Linear Regression

Learning outcomes:

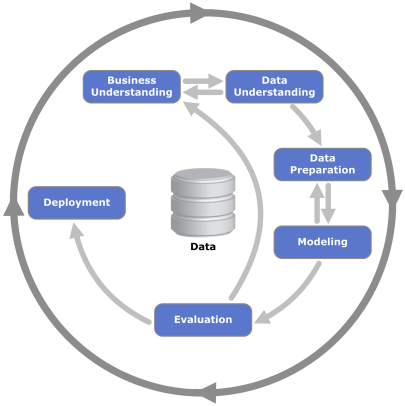
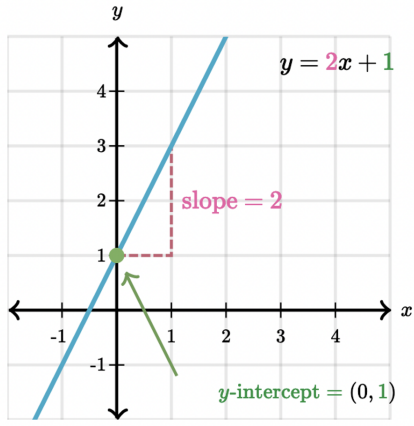


Figure: The CRISP-DM process.

- State the type of response variable linear regression is appropriate for;
- Write down the linear regression model;
- Interpret a regression table;
- Describe how linear regression learns, when it works, and why we use it;
- Define and compute two metrics that measure linear regression performance.

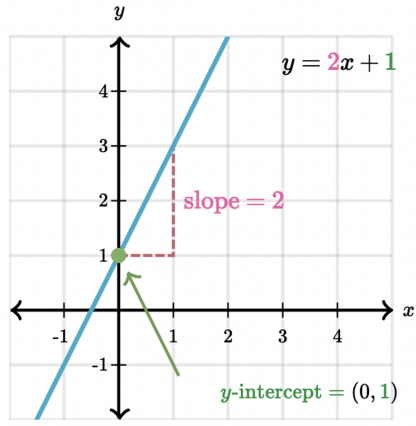
Modeling a linear relationship

- In high school algebra, you learned how to find the line that passed through two points;
- How?



Modeling a linear relationship

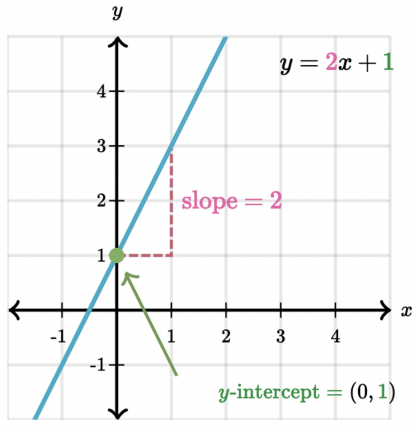
- In high school algebra, you learned how to find the line that passed through two points;
- How? Work with $y = mx + b$. Given two points (x_1, y_1) and (x_2, y_2):



Modeling a linear relationship

- In high school algebra, you learned how to find the line that passed through two points;
- How? Work with $y = mx + b$. Given two points (x_1, y_1) and (x_2, y_2):

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{3 - 1}{1 - 0} = 2;$$



Modeling a linear relationship

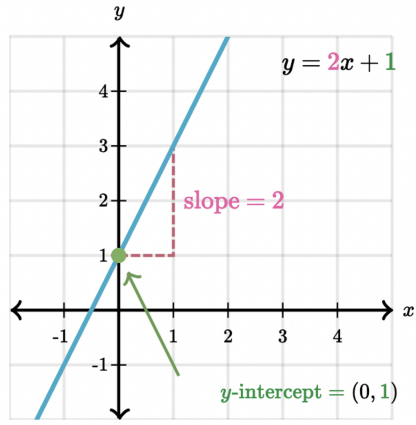
- In high school algebra, you learned how to find the line that passed through two points;

- How? Work with $y = mx + b$. Given two points (x_1, y_1) and (x_2, y_2):

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{3-1}{1-0} = 2;$$

$$y_1 = mx_1 + b \Rightarrow b = y_1 - \left(\frac{y_2 - y_1}{x_2 - x_1}\right) x_1;$$

