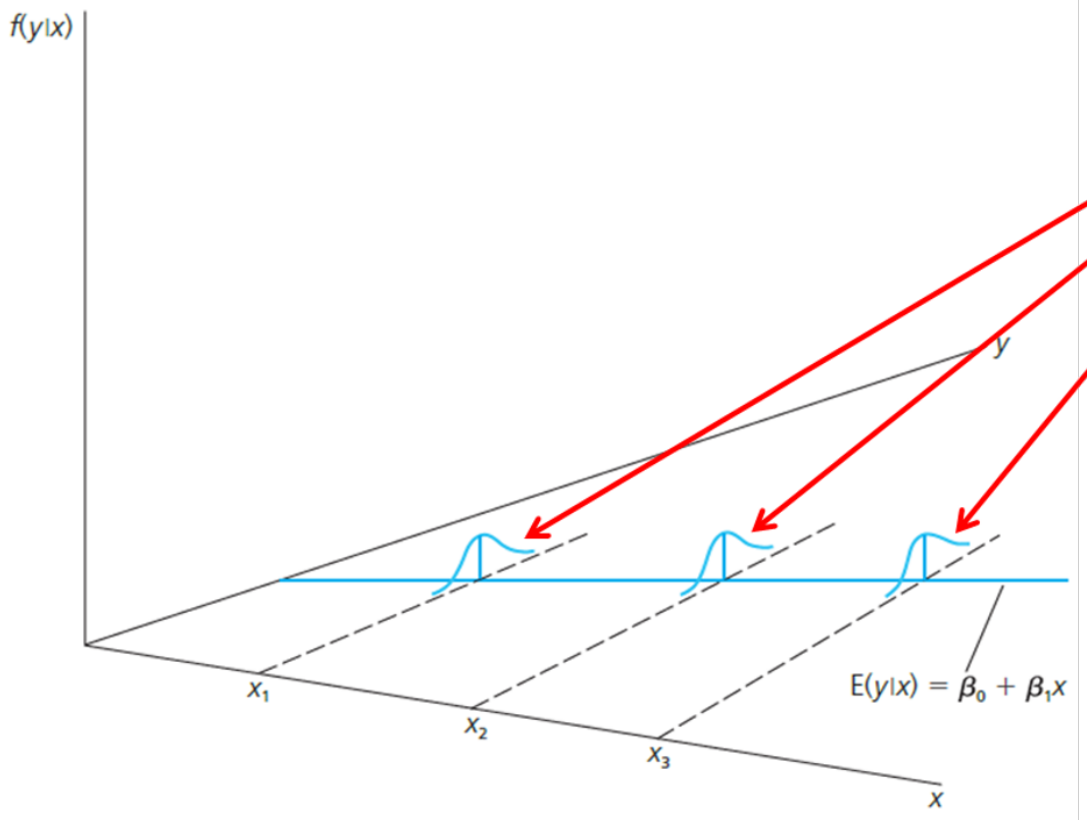
OLS Assumption:

5. **SLR.5 Homoskedasticity:** The error term has the same variance conditional on the independent variable:

