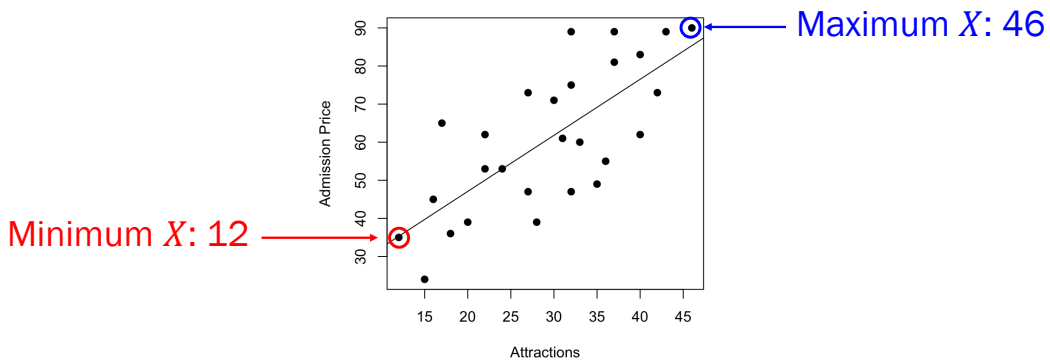
Prediction Intervals and Confidence Intervals

- Extrapolation
- Confidence Intervals
- Prediction Intervals
- Comparing Confidence and Prediction Intervals
- Confidence and Prediction Bands

University of Pittsburgh

Extrapolation

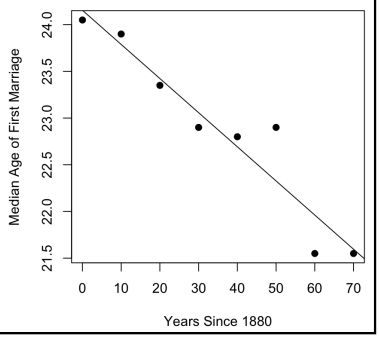
- **Extrapolation:** the process of making a prediction outside of the observed range of predictor values used to construct the original regression line



Predicting the admission price when the number of attractions is less than 12 or greater than 46 would be extrapolating.

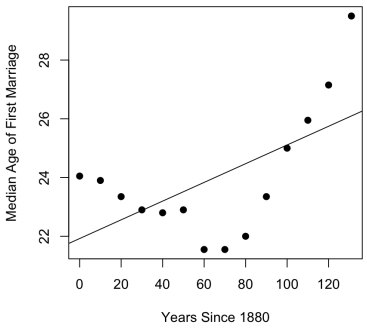
Example: Extrapolation

- **Scenario:** Use the number of years since 1890 to predict median age at the time of a person's first marriage.
  - **Regression Line:** $\hat{Y} = 24.15 - 0.0365x$
- **Question:** What would be the predicted median age of a person's first marriage in 1960? What about 2021?
- **Answer:**
  - **1960:** $\hat{Y} = 24.15 - 0.0365(70) =$ _____
  - **2021:** $\hat{Y} = 24.15 - 0.0365(131) =$ _____
- **Actual Median Ages and Residuals:**
  - **1960:** 21.55 → $e = 21.55 - 21.60 =$ _____
  - **2021:** 29.50 → $e = 29.50 - 19.37 =$ _____



Example: Extrapolation

- **Scenario:** Use the number of years since 1890 to predict median age at the time of a person's first marriage.
- **Question:** What is the problem with using this regression line to predict future observations?
- **Answer:** This trend obviously _____ and had to _____ or _____ at some point
  - Eventually we would predict _____
- **True Relationship:** _____
  - Our perception changed as time passed and we _____
  - Linear regression was _____



Extrapolation Takeaways

- Extrapolating makes the assumption that the relationship between the predictor and response _____ at all X values, even at _____ values.
- In reality, we only have an understanding of how the data behaves within the _____ values in the sample.
  - Even this is only a _____ because we only have a _____.
- If we believe that the relationship between the predictor and response is similar beyond the scope of the collected data, we can _____.
  - However, this comes at the cost of having much _____ and _____ of plausible values of our predictions.

Two Types of Intervals in Regression

- **Recall:** A confidence interval can be used to estimate the slope parameter, gauging if a significant relationship exists.
- There are two types of intervals that can be used to perform inference on the **predictions** themselves:
  - Confidence Interval
  - Prediction Interval
- Before calculating either interval, need to choose a specific value of the predictor that we are interested in studying.
  - *Reason: In regression, the response _____ depending on the value of the _____ that is selected.*

Confidence Interval

- **Confidence interval:** an interval estimate that provides a range of plausible values for where the mean of all responses for a chosen value of the predictor will fall

$$\hat{Y} \pm t_{n-2} \times S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)s_X^2}}$$

  - Choose the value of the predictor X_g to perform inference at
  - Center the interval at _____
  - Confidence interval yields a range of plausible values for the _____ _____ when the value of the predictor is X_g

