Intermediate Econometrics

Lecture 1

Vincenzo Gioia

14th October 2026

Regression Models: Examples

Some uses of regression models:

1. To **predict** an individual's **income** based on **gender**, holding other conditions constant (such as **education level, age**, etc.)
2. To **predict** the **number of exams** taken by a first-year student based on **demographic data, income, school background**, etc.
3. To **predict** the **number of claims** made by an insured person based on their **individual characteristics and past history**
4. To **assess** whether **blood pressure** decreases following the administration of a **drug**, taking **individual characteristics** into account
5. To **evaluate** how **mortality** in the population varies according to the concentration of **air pollutants**
6. To **decide** whether a **credit card payment is fraudulent**, based on certain characteristics of the transaction

Regression Models: A Common Framework

Framework:

- a quantity of interest (income, number of exams, number of claims, blood pressure, mortality, fraudulence): **the response (or outcome or dependent variable)**
- other quantities: **the explanatory (independent) variables (also called covariates or regressors)**

Question: **the first is related to the latter, and, if so, how?**

Note: The latter (covariates) are of practical interest only insofar as they are connected to the first (outcome)

Regression Models: Goals

Predictive

Obtain a tool for predicting the value a variable of interest given the values of the explanatory variables (e.g., because these are easier to measure or can be observed in advance with respect to the response)

- Example 3: when the goal is to determine the insurance premium that the policyholder should pay
- Example 6: to block a transaction before it is carried out

Interpretative

The main interest is to determine which explanatory variables have the strongest relationship with the response, and in which direction that relationship goes

- Example 1: when the goal is to determine whether there is gender-based discrimination
- Example 4: when the goal is to determine whether the drug is effective

Regression Models: General Form

The probability distribution of the outcome depends on the covariates

$$([\text{outcome}]) \sim f(y; [\text{covariates}])$$

- Asymmetric relationship
- $f(\cdot;\cdot)$ is specified up to a parameter

General structure depending on

1. **The type of outcome variable**
2. **The functional form of the relationship**

Regression Models: Type of Outcome

Different models

- Binary: (GLM) logistic/probit/... regression
- (Qualitative) Categorical: (GLM) multinomial regression (contingency table, log-linear models)
- Counts (Quantitative discrete): (GLM) Poisson (Negative Binomial) regression
- (Quantitative) Continuous: **Linear regression model**

Note: Under certain conditions the linear regression model can be used for quantitative discrete variables

Regression Models: Overview

Different models:

Covariates (X)	Outcome (Y)			
	Binary	**Nominal (>2)**	**Quantitative discrete**	**Quantitative continuous**
Nominal	**contingency tables** log-linear models **logistic regression**	**contingency tables** log-linear models **multinomial regression**	**Poisson regression** log-linear models	**t-test** / ANOVA (**multiple regression**)
Quantitative	**logistic regression**	**multinomial regression**	**Poisson regression**	**multiple regression**
Mixed	**logistic regression**	**multinomial regression**	**Poisson regression**	ANCOVA (**multiple regression**)

Regression Models: Data

Data Matrix

General structure including outcome and covariates

Unit	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	y_i	x_{i1}	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Regression Models: Linear Regression Model

$$Y_i \sim f(y_i; x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n$$

Additive structure

$$h(Y_i) = g(x_{i1}, \ldots, x_{ip}) + \varepsilon_i$$

- $h(\cdot)$: known function
- $g(x_{i1}, \ldots, x_{ip})$: systematic component
- ε_i: error component

Regression Models: Linear Regression Model

Linearity of $g(\cdot)$

$$g(x_{i1}, \ldots, x_{ip}) = \beta_1 g_1(x_{i1}) + \ldots + \beta_p g_p(x_{ip})$$

- $g_j(\cdot)$: known function

Linear model

$$h(Y_i) = \beta_1 g_1(x_{i1}) + \ldots + \beta_p g_p(x_{ip}) + \varepsilon_i$$

- $h(\cdot)$ and $g_j(\cdot)$: known functions
- ε_i: random variables with mean zero (whose distribution is specified up to a parameter)
- $\beta_1, \ldots, \beta_p$: parameters to estimate

