

Lecture 4

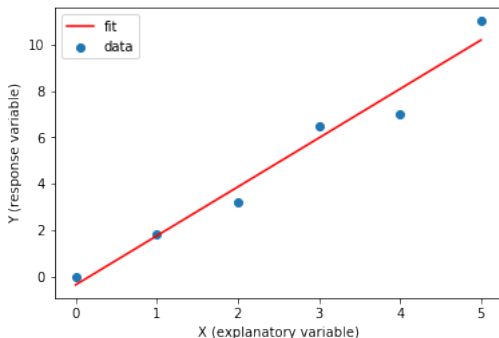
LINEAR REGRESSION I

Dr Adam Mahdi
University of Oxford

Oxford, 28 October 2021

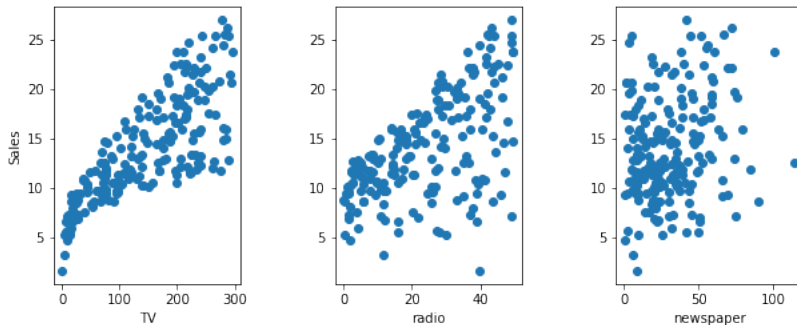
What is linear regression?

- It is a linear approach to modelling the relationship between a response variable Y and one or more explanatory variables X_1, \dots, X_p .



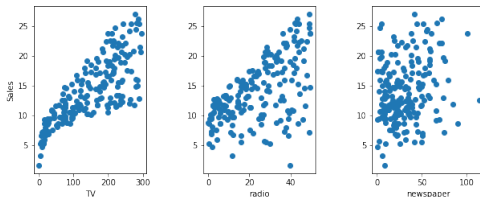
- In practice, the regression functions are never linear!!

Example: Advertising data



Source: G James, D Witten, T Hastie, R Tibshirani *An Introduction to Statistical Learning*, Springer 2013.

Example: Advertising data



Some questions we might want to ask about the **Advertising Data**:

- Is there a relationship between advertising budget and sales?
- If yes, how strong is this relationship?
- Is this relationship linear?
- Which media (TV, radio, newspaper) contribute to sales most?
- Is there a synergy (interaction) between the media?

Simple Linear Regression

Simple Linear Regression

Simple means that we consider only one scalar explanatory variable and one scalar response variable.

In **Simple Linear Regression** we assume the following statistical model:

$$Y = \alpha + \beta X + \varepsilon$$

- X is the **explanatory variable**; the constant coefficients α is the **intercept** and β is the **slope**; and ε is the **error** term.
- ε represents all the omitted causes of Y beyond X .
- one unit increase in X gives β units increase in Y .

Simple Linear Regression

- For specific x_i , it can be written as:

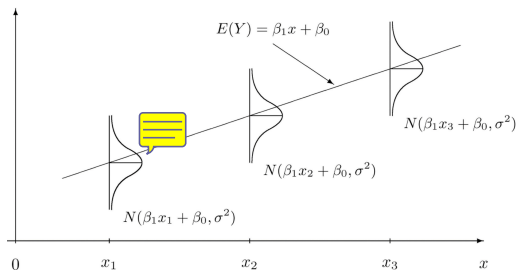
$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n$$

- ε_i is the **statistical error**.
- Note that y_i almost equals $\alpha + \beta x_i$; the difference is the random quantity ε_i .
- The statistical errors ε_i are *unobserved*.

Assumptions

For $i = 1, \dots, n$ we assume

- **Linearity:** $E(\varepsilon_i) = 0$
- **Constant variance:** $\text{Var}(\varepsilon_i) = \sigma^2$
- **Independence:** $\varepsilon_i, \varepsilon_j$ are independent for $i \neq j$
-



More about it later...

Distinction Between *Estimators* and *Estimates*

- **Estimator:** formula or rule or procedure by which a numerical estimate of an unknown population parameter is computed from any given sample data.
- **Estimate:** is the value of the estimator obtained when the formula is evaluated for a particular set of sample values of the observable variables.