Prediction Interval

- **Prediction interval:** an interval estimate that provides a range of plausible values for where a single response for a chosen value of the predictor will fall

$$\hat{Y} \pm t_{n-2} \times S_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)s_X^2}}$$

  - Choose the value of the predictor X_g to perform inference at
  - Center the interval at $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_g$
  - Prediction interval yields a range of plausible values for where a _____ _____ is likely to fall when the value of the predictor is X_g

Four Types of Inference

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price

Value	Mean	Std. Dev.	Sample Size
Price (Y)	61	18.80	27
Attractions (X)	30	9.33	27

- **Question:** How can we perform inference on…
  1. Average admission price for parks with 30 attractions?
  2. Admission price for a single park with 30 attractions?
  3. Average admission price for parks with 40 attractions?
  4. Admission price for a single park with 40 attractions?

Four Types of Inference

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.6233     8.5849   1.936   0.0642 .
attractions   1.4792     0.2737   5.404 1.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.02 on 25 degrees of freedom
Multiple R-squared:  0.5388,    Adjusted R-squared:  0.5204
F-statistic: 29.21 on 1 and 25 DF,  p-value: 1.314e-05
```

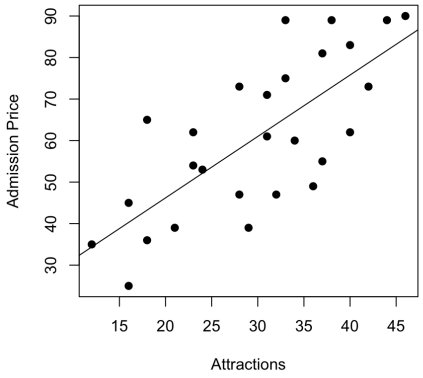
Degrees of Freedom	0.25	0.20	0.15	0.10	0.05	0.025
1	1.000	1.376	1.963	3.078	6.314	12.71
2	0.816	1.061	1.386	1.886	2.920	4.303
3	0.765	0.978	1.250	1.638	2.353	3.182
4	0.741	0.941	1.190	1.533	2.132	2.776
5	0.727	0.920	1.156	1.476	2.015	2.571
6	0.718	0.906	1.134	1.440	1.943	2.447
7	0.711	0.896	1.119	1.415	1.895	2.365
8	0.706	0.889	1.108	1.397	1.860	2.306
9	0.703	0.883	1.100	1.383	1.833	2.262
10	0.700	0.879	1.093	1.372	1.812	2.228
11	0.697	0.876	1.088	1.363	1.796	2.201
12	0.695	0.873	1.083	1.356	1.782	2.179
13	0.694	0.870	1.079	1.350	1.771	2.160
14	0.692	0.868	1.076	1.345	1.761	2.145
15	0.691	0.866	1.074	1.341	1.753	2.131
16	0.690	0.865	1.071	1.337	1.746	2.120
17	0.689	0.863	1.069	1.333	1.740	2.110
18	0.688	0.862	1.067	1.330	1.734	2.101
19	0.688	0.861	1.066	1.328	1.729	2.093
20	0.687	0.860	1.064	1.325	1.725	2.086
21	0.686	0.859	1.063	1.323	1.721	2.080
22	0.686	0.858	1.061	1.321	1.717	2.074
23	0.685	0.858	1.060	1.319	1.714	2.069
24	0.685	0.857	1.059	1.318	1.711	2.064
25	0.684	0.856	1.058	1.316	1.708	2.060

Example: Point Estimates

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price. Regression line is $\hat{Y} = 16.62 + 1.479X$.

- **Question:** What is the predicted admission price for a park with 30 attractions?
- **Answer:** $\hat{Y} =$ _____ $=$ _____

- **Question:** What is the predicted admission price for a park with 40 attractions?
- **Answer:** $\hat{Y} =$ _____ $=$ _____



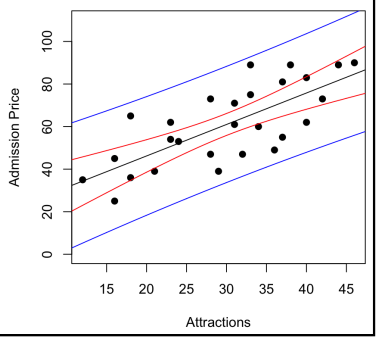
Example: Approximate Mean Response at $X_g = 30$

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price.
- **Question:** What interval approximates the average admission price for all parks with 30 attractions?
- **Answer:** _____

_____

_____

_____



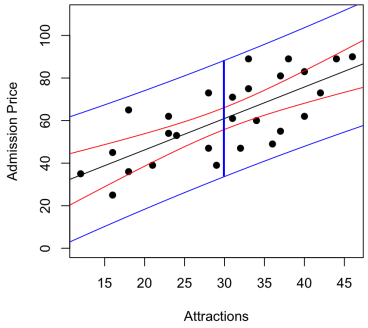
Example: Predict Individual Response at $X_g = 30$

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price.
- **Question:** What interval approximates the admission price for a single park with 30 attractions?
- **Answer:** _____

