

Supplement for Lecture 12: Intervals for Prediction

Load and Clean Data

Variables of Interest in `fatal` - `adj_fatal` = Number of Vehicle Fatalities Per 1,000 People - `yd` = Percent of Drivers 15 - 24

```
data("Fatalities") # Load Data

fatal = Fatalities[,c("fatal","pop","youngdrivers")]
fatal$adj_fatal = (fatal$fatal/fatal$pop)*1000
fatal$yd=fatal$youngdrivers*100

fatal$fatal=NULL
fatal$pop=NULL
fatal$youngdrivers=NULL

head(fatal)
```

```
##   adj_fatal    yd
## 1  0.212836 21.1572
## 2  0.234848 21.0768
## 3  0.233643 21.1484
## 4  0.219348 21.1140
## 5  0.266914 21.3400
## 6  0.271859 21.5527
```

Fit Linear Regression Model

```
#Fit Linear Regression Model
mod = lm(adj_fatal~yd,data=fatal)

#Results from t-test for slope
summary(mod)

#Results from ANOVA F-test for Effectiveness of SLR Model
anova(mod)

#Relationship between test statistics between both tests
```

Correlation Test and R^2

```
# Correlation Test (Compare p-value to previous p-values)

# R-squared
```

Interpretation: We are able to explain 5% of the variation in the number of fatalities per 1000 people in a

state by the simple linear regression model based on the percent of young drivers (ages 15-24) in state.

Prediction for a State Where 20% of the drivers are ages 15 to 24

```
#Are We Extrapolating?
quantile(x=fatal$yd)
boxplot(x=fatal$yd)

#Predict when yd=0.2
xstar=20
fit.yint =
fit.slope =

fit.yint+fit.slope*xstar

#Alternative Way to predict when yd=20 or yd=15
unknown = data.frame()
predict(mod,newdata=unknown)

#Plot of predictions
plot(adj_fatal~yd,data=fatal)
abline(mod)
points(x=c(20,15),y=predict(mod,newdata=unknown),pch=17,cex=3)
```

Interpretation of Prediction: If 20% of the drivers in a state were between the ages of 15 to 24, then we would predict or expect the number of vehicle fatalities to be AROUND 0.21 per 1,000 people.

Alternative Interpretation: If an infinite number of states each had 20% of their drivers between the ages of 15 and 24, the average number of vehicle fatalities per 1,000 people across all those states is ESTIMATED to be 0.21.

Confidence Interval

```
predict()
```

Interpretation of Confidence Interval: We are 95% confident that the average number vehicle fatalities per 1000 people in state is somewhere between 0.20 and 0.22 whenever a state's proportion of young drivers is 20%.

Prediction Interval

Interpretation of Prediction Interval: For a state with 20% of the drivers being young (ages 15-24), I am 95% confident that the actual number of vehicle fatalities per 1000 people in that specific state is somewhere between 0.10 and 0.32.

Comparison of Both Intervals

```
CIPIPlot
CIPIPlot(x=fatal$yd, y=fatal$adj_fatal)
```

The red dashed lines represents are uncertainty in the fitted line itself. We are 95% confident that the fitted line to the population lies somewhere in this interval. This is based on our assumptions for the linear

regression model to be reasonable.

The blue dashed lines represent are uncertainty in using the line to predict. We are 95% confident that a prediction of y for any x value is somewhere in this interval. If are the assumptions of the linear regression model are reasonable, you would expect approximately 95% of the data points in the population to lie within the blue interval.