$$\sigma^2 = var(u_i|x_i).$$

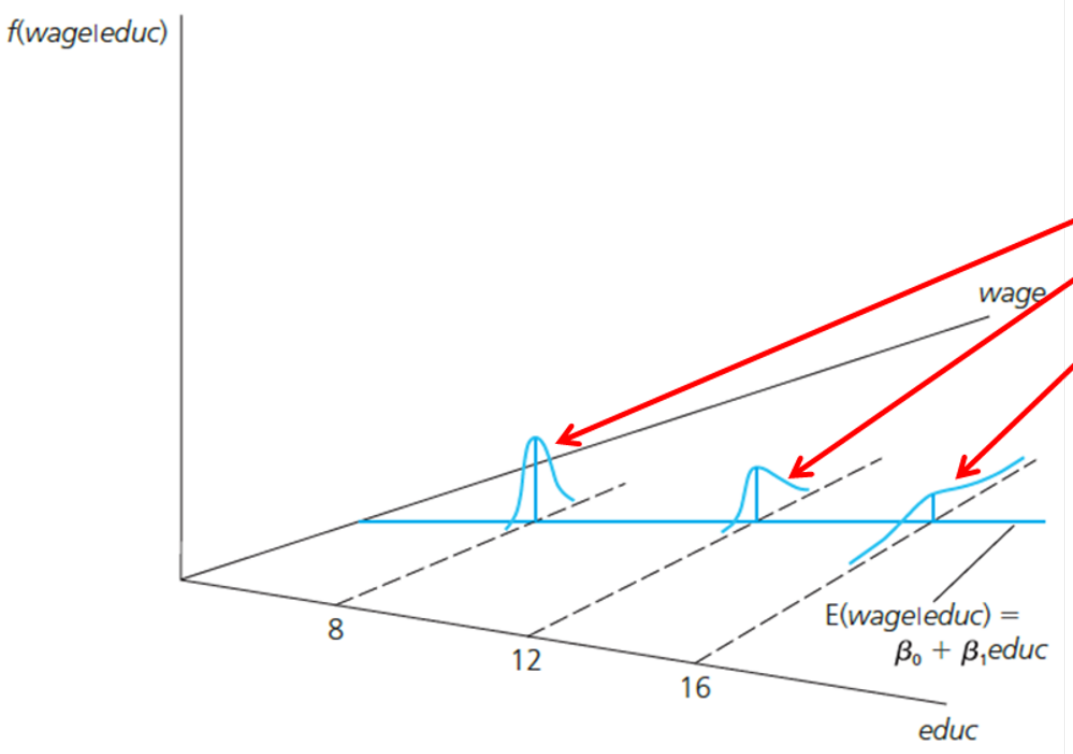
That is, the value of the xs must contain no information about the variability of the unobserved factors (variables in u).

Graphical illustration of homoskedasticity



The variability of the unobserved influences does not depend on the value of the explanatory variable

Graphical illustration of heteroskedasticity



The variance of the unobserved determinants of wages increases with the level of education

Variance of the OLS Estimators

SLR.1 – SLR.5

$$\implies var(\hat{\beta}_0) = \frac{\sigma^2}{n} \cdot \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Key Takeaway: The sampling variability of the estimated regression coefficients will be the higher the larger the variability of the unobserved factors, and lower, the higher the variation in the explanatory variable.

Variance of the OLS Estimators

Proof of $var(\hat{\beta}_1)$:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} u_i$$

running the variance operator through:

$$var(\hat{\beta}_1) = var\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}\left((x_i - \bar{x})^2\right)^2} var(u_i)\right)$$

$$var(\hat{\beta}_1) = 0 + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

Proof of $var(\hat{\beta}_0)$:

I leave this to you!

Estimating the Error Variance

Like the values of the parameters, the value of σ^2 is unknown and must be estimated from the data.

$$var(u_i|x_i) = \sigma^2 = E(\hat{u}_i - \bar{\hat{u}})^2 = E(\hat{u}_i)^2$$

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i - \bar{\hat{u}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 \tag{12}$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2} \tag{13}$$

- (12) is the sample variance of the residuals, however it **is biased**.
- (13) is the **unbiased** estimator of the population variance of the error term and is obtained by subtracting the number of estimated regression parameters from the sample size (number of obs).

Variance of the OLS Estimators

Substituting (13) into the respective formulas for the variances of the OLS estimators, we get:

$$\widehat{var(\hat{\beta}_0)} = \frac{\hat{\sigma}^2}{n} \cdot \frac{\sum_{i=1}^{n} x_i^2}{SST_x}$$

$$\widehat{var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{SST_x}$$

$$se(\hat{\beta}_0) = \sqrt{var(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2/n \cdot \sum_{i=1}^{n} x_i^2/SST_x}$$

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2/SST_x}$$

The estimated standard deviations of the regression coefficients are called “standard errors”. They measure how precisely the regression coefficients are estimated.