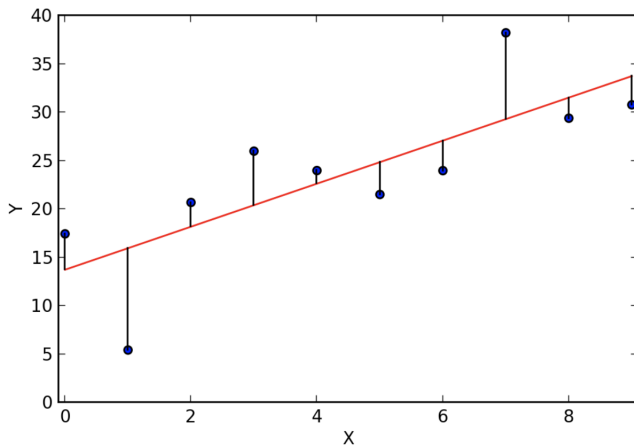
Simple Linear Regression

- The population parameters α , β and σ are unknown.
- Our goal is to estimate them: $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}$.
- $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ is the **fitted** (or **predicted**) **value**.
- $e_i = \hat{y}_i - y_i$ is called the ***i*th residual**.
- Thus given the data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ we can write

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i, \quad i = 1, \dots, n$$


- The residuals are observable, and can be used to check the assumptions on the statistical errors ε_i .

Simple Linear Regression



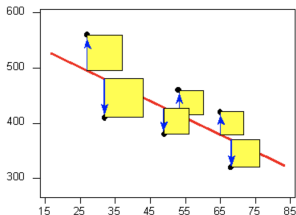
Simple Linear Regression

Goal: We want the line to be 'as close' to the data as possible. A line that fits the data well has small residuals.

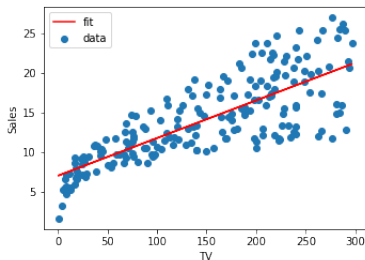
- Can we simply minimise $\sum_{i=1}^n e_i$?
 - * No, because for any line through the means (\bar{x}, \bar{y}) we have $\sum_{i=1}^n e_i = 0$.
- How about instead minimising $\sum_{i=1}^n |e_i|$ or $\sum_{i=1}^n e_i^2$?
 - * $\sum_{i=1}^n |e_i|$ although this is good for outliers, it is not very convenient to work with mathematically (why?). 
 - * $\sum_{i=1}^n e_i^2$ is our choice! Seems the best trade-off.

Least squares in simple linear regression

- minimising residual sum of squares



- the linear regression of sales onto TV



Estimating the coefficients α and β

Find $\hat{\alpha}, \hat{\beta}$ so the following **residual sum of squares (RSS)** is minimised:

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

We can find $\hat{\alpha}$ and $\hat{\beta}$ by applying simple calculus:

$$\begin{aligned}\partial RSS / \partial \hat{\alpha} &= 0, \quad \Rightarrow \quad \hat{\alpha}n + \hat{\beta} \sum x_i = \sum y_i \\ \partial RSS / \partial \hat{\beta} &= 0, \quad \Rightarrow \quad \hat{\alpha} \sum x_i + \hat{\beta} \sum x_i^2 = \sum x_i y_i\end{aligned}$$

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

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ are the sample means.

Properties of residuals

We have the following useful properties of the residuals:

Properties (residuals)

- $\sum_{i=1}^n e_i = 0$
- $\sum_{i=1}^n x_i e_i = 0$
- $\sum_{i=1}^n \bar{y}_i e_i = 0.$

Properties of the least-square estimators (1)

Exercise (important)

Show that the slope $\hat{\beta}$ can be written as the linear combination of y_i :

$$\hat{\beta} = \sum_{i=1}^n w_i y_i, \quad \text{where} \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Exercise (useful properties)

Show the following useful equalities involving the coefficients w_i 's:

- (a) $\sum_{i=1}^n w_i = 0$
- (b) $\sum_{i=1}^n w_i (x_i - \bar{x}) = 1$
- (c) $\sum_{i=1}^n w_i x_i = 1$

Properties of the least-square estimators (2)

Exercise (unbiased estimators)

Assume $E[\varepsilon_i] = 0$ and $\text{Var}[\varepsilon_i] = \sigma^2$. Show that

(a) $E[\hat{\alpha}] = \alpha, \quad E[\hat{\beta}] = \beta$

(b) $\text{Var}[\hat{\alpha}] = \frac{\sigma^2 S_x}{n S_{xx}}, \quad \text{Var}[\hat{\beta}] = \frac{\sigma^2}{S_{xx}}$

where

$$S_{xx} := \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{and} \quad S_x := \sum_{i=1}^n x_i^2.$$

