Unit 6: Introduction to linear regression
2. Outliers and inference for regression

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at *https://www2.stat.duke.edu/courses/Spring19/sta101.001/*

Outline

1. Housekeeping

2. Main ideas
    1. **Assessing the Fit:** ***Simple Linear Regression Model***
        1. USING the Model:
            1. Predictions: Predicted values also have uncertainty around them
        2. UNDERSTANDING Relationships in the Model:
            1. Overall Fit of Model: R^2 assesses model fit -- higher the better
            2. Individual Coefficients: Inference for regression uses the t-distribution
            3. Conditions/Diagnostic Checking
            4. Outliers: Type of outlier determines how it should be handled

Coming up…

- Peer Evaluation 2 due Tuesday 4/9 11:55 pm
- Problem Set 6 due Wednesday 4/10
- Readiness Assessment 7 Wednesday 4/10
- Lab Assignment 10 due Friday 4/12 11:55pm (extension)
- Performance Assessment 6 due Sunday 4/14 11:55pm (opens today)
- Don't forget the Project Stage 2 due in ~1.5 weeks

Outline

1. Housekeeping

2. Main ideas
    1. **Assessing the Fit:** ***Simple Linear Regression Model***
        1. USING the Model:
            1. Predictions: Predicted values also have uncertainty around them
        2. UNDERSTANDING Relationships in the Model:
            1. Overall Fit of Model: R^2 assesses model fit -- higher the better
            2. Individual Coefficients: Inference for regression uses the t-distribution
            3. Conditions/Diagnostic Checking
            4. Outliers: Type of outlier determines how it should be handled

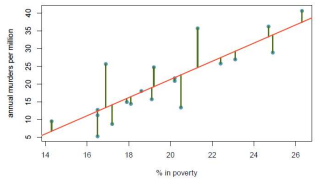
Regression Models:

Using vs. Understanding

Regression Models:

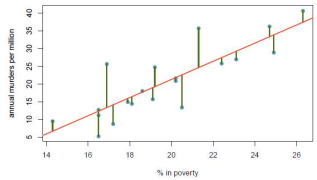
Using: Make Predictions

What's one way we can assess how well our regression line predicted y with a given x^*, *when we know the observed y?*



What's one way we can assess how well our regression line predicted y with a given x^*, *when we know the observed y?*

Residual of x^*: $e = y - \hat{y}$



Uncertainty of predictions

- Regression models are useful for making predictions for **new observations not include in the original dataset**.

Uncertainty of predictions

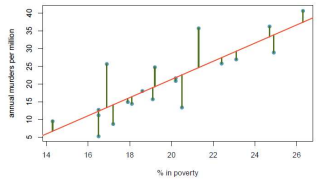
- Regression models are useful for making predictions for new observations not include in the original dataset.
- If the model is good, the predictions should be close to the true value of the response variable for this observation, however it may not be exact, i.e. $\hat{y}$ might be different than y.

Uncertainty of predictions

- Regression models are useful for making predictions for new observations not include in the original dataset.
- If the model is good, the predictions should be close to the true value of the response variable for this observation, however it may not be exact, i.e. $\hat{y}$ might be different than y.
- With any prediction we can (and should) also report a measure of uncertainty of the prediction.

Outline

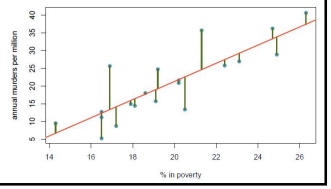
What's one way we can quantify the uncertainty around our predicted y with a given x^*, *when we DON'T KNOW the observed y?*



Outline

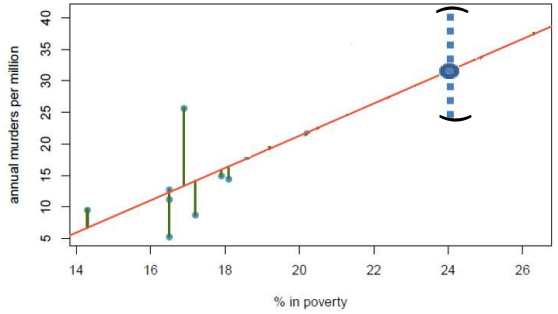
What's one way we can quantify the uncertainty around our predicted y with a given x^*, *when we DON'T KNOW the observed y?*

Make a prediction interval for y for the given x^*



Outline

How do we calculate a prediction interval for y for the given x^*?