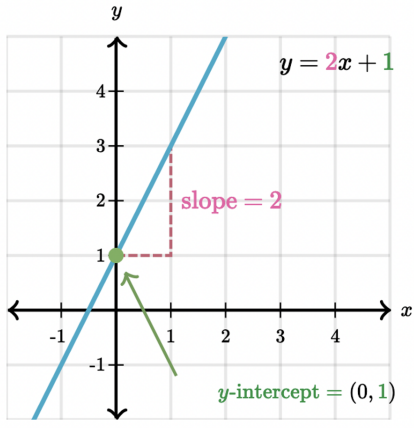
$$\Rightarrow b = 1 - 2 * 0 = 1.$$



Modeling a linear relationship

- We've got $y = 2x + 1$;
- For any value of x, we can now calculate the associated value of y on the line:

$$y = 2x + 1 = 2(1) + 1 = 2;$$
$$y = 2x + 1 = 2(0) + 1 = 1;$$
$$y = 2x + 1 = 2(2) + 1 = 5.$$



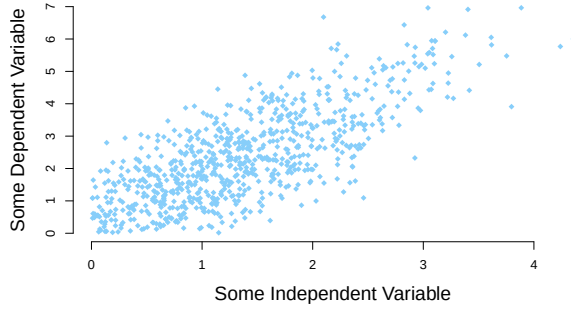
Ok, but what if we have MANY points?

- We can no longer exactly the predict the y value for every x – we'll need **approximation**;
- We will use the same equation, but with one modification:

$$\widehat{y} = mx + b,$$

where the $\widehat{y}$ indicates that the output is an approximation;

- We call this approximation linear regression.



What is linear regression?

- **Linear regression** is a method that answers three questions simultaneously:
  1. What is the effect of one variable x upon another variable y?
  2. How sure are we that x affects y?
  3. Do the answers to the first two questions depend on other independent variables?

- Typically used to model a dependent variable that is **continuous** (so interval or ratio data) and **unbounded** (so has no theoretical max or min);

- The essence of the linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

What is linear regression?

- **Linear regression** is a method that answers three questions simultaneously:
  1. What is the effect of one variable x upon another variable y?
  2. How sure are we that x affects y?
  3. Do the answers to the first two questions depend on other independent variables?

- Typically used to model a dependent variable that is **continuous** (so interval or ratio data) and **unbounded** (so has no theoretical max or min);

- The essence of the linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

What is linear regression?

Prediction:

- In machine learning, regression is primarily used for prediction;
- We fit a model to estimate a function that predicts the unknown output give a vector of inputs:

$$\widehat{y} = f(x);$$

- We treat the model as a black box.

Explanation:

- Statisticians and social scientists use regression to investigate relationships between input and output variables;
- They take advantage of the hypothesis testing built into regression;
- For example, how much of a college student's graduating GPA be predicted by SAT scores, HS GPA, number of HS math or science classes at admission time?

What is linear regression?

- Input variables;
  - Independent variables;
  - Regressors
  - Predictors;
  - Features;
  - Exogenous variables;

- Output variable;
  - Dependent variable
  - Regressand;
  - Response;
  - Target;
  - Endogenous variable.

What is linear regression?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- y is the **dependent variable** – this is part of the data set;
- The x's are **independent variables** that explain y – also part of the data set;
- The β's are called **coefficient effects** that control how the x's affect y – they are **learned** from the data;
- Finally the ε is a **noise** term that models the fact that y may also depend on random stuff (we don't observe this but use it to learn the β's – that'll be for next time).