Regression Models: Linear Regression Model

Examples

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

$$Y_i = \beta_1 + \beta_2 x_{i2}^2 + \beta_3 \sqrt{x_{i3}} + \varepsilon_i$$

$$\log(Y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

$$\log(Y_i) = \beta_1 + \beta_2 \log(x_{i2}) + \beta_3 \log(x_{i3}) + \varepsilon_i$$

Regression Models: Interpretation - Be Careful!

Relationship should not be interpreted as Cause-and-Effect

When we write a model in which one variable (y) is a function of another (x), it is very tempting to interpret it as x causes y

- **A statistical relationship — even a strong one — between y and x does not imply a cause-and-effect relationship**
- For example, both variables might be related to a third variable that causes them both
- There are statistical methods for making inferences about cause-and-effect relationships, but they require greater sophistication or a sample constructed in a specific way

Regression Models: Linear Model

Linear model: $Y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$

Matrix representation $Y = X\beta + \varepsilon$

- Y: n-dimensional outcome vector
- X: $n \times p$ model matrix

Unit	x_1	x_2	$\cdots$	x_p
1	1	x_{12}	$\cdots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	1	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	1	x_{n2}	$\cdots$	x_{np}

Linear Model: Error Term

Linear model: $Y = X\beta + \varepsilon$

- ε: n-dimensional error vector (**this component introduces casuality in the model**, Y is random beacuse ε is random)

Assumptions

1. **Linearity**
2. **Errors having mean 0, homoschedastic, and uncorrelated**
3. **Linear indepencence between explanatory variables**

We do not make distributional assumptions on the error components: these are called **second-order hypotheses**

Linear Model: Assumption 1

Linearity

1. Linearity: $Y = X\beta + \varepsilon$

$$\underset{(n\times 1)}{\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}} = \underset{(n\times p)}{\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}} \underset{(p\times 1)}{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}} + \underset{(n\times 1)}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}}$$

Linear Model: Assumption 2

Second-order hypothesis for the error term

2. Errors having mean 0, homoschedastic, and uncorrelated

$$\mathbb{E}(\varepsilon) = 0 \qquad V(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I$$

- $E[Y|X] = X\beta \qquad V(Y|X) = \sigma^2 I$

Linear Model: Assumption 3

Linear indepencence between explanatory variables

3. The vectors x_j, $j = 1, \ldots, p$, are linearly independent

- **This guarantees the identifiability of the model**
- **It translates into a matrix** X (which is non-stochastic because we are working conditional to the values observed for the covariates) **that is of full rank** $(\mathrm{rank}(X) = p)$

Linear Model: Least Square Estimator

Least Square (LS) Estimator (OLS: ordinary least square)

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$$

LS estimator is obtained by minimizing the residual sum of squares

$$\mathrm{RSS}(\beta) = (Y - X\beta)^\top (Y - X\beta) = Y^t Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta$$

$$\frac{\partial}{\partial \beta}\mathrm{RSS}(\beta) = -2X^\top Y + 2X^\top X\beta$$

$$\frac{\partial}{\partial \beta}\mathrm{RSS}(\beta) = 0 \implies \hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$$

Linear Model: Important quantities

LS estimate

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$

Predicted values

$$\hat{y} = X\hat{\beta}_{OLS} = X(X^\top X)^{-1} X^\top y = Py$$

where $P = X(X^\top X)^{-1} X^\top$ is called projection matrix (symmetric and idempotent)

Residuals

$$e = y - \hat{y} = (I - P)y$$

Linear Models: OLS properties

Properties of $\hat{\beta}_{OLS}$

- Unbiasedness: $E(\hat{\beta}_{OLS}) = \beta$
- $V(\hat{\beta}_{OLS}) = \sigma^2(X^\top X)^{-1}$

