JSC 270 - Lecture 3
Regression: Modelling, Fit, Regularization

https://jsc270.github.io/

Announcements

1. Homework 2 is out
2. Perusall 2 is out
3. Next week we have our first invited speaker: Fanny Chevalier. She will talk about visualization 2-3pm, right after class

Linear Regression

Linear regression is a simple approach to supervised learning

Assumes that outcome Y depends on the inputs $X_1, X_2, X_3, \ldots X_p$ *linearly*

Extremely useful conceptually and practically

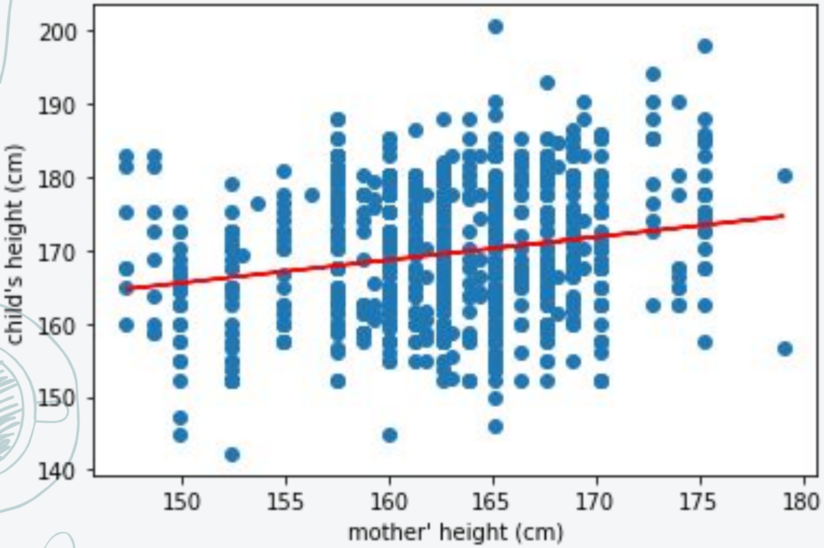
Example Data

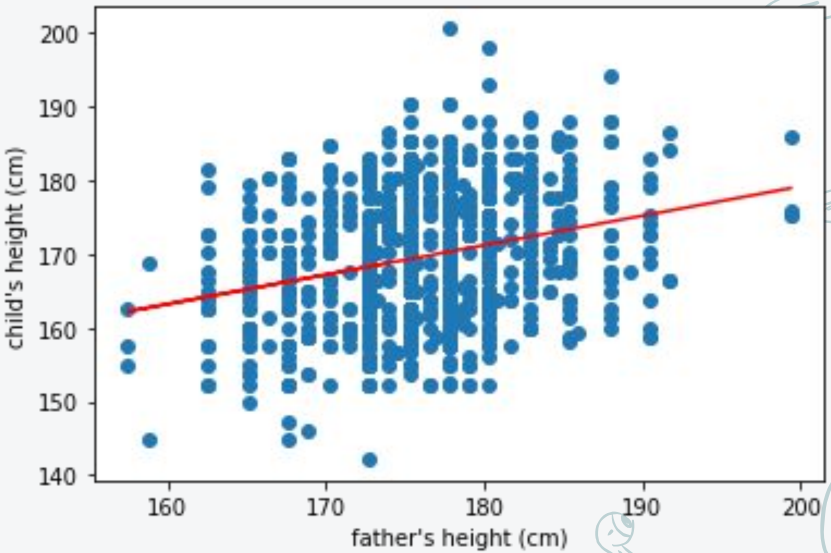
In 1880 Galton was developing a way to quantify heritability of traits