An Example: the US Census American Community Survey, 2012.

income	hrs	race	age	gender	cmte	lang	married	edu	disability
1700	40	other	35	female	15	other	yes	hs or lower	yes
45000	84	white	27	male	40	english	yes	hs or lower	no
8600	23	white	69	female	5	english	no	hs or lower	no
33500	55	white	52	male	20	english	yes	hs or lower	no
4000	8	white	67	female	10	english	yes	hs or lower	no
19000	35	white	36	female	15	english	yes	college	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

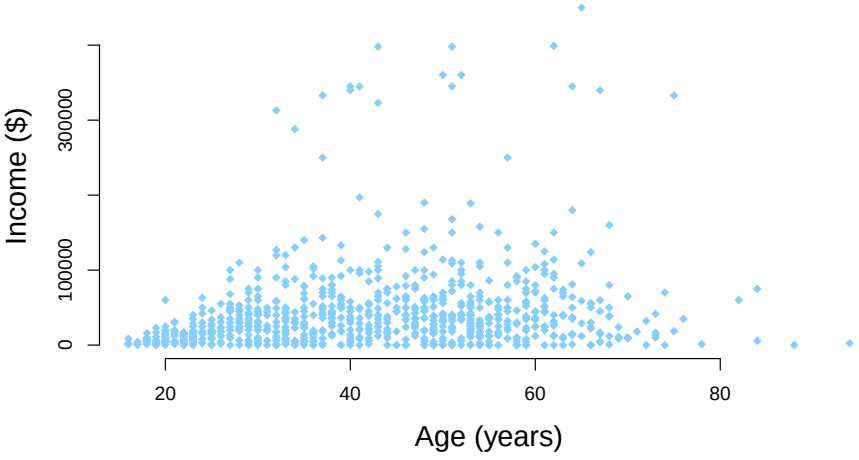
Question: do people earn more money as they get older?

An Example: the US Census American Community Survey, 2012.

income	hrs	race	**age**	gender	cmte	lang	married	edu	disability
1700	40	other	35	female	15	other	yes	hs or lower	yes
45000	84	white	27	male	40	english	yes	hs or lower	no
8600	23	white	69	female	5	english	no	hs or lower	no
33500	55	white	52	male	20	english	yes	hs or lower	no
4000	8	white	67	female	10	english	yes	hs or lower	no
19000	35	white	36	female	15	english	yes	college	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Question: do people earn more money as they get older?

Do people earn more money as they get older?



Do people earn more money as they get older?

- Let's answer this question by building a simple regression model;
  - The dependent variable y will be **income** from the 2012 US Census American Community Survey;
  - The independent variable x will be **age** from the 2012 US Census American Community Survey;
  - We will add an intercept or constant;

- This leads to the regression equation:

$$income = \beta_0 + \beta_{age} * age + \varepsilon;$$

- When we run the linear regression we will learn β_0 and β_{age}.

Do people earn more money as they get older – results!

OLS Regression Results

Dep. Variable:	income	**R-squared:**	0.041
Model:	OLS	**Adj. R-squared:**	0.039
Method:	Least Squares	**F-statistic:**	33.12
Date:	Tue, 02 Apr 2024	**Prob (F-statistic):**	1.24e-08
Time:	17:07:15	**Log-Likelihood:**	-9683.7
No. Observations:	783	**AIC:**	1.937e+04
Df Residuals:	781	**BIC:**	1.938e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
constant	1.067e+04	6333.943	1.684	0.093	-1767.918	2.31e+04
age	804.1380	139.720	5.755	0.000	529.867	1078.409

Omnibus:	631.125	**Durbin-Watson:**	1.917
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11067.135
Skew:	3.654	**Prob(JB):**	0.00
Kurtosis:	19.906	**Cond. No.**	141.

An Example: the US Census American Community Survey, 2012.

income	**hrs**	race	age	**gender**	**cmte**	lang	married	**edu**	disability
1700	40	other	35	female	15	other	yes	hs or lower	yes
45000	84	white	27	male	40	english	yes	hs or lower	no
8600	23	white	69	female	5	english	no	hs or lower	no
33500	55	white	52	male	20	english	yes	hs or lower	no
4000	8	white	67	female	10	english	yes	hs or lower	no
19000	35	white	36	female	15	english	yes	college	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Question: do people earn more money as they get older?

Do people earn more money as they get older – results!

