UNIVERSITY OF
MICHIGAN

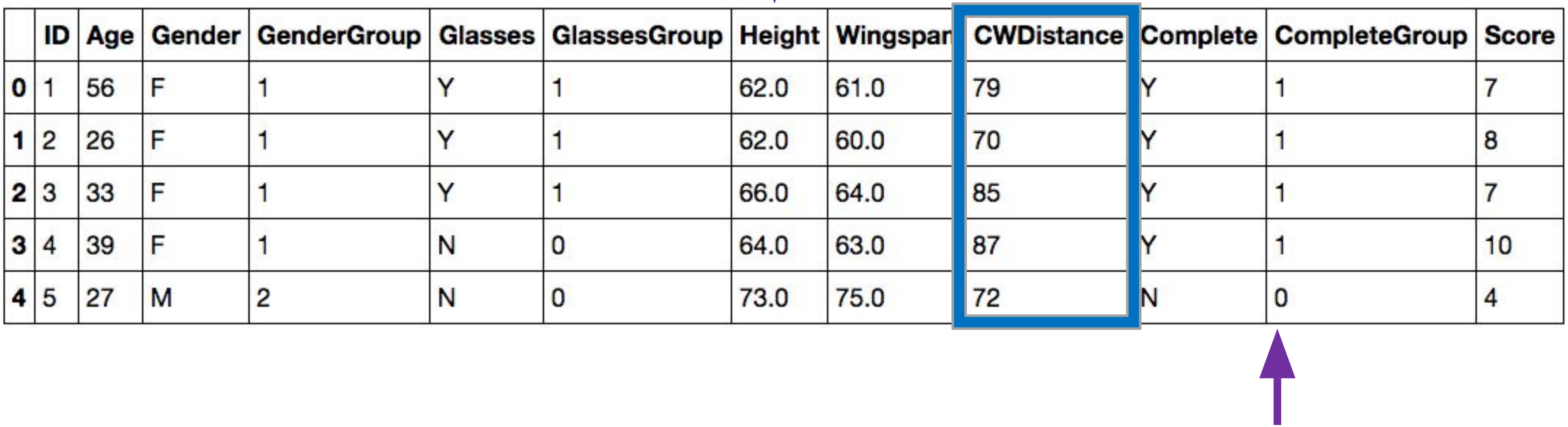
Linear Regression Introduction

Brenda Gunderson

Cartwheel Study

- 25 team members/colleagues (all adults) asked to perform a cartwheel
- **Many Variables recorded**:
  Primary outcome of interest = Cartwheel Distance (inches)

Cartwheel Study Data



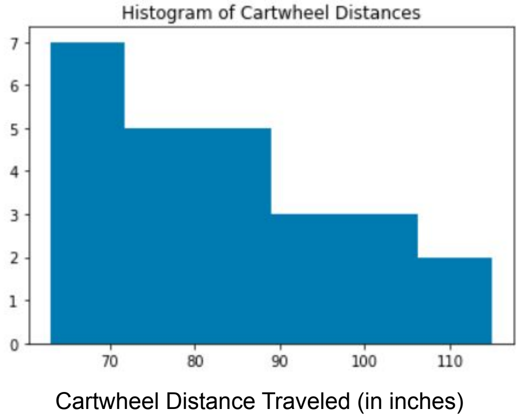
	ID	Age	Gender	GenderGroup	Glasses	GlassesGroup	Height	Wingspan	CWDistance	Complete	CompleteGroup	Score
0	1	56	F	1	Y	1	62.0	61.0	79	Y	1	7
1	2	26	F	1	Y	1	62.0	60.0	70	Y	1	8
2	3	33	F	1	Y	1	66.0	64.0	85	Y	1	7
3	4	39	F	1	N	0	64.0	63.0	87	Y	1	10
4	5	27	M	2	N	0	73.0	75.0	72	N	0	4

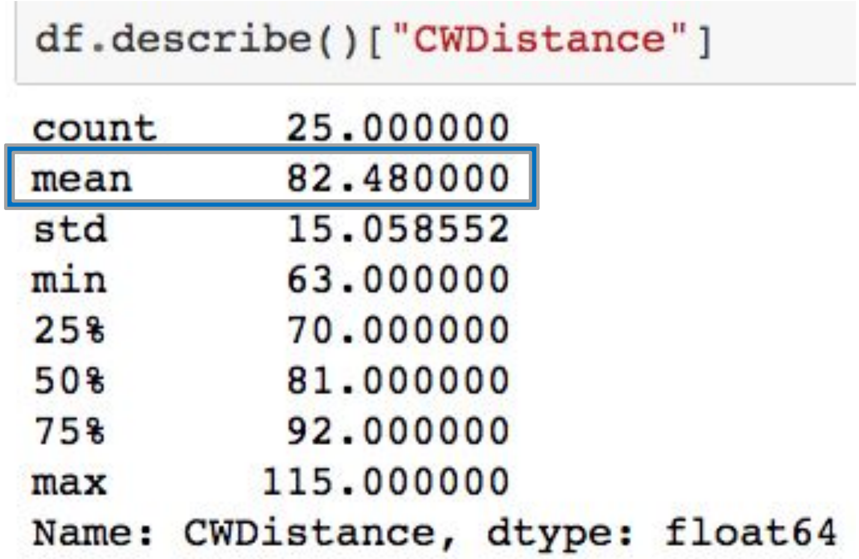
Possible Research Goals/Questions

Develop a model to predict the (mean) cartwheel distance for the population of all such adults…

- Is a person’s height a useful predictor for cartwheel distance?
- Does knowing if they actually *completed* the cartwheel make a difference in terms of cartwheel distance?