He collected heights of parents and children

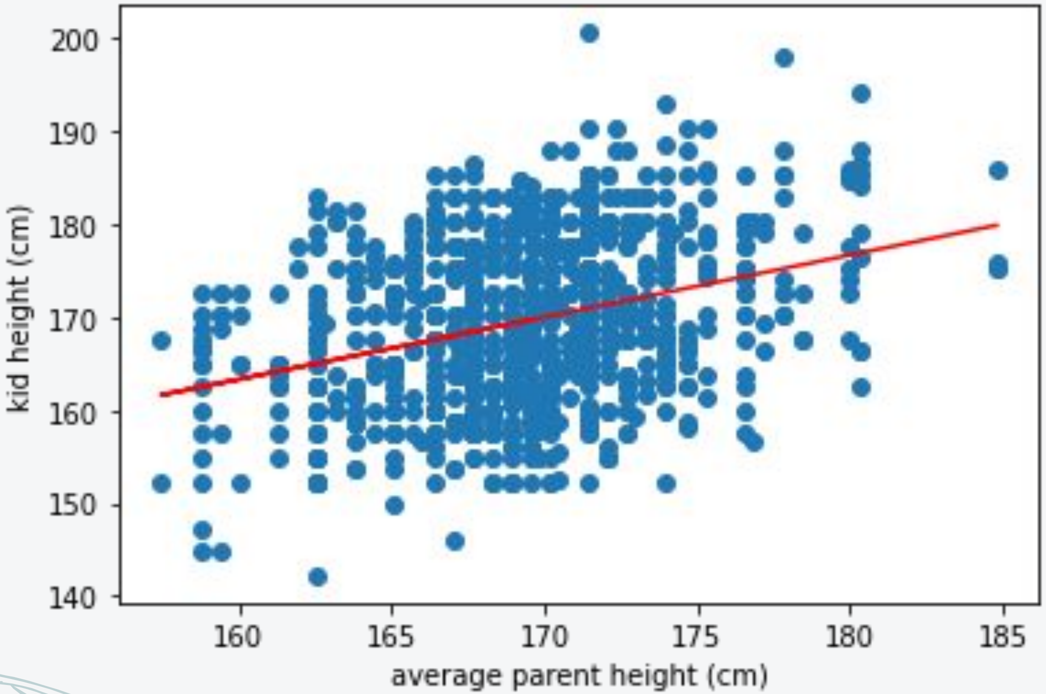
Height of mother, father, child

Galton Data


200
190
180
170
160
150
140
child's height (cm)
150
155
160
165
170
175
180
mother' height (cm)


200
190
180
170
160
150
140
child's height (cm)
160
170
180
190
200
father's height (cm)

Galton Data



200
190
180
170
160
150
140
kid height (cm)
160
165
170
175
180
185
average parent height (cm)

Linear Regression

What questions is regression helping to answer in this case?

Linear Regression: Galton Data

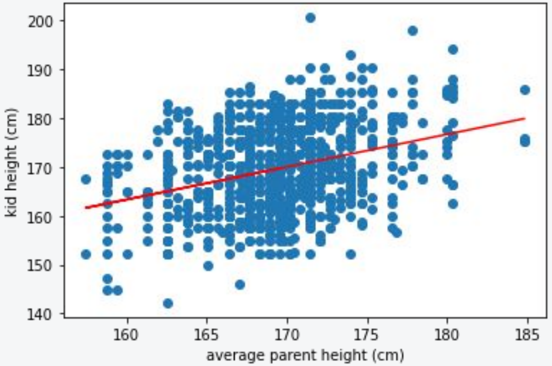
What questions is regression helping to answer in this case?

- How well can parents' height predict child's height?
- Is father's height a better predictor than mother's height?
- Is this a linear or a nonlinear relationship?

Linear Regression

$$Y = \beta_0 + \beta_1 X$$

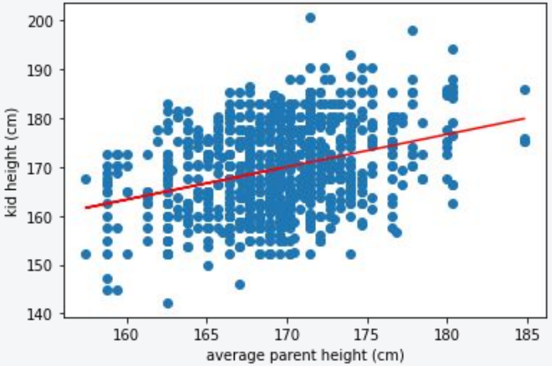
X – avg parent height
Y – child height



Linear Regression

$$Y = \beta_0 + \beta_1 X$$

X – avg parent height
Y – child height



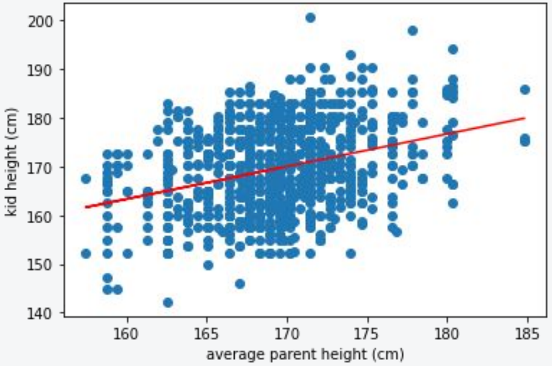
$$Y = \beta_0 + \beta_1 X + \epsilon$$

ϵ – error term (everything we didn't measure)