We need an estimate of σ^2 (which is unknown)

- **The idea is to use the residuals as substitutes for the errors and to use their variance as an estimator of σ^2**
- $\hat{\sigma}^2 = \frac{1}{n} e^\top e$, which is biased
- A consistent estimate is given by

$$S^2 = \frac{1}{n-p} e^\top e$$

- This implies that the variance/covariance of the $\hat{\beta}_{OLS}$ estimator is

$$\hat{V}(\hat{\beta}_{OLS}) = S^2(X^\top X)^{-1}$$

Linear Models: R^2 coefficient

How well the pedicted values $\hat{y}$ are able to represent the observed data y

- Measure of goodness of fit: R^2 coefficient

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $R^2 \in [0, 1]$
- It represents the fraction of variability of Y explained by the model

Deviance decomposition

$$\text{Total deviance} = \text{Model deviance} + \text{Residual deviance}$$

- $\text{Total deviance} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $\text{Model deviance} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $\text{Residual deviance} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Linear Models: In practice

Wage Data

The dataset Wage is included in the ISLR package

- Description: Wage and other data for a group of 3000 male workers in the Mid-Atlantic region.
- Outcome: wage (or lwage)
- Available Covariates: year, age, maritl, race, education, region, jobclass, health, health_ins

We just considered a simple example regressing wage on age, year and education

```
library(ISLR)
```

```
## Warning: il pacchetto 'ISLR' è stato creato con R versio
```

```
data("Wage")
#help(Wage) you can see the help
```

Linear Models: In practice

Data structure

Using `str()` we can see the structure of the dataset (number of instances and variables), by identifying the type of a variable:

- Integer: year and age
- Factor: maritl, race, education, region, jobclass, health, health_ins
- Numeric: wage and lwage

```
str(Wage)
```

```
## 'data.frame':    3000 obs. of  11 variables:
##  $ year       : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 200
##  $ age        : int  18 24 45 43 50 54 44 30 41 52 ...
##  $ maritl     : Factor w/ 5 levels "1. Never Married",..: 1 1 2 2 4 2
##  $ race       : Factor w/ 4 levels "1. White","2. Black",..: 1 1 1 3
##  $ education  : Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3
##  $ region     : Factor w/ 9 levels "1. New England",..: 2 2 2 2 2 2 2
##  $ jobclass   : Factor w/ 2 levels "1. Industrial",..: 1 2 1 2 2 2 1
##  $ health     : Factor w/ 2 levels "1. <=Good","2. >=Very Good": 1 2
##  $ health_ins: Factor w/ 2 levels "1. Yes","2. No": 2 2 1 1 1 1 1 1
##  $ logwage    : num  4.32 4.26 4.88 5.04 4.32 ...
##  $ wage       : num  75 70.5 131 154.7 75 ...
```

Linear Models: In practice

Exploratory purposes (continuous variables)

- Univariate: summary

To get the summary (minimum, 1st quartile, mean, median, 3rd quartile, maximum) we use the function `summary`, that can be applied to the a set of variables (or the entire datset), as well as to the single variables. Other numerical information of interest can be the standard deviation, the IQR, etc.

```
summary(Wage$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.09   85.38  104.92  111.70  128.68  318.34
```

```
sd(Wage$wage)
```

```
## [1] 41.7286
```

```
IQR(Wage$wage)
```