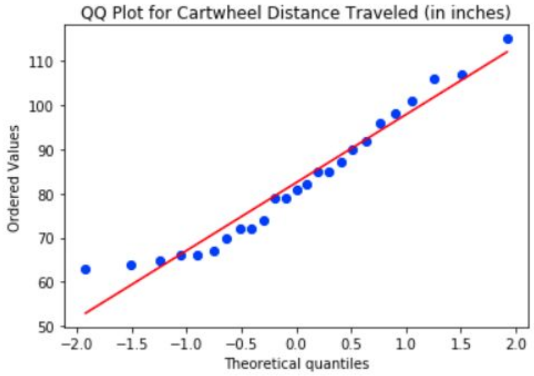
Cartwheel Distance Summary


Histogram of Cartwheel Distances
0
1
2
3
4
5
6
7
70
80
90
100
110
Cartwheel Distance Traveled (in inches)



```
df.describe()["CWDistance"]
```

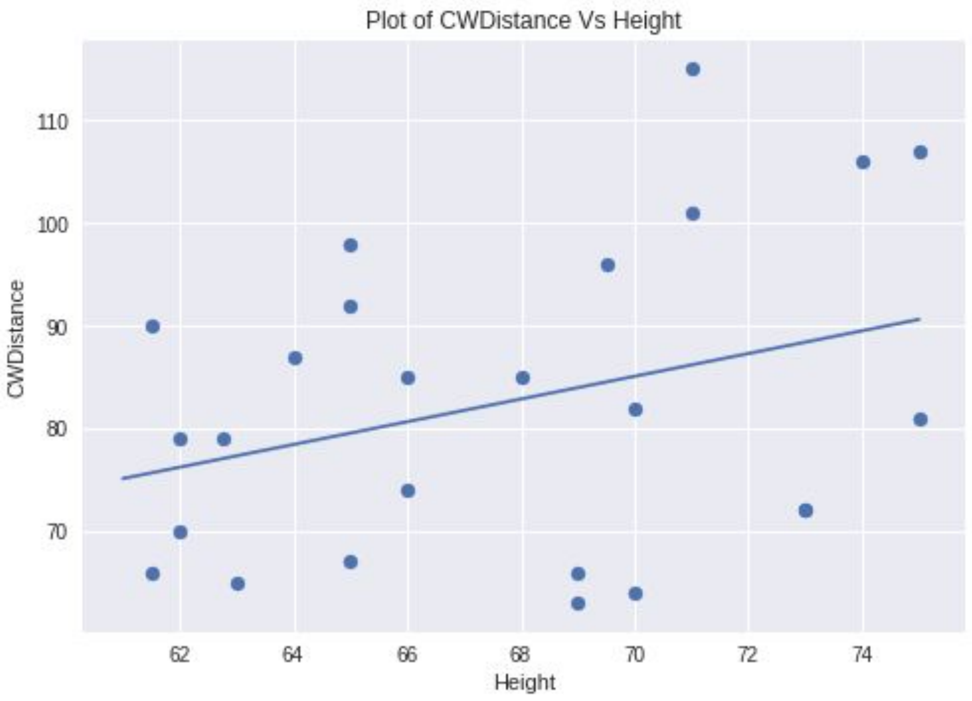
```
count     25.000000
mean      82.480000
std       15.058552
min       63.000000
25%       70.000000
50%       81.000000
75%       92.000000
max      115.000000
Name: CWDistance, dtype: float64
```



Is there a Relationship?

- **Is HEIGHT a useful predictor for cartwheel distance?**
- **Do taller people generally have larger cartwheel distances?**
- **Is there a significant (positive) relationship between the height and cartwheel distance?**