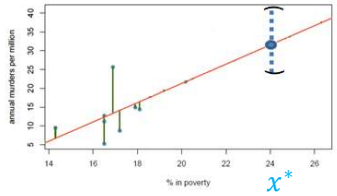


Prediction intervals for specific predicted values

A *prediction interval* for y **for a given x^*** is

$$\hat{y} \pm t^{\star}_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^{\star} - \bar{x})^2}{(n-1)s_x^2}}$$

- x^* is a new observation that you plug into the regression equation *(don't usually know the corresponding observed y)*

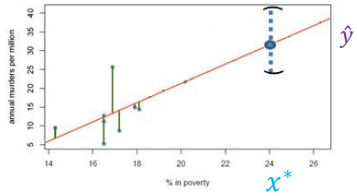


Prediction intervals for specific predicted values

A *prediction interval* for y **for a given x^*** is

$$\hat{y} \pm t^{\star}_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^{\star} - \bar{x})^2}{(n-1)s_x^2}}$$

- x^* is a new observation that you plug into the regression equation *(don't usually know the corresponding observed y)*
- $\hat{y}$ is the predicted response you get by plugging in x^* into the regression equation

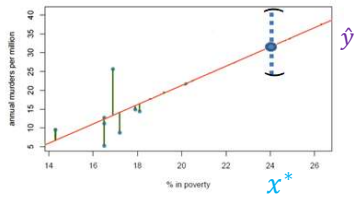


Prediction intervals for specific predicted values

A *prediction interval* for y **for a given $x^\star$** is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

- x^* is a new observation that you plug into the regression equation *(don't usually know the corresponding observed y)*
- $\hat{y}$ is the predicted response you get by plugging in x^* into the regression equation
- s is the standard deviation of the residuals
- n is the number of observations that were used to calculate the regression coefficients (ie: trained the model)



Outline

How does the prediction interval for y for the given x^* change when:

- **$x^\star$ moves farther away from the center (ie.($x^\star$-$\bar{x}$) increases)?**
- **s (variability of residuals) increases?**

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

Prediction intervals for specific predicted values

A *prediction interval* for y **for a given $x^\star$** is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

s = the variability of residuals

▶ Relationship: The width of the prediction interval for $\hat{y}$ increases as

- $x^\star$ moves away from the center
- s (the variability of residuals), i.e. the scatter, increases

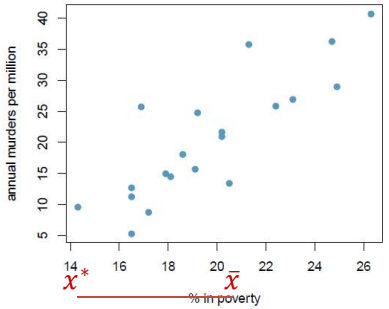
Remember from Deck 6.1….

Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. *For which of the following districts would you be most comfortable with your prediction?*

A district where % in poverty =

(c) 20%

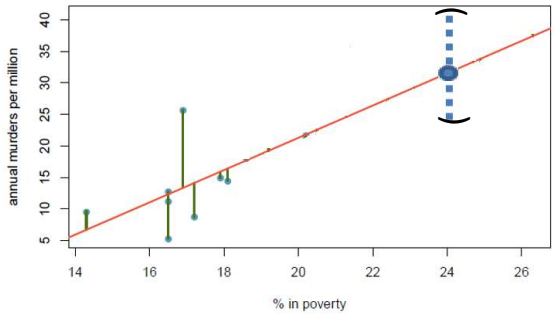
$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$



Larger distance x is from $\bar{x}$
… prediction interval larger… more uncertainty.

Outline

How do we interpret a prediction interval for y for the given x^*?