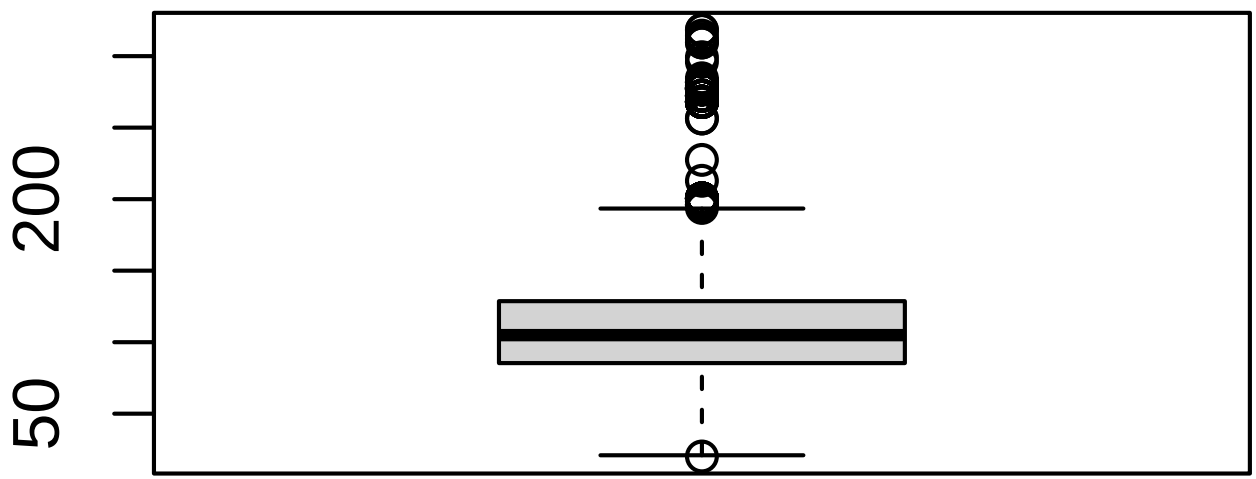
```
## [1] 43.29655
```

Linear Models: In practice

Exploratory purposes (continuous variables)

- Univariate: Boxplot To get a boxplot of a continuous variables we can use the function `boxplot`

```
boxplot(Wage$wage)
```