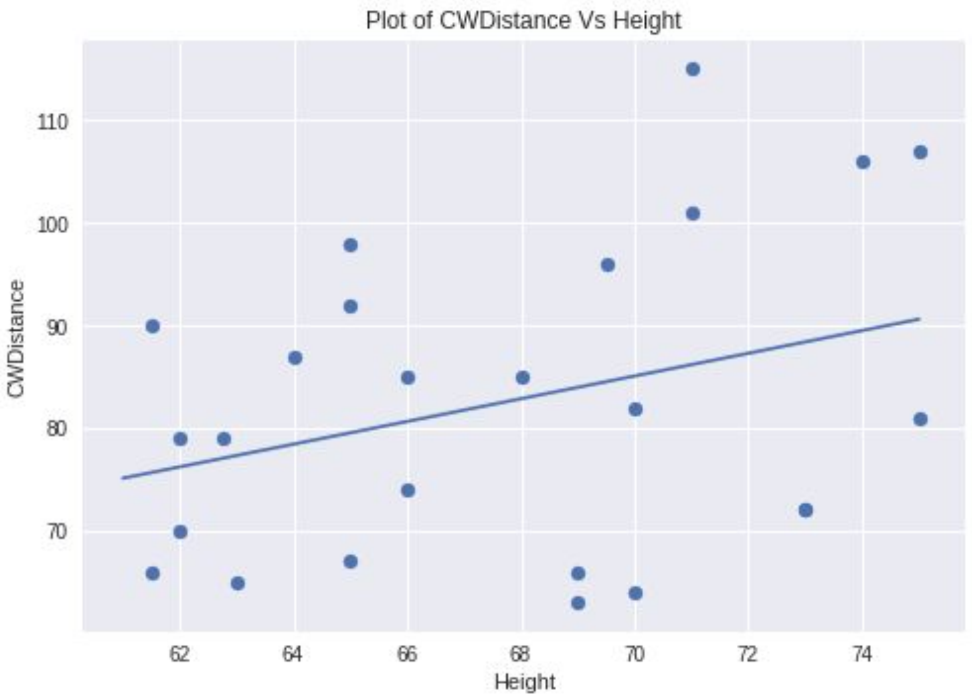
Visualizing the Relationship



Dependent Variable (DV) = CWDistance

Independent Variable (IV) = Height

Visualizing the Relationship



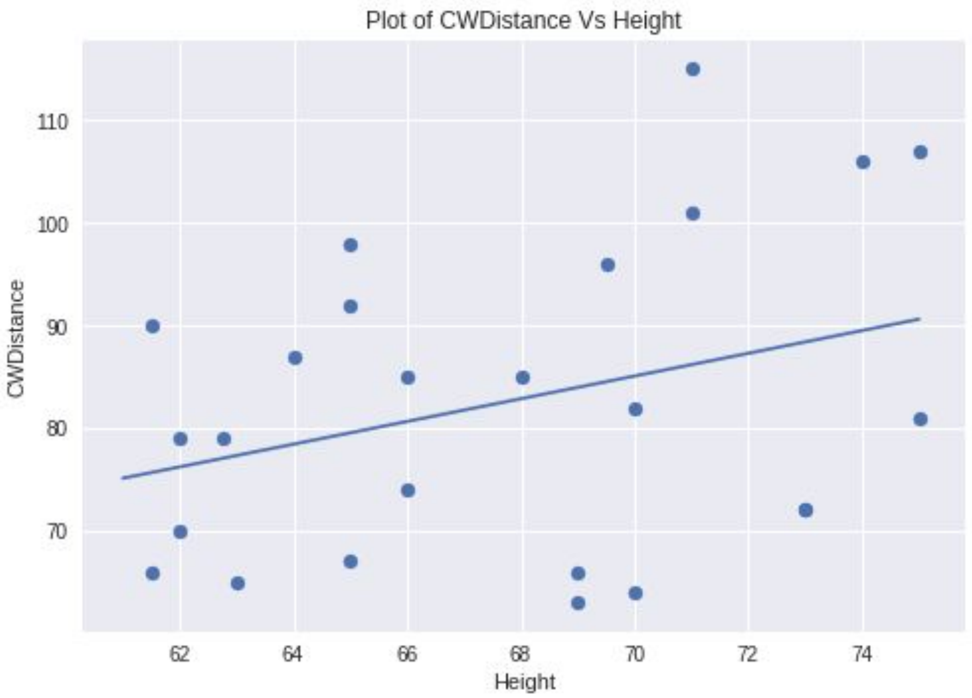
Dependent Variable (DV) = CWDistance

Independent Variable (IV) = Height

- **Form**: ____________________
- **Direction**: ________________
- **Strength**: _________________
- **Outliers**: __________________

PAUSE HERE to provide time for IVQ

Visualizing the Relationship

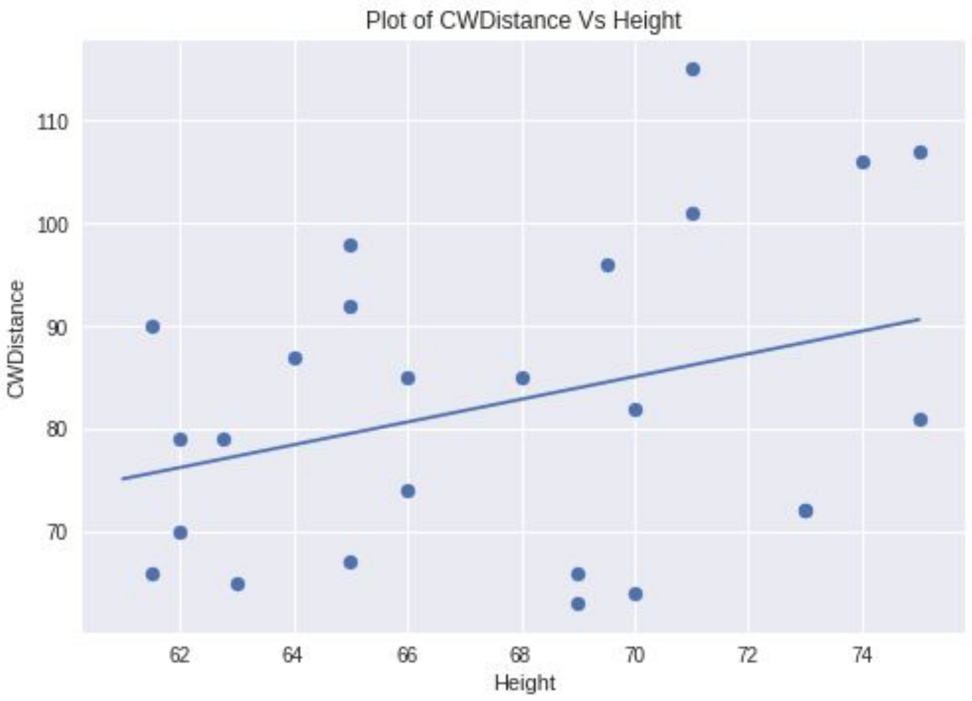


Dependent Variable (DV) = CWDistance

Independent Variable (IV) = Height

- **Form**: approximately linear
- **Direction**: positive
- **Strength**: weak to moderate
- **Outliers**: none apparent