Prediction intervals for specific predicted values

A *prediction interval* for y **for a given x^*** is

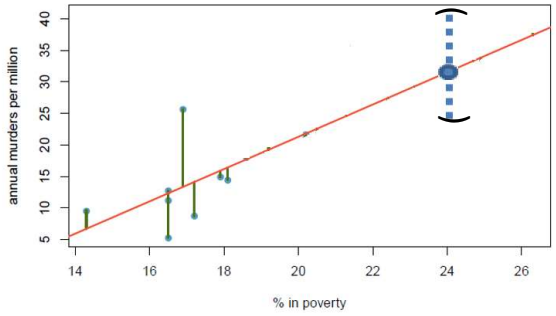
$$\hat{y} \pm t^*_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

- **Interpretation**: "We are XX% confident that y for given x^* is within this interval."

Outline

What does "XX% confident" mean when it is used in a prediction interval?

(Ie: what does the "prediction level" mean?)



Prediction intervals for specific predicted values

- Prediction level meaning:

  If we repeat the process of:
  - obtaining a regression data set $(x_1, y_1), \ldots (x_n, y_n)$ *(random sampling)*
  - calculating a regression line for this data set,
  - calculating a prediction for $\hat{y}$ given x* and
  - forming a XX% prediction interval at x^* with using $\hat{y}$ and this regression line

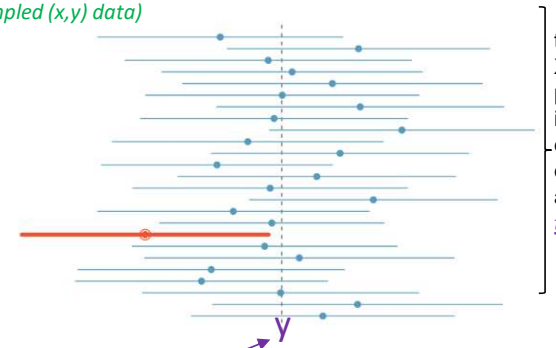
  many times and wait to see what the future value of y is at x^* …

  …then roughly XX% of the prediction intervals will contain the corresponding actual value of y.

Prediction intervals for specific predicted values

- Prediction level meaning:

First, make many prediction XX% intervals for y given x**(each prediction interval uses a new regression line… which was calculated from n new random sampled (x,y) data)*



then roughly XX% of the prediction intervals will contain the corresponding actual value of *y*.

Then, randomly select a value with x=x* from the population, note the corresponding y.

Calculating the prediction interval

By hand:

Don't worry about it...

Calculating the prediction interval

By hand:

Don't worry about it...

In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata, interval = "prediction", level = 0.95)
```

Calculating the prediction interval

By hand:

Don't worry about it...

In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata, interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 21.28663 9.418327 33.15493
```

$\hat{y}$ prediction(s) for the x* value(s) in newdata

Prediction interval(s) for y given the x* value(s) in newdata

Calculating the prediction interval

By hand:
Don't worry about it...

In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata, interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 21.28663 9.418327 33.15493
```

"We are 95% confident that the annual murders per million for a county with 20% poverty rate is between 9.52 and 33.15."

Outline

1. Housekeeping

2. Main ideas
   1. **Assessing the Fit:** ***Simple Linear Regression Model***
      1. USING the Model:
         1. Predictions: Predicted values also have uncertainty around them
      2. UNDERSTANDING Relationships in the Model:
         1. Overall Fit of Model: R^2 assesses model fit -- higher the better
         2. Individual Coefficients: Inference for regression uses the t-distribution
         3. Conditions/Diagnostic Checking
         4. Outliers: Type of outlier determines how it should be handled

Outline

Regression Models:

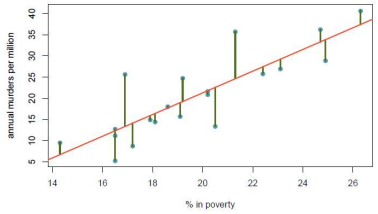
Using vs. Understanding

Outline

Regression Models:
Understanding:
Relationships between Variables

How do we assess the overall fit of the whole model?

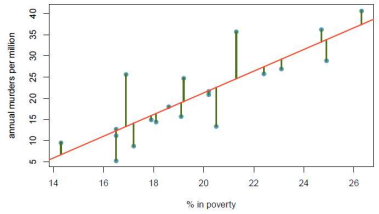
$$\hat{y} = b_0 + b_1 x$$



How