Linear Regression

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

X – avg parent height
Y – child height

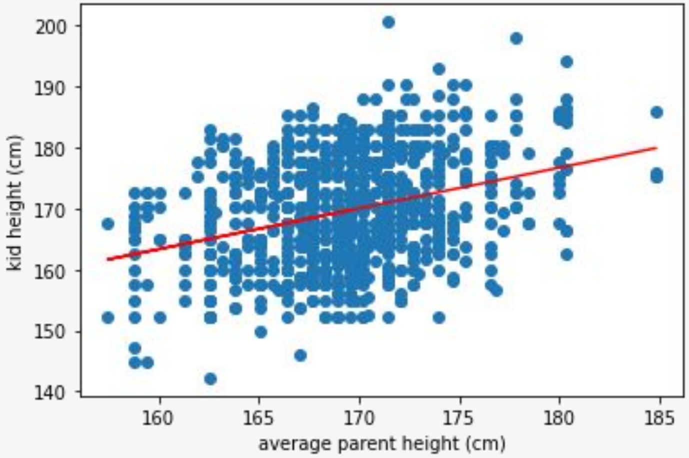


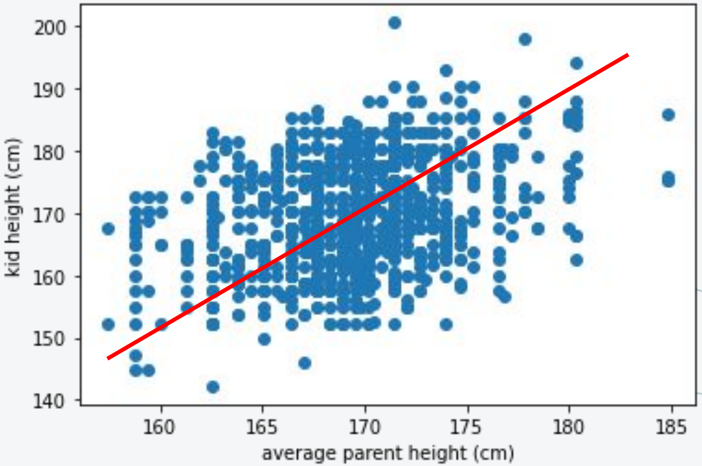
β_0 – intercept

β_1 – slope

$\{\beta_0, \beta_1\}$ – parameters, coefficients

How do we pick the best line?





Estimate of the outcome

Given some estimates for the coefficients $\hat{\beta}_0, \hat{\beta}_1$

The estimate of outcome y_i for x_i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Estimate of the outcome

Given some estimates for the coefficients $\hat{\beta}_0, \hat{\beta}_1$

The estimate of outcome y_i for x_i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Errors are $e_i = y_i - \hat{y}_i$ - residual error

Best Fit - least squared error

The sum of all errors should be as small as possible!

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Residual Sum Square

Best Fit - least squared error

The sum of all errors should be as small as possible!