Properties of the least-square estimators (3)

Exercise (distributions)

Assume $E[\varepsilon_i] = 0$, $\text{Var}[\varepsilon_i] = \sigma^2$ and $\varepsilon_i \sim N(0, \sigma^2)$. Show that

(a) $\hat{\alpha} \sim N(\alpha, \sigma^2 S_x / (n S_{xx}))$

(b) $\hat{\beta} \sim N(\beta, \sigma^2 / S_{xx})$

where

$$S_{xx} := \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{and} \quad S_x := \sum_{i=1}^n x_i^2.$$

Parameters α, β and estimators $\hat{\alpha}, \hat{\beta}$

α and β

- parameters
- there is no formula for α and β
- their values are unknown (not computed)

$\hat{\alpha}$ and $\hat{\beta}$

- estimators
- there are formulas for $\hat{\alpha}$ and $\hat{\beta}$
- their values are computed from the data

- correlation coefficient
- residual standard error
- coefficient of determination
- confidence intervals
- hypothesis test
- F-statistics

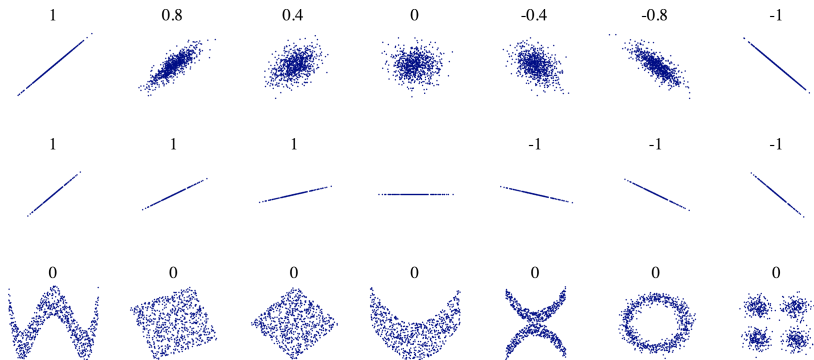
Assessing the model: correlation coefficient (r)

The **correlation coefficient** (also called **Pearson correlation coefficient**) applied to a sample $\{(x_i, y_i), i = 1, \dots, n\}$ is defined as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- *Relative* measure of “fit” and linear correlation between X and Y
- Assumes the values $-1 \leq r \leq 1$, where
 - * +1 total positive linear correlation
 - * 0 no linear correlation
 - * -1 total negative linear correlation
- Note that $(x_i - \bar{x})(y_i - \bar{y})$ is positive if and only if x_i and y_i lie on the same side of their respective means. Thus the correlation coefficient is positive if x_i and y_i tend to be simultaneously greater than, or simultaneously less than, their respective means.

Assessing the model: correlation coefficient (r)



Assessing the model: standard error of the regression ($\hat{\sigma}$)

The **standard error of the regression** (also called **residual standard error**) is defined as:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \sqrt{\frac{RSS}{n-2}}$$

- $n - 2$ are the degrees of freedom (because we estimate α and β)
- *Absolute* measure of fit
- Intuition: type of “average” residual
- Measured in the units of the dependent variable
- Caution: This is not a standard error, i.e. estimated standard deviation of the sampling distribution of a statistic !!!

Assessing the model: standard error of the regression ($\hat{\sigma}$)

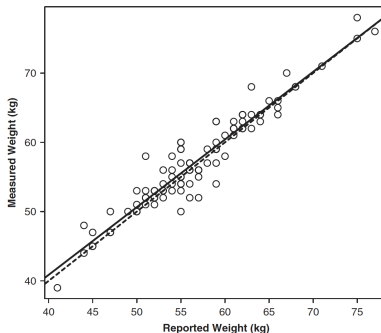


Figure: Scatterplot of Davis's data on the measured and reported weight (given to the nearest kilogram) of 101 people.