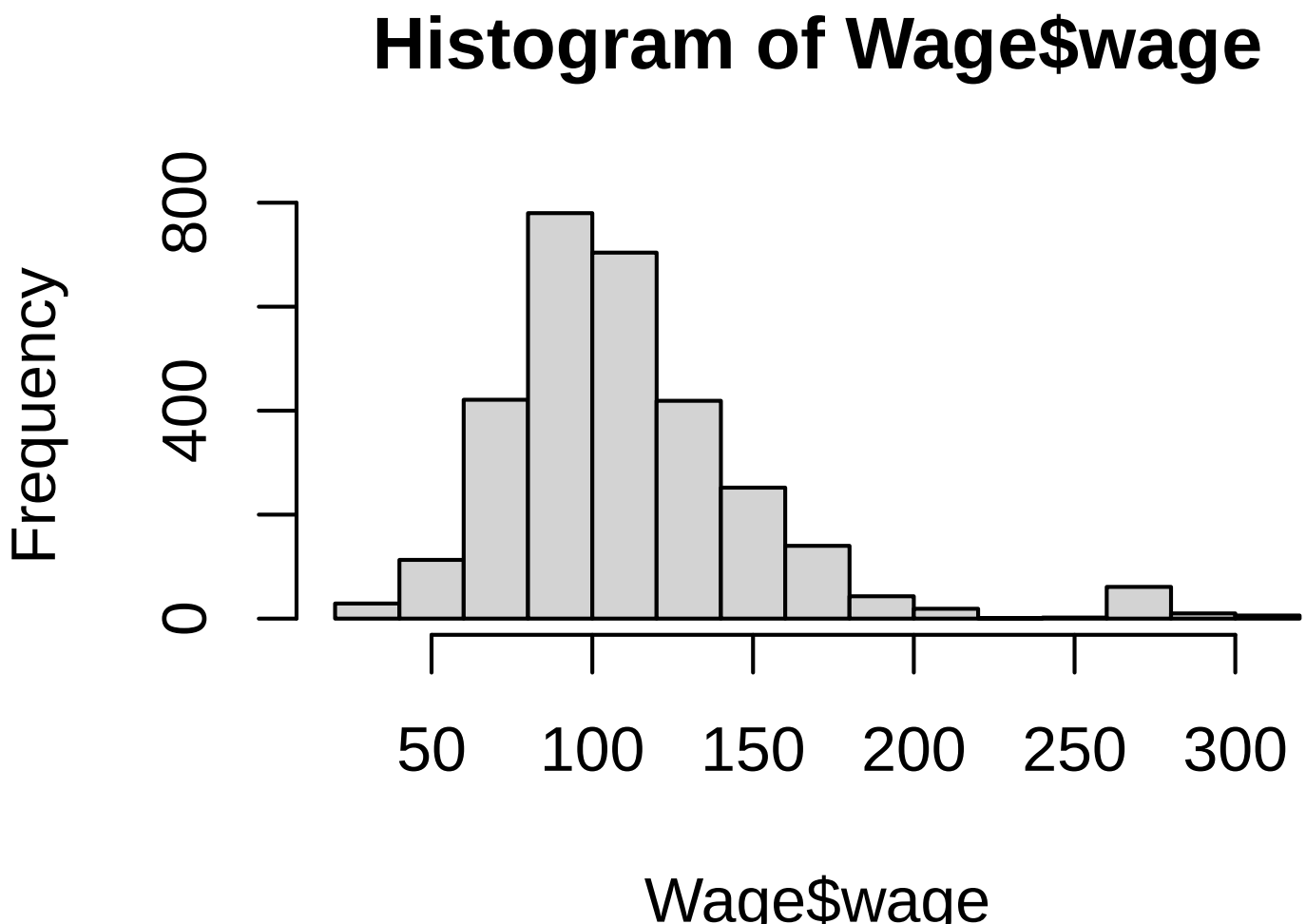


Linear Models: In practice

Exploratory purposes (continuous variables)