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

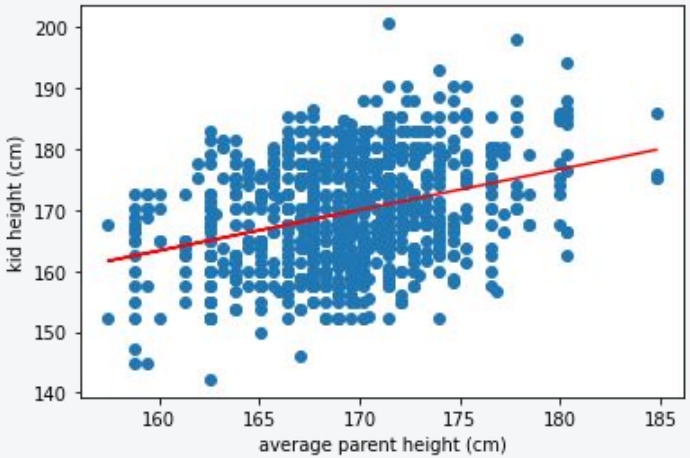
Residual Sum Square

Turns out that there is a closed form solution for the coefficients of this line

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

sample means

Back to Galton

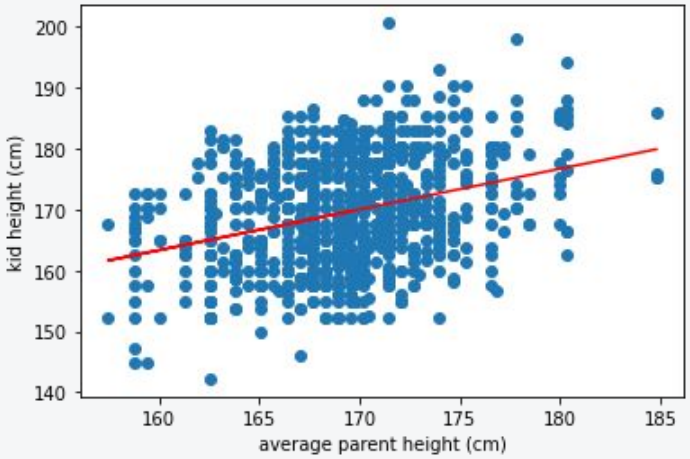


$$\hat{\beta}_0 = 56.26 \quad \hat{\beta}_1 = 0.67$$

$$CH = 56.26 + 0.67 \cdot APH + \epsilon$$

What does it mean?

Back to Galton

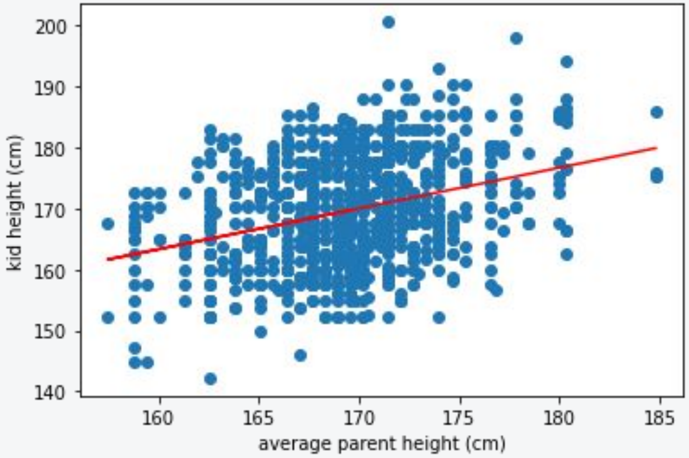


$$\hat{\beta}_0 = 56.26 \quad \hat{\beta}_1 = 0.67$$

What does it mean? If the avg of parents' heights is 1 unit (1 cm) bigger than the child's height is expected to be 0.67 cm bigger

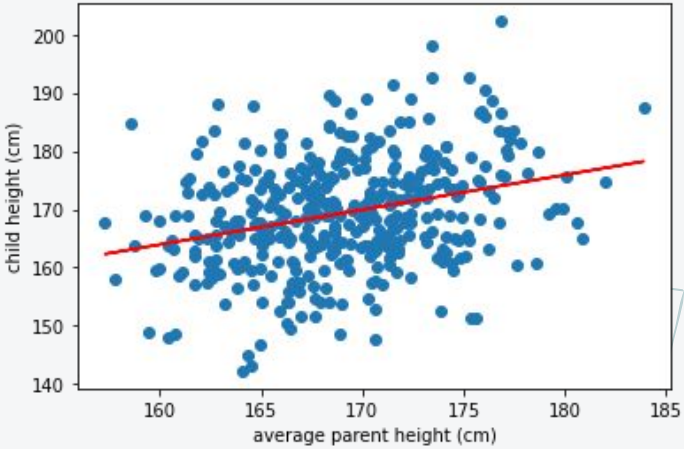
But we just had a sample...