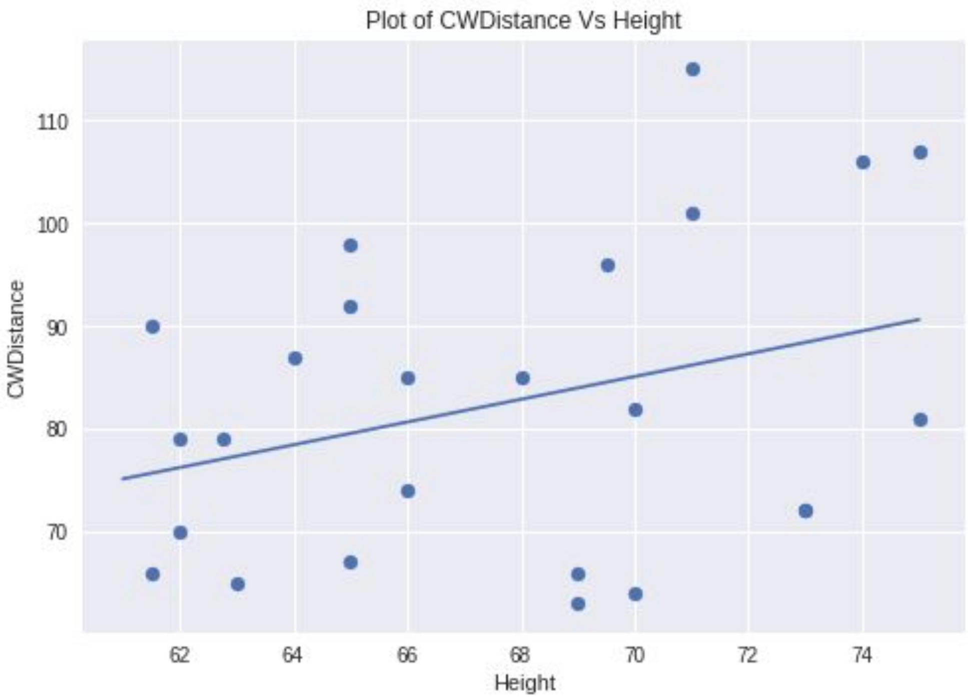
Visualizing the Relationship



- Strength:

  r = 0.33

Visualizing the Relationship



- Strength:

r = 0.33

coefficient

r^2= 0.107 □

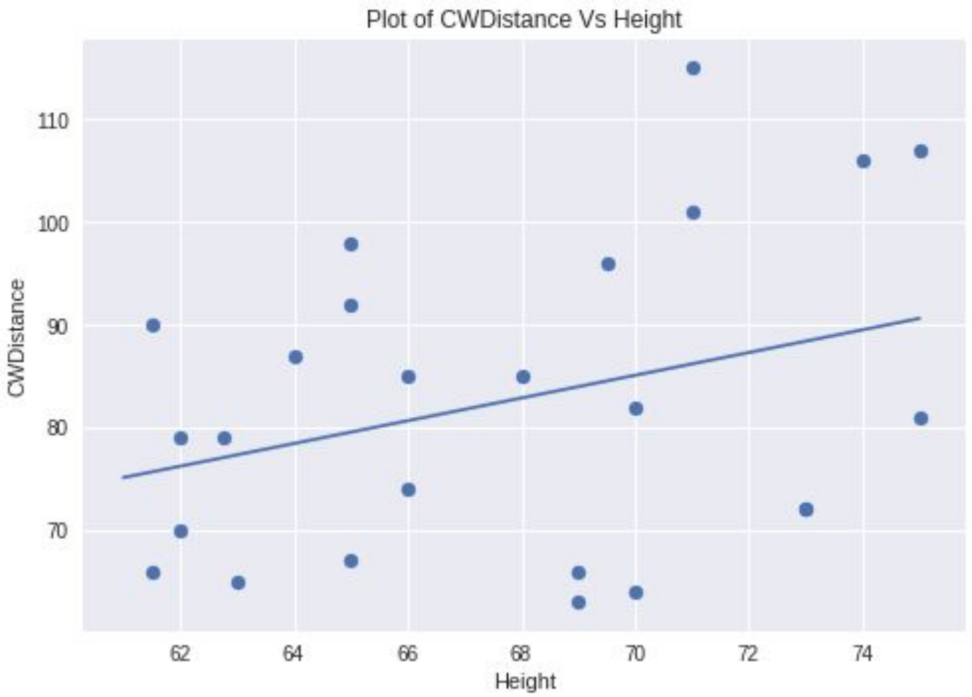
↓ R^2- squared

Only about 11% of the variation in CW Distance is explained by the linear relationship with height

Best Fitting Line

- General Line:

$$y = mx + b$$



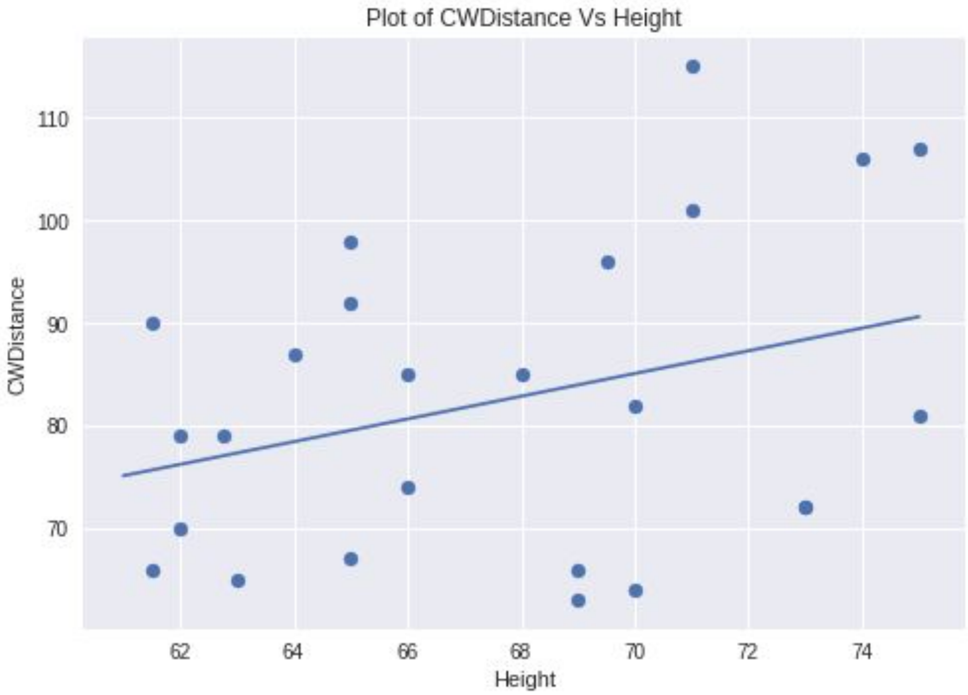
Best Fitting Line

- General Line:

$$y = mx + b$$

- Estimate Regression Line:

$$\hat{y} = b_0 + b_1 x$$



Best Fitting Line

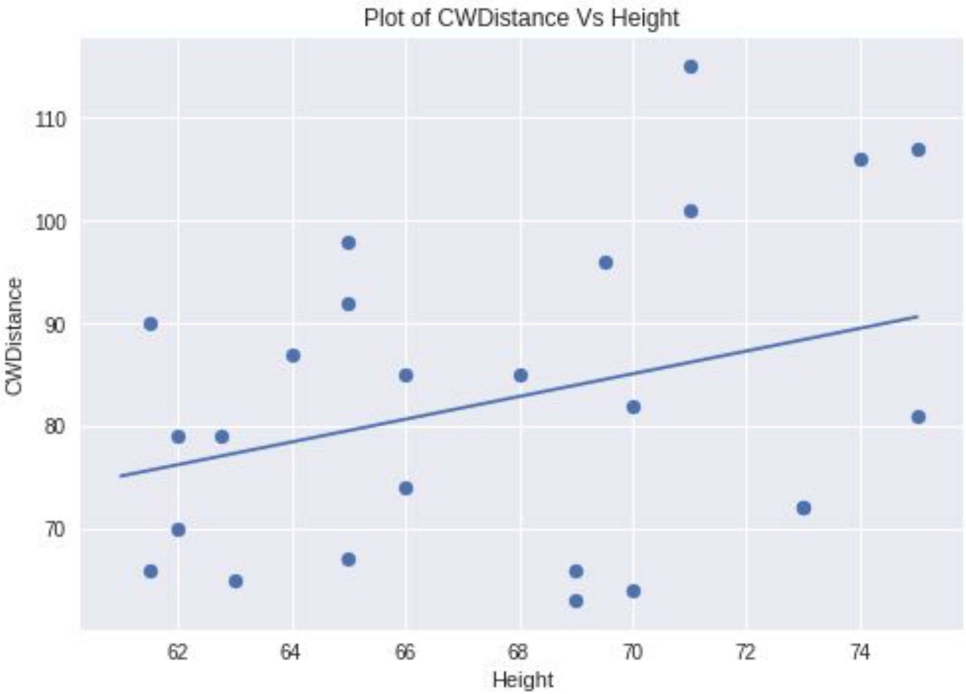
- General Line:

$$y = mx + b$$

- Estimate Regression Line:

$$\hat{y} = b_0 + b_1 x$$

y-intercept:
estimated y when
x = 0
(not always meaningful)



Best Fitting Line

- General Line:

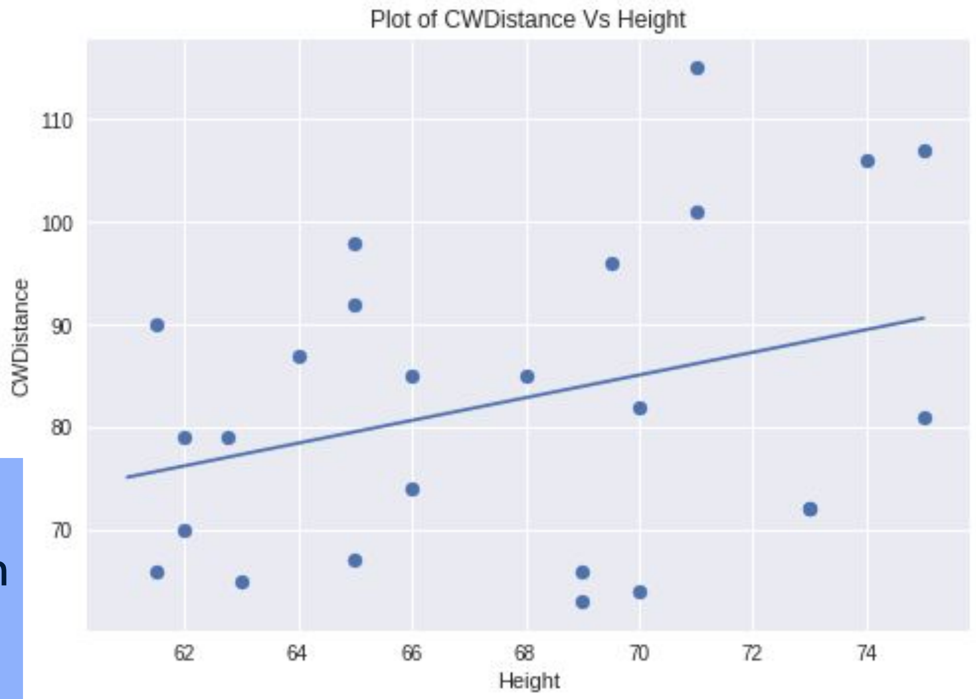
$$y = mx + b$$

- Estimate Regression Line:

$$\hat{y} = b_0 + b_1 x$$

y-intercept: estimated y when x = 0 *(not always meaningful)*

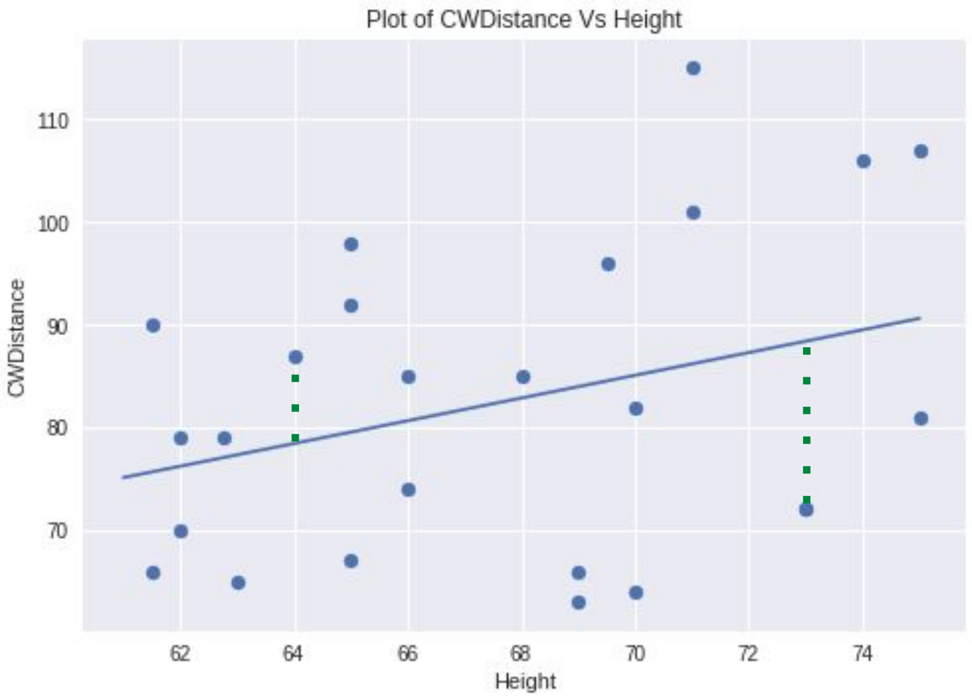
slope: estimated change in y for one unit increase in x



Best Fitting Line

- Estimate Regression Line:

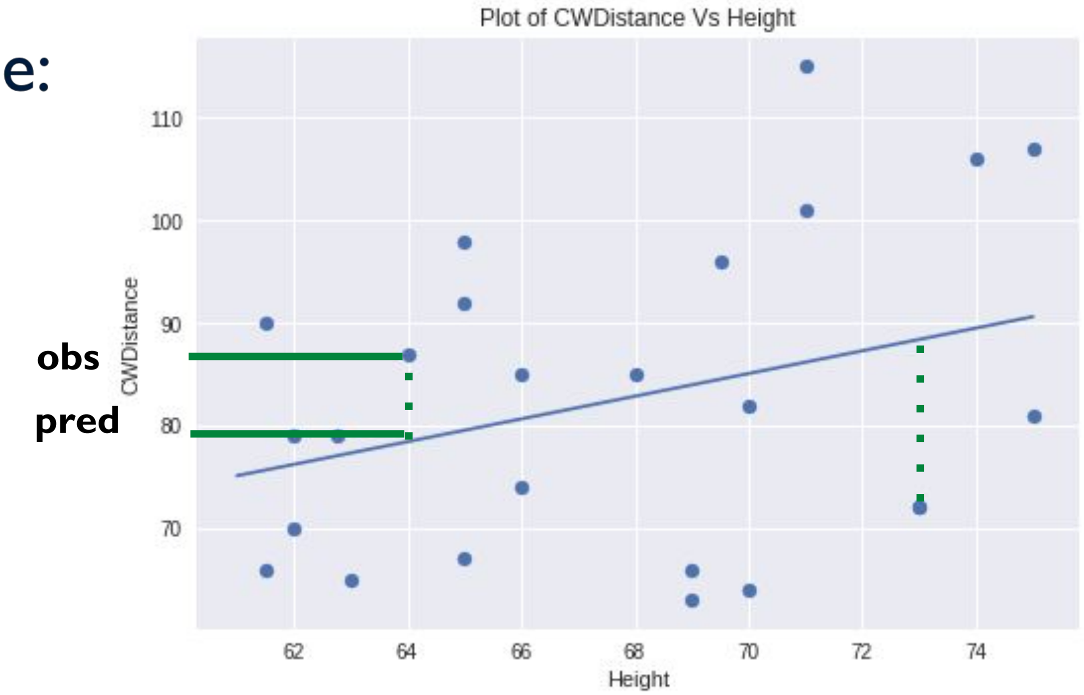
$$\hat{y} = b_0 + b_1 x$$



Best Fitting Line

- Estimate Regression Line:

$$\hat{y} = b_0 + b_1 x$$



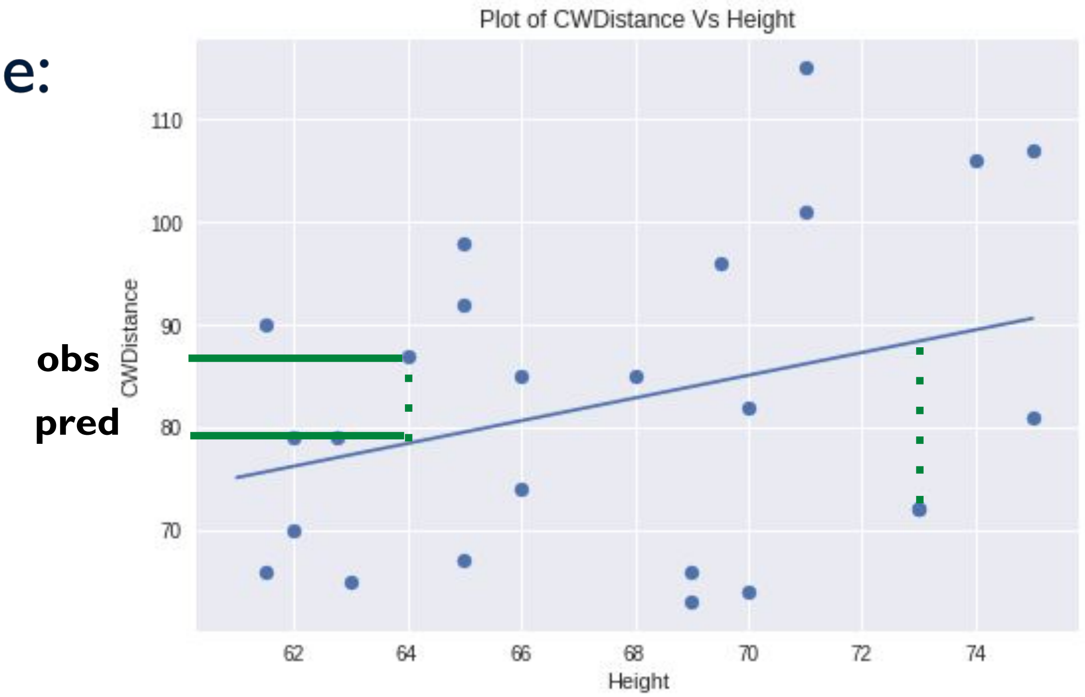
Best Fitting Line

- Estimate Regression Line:

$$\hat{y} = b_0 + b_1 x$$

Goal:

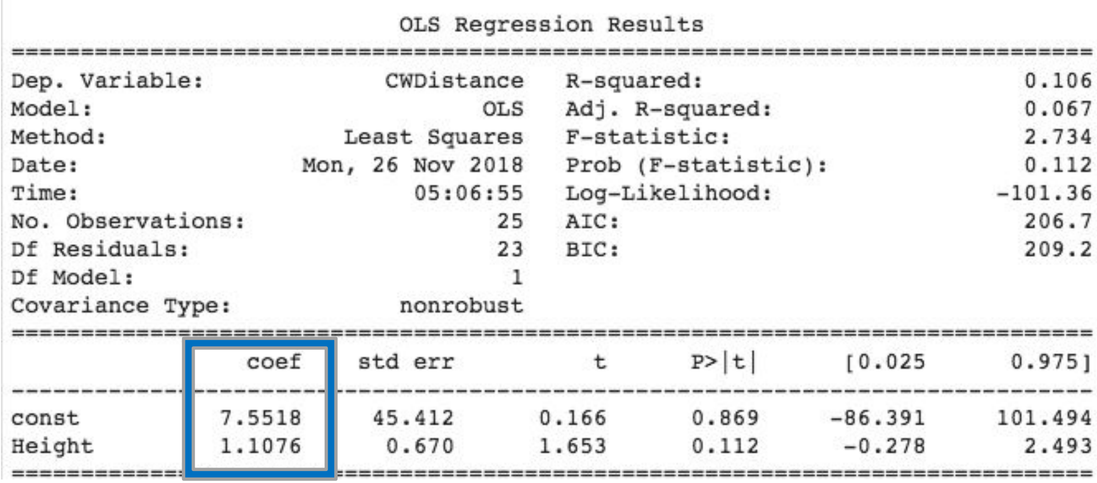
Find line that minimizes total squared (observed) error □ **Least Squares Regression**



Best Fitting Line

Predicted CWDist

= 7.5518 + 1.1076(height)

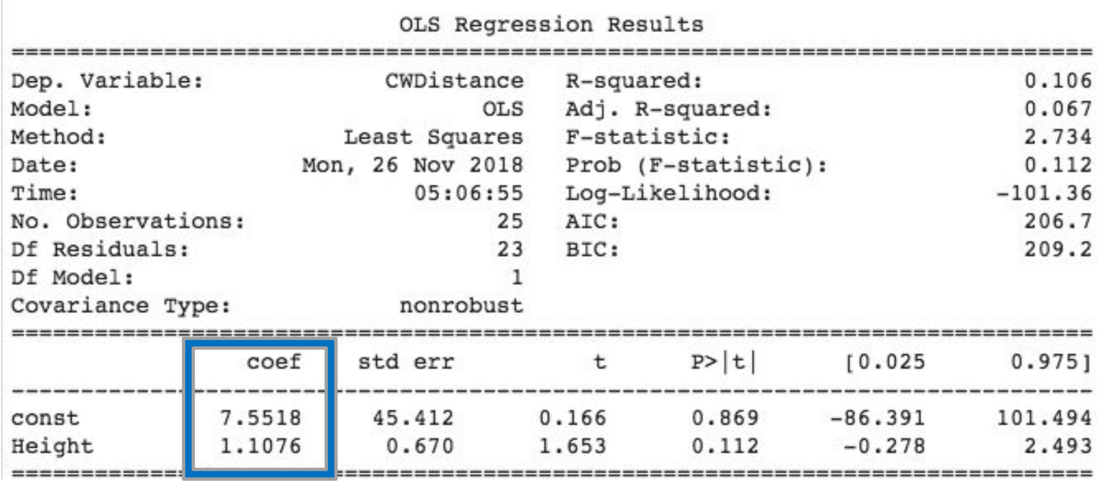


```
                            OLS Regression Results
==============================================================================
Dep. Variable:             CWDistance   R-squared:                       0.106
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     2.734
Date:                Mon, 26 Nov 2018   Prob (F-statistic):              0.112
Time:                        05:06:55   Log-Likelihood:                -101.36
No. Observations:                  25   AIC:                             206.7
Df Residuals:                      23   BIC:                             209.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.5518     45.412      0.166      0.869     -86.391     101.494
Height         1.1076      0.670      1.653      0.112      -0.278       2.493
==============================================================================
```