- Interpretation: For the Davis's data we have $\hat{\sigma} \approx 2$.
 - on average, using the least squares regression line to predict weight from reported weight, results in an error of about 2 kg:

Sums of Squares

We define the following sum of squares:

- Total Sum of Squares :
- Regression Sum of Squares:
- Residual Sum of Squares:

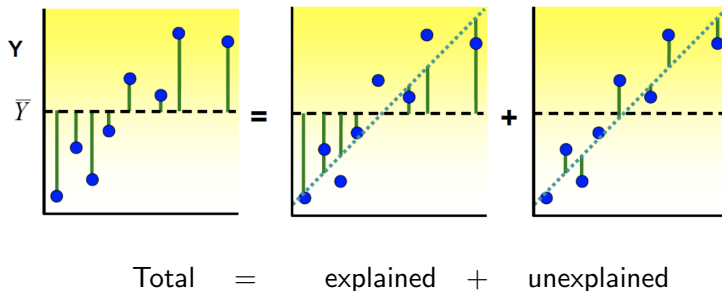


$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

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

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

What do they represent?



Variance Decomposition

We can decompose total variation into **explained** and **unexplained** components:


$$TSS = RegSS + RSS$$

Intuition: think of decomposition of each observation into a fitted value and a residual.

$$\begin{aligned} TSS &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y} + \hat{y} - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y})^2 + \sum_{i=1}^n (\hat{y} - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y})(\hat{y} - \bar{y}) \\ &= RSS \quad + \quad RegSS \quad + \quad 0 \end{aligned}$$

Assessing the model: Coefficient of Determination

The **coefficient of determination** is defined as


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

Note some properties:

- R^2 always takes the values between 0 and 1 ($0 \leq R^2 \leq 1$).
- A value close to 1 indicates that a significant portion of the variability in the response variable has been explained by the linear regression and 0 the opposite.
- It is impossible to give rigid rules as to what values of R^2 constitute a good and poor fit and, in general, it will depend on the specific problem.

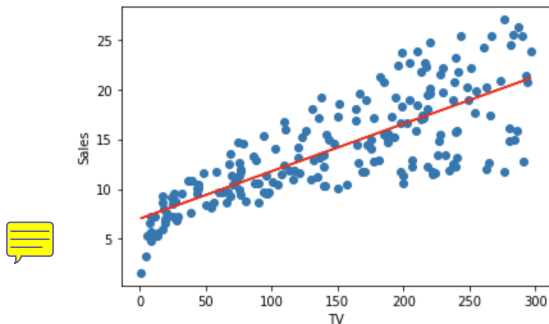


Exercise

Show that in Simple Linear Regression we have $r^2 = R^2$.

Assessing the model: confidence intervals

Why computing **confidence intervals** on the regression coefficients?



- Recall that the standard error of an estimator reflects how it varies under repeated sampling

$$SE(\hat{\beta})^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad SE(\hat{\alpha})^2 = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

- $\sigma^2 \approx \hat{\sigma}^2 = \frac{RSS}{n-2}$ (residual standard error)

Assessing the model: confidence intervals

A $100(1 - \alpha)\%$ **confidence interval** for the coefficients α , β :

$$\hat{\beta} \pm t_{\alpha/2, n-2} \text{SE}(\hat{\beta}) \quad \text{and} \quad \hat{\alpha} \pm t_{\alpha/2, n-2} \text{SE}(\hat{\alpha})$$

where $t_{\alpha/2, n-2}$ is the critical value of the t-distribution with $n - 2$ degrees of freedom; and $\text{SE}(\hat{\beta})$ is the standard error of $\hat{\beta}$.

- For **95%** CI it is approx. $\hat{\beta} \pm 2 \cdot \text{SE}(\hat{\beta})$ and $\hat{\alpha} \pm 2 \cdot \text{SE}(\hat{\alpha})$

Assessing the model: hypothesis test

- The question of the relationship between X and Y involves checking whether the slope is different from zero, i.e. $\beta \neq 0$.
- Note that if $\beta = 0$, then the simple linear regression model becomes $Y = \alpha + \epsilon$ showing no association of X with Y .
- Thus we are constructing the following **hypothesis test**:

$H_0 : \beta = 0$ (no relationship between X and Y)

$H_A : \beta \neq 0$ (there is a relationship between X and Y)

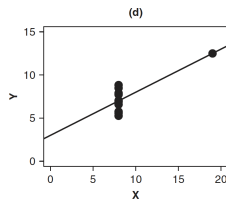
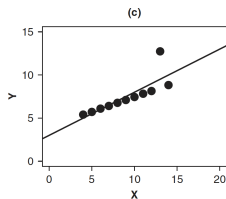
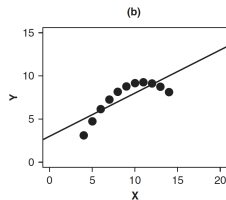
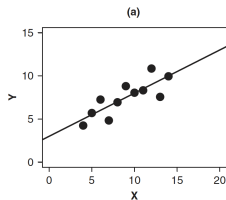
- In practice we use t-statistics

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

the probability (p-value) of observing any value equal to $|t|$ or larger.