Sample 1



$$\hat{\beta}_0 = 56.26 \quad \hat{\beta}_1 = 0.67$$

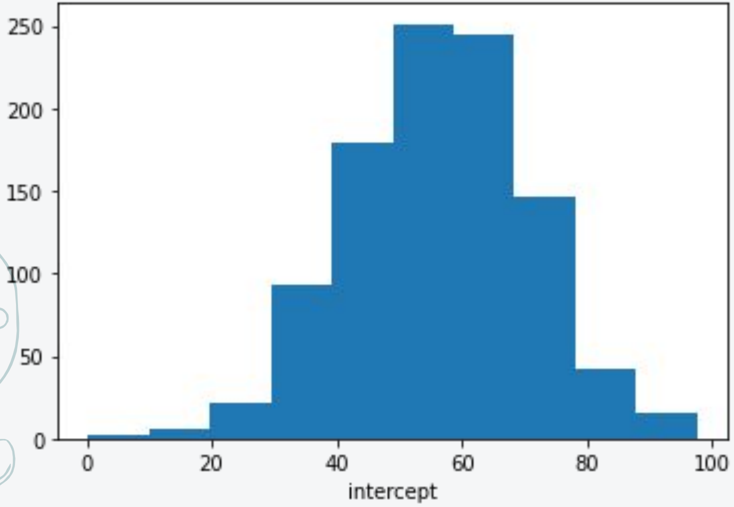
Sample 2



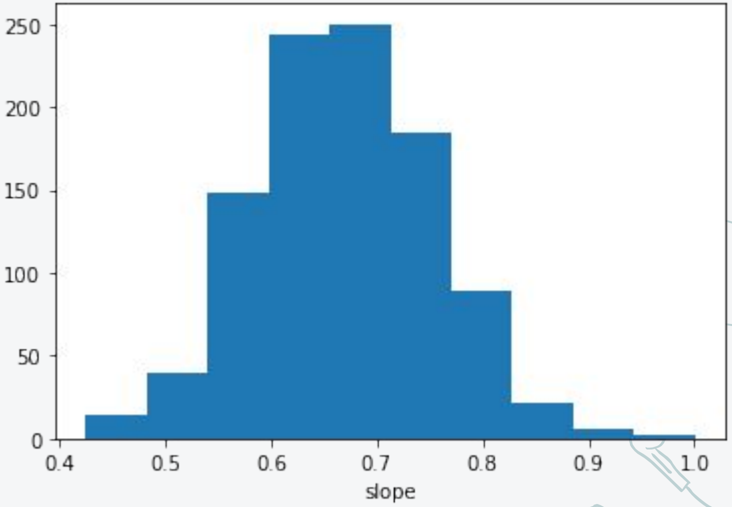
$$\hat{\beta}_0 = 68.3 \quad \hat{\beta}_1 = 0.6$$

Uncertainty of the coefficients

Across many samples....



$$\hat{\beta}_0 = N(56.16, 14.52)$$



$$\hat{\beta}_1 = N(0.67, 0.086)$$

Why is it important?

Q: is the average height of parents predictive of the child's height?

WHY IS IT IMPORTANT?

Q: is the average height of parents predictive of the child's height?

95% confidence interval: $[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$

Why is it important?

Q: is the average height of parents predictive of the child's height?

95% confidence interval: $[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$

In the current model: $\hat{\beta}_1 \in [0.5393, 0.7993]$

Why is it important?

Q: is the average height of parents predictive of the child's height?

95% confidence interval: $[\hat{\beta}_1 - 2 * SE(\hat{\beta}_1), \hat{\beta}_1 + 2 * SE(\hat{\beta}_1)]$

In the current model: $\hat{\beta}_1 \in [0.5393, 0.7993]$

Yes! Parent's height is predictive

More formally

H_0: There is no relationship between X and Y vs the alternative:

H_A: There is some relationship between X and Y

Otherwise can be stated as

$$H_0 : \quad \beta_1 = 0$$
$$H_A : \quad \beta_1 \neq 0$$

More Formally

To test the null hypothesis we compute t-statistic

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- t-statistic has t distribution with n-2 degrees of freedom
- compute probability of seeing |t| value or larger (p-value)

Back to Galton

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept (beta0)	56.2580	10.941	5.142	0.000	34.785	77.731
Slope (beta1)	0.6693	0.065	10.360	0.000	0.542	0.796

What do we do with the null hypothesis given these numbers?

Back to Galton

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept (beta0)	56.2580	10.941	5.142	0.000	34.785	77.731
Slope (beta1)	0.6693	0.065	10.360	0.000	0.542	0.796

What do we do with the null hypothesis given these numbers?

