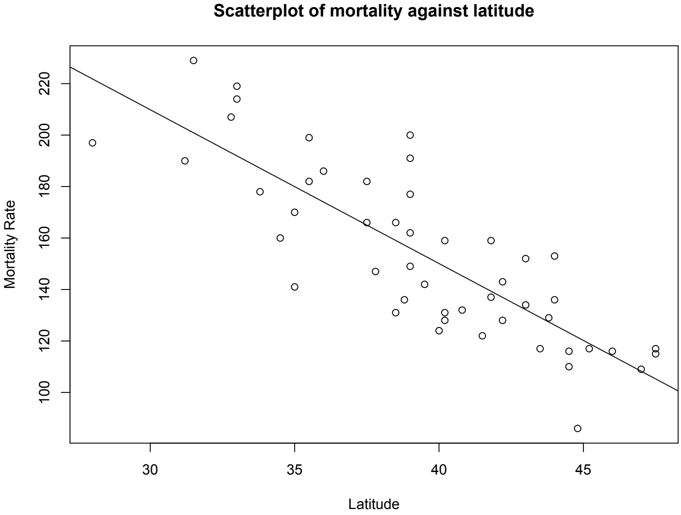
Example from Dr. Gulessarian (Stat 110 last year)

Example: the skincancer.txt dataset contains the mortality rates due to skin cancer from 48 continental states + Washington DC. The goal is to assess if the latitude of the state predicts (or explains) the mortality rate of skin cancer.

- Y = mortality rate.
- X = latitude.

Simple Linear Regression



Simple Linear Regression: Fitting the Model

Fitting the model using the data, the output is:

```
lm(formula = Mort ~ Lat, data = skincancer)
Residuals:
    Min      1Q  Median      3Q     Max
-38.972 -13.185   0.972  12.006  43.938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
---

Residual standard error: 19.12 on 47 degrees of freedom
Multiple R-squared:  0.6798, Adjusted R-squared:  0.673
F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

Simple Linear Regression

Example: the skincancer.txt dataset contains the mortality rates due to skin cancer from 48 states + DC. The goal is to assess if the latitude of the state predicts (or explains) the mortality rate of skin cancer.

- The regression equation is $\hat{Y}_i = 389.18 - 5.97X_i$.
- The residual standard error, $\hat{\sigma_\varepsilon}$ is 19.12.
- Degrees of freedom, $n - 2$, is equal to 47. Therefore $n = 47 + 2 = 49$ observations.

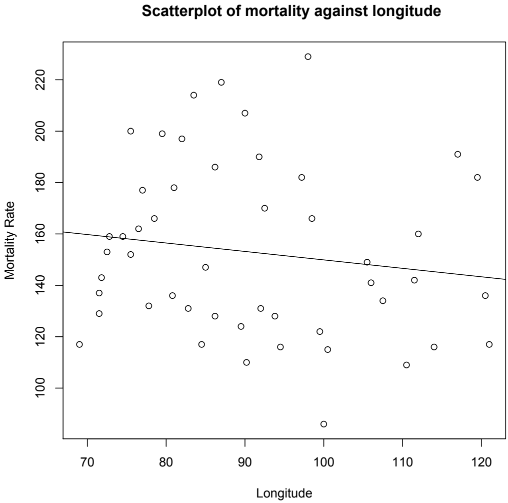
Simple Linear Regression

For now, we will look at the multiple R-squared value for R^2.

- $R^2 = 0.6798$.
- This to say that 68% of the variation in Y is explained by X.
    - 68% of the variation in mortality is explained by the latitude.

Simple Linear Regression

Using the skin cancer data again, lets look at the case where $X=$ Longitude (instead of latitude).



Simple Linear Regression: Fitting the Model

Fitting the model using the data, the output is:

```
  lm(formula = Mort ~ Long, data = skincancer)
  Residuals:
      Min      1Q  Median      3Q     Max
  -63.898 -25.995  -5.952  21.856  78.444

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) 182.7696    29.8893   6.115  1.8e-07 ***
  Long         -0.3287     0.3245  -1.013    0.316
  ---

Residual standard error: 33.42 on 47 degrees of freedom
Multiple R-squared:  0.02137, Adjusted R-squared:  0.0005491
F-statistic: 1.026 on 1 and 47 DF,  p-value: 0.3162
```

Simple Linear Regression

- Can see that longitude is not nearly as good a predictor as latitude.
- $R^2 = 0.02$.
- This to say that 2% of the variation in Y is explained by X.
  - 2% of the variation in mortality is explained by the longitude.

Simple Linear Regression

Quick review using the skin cancer dataset.

- Hospital records were used to record the mortality rate for each state.
- This is an observational study, since subjects were not randomized to live in a state.
- Can we say that latitude causes mortality rates to increase?
- Any possible confounders?