Labs: Simple Linear Regression

Lab 3a: Anscombe plot (1973)



Multiple Linear Regression

Multiple Linear Regression

Multiple means that we consider ≥ 2 independent variables.

In **Multiple Linear Regression** we assume the following statistical model:

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

- α and β_1, \dots, β_p are unknown **population** coefficients that must be **estimated**.
- ε represent all the omitted causes of Y beyond the explanatory variables X_1, \dots, X_p .
- the model coefficient β_j can be thought of as the average effect on Y of a unit increase in X_j keeping all other predictors fixed.

Multiple Linear Regression

- For the **data**:

$$\{(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{np}, \dots, x_{np}, y_n)\}$$

- the **statistical model** is:

$$y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n$$

- ε_i are called **statistical errors**.
- Note that y_i almost equals $\alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$; the difference is the random quantity ε_i .
- The statistical errors ε_i are *unobserved*.

Multiple Linear Regression

- The population parameters $\alpha, \beta_1, \dots, \beta_p$ and σ are unknown
- Our goal is to estimate them: $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_p$ and $\hat{\sigma}$
- $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}$ is the **fitted** (or **predicted**) **value**
- $e_i = \hat{y}_i - y_i$ is called the ***i*th residual**

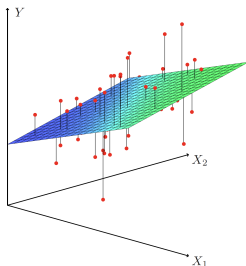
Thus we can write

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip} + e_i, \quad i = 1, \dots, n$$

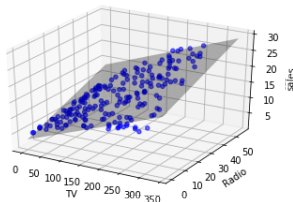
- The residuals are observable
- Note the difference between statistical errors ε_i (unobserved) and residuals e_i (observed)

Least squares in multiple linear regression (two variables)

- minimizing residual sum of squares



- Least square regression of sales onto TV and radio



Estimating the coefficients α and β_1, \dots, β_p

$$\text{minimise } RSS = \text{minimise } \sum_{i=1}^n \left(y_i - \hat{\alpha} - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_p x_{ip} \right)^2.$$

Define the following matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

We can rewrite the equation $y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip} + e_i$, $i=1, \dots, n$ as

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}.$$

Minimising RSS with respect to $\hat{\boldsymbol{\beta}}$ gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

where \mathbf{X}^\top is the transpose and \mathbf{X}^{-1} is the inverse of the matrix \mathbf{X} .

Properties of residuals

We have the following useful properties of the residuals:

- $\sum_{i=1}^n e_i = 0$
- The residuals e_i are uncorrelated with the fitted values \hat{y}_i and with each of the independent variables x_1, \dots, x_p .
- The **standard error of the regression** (or **residual standard error**) $\hat{\sigma} = \sqrt{\sum_{i=1}^n e_i^2 / (n - p - 1)}$ is an "average" size of the residuals.
- $n - p - 1$ is the **degree of freedom** (since we estimate $p + 1$ parameters $\alpha, \beta_1, \dots, \beta_p$).

We have again:

- $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$, $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, $ReggSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- $R^2 = ReggSS / TSS = 1 - RSS / TSS$ is the proportion of the variation in Y that is captured by its linear regression the X 's.
- Problems with R^2 :
 - R^2 never decreases as we keep on adding more variables (why?).
 - insensitive to overfitting
- Adjusted R^2 denote (R^2_{adj} or \bar{R}^2) is defined as

$$R^2_{adj} = 1 - (1 - R^2) \frac{RSS(n-1)}{TSS(n-p-1)}$$

- We want to compare the coefficients of different variables.
- This is straightforward when the independent variables are measured **in the same units**.
- If this is not the case, we might only offer a **limited comparison** by rescaling the regression coefficients

Rescaling Coefficients: standard deviation

Prodedure:

- Start with: $y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip} + e_i$
- Let S_y be the standard deviation of y , and let S_1, \dots, S_p be the standard deviations of x_1, \dots, x_p .
- Rewrite this as (why? Where is $\hat{\alpha}$?):

$$\frac{y_i - \bar{y}}{S_y} = \left(\hat{\beta}_1 \frac{S_1}{S_y} \right) \frac{x_{i1} - \bar{x}_1}{S_1} + \dots + \left(\hat{\beta}_p \frac{S_p}{S_y} \right) \frac{x_{ip} - \bar{x}_p}{S_p} + \frac{e_i}{S_y}$$

- Let $\hat{\beta}_j^* = \hat{\beta}_j \frac{S_j}{S_y}$ and $Z_{iy} = \frac{y_i - \bar{y}}{S_y}$, $Z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$, then

$$Z_{iy} = \hat{\beta}_1^* Z_{i1} + \dots + \hat{\beta}_p^* Z_{ip} + e_i^*$$

where $\hat{\beta}_j^*$ are called the **standardised regression coefficients**.