- Univariate: histogram To get a histogram `boxplot`

```
hist(Wage$wage)
```

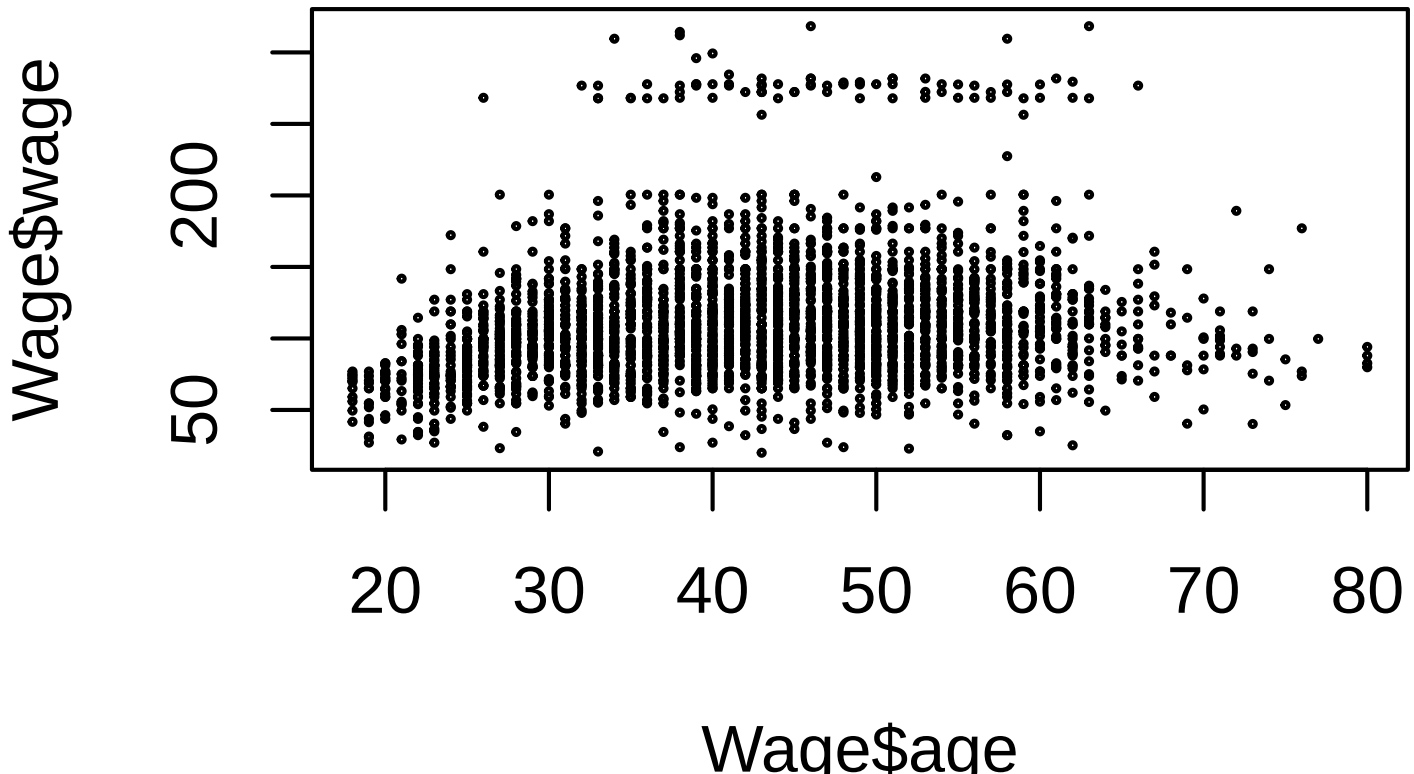


Linear Models: In practice

Exploratory purposes (continuous variables)

- Bivariate: Scatterplot (age, wage)

```
plot(Wage$age, Wage$wage, cex = 0.2)
```

