Linear Regression Inference
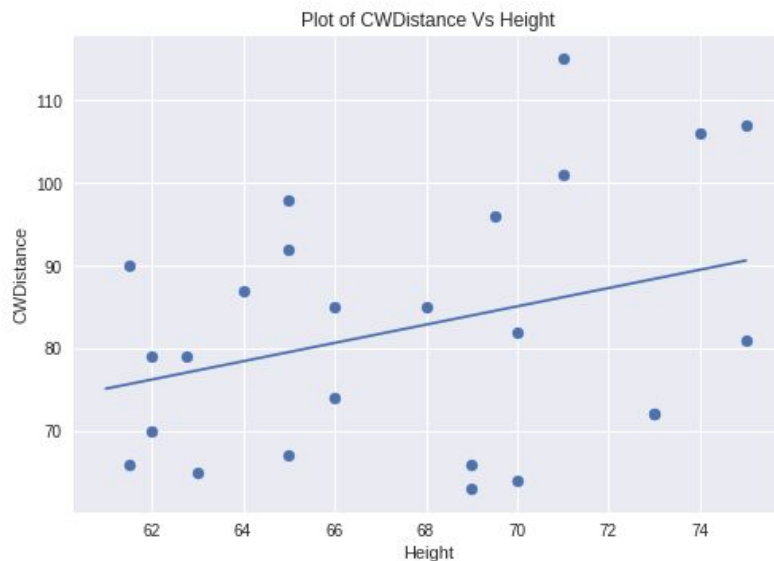
Brenda Gunderson

Regression of CWDistance on Height

Predicted CWDist = 7.5518 + 1.1076(height)
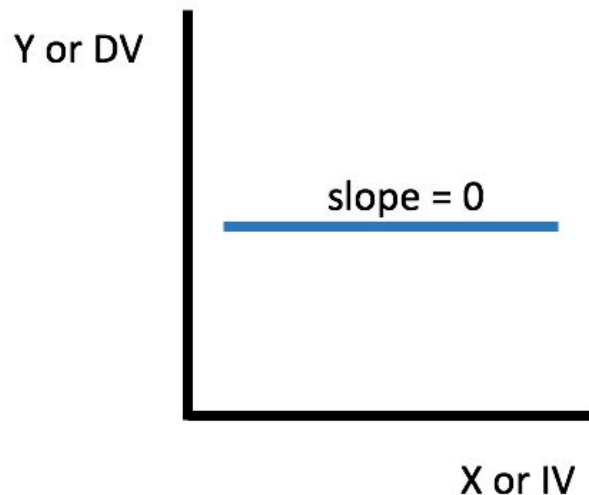

Plot of CWDistance Vs Height

Regression Inference

Is there a significant (positive) linear relationship between CW Distance and Height?

Regression Inference

Is there a significant (positive) linear relationship
between CW Distance and Height?

- **Think about it:**
  What would a slope = 0 imply?
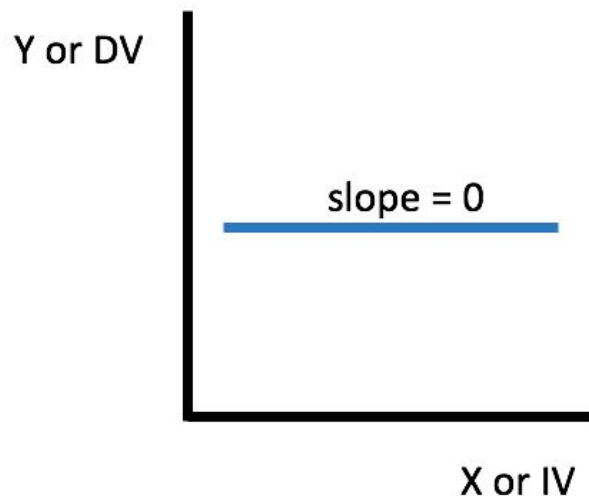
Regression Inference

Is there a significant (positive) linear relationship between CW Distance and Height?

- **Think about it:**

  What would a slope = 0 imply?