Interpretation:

- Each variable Z_j has mean 0 and standard deviation 1.
- Increasing Z_j by 1 and keeping the other variables Z_i ($i \neq j$) constant gives, on average, an increase of $\hat{\beta}_j^*$ in Z_y .
- Standardised coefficients tell you how increases in the independent variables affect relative position within the group. You can determine whether a 1 standard deviation change in one independent variable produces more of a change in relative position than a 1 standard deviation change in another independent variable.

Food for thought:

- It is often difficult to say which of the dependent variables is most important in determining the value of the dependent variable, since the value of the regression coefficients depends on the **choice of units to measure**.
- For example, if we consider education and job experience, which are measured in years this is not so problematic.
- Suppose instead that our independent variables were education and IQ - how would we determine which variable was more important? The values of the metric coefficients would tell us little, since IQ and education are measured in very different ways.

Model Extension

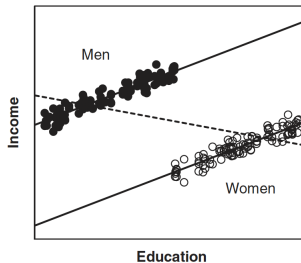
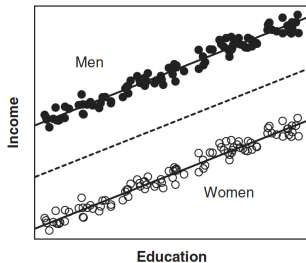
- categorical variable (dichotomous) -

Categorical variables

- We are interested in the effect of a qualitative independent variable (for example: do men earn more than women?)
- We want to better predict/describe the dependent variable. We can make the errors smaller by including variables like gender, race, etc
- Omitting some categorical variables may cause biased estimates of other coefficients.
- A **dichotomous** variable can take on one of only two possible values when observed or measured.

Example

- Consider
 - Dependent variable: **income**
 - One categorical independent variable: **education**
 - One dichotomous independent variable: **gender**
- Suppose that we are interested in the effect of education on income:



Independent variable vs regressor

- Y =income, X =education, D =regressor for gender

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

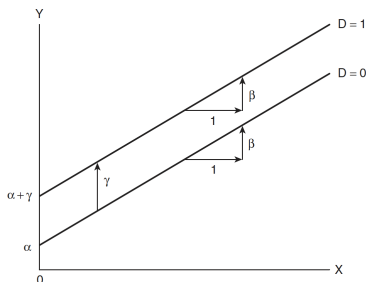
- Independent variable = real variable of interest
- Regressor = variable we include in the regression model
- Typically, regressors are functions of the independent variables. Sometimes regressors are equal to the independent variables.

Additive models with a dichotomous factor

- Consider an additive model with 2 categorical variables:

$$y_i = \alpha + \beta x_i + \gamma D_i + \epsilon_i$$

- For women ($D_i = 0$): $y_i = \alpha + \beta x_i + \gamma \cdot 0 + \epsilon_i = \alpha + \beta x_i + \epsilon_i$
- For men ($D_i = 1$): $y_i = \alpha + \beta x_i + \gamma \cdot 1 + \epsilon_i = (\alpha + \gamma) + \beta x_i + \epsilon_i$
- what are the interpretations of α , β and γ ?



- Test the partial effect of gender (i.e. effect of gender when education is in the model)
 - $H_0 : \gamma = 0, \quad H_a : \gamma \neq 0$
 - Compute t -statistic
- Test the partial effect of education (i.e. effect of education when gender is in the model)
 - $H_0 : \gamma = 0, \quad H_a : \gamma \neq 0$
 - Compute t -statistic

More general models

- Consider the following additive model:

$$y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \gamma D_i + \epsilon_i$$

- For women ($D_i = 0$): $y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$
- For men ($D_i = 1$): $y_i = (\alpha + \gamma) + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$
- A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the factor and 0 for the other category. A model incorporating a dummy regressor represents **parallel regression (hyper)surfaces**, with the constant vertical separation between the surfaces given by the coefficient of the dummy regressor.

