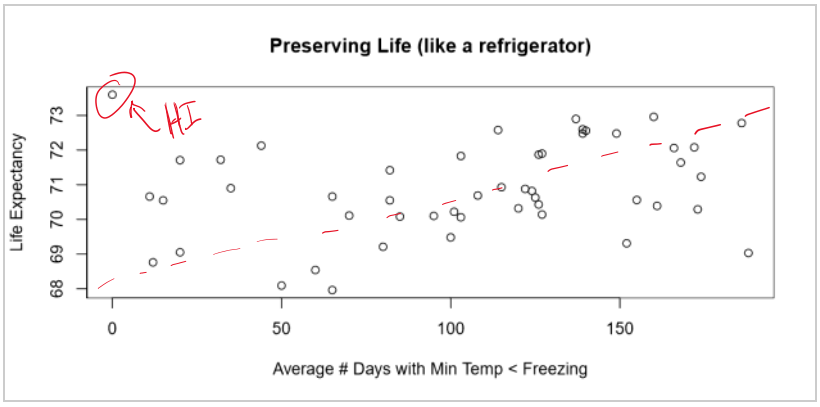
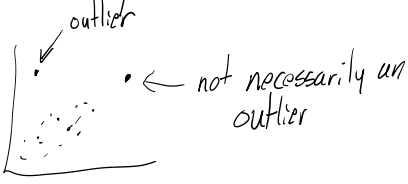
Inference for Regression: CIs and PIs

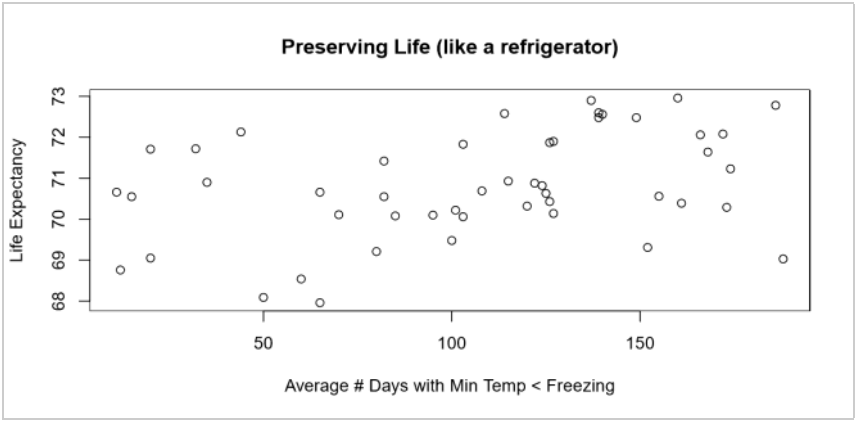
(Also Outliers....)



Data on US states, looking at life exp and Avg # of days with min temp $< 0^\circ C$, or $< 32^\circ F$

In the context of regression outliers are points that don't fit the trend of the data.

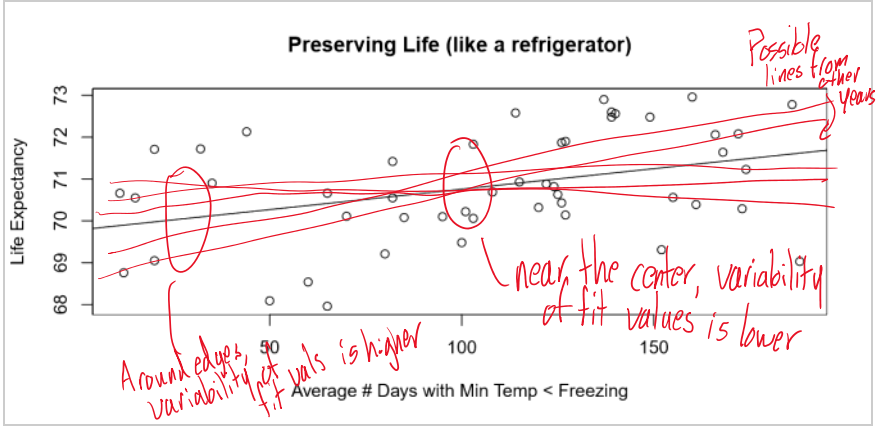




$\hat{y} = 69.78 \text{ years} + 0.009781 \frac{\text{years}}{\text{day}} \times X$

$\hat{b}_0 = 69.78$

$\hat{b}_1 = 0.009781$



What if I want to predict life expectancy for a state with $x = 125$ days.