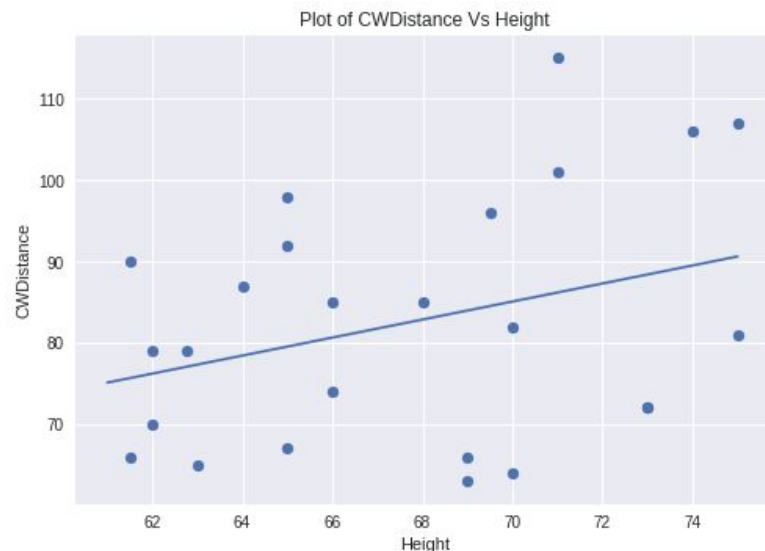 ☐ knowing x does not help to predict y

- Our slope b_1 = 1.1 ~ only an estimated slope

Regression Inference

Is there a significant (positive) linear relationship between CW Distance and Height?

- **Imagine:** have population data on CW Distance and Height of all adults

Plot of CWDistance Vs Height

Regression Inference

Is there a significant (positive) linear relationship between CW Distance and Height?

- **Imagine:** have population data on CW Distance and Height of all adults

- So there is an **underlying *true* slope b_1** want to assess if the true slope is 0 or not *(in our case is it positive > 0)*

Plot of CWDistance Vs Height

Regression Inference

Test H_0: True slope (β_1) = 0

```
=========================================================================
                 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
=========================================================================
```

Regression Inference

Test H_0: True slope (β_1) = 0

```
============================================================================
                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
============================================================================
```

Regression Inference

Test H_0: True slope (β_1) = 0

```
============================================================================
              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
============================================================================
```

Regression Inference

Test H_0: True slope (β_1) = 0

```
==============================================================================
                 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
==============================================================================
```

t-statistic   p-value (two-sided)

Regression Inference

Test H_0: True slope (β_1) = 0

```
==============================================================================
                 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
==============================================================================
```

Two-sided p-value of 0.112 is for testing H_a: True slope (β_1) ≠ 0
For significant positive association test H_a: True slope (β_1) > 0
p-value would be 0.112/2 = 0.056 *(marginally significant)*

Regression Inference

95% Confidence Interval for True slope (β_1)

```
==============================================================================
                 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
==============================================================================
```

With **95% confidence**, the population mean change in cartwheel distance
for one inch increase in height
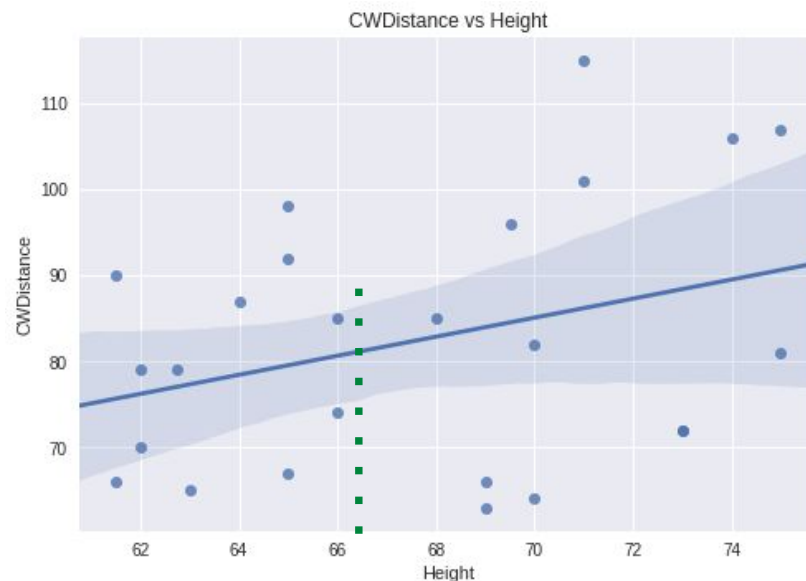is <u>estimated</u> to be anywhere from 0.2 inches shorter to 2.5 inches longer.