OLS Regression Results

Dep. Variable:	income	**R-squared:**	0.041
Model:	OLS	**Adj. R-squared:**	0.039
Method:	Least Squares	**F-statistic:**	33.12
Date:	Tue, 02 Apr 2024	**Prob (F-statistic):**	1.24e-08
Time:	17:07:15	**Log-Likelihood:**	-9683.7
No. Observations:	783	**AIC:**	1.937e+04
Df Residuals:	781	**BIC:**	1.938e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
constant	1.067e+04	6333.943	1.684	0.093	-1767.918	2.31e+04
age	804.1380	139.720	5.755	0.000	529.867	1078.409

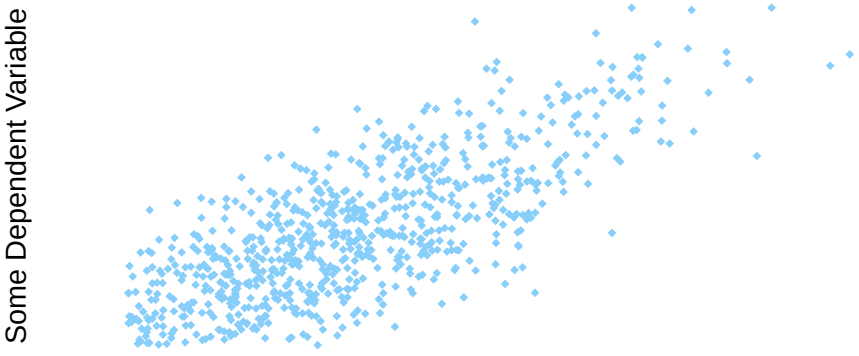
Omnibus:	631.125	**Durbin-Watson:**	1.917
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11067.135
Skew:	3.654	**Prob(JB):**	0.00
Kurtosis:	19.906	**Cond. No.**	141.

Dep. Variable:	income	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	51.92
Date:	Tue, 02 Apr 2024	**Prob (F-statistic):**	8.54e-54
Time:	17:07:40	**Log-Likelihood:**	-9567.9
No. Observations:	783	**AIC:**	1.915e+04
Df Residuals:	776	**BIC:**	1.918e+04
Df Model:	6		
Covariance Type:	nonrobust		

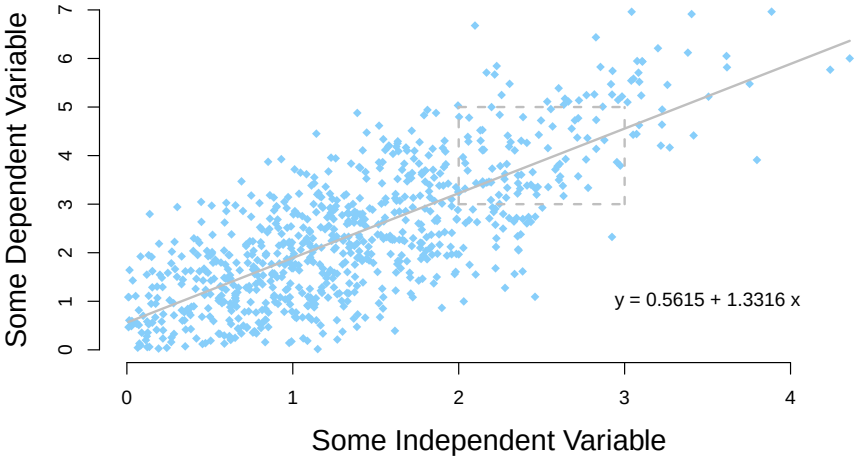
	coef	std err	t	P>|t|	[0.025	0.975]
constant	-2.695e+04	8232.570	-3.273	0.001	-4.31e+04	-1.08e+04
age	540.3949	122.267	4.420	0.000	300.381	780.408
hrs_work	1061.8244	149.483	7.103	0.000	768.386	1355.263
gendermale	1.948e+04	3688.594	5.282	0.000	1.22e+04	2.67e+04
time_to_work	93.0583	80.098	1.162	0.246	-64.175	250.292
edugrad	4.473e+04	6140.196	7.285	0.000	3.27e+04	5.68e+04
edu hs or lower	-1.852e+04	4077.024	-4.542	0.000	-2.65e+04	-1.05e+04

Omnibus:	578.247	**Durbin-Watson:**	1.919
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9838.046
Skew:	3.210	**Prob(JB):**	0.00
Kurtosis:	19.135	**Cond. No.**	320.