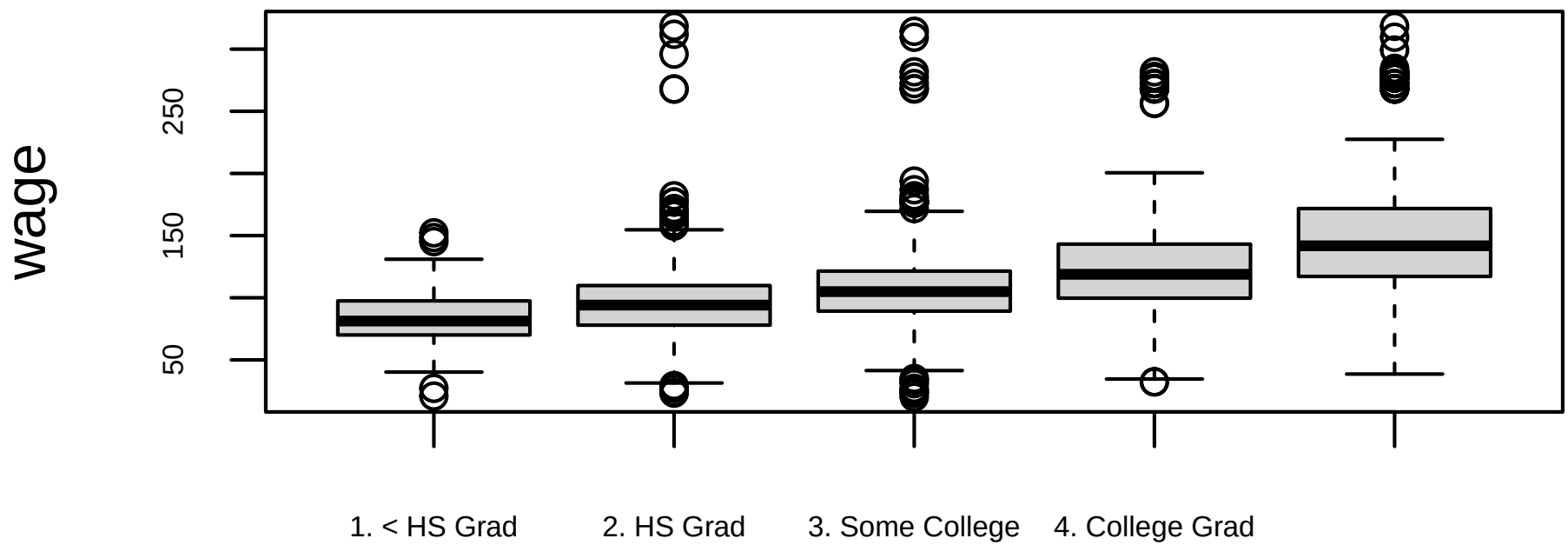


Linear Models: In practice

Exploratory purposes (conditional distribution)

- Bivariate: Boxplot (wage|education)

```
boxplot(wage ~ education, data = Wage,
        cex.axis = 0.5)
```



Linear Models: In practice

The lm() function: fit a linear model on your dataset

- Model formula
- Specifying the dataset (where the variables can be found)
- Assign it to an object (digitizing the object you can see only the estimate)

```
lmFit <- lm(wage ~ age + year + education, data = Wage)
lmFit
```

```
##
## Call:
## lm(formula = wage ~ age + year + education, data = Wage)
##
## Coefficients:
##                  (Intercept)                          age
##                    -2058.0770                       0.5621
##                         year           education2. HS Grad
##                       1.0563                      11.3976
##    education3. Some College      education4. College Grad
##                      24.2345                      39.7408
## education5. Advanced Degree
##                      64.8537
```

Linear Models: In practice

To see an extensive summary you can apply the `summary` function to the `lm` object.

```
summary(lmFit)
```

```
##
## Call:
## lm(formula = wage ~ age + year + education, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.323  -19.521   -3.964   14.438  219.172
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.058e+03  6.493e+02  -3.169  0.00154 **
## age                          5.621e-01  5.714e-02   9.838  < 2e-16 ***
## year                         1.056e+00  3.238e-01   3.262  0.00112 **
## education2. HS Grad          1.140e+01  2.476e+00   4.603 4.34e-06 ***
## education3. Some College     2.423e+01  2.606e+00   9.301  < 2e-16 ***
## education4. College Grad     3.974e+01  2.586e+00  15.367  < 2e-16 ***
## education5. Advanced Degree  6.485e+01  2.804e+00  23.128  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.89 on 2993 degrees of freedom
## Multiple R-squared:  0.2619, Adjusted R-squared:  0.2604
## F-statistic:   177 on 6 and 2993 DF,  p-value: < 2.2e-16
```

Linear Models: In practice

- You can extract the table coefficients or simply the coefficients easily

```
summary(lmFit)$coefficients
```

```
##                                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)                -2058.0770290 649.34597951 -3.169461  1.542629e-03
## age                            0.5621005   0.05713794  9.837604  1.694287e-22
## year                           1.0562995   0.32377790  3.262420  1.117059e-03
## education2. HS Grad           11.3976404   2.47630123  4.602687  4.344327e-06
## education3. Some College      24.2344792   2.60564592  9.300757  2.611570e-20
## education4. College Grad      39.7408035   2.58616318 15.366704  2.377262e-51
## education5. Advanced Degree   64.8536598   2.80416339 23.127632 5.135751e-109
```

```
lmFit$coefficients
```

```
##                  (Intercept)                          age
##                -2058.0770290                    0.5621005
##                         year          education2. HS Grad
##                    1.0562995                   11.3976404
##     education3. Some College      education4. College Grad
##                   24.2344792                   39.7408035
## education5. Advanced Degree
##                   64.8536598
```

Linear Models: In practice

Obtaining the parameter LS estimates by hand

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$

```
X <- model.matrix(wage ~ age + year + education, data = Wage)
y <- Wage$wage
beta_OLS <- solve(t(X)%*%X)%*%t(X)%*%y
beta_OLS
```

```
##                                      [,1]
## (Intercept)                -2058.0770291
## age                            0.5621005
## year                           1.0562995
## education2. HS Grad           11.3976404
## education3. Some College      24.2344792
## education4. College Grad      39.7408035
## education5. Advanced Degree   64.8536598
```