More Regression Inference

Used our regression line to **estimate mean cartwheel distance for all adults who are 64 inches tall** to be 78.4 inches

More Regression Inference

Used our regression line to **estimate mean cartwheel distance for all adults who are 64 inches tall** to be 78.4 inches

95% Confidence Interval Bands for Mean CW Distance based on Height □

More Regression Inference

Used our regression line to **estimate mean cartwheel distance for all adults who are 64 inches tall** to be 78.4 inches

95% Confidence Interval Bands
for Mean CW Distance based on Height

Notes:

1. **Intervals** are **narrower** for **values closer to sample mean** height of 67.6 inches

CWDistance vs Height

More Regression Inference

Used our regression line to **estimate mean cartwheel distance for all adults who are 64 inches tall** to be 78.4 inches

95% Confidence Interval Bands
for Mean CW Distance based on Height 🡒

Notes:

1. **Intervals** are **narrower** for **values closer to sample mean** height of 67.6 inches

2. **Prediction Interval** for Individual Response (wider than corresponding CI for mean)

CWDistance vs Height
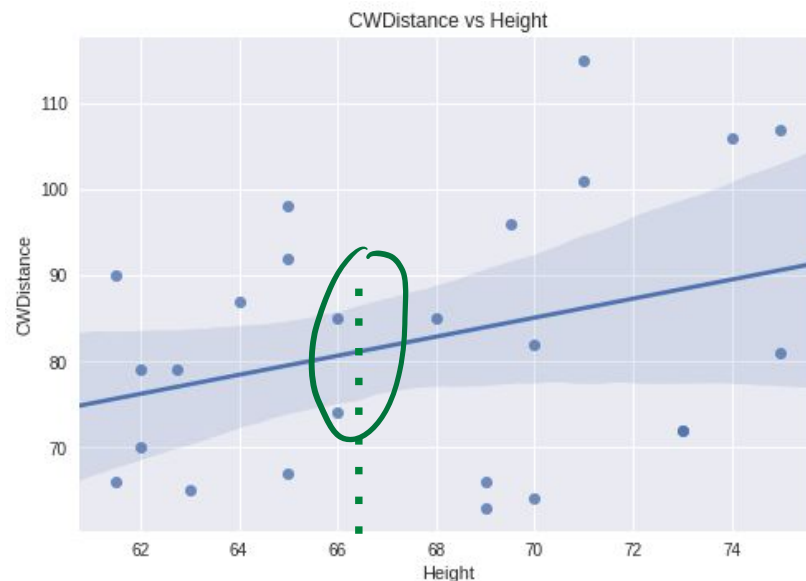
Underlying Assumptions

Fit (population) **regression model:** regressed cart wheel distance on height

$$\textbf{CWDist = b}_0 \textbf{ + b}_1 \textbf{ (height) + e,} \quad \textbf{where e} \sim \textbf{N(0, } \sigma^2\textbf{)}$$

Need to check this!

b_0 **and** b_1 **are two parameters**

e = random error

Errors are **normally distributed**

Checking Assumptions

True errors e ~ N(0, σ^2)

See if residuals (realized values of e):

- appear to be **normally distributed**
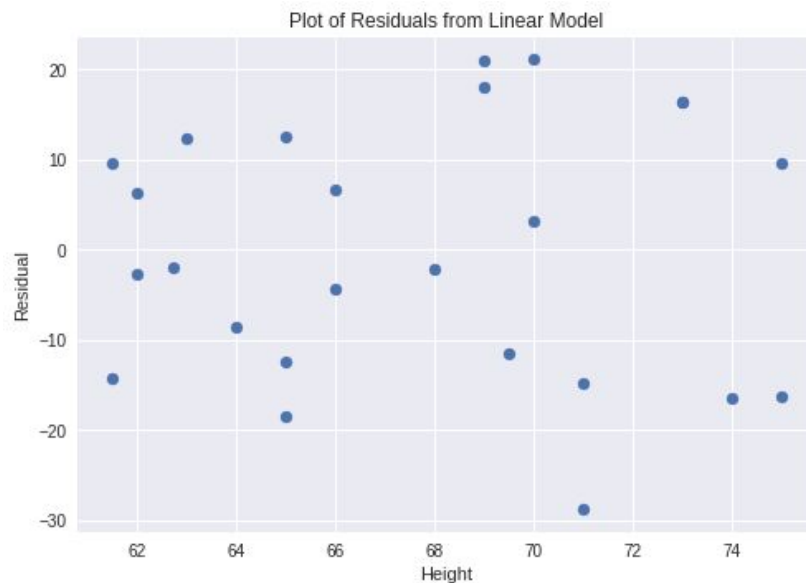

QQPlot of Residuals of CWDistance vs Height

Checking Assumptions

True errors e ~ N(0, σ^2)