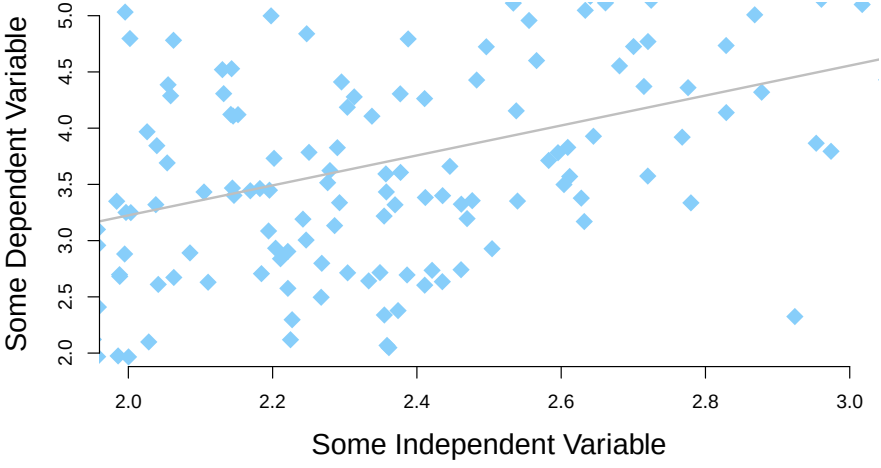
How does linear regression work?



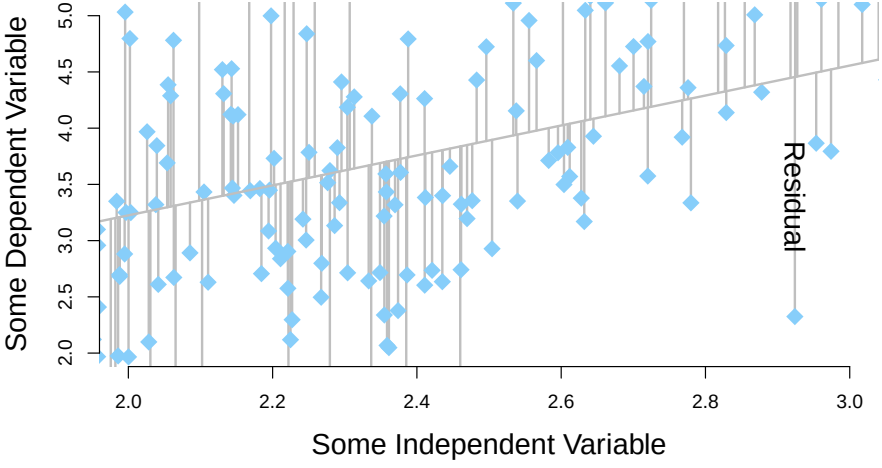
How does linear regression work?



How does linear regression work?



How does linear regression work?



How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...

y_i

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...

$$y_i \qquad \beta_0 + \beta_1 x_i$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...

$$y_i - (\beta_0 + \beta_1 x_i)$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...
  4. Square the difference...

$$(y_i - (\beta_0 + \beta_1 x_i))^2$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...
  4. Square the difference...
  5. Add it up for all the observations in the data...

$$\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...
  4. Square the difference...
  5. Add it up for all the observations in the data...
  6. Choose the βs that make this as small as possible...

$$\min_{\beta_0,\beta_1} \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2.$$

How does linear regression work?

- In general if we have:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p = X\beta$$

we learn the βs by solving:

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \underbrace{\sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2}_{\text{residual sum of squared errors (RSS)}} ;$$

- This is an optimization problem – RSS is our objective function and we need to choose the βs that minimize it.