Linear Models: In practice

Obtaining the estimate of the standard deviation of the LS estimator (second column of the table of coefficients)

$$\sqrt{[\hat{V}(\hat{\beta}_{OLS})]_{jj}} = \sqrt{[s^2(X^\top X)^{-1}]_{jj}}$$

```
p <- ncol(X)
hat_s2 <- sum(residuals(lmFit)^2)/(nrow(Wage) - p)
hat_Vb <- hat_s2*solve(t(X)%*%X)
sqrt(diag(hat_Vb))
```

```
##                  (Intercept)                                  age
##                649.34597952                           0.05713794
##                         year               education2. HS Grad
##                  0.32377790                           2.47630123
##    education3. Some College          education4. College Grad
##                  2.60564592                           2.58616318
## education5. Advanced Degree
##                  2.80416339
```

Linear Models: Residuals summary

Residuals: $e = y - \hat{y}$

- A summary of the residuals
- Residual standard error (the square root of the unbiased estimate of the variance of the error term, σ^2)

```
head(residuals(lmFit))
```

```
##     231655      86582     161300     155159      11443     376662
##   4.065485 -41.502457  23.762215  33.083207 -24.263071  -5.950945
```

```
head(Wage$wage - predict(lmFit))
```

```
##     231655      86582     161300     155159      11443     376662
##   4.065485 -41.502457  23.762215  33.083207 -24.263071  -5.950945
```

```
summary(residuals(lmFit))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -113.323  -19.521   -3.964    0.000   14.438  219.172
```

```
summary(lmFit)$sigma
```

```
## [1] 35.88547
```

```
sqrt(sum(residuals(lmFit)^2)/(nrow(Wage)-ncol(X)))
```

```
## [1] 35.88547
```

Linear Models: Residuals summary

R^2 coefficient and the adjusted R-squared (R_c^2)

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R_c^2 = R^2 - \frac{p-1}{n-p}(1 - R^2)$$

Exercise: try to obtain the R^2 coefficient and the adjusted R-squared (R_c^2) by hand

Linear Models: Further steps

Remember: This is just a toy example

- We are just using this data and this model setting to show how some quantities are obtained

Obtaining the remaining quantitities of the summary

- t value
- $\Pr(>|\mathrm{t}|)$
- F-statistic

Explore violations of the linear model assumptions

- Residual analysis
- ...

What we need?

- A further assumption

Linear Models: Statistical Inference

Problems of Statistical Inference

- Estimation: **point** and interval estimation
- Hypothesis test
- Prediction (**point** or interval)

By deriving the OLS estimator we have just obtained a point estimator and derived its variance (which provide information on how the estimator is far from the unknown parameter β)

To carry out the remaining inferential results (interval estimation, hypothesis test), we need to

1. Use the asymptotic theory of the least squares or resampling techinque
2. **Introduce a further assumption: the errors are distributed according to a $\mathcal{N}(0, \sigma^2)$ and they are independent**

Normal Linear Model

Assumptions

1. Linearity
$$Y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$
2. Errors having mean zero, homoschedastic, normally distributed and independent
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \text{independent}, \quad i = 1, \ldots, n$$
$$\varepsilon \sim \mathcal{N}_n(0, \Sigma), \quad \text{where} \quad \Sigma = \sigma^2 I$$
3. Linear independence between explanatory variables

From 1) and 2), we get that the Y_i are independent with

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{where} \quad \mu_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, n$$

Normal Linear Model: What we need?

By introducing the normality assumption for the errors we can derive

- Distribution of the estimators ($\hat{\beta}$ and $\hat{\sigma}^2$)
- Joint distribution of $(\hat{\beta}, \hat{\sigma}^2)$
- Pivotal quantity to make inference on a single coefficient (confidence interval, hypothesis test)
- Procedure for testing hypothesis on a group of coefficient and make predictions