See if residuals (realized values of e):

- appear to be **normally distributed**
- are symmetrically distributed around **zero** with **constant variance**
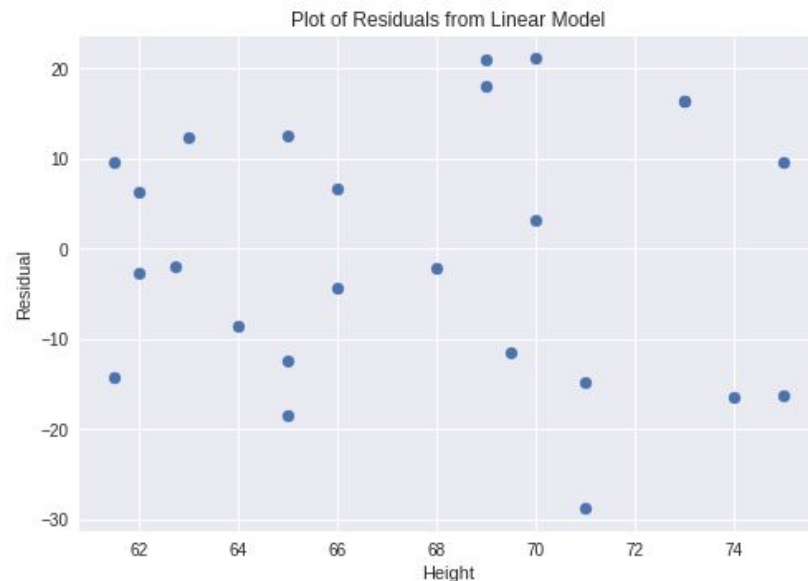- Estimate of σ = 14.5 inches

Plot of Residuals from Linear Model

Checking Assumptions

True errors e ~ N(0, σ^2)

See if residuals (realized values of e):

- appear to be **normally distributed**

- are symmetrically distributed around **zero** with **constant variance**

- Estimate of σ = 14.5 inches

Model fit looks fine
... can we do better?

Plot of Residuals from Linear Model

Adding a second variable

Does knowing if they actually *completed* the cartwheel make a difference in terms of cartwheel distance?

	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

Regression Results

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

```
                        OLS Regression Results
==============================================================================
Dep. Variable:              CWDistance   R-squared:                       0.135
Model:                             OLS   Adj. R-squared:                  0.056
Method:                  Least Squares   F-statistic:                     1.712
Date:                 Mon, 26 Nov 2018   Prob (F-statistic):              0.204
Time:                         05:06:55   Log-Likelihood:                 -100.95
No. Observations:                   25   AIC:                             207.9
Df Residuals:                       22   BIC:                             211.6
Df Model:                            2
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -7.0457     48.805     -0.144      0.887    -108.261      94.170
Height         1.2557      0.696      1.804      0.085      -0.188       2.699
Complete       6.0190      7.077      0.851      0.404      -8.657      20.695
==============================================================================
Omnibus:                         1.786   Durbin-Watson:                   1.876
Prob(Omnibus):                   0.409   Jarque-Bera (JB):                1.078
Skew:                            0.137   Prob(JB):                        0.583
Kurtosis:                        2.020   Cond. No.                     1.13e+03
==============================================================================
```

Regression Results: Interpreting Coefficients

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

Two adults with same completion status whose height differ by 1 inch tend to have cart wheel distances differing by 1.26 inches.

Regression Results: Interpreting Coefficients

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

Two adults with same completion status whose height differ by 1 inch tend to have cart wheel distances differing by 1.26 inches.

Comparing adult who completed cartwheel with one of same height who did not: completer will on average have a CW Distance of 6 inches longer.

Regression Results: Interpreting Coefficients

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

Two adults with same completion status whose height differ by 1 inch tend to have cart wheel distances differing by 1.26 inches.

Comparing adult who completed cartwheel with one of same height who did not: completer will on average have a CW Distance of 6 inches longer.