Model Extension

- categorical variable (polytomous) -

Example (prestige data)

	type	inc	edu	pres
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89
physician	prof	76	97	97
welfare.worker	prof	41	84	59
teacher	prof	48	91	73
conductor	wc	76	34	38
contractor	prof	53	45	76
factory.owner	prof	60	56	81
store.manager	prof	42	44	45
banker	prof	78	82	92
bookkeeper	wc	29	72	39
mail.carrier	wc	48	55	34
insurance.agent	wc	55	71	41
store.clerk	wc	29	50	16
carpenter	bc	21	23	33
electrician	bc	47	39	53
RR.engineer	bc	81	28	67

Description

The ``Duncan`` data frame has 45 rows and 4 columns. Data on the prestige and other characteristics of 45 U. S. occupations in 1950.

Usage

::

Duncan

Format

This data frame contains the following columns:

type

Type of occupation. A factor with the following levels: ``prof``, professional and managerial; ``wc``, white-collar; ``bc``, blue-collar.

income

Percentage of occupational incumbents in the 1950 US Census who earned \$3,500 or more per year (about \$36,000 in 2017 US dollars).

education

Percentage of occupational incumbents in 1950 who were high school graduates (which, were we cynical, we would say is roughly equivalent to a PhD in 2017)

prestige

Percentage of respondents in a social survey who rated the occupation as "good" or better in prestige

Example (prestige data)

- **Polytomous variable:** qualitative variable with > 2 categories
- Consider an example (Duncan data):
 - Dependent variable: $Y = \text{prestige}$
 - Quantitative independent variables: $X_1 = \text{income}$, $X_2 = \text{education}$
 - Categorical independent variable: blue collar, professional, white collar
 - Define regressors D_1 and D_2 .

<i>Category</i>	D_1	D_2
Professional and managerial	1	0
White collar	0	1
Blue collar	0	0

- For p categories, use $p - 1$ dummy regressors.

Example (prestige data)

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon$$

- Blue collar ($D_{i1} = 0$ and $D_{i2} = 0$):

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 + \varepsilon_i \\ &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \end{aligned}$$

- Professional ($D_{i1} = 1$ and $D_{i2} = 0$):

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 + \varepsilon_i \\ &= (\alpha + \gamma_1) + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \end{aligned}$$

- White collar ($D_{i1} = 0$ and $D_{i2} = 1$):

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 + \varepsilon_i \\ &= (\alpha + \gamma_2) + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \end{aligned}$$

Example (prestige data)

- The mean prestige

<i>Category</i>	<i>Number of Cases</i>	<i>Mean Prestige</i>
Professional and managerial	31	67.85
White collar	23	42.24
Blue collar	44	35.53
All occupations	98	47.33

- Regressing occupational prestige on income and education produces the fitted regression equation ($R^2 = 0.814$):

$$\hat{y} = -7.621 + 0.001241x_1 + 4.292x_2$$

- and with dummy variables ($R^2 = 0.835$):

$$\hat{y} = -0.6229 + 0.001013x_1 + 3.673x_2 + 6.039D_1 - 2.737D_2$$

Example (prestige data)

$$\hat{y} = -7.621 + 0.001241x_1 + 4.292x_2$$

$$\hat{y} = -0.6229 + 0.001013x_1 + 3.673x_2 + 6.039D_1 - 2.737D_2$$

- Note that the coefficients for both income and education become slightly smaller when type of occupation is controlled.
- When income and education levels are held constant, the difference in average prestige declines greatly:
 - $67.85 - 35.53 = 32.32$ to 6.04 points (professional and blue-collar)
 - $42.24 - 35.53 = 6.71$ to -2.74 points (white-collar and blue-collar)
- The greater prestige of professional occupations compared to blue-collar occupations appears to be due mostly to differences in education and income between these two classes of occupations.
- While white-collar occupations have greater prestige, on average, than blue-collar occupations, they have lower prestige than blue-collar occupations of the same educational and income levels

Model Extension

- interactions -

- Two variables are said to **interact** in determining a dependent variable if the partial effect of one depends on the value of the other.
- Interaction between two qualitative variables means that the effect of one of the variables depends on the value of the other variable.
- Example: the effect of type of job on prestige is bigger for men than for women.

Interaction vs correlation

- Note that in general, the independent variables are not independent of each other.
- Interaction and correlation of explanatory variables are empirically and logically distinct phenomena.
- **Correlation:** Independent variables are statistically related
- **Interaction:** Effect of one independent variable on the dependent variable depends on the value of the other independent variable.
- **Example:** The relationship between height (X_1) and weight (Y) in male ($X_2 = 1$) and female ($X_2 = 0$) teenagers. There is a relationship between height (X_1) and gender (X_2), but for both genders, the relationship between height and weight is the same.

Constructing regressors

- Y = income, X =education, D =dummy for gender
- Statistical model:

$$y_i = \alpha + \beta x_i + \gamma D_i + \delta(x_i D_i) + \varepsilon_i$$

- Note $x_i D_i$ is a new regressor. It is a function of x_i and D_i , but not a linear function.
- Women ($D_i = 0$):

$$y_i = \alpha + \beta x_i + \gamma \cdot 0 + \delta(x_i \cdot 0) + \varepsilon_i = \alpha + \beta x_i + \varepsilon_i$$

- Men ($D_i = 1$):

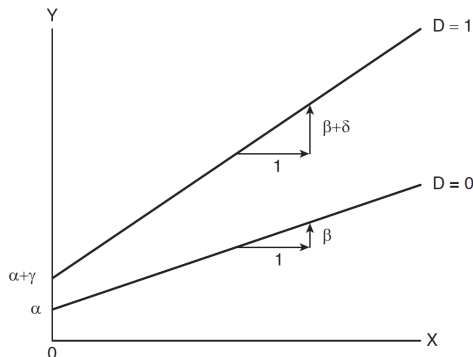
$$y_i = \alpha + \beta x_i + \gamma \cdot 1 + \delta(x_i \cdot 1) + \varepsilon_i = (\alpha + \gamma) + (\beta + \delta)x_i + \varepsilon_i$$

Interpretation

- How to interpret α , β , γ , δ ?

Women ($D_i = 0$) $y_i = \alpha + \beta x_i + \varepsilon_i$

Men ($D_i = 1$) $y_i = (\alpha + \gamma) + (\beta + \delta) x_i + \varepsilon_i$



Principle of marginality

- If interaction is significant, do not test or interpret main effects:
 - First test for interaction effect.
 - If no interaction, test and interpret main effects.
- If interaction is included in the model, main effects should also be included.

Polytomous variables

- Create interaction regressors by taking the products of all dummy variable regressors and the quantitative variable.
- Back to the example on 'prestige':
 - $Y = \text{prestige}$, $x_1 = \text{education}$, $x_2 = \text{income}$
 - $D_1, D_2 = \text{dummies for the type of job}$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 D_1 + \gamma_2 D_2 + \delta_{11} x_1 D_1 + \delta_{12} x_1 D_2 + \delta_{21} x_2 D_1 + \delta_{22} x_2 D_2 + \varepsilon$$

Example: fit

- The model permits different intercepts and slopes for the three types of occupations:
- Blue-collar occupations, which are coded 0 for both dummy regressors, serve as the baseline

$$\text{Professional:} \quad y_i = (\alpha + \gamma_1) + (\beta_1 + \delta_{11})x_{i1} + (\beta_2 + \delta_{21})x_{i2} + \varepsilon_i$$

$$\text{White collar:} \quad y_i = (\alpha + \gamma_2) + (\beta_1 + \delta_{12})x_{i1} + (\beta_2 + \delta_{22})x_{i2} + \varepsilon_i$$

$$\text{Blue collar:} \quad y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

- Fitting the model to the data: occupations:

$$\begin{aligned} \hat{y} = & 2.276 + 0.003522x_1 + 1.713x_2 + 15.35D_1 - 15.35D_1 - \\ & 0.002903x_1D_1 - 0.002072x_1D_2 + 1.388x_2D_1 + 4.291x_2D_2 \end{aligned}$$

Example: interpretation

- Caution: It is difficult in dummy-regression models with interactions to understand what the model is saying about the data simply by examining the regression coefficients.
- One approach is to write out the regression equation for each group:

$$\text{Professional: } \widehat{\text{Prestige}} = 17.63 + 0.000619 \times \text{Income} + 3.101 \times \text{Education}$$

$$\text{White collar: } \widehat{\text{Prestige}} = -31.26 + 0.001450 \times \text{Income} + 6.004 \times \text{Education}$$

$$\text{Blue collar: } \widehat{\text{Prestige}} = -2.276 + 0.003522 \times \text{Income} + 1.713 \times \text{Education}$$

- Income seems to make much more difference to prestige in blue-collar occupations than in white-collar occupations
- Education has the largest impact on prestige among white-collar occupations and has the smallest effect in blue collar occupations.

Examples !!

Practise, practise, practise..