Best Fitting Line

Predicted CWDist

= 7.5518 + 1.1076(height)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             CWDistance   R-squared:                       0.106
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     2.734
Date:                Mon, 26 Nov 2018   Prob (F-statistic):              0.112
Time:                        05:06:55   Log-Likelihood:                -101.36
No. Observations:                  25   AIC:                             206.7
Df Residuals:                      23   BIC:                             209.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.5518     45.412      0.166      0.869     -86.391     101.494
Height         1.1076      0.670      1.653      0.112      -0.278       2.493
==============================================================================
```

slope: estimated change in y for one unit increase in x □

We would estimate that an adult who is **one inch taller** than another adult would have a **CW distance** that is **1.1 inch longer**, *on average.*

Making Predictions

What would you predict the cartwheel distance to be for an adult who is 64 inches tall?

Predicted CWDist = 7.5518 + 1.1076(height)

PAUSE HERE to provide time for IVQ

Making Predictions

What would you predict the cartwheel distance to be for an adult who is 64 inches tall?

Predicted CWDist = 7.5518 + 1.1076(height)
= 7.5518 + 1.1076(64)
= 78.4382 ~ 78.4 inches

Making Predictions

What would you predict the cartwheel distance to be for an adult who is 64 inches tall?

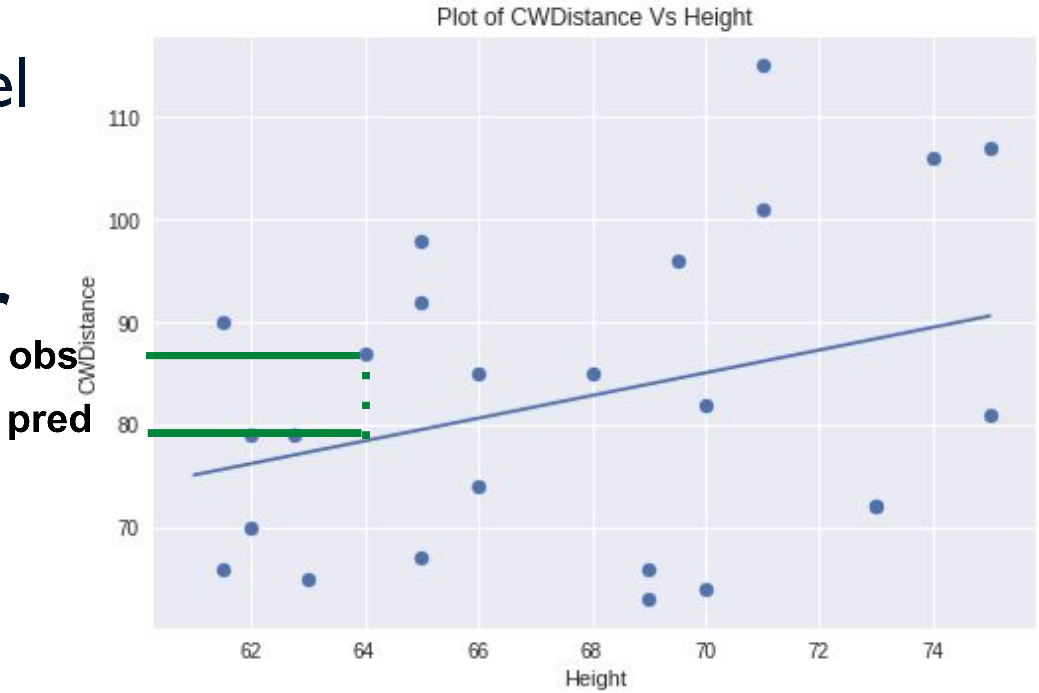
Predicted CWDist = 7.5518 + 1.1076(height)
= 7.5518 + 1.1076(64)
= 78.4382 ~ 78.4 inches

We would also **estimate the mean cartwheel distance for all adults who are 64 inches tall** to be 78.4 inches

Observed Errors (Residuals)

64 inch tall adult had cartwheel distance of 87 inches

What is the **observed error (residual)** for this adult?

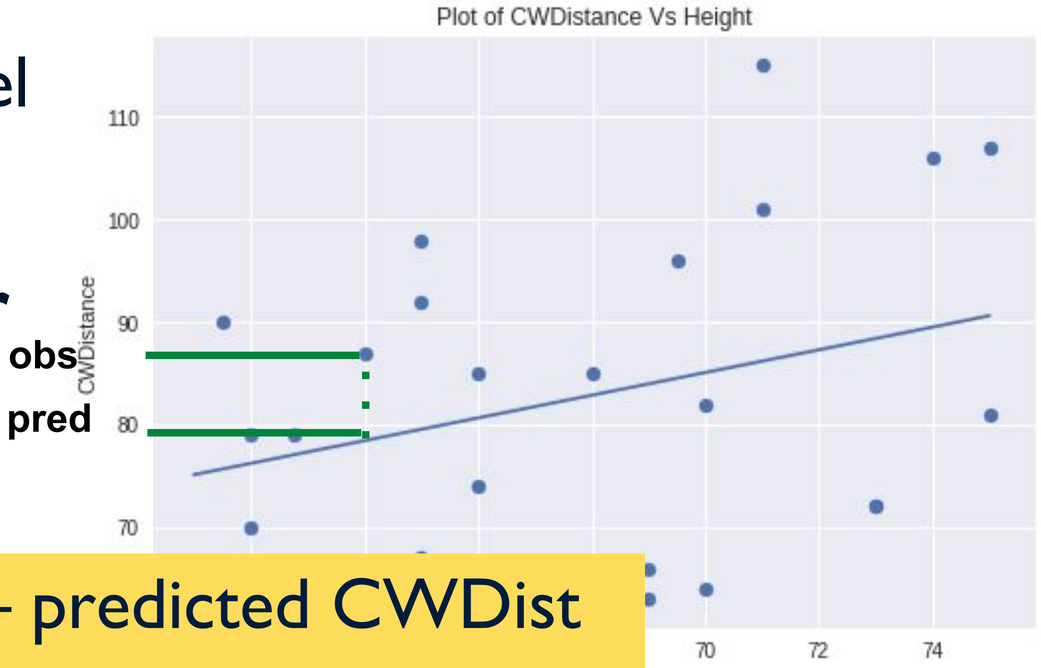


PAUSE HERE to provide time for IVQ

Observed Errors (Residuals)

64 inch tall adult had cartwheel distance of 87 inches

What is the **observed error (residual)** for this adult?



Residual = observed CWDist – predicted CWDist

= 87 inches – 78.4 inches = 8.6 inches

What's Next?

Now that we have worked with the **descriptive** side of regression, we turn to **drawing inferences** from regression:

- **Assessing significance** of the relationship
- **Checking** underlying **assumptions**
- **Extending** regression model to include **more predictors**