Height coefficient of 1.26 is only meaningful
when comparing two adults of the same completion status.
Complete coefficient of 6 is only meaningful
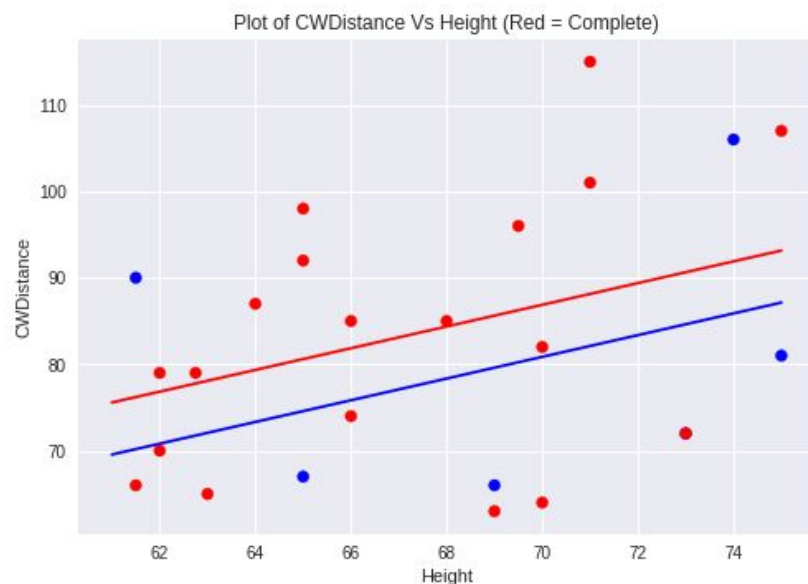when comparing two adults of the same height.

Visualizing Regression Results

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

Plot of CWDistance Vs Height (Red = Complete)
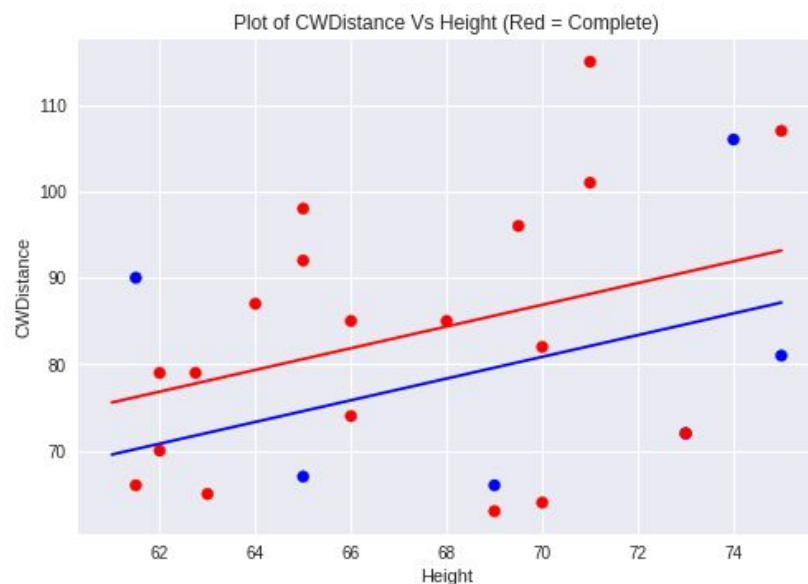
Visualizing Regression Results

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

Regression Results

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

```
==============================================================
              coef      std err          t      P>|t|
--------------------------------------------------------------
Intercept   -7.0457      48.805     -0.144      0.887
Height       1.2557       0.696      1.804      0.085
Complete     6.0190       7.077      0.851      0.404
==============================================================
```

After adjusting for completion status, does there appear to be a significant positive linear relationship between CW Distance and Height?

PAUSE HERE to provide time for IVQ

Regression Results

Predicted CWDist = -7.0457 + 1.2557(Height) + 6.0190(Complete)

```
================================================
              coef    std err         t     P>|t|
------------------------------------------------
Intercept   -7.0457    48.805    -0.144    0.887
Height       1.2557     0.696     1.804    0.085
Complete     6.0190     7.077     0.851    0.404
================================================
```

After adjusting for completion status,

Estimate of Height coefficient = 1.26 (SE = 0.7) *two-sided*

p-value for assessing significant positive association = 0.085/2 = 0.0425

Estimate of σ = 14.6 inches

Summary

- **Regression** for predicting a quantitative response (DV) based on one or more explanatory variables (IV) (quantitative or categorical)

- **Inference** side: Confidence Intervals and Hypothesis Tests

- **Assumptions** for Inference

- *Coming up next:* Regression models when the **response (DV) is binary** called Logistic Regression