How do we find the βs?

$$\text{RSS} = \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2$$

How do we find the βs?

$$\begin{aligned} \text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\ &= (y - X\beta)^T (y - X\beta) \end{aligned}$$

How do we find the βs?

$$
\begin{aligned}
\text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta
\end{aligned}
$$

How do we find the βs?

$$
\begin{aligned}
\text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
\Rightarrow \frac{\partial RSS}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \right)
\end{aligned}
$$

How do we find the βs?

$$
\begin{aligned}
\text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
\Rightarrow \frac{\partial RSS}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \right) \\
&= 2\beta^T X^T X - 2y^T X
\end{aligned}
$$

How do we find the βs?

$$
\begin{aligned}
\text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
\Rightarrow \frac{\partial RSS}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \right) \\
&= 2\beta^T X^T X - 2y^T X = 0 \text{ (why?)}
\end{aligned}
$$

How do we find the βs?

$$
\begin{aligned}
\text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
\Rightarrow \frac{\partial RSS}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \right) \\
&= 2\beta^T X^T X - 2y^T X = 0 \text{ (why?)} \\
&\Rightarrow \beta^T X^T X = y^T X
\end{aligned}
$$

How do we find the βs?

$$
\begin{aligned}
\text{RSS} &= \sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right)^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
\Rightarrow \frac{\partial RSS}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \right) \\
&= 2\beta^T X^T X - 2y^T X = 0 \text{ (why?)} \\
&\Rightarrow \beta^T X^T X = y^T X \\
&\Rightarrow \beta^T = y^T X (X^T X)^{-1} \\
&\Rightarrow \beta = (X^T X)^{-1} X^T y
\end{aligned}
$$

Fitting the model

- So we derived that $\beta = (X^T X)^{-1} X^T y$ – this is called the **normal equation**;
- You can use the normal equation and implement it in numpy yourself (if you do, don't forget to add a column of 1s to your data or it won't include β_0);
- There are libraries available that will perform this functionality and give us additional information (hypothesis tests);
  - OLS() in statsmodels.api;
  - LinearRegression() in sklearn.linear_model.

Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

2. Independent variables have **no measurement error**;

3. Independent variables are not related to each other – **no multicollinearity**;

4. Noise term is a random variable following the **normal distribution**.

Why do we use Linear Regression?

- **Theorem (Gauss-Markov)** When the assumptions of linear regression are met then it is the **Best Linear Unbiased Estimator** of the relationship between the dependent and independent variables;
- 'Unbiased' means that it will give you the correct βs on average;
- 'Best' means that it will give you the most precise estimates of those βs possible;
- This is usually called BLUE for short.

Metrics

- Metrics in linear regression typically make use of RSS – we'll focus on the **coefficient of determination** R^2;

- R^2 is often interpreted as the proportion of response variation "explained" by the features in the model;
  - $R^2 = 1$ indicates that the fitted model explains all variability in y;
  - $R^2 = 0$ indicates no "linear" relationship;
  - For example, if $R^2 = 0.7$ we could say: "Seventy percent of the variance in the response variable can be explained by the features. The remaining thirty percent can be attributed to unknown, lurking variables or irreducible error."

Metrics

- Metrics in linear regression typically make use of RSS – we'll focus on the **coefficient of determination** R^2;

- R^2 is often interpreted as the proportion of response variation "explained" by the features in the model;
  - $R^2 = 1$ indicates that the fitted model explains all variability in y;
  - $R^2 = 0$ indicates no "linear" relationship;
  - For example, if $R^2 = 0.7$ we could say: "Seventy percent of the variance in the response variable can be explained by the features. The remaining thirty percent can be attributed to unknown, lurking variables or irreducible error."

- It's easy to compute R^2:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad \text{where} \quad \begin{array}{l} \text{RSS is as above} \\ \text{TSS} = \sum_i (y_i - \overline{y}) \\ \overline{y} = \text{the sample mean} \end{array}$$

Metrics

- It turns out that R^2 weakly increases with the number of features in the model;
- It is common to adjust R^2 in an attempt to account for this:

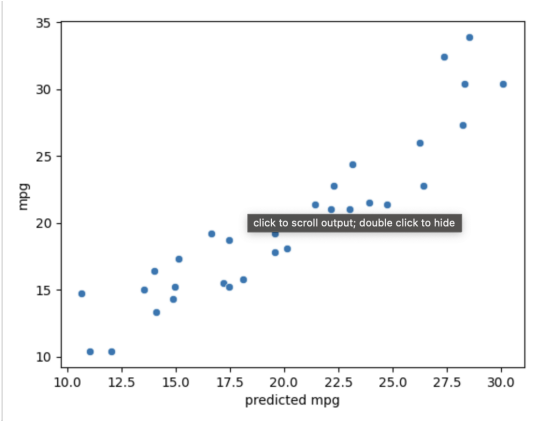
$$1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right) \quad \text{where} \quad \begin{array}{l} n \text{ is the number of observations in your data} \\ p \text{ is the number of features in the model.} \end{array}$$