What if I want to predict life expectancy for a state with $x = 125$ days.

$\hat{y}$ Average? Reasonable

Our model is wrong.

$\hat{y}$ is close to the average, but is subject to variability

We can estimate the value of $\hat{y}$ for a specific x value while acknowledging the variability and uncertainty of our regression model using confidence intervals

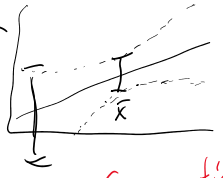
The variability of $\hat{y}$ depends on the accuracy of our line fit.

Predicting the true value of $\hat{y}$ is "easier" (less variability) for points closer to $\bar{x}$, and "harder" (more variation) for points further from $\bar{x}$.

$$SE \text{ of } \hat{y}(x_0) = \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

($x = x_0$; highest when x_0 is far from $\bar{x}$)

MSE: Mean Squared Error of residuals $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



95% CI has degrees of freedom of $n-2$

$df = 49 - 2 = 47$

$$\hat{y}_0 \pm T_{.975,\, df=47} \times SE_{\hat{y}(x_0)} \quad \text{(or } SE_{\mu_y}\text{)}$$

($x = x_0$)

Correction from video, should be $\frac{1}{n-2}$ for MSE, not $\frac{1}{n}$

We are 95% confident that the true mean life expectancy for a state with $x_0 = 125$ days freezing is between (70.64y, 71.37y)

A Prediction Interval is an interval in which we expect some percentage of individual values to fall (We are not discussing average, but the range and likely values)

For CI we ask: What the average life exp for a state with $x_0 = 125$ days freezing?

For PI we ask: What is the range of typical values for a state with $x = 125$ days freezing?

For PI we ask: What is the range of typical values for a state with $X_0 = 125$ days freezing?

df is same for PI and CI

95% PI / $\hat{y} \pm T_{.975, df=47} \cdot SE_{\hat{y}(X_0)}$

(This SE is the only difference)

$$SE_{\hat{y}_0(X_0)} = \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)}$$

(Almost identical except for....)

$$= \sqrt{MSE + \left(SE_{\hat{y}(X_0)}\right)^2}$$

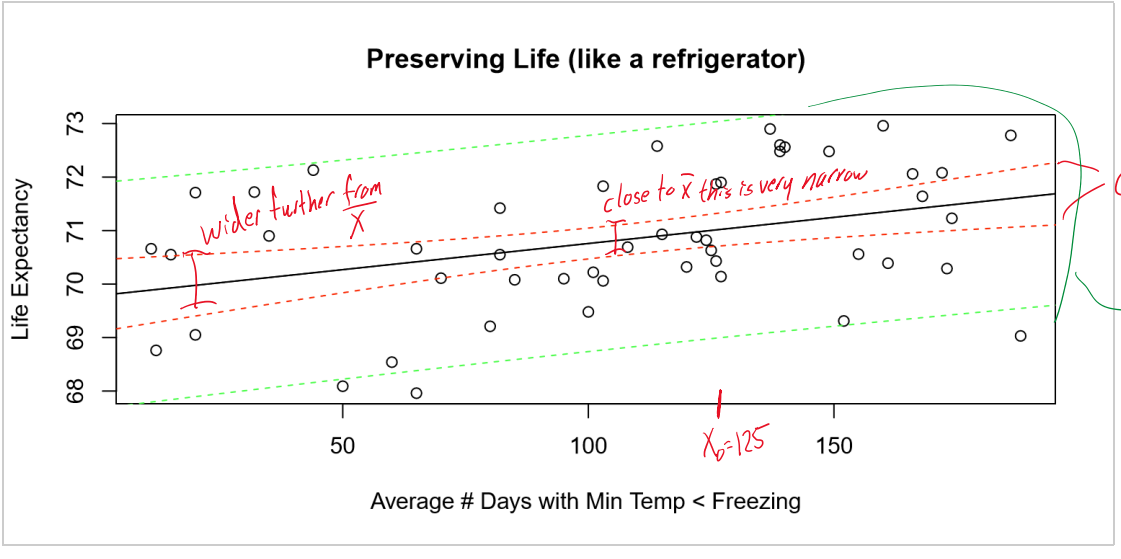
MSE: Variability of residuals

$SE_{\hat{y}(X_0)}$: This is uncertainty in the mean (same as for CI)

For $X_0 = 125$

using R as PI = (68.59y, 73.42y)

We predict that 95% of points for $X = 125$ days will fall between 68.6 years and 73.4 years.



90% CI and PI for the life exp data.