_____

_____

_____



Example: Interpreting Intervals

- **Scenario:** 95% confidence interval and prediction intervals were:
  - **Confidence Interval:** (55.84, 66.16)
  - **Prediction Interval:** (33.69, 88.31)
- **Question:** What is the interpretation of the confidence interval?
- **Answer:** We are _____ that the _____ admission price for _____ with ____ attractions is between _____ and _____.

- **Question:** What is the interpretation of the prediction interval?
- **Answer:** We are _____ that the admission price for a _____ with ____ attractions is between _____ and _____.

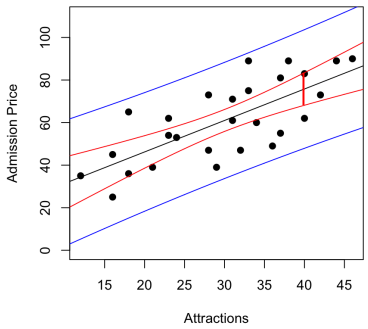
Example: Approximate Mean Response at $X_g = 40$

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price.
- **Question:** What interval approximates the average admission price for all parks with 40 attractions?
- **Answer:** _____

_____

_____

_____



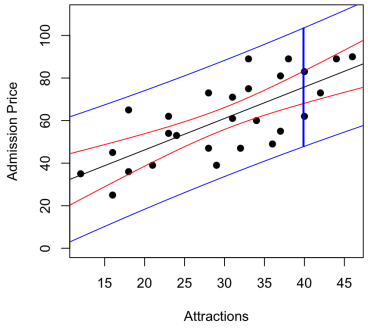
Example: Predict Individual Response at $X_g = 40$

- **Scenario:** Use number of attractions at 27 amusement parks to predict admission price.
- **Question:** What interval approximates the admission price for a single park with 40 attractions?
- **Answer:** _____

_____

_____

_____



Confidence Interval vs. Prediction Interval at Same X_g

- **Question:** Which interval is wider for 30 attractions? What about for 40?

Interval	30 Attractions	40 Attractions
Confidence	(55.84, 66.16)	(68.14, 83.42)
Prediction	(33.69, 88.31)	(47.89, 103.67)

- **Answer:** _____ for both numbers of attractions
- **Reason:** Prediction intervals predict a _____ while confidence intervals approximate the _____
  - More variability when sampling _____ observation compared to _____

CI: $\sqrt{\frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)s_X^2}}$

Extra '1' makes margin of error _____

PI: $\sqrt{1 + \frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)s_X^2}}$

Comparing Intervals at Different Values of X_g

- **Question:** Which confidence interval is wider?

Interval	30 Attractions	40 Attractions
Confidence	(55.84, 66.16)	(68.14, 83.42)

- **Answer:** Confidence interval for ____ attractions
  - Width of _____ vs. _____

- **Question:** Which prediction interval is wider?

Interval	30 Attractions	40 Attractions
Prediction	(33.69, 88.31)	(47.89, 103.67)

- **Answer:** Prediction interval for ____ attractions
  - Width of _____ vs. _____

Comparing Intervals at Different Values of X_g

- **Question:** Why were the intervals for $X_g = 40$ wider than $X_g = 30$?
- **Answer:** Both confidence intervals and prediction intervals will get _____
  - Tend to have _____ about values _____ than those _____ away so the predictions are _____

If X_g is farther from $\bar{X}$, then _____ becomes larger, which makes the entire _____ larger.

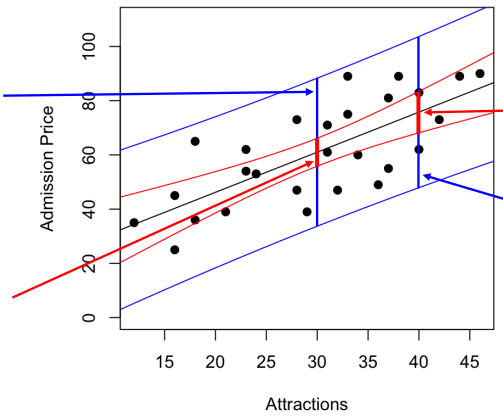
$$\text{CI: } \sqrt{\frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)s_X^2}} \qquad \text{PI: } \sqrt{1 + \frac{1}{n} + \frac{(X_g - \bar{X})^2}{(n-1)s_X^2}}$$

Confidence Bands and Prediction Bands

- **Confidence band:** a curved line drawn on a scatterplot representing the set of all confidence intervals for a regression
- **Prediction band:** a curved line drawn on a scatterplot representing the set of all prediction intervals for a regression

95% PI @ $X = 30$: $(33.69, 88.31)$

95% CI @ $X = 30$: $(55.84, 66.16)$



95% CI @ $X = 40$: $(68.14, 83.42)$

95% PI @ $X = 40$: $(47.89, 103.67)$