Yes! We reject the null hypothesis

Assessing model fit

1. Residual Standard Error (RSE)

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

Assessing model fit

1. Residual Standard Error (RSE)

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

2. Root Mean Squared Error (RMSE)

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

ASSESSING MODEL FIT

1. Residual Standard Error (RSE)

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

= 8.6

2. Root Mean Squared Error (RMSE)

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

= 8.59

Assessing model fit

3. Fraction of variance explained (R^2)

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

Total Sum Squares

Assessing model fit

3. Fraction of variance explained (R^2)

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

Total Sum Squares

Turns out R2 = correlation(x,y) = 0.107

What does it mean?

Assessing model fit

3. Fraction of variance explained (R^2)

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

Total Sum Squares

Turns out R2 = correlation(x,y) = 0.107

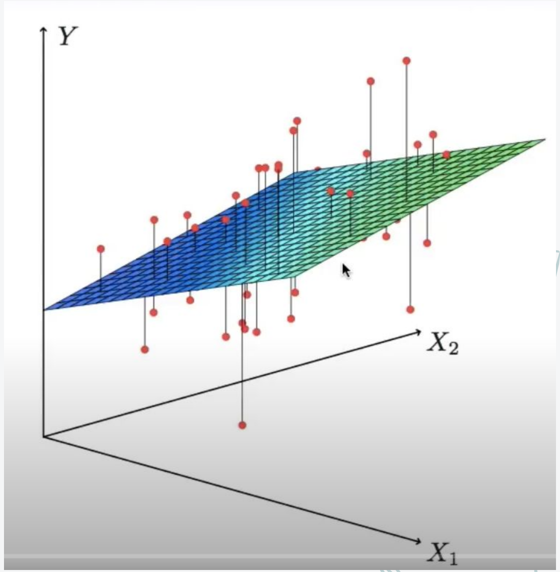
What does it mean? Variance in average parent's height explains ~10% of variance in child's height

Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Coefficient β_j is an average effect of a one unit increase in X_J on Y, holding all other predictors fixed

child's height = β_0 + β_1 * mother's height + β_2 * father's height + ε



Back to Galton Data

FH – Father's Height
MH – Mother's Height

$$Y = 56.26 + 0.67 \cdot (FH + MH)/2 + \epsilon$$

RSE	8.605
RMSE	8.595
R^2	0.107

$$Y = 56.67 + 0.38 \cdot FH + 0.28 \cdot MH + \epsilon$$

RSE	8.596
RMSE	8.586
R^2	0.109

Which model would you pick and why?

Model selection: you don't have to guess!

Which factors do we want to incorporate?

Model selection: you don't have to guess!

Akaike Information Criterion (AIC)

Bayesian Information Criterion (BIC)

Cross Validation

Note: this is a **non**-exhaustive list
We will revisit the topic of model selection later

What do coefficients mean in MLR?

If variables were independent, could analyze one by one

But heights of parents are not independent (7% correlation)

Claims of causality should be avoided

Variable selection

- Is at least one of the predictors useful at predicting response?
- Are all of the predictors needed?
- Given a set of values for input variables – what should we predict and how accurate is that prediction?

Deciding on the Best variables

All subsets - consider models with all possible subsets of variables. Prohibitively expensive for large number of variables =((e.g. 40 variables -> over a billion subsets!)

Forward selection

Backward selection

Regularization

Forward selection

- Start with a *null* model
- Fit p models and add to the *null* model the one with the lowest RSS
- Add to that model the second variable, that results in the lowest RSS amongst all two-variable models
- Continue until some stopping criterion is satisfied

Forward Selection: Galton Data

Step 1

Fit $y = \beta_0 + \beta_1 *$ mother's height $+ \epsilon$

$RSS_{mother} = 71269.57$

Fit $y = \beta_0 + \beta_1 *$ father's height $+ \epsilon$ ✓

$RSS_{father} = 68657.8$

Step 2

$y = \beta_0 + \beta_1 *$ father's height $+ \beta_2 *$ mother's height $+ \epsilon$

$RSS = 66200.7$

Summary

- Defined linear regression
- Fit linear regression to real data
- Learned to answer questions about coefs
- Compared goodness of fit metrics
- Fit multivariate linear regression
- Intro to variable selection