Process for doing regression

- **Fit model;**
  - Use OLS() in statsmodels.api if you are explaining;
  - Use LinearRegression() in sklearn.linear_model if you are predicting.

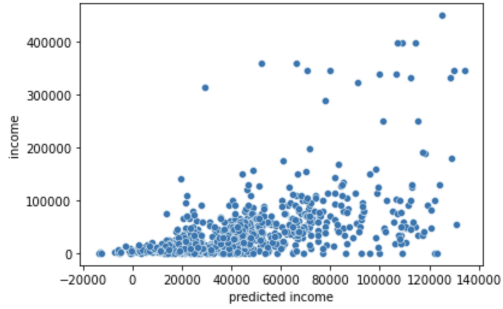
Process for doing regression

- Fit model;
- **Evaluate model fit (visualize, metrics, statistical test)**;



Process for doing regression

- Fit model;
- **Evaluate model fit (visualize, metrics, statistical test)**;



Process for doing regression

- Fit model;
- **Evaluate model fit (visualize, metrics, statistical test)**;

Use R^2 or Adjusted R^2;

Dep. Variable:	income	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	51.92
Date:	Tue, 02 Apr 2024	**Prob (F-statistic):**	8.54e-54
Time:	17:07:40	**Log-Likelihood:**	-9567.9
No. Observations:	783	**AIC:**	1.915e+04
Df Residuals:	776	**BIC:**	1.918e+04
Df Model:	6		
Covariance Type:	nonrobust		

Process for doing regression

- Fit model;
- **Evaluate model fit (visualize, metrics, statistical test)**;

Is there at least one $\beta_i \neq 0$?

- $H_0 : \beta_i = 0$ for all $i = 1, \ldots, p$;
- H_A : "At least one" $\beta_i \neq 0$;
- The F-test is used to compare the fitted model with a null model;

Dep. Variable:	income	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	51.92
Date:	Tue, 02 Apr 2024	**Prob (F-statistic):**	8.54e-54
Time:	17:07:40	**Log-Likelihood:**	-9567.9
No. Observations:	783	**AIC:**	1.915e+04
Df Residuals:	776	**BIC:**	1.918e+04
Df Model:	6		
Covariance Type:	nonrobust		

Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- **Confirm model assumptions (residuals)**;

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

2. Independent variables have **no measurement error**;
3. Independent variables are not related to each other – **no multicollinearity**;
4. Noise term is a random variable following the **normal distribution** with 0 mean.

Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- **Confirm model assumptions (residuals)**;

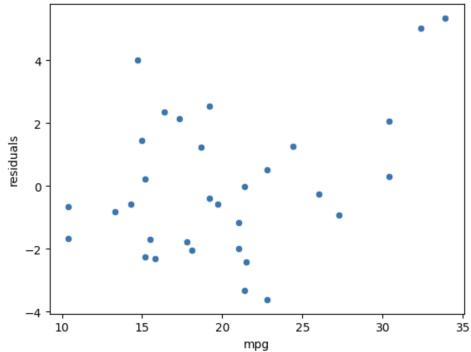
Hypothesis tests:

Omnibus:	578.247	**Durbin-Watson:**	1.919
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9838.046
Skew:	3.210	**Prob(JB):**	0.00
Kurtosis:	19.135	**Cond. No.**	320.

Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- **Confirm model assumptions (residuals)**;

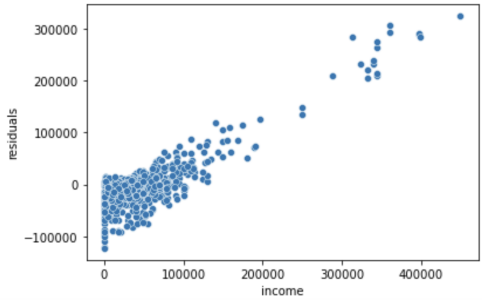
Residual plot (from the cars data):



Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- **Confirm model assumptions (residuals)**;

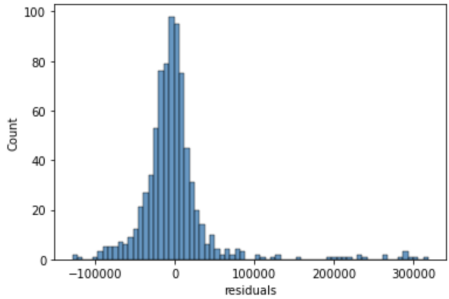
Residual plot (from the American Community Survey):



Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- **Confirm model assumptions (residuals)**;

Residual distribution (from the American Community Survey):



Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- Confirm model assumptions (residuals);
- **Identify significant predictors**;

For each feature is $\beta_i \neq 0$?

- $H_0 : \beta_i = 0$;
- $H_A : \beta_i \neq 0$;
- Apply the t-test;

	coef	std err	t	P>|t|	[0.025	0.975]
constant	-2.695e+04	8232.570	-3.273	0.001	-4.31e+04	-1.08e+04
age	540.3949	122.267	4.420	0.000	300.381	780.408
hrs_work	1061.8244	149.483	7.103	0.000	768.386	1355.263
gendermale	1.948e+04	3688.594	5.282	0.000	1.22e+04	2.67e+04
time_to_work	93.0583	80.098	1.162	0.246	-64.175	250.292
edugrad	4.473e+04	6140.196	7.285	0.000	3.27e+04	5.68e+04
edu hs or lower	-1.852e+04	4077.024	-4.542	0.000	-2.65e+04	-1.05e+04

Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- Confirm model assumptions (residuals);
- Identify significant predictors;
- **Interpret significant predictors**.

Each additional year of life increases income by $540;

	coef	std err	t	P>|t|	[0.025	0.975]
constant	-2.695e+04	8232.570	-3.273	0.001	-4.31e+04	-1.08e+04
age	540.3949	122.267	4.420	0.000	300.381	780.408
hrs_work	1061.8244	149.483	7.103	0.000	768.386	1355.263
gendermale	1.948e+04	3688.594	5.282	0.000	1.22e+04	2.67e+04
time_to_work	93.0583	80.098	1.162	0.246	-64.175	250.292
edugrad	4.473e+04	6140.196	7.285	0.000	3.27e+04	5.68e+04
edu hs or lower	-1.852e+04	4077.024	-4.542	0.000	-2.65e+04	-1.05e+04

Process for doing regression

- Fit model;
- Evaluate model fit (visualize, metrics, statistical test);
- Confirm model assumptions (residuals);
- Identify significant predictors;
- **Interpret significant predictors**.

Each additional hour worked increases income by $1062;

	coef	std err	t	P>|t|	[0.025	0.975]
constant	-2.695e+04	8232.570	-3.273	0.001	-4.31e+04	-1.08e+04
age	540.3949	122.267	4.420	0.000	300.381	780.408
hrs_work	1061.8244	149.483	7.103	0.000	768.386	1355.263
gendermale	1.948e+04	3688.594	5.282	0.000	1.22e+04	2.67e+04
time_to_work	93.0583	80.098	1.162	0.246	-64.175	250.292
edugrad	4.473e+04	6140.196	7.285	0.000	3.27e+04	5.68e+04
edu hs or lower	-1.852e+04	4077.024	-4.542	0.000	-2.65e+04	-1.05e+04

Summary

- This is one type of regression model – multiple linear regression;
  - We can use it for prediction and explanation;
  - Statistical backing gives us confidence in relationships;
  - BLUE;
- There are other regression models available;
  - Residuals are useful in these model interpretations – so are metrics like MSE;
  - Some models provide feature selection;
  - If we want confidence intervals we may have to bootstrap / use sampling techniques;
  - Models include:
    - Polynomial Regression;
    - Multivariate Adaptive Regression Splines (MARS);
    - k-Nearest Neighbors (kNN) Regression;
    - Random Forest Regression;
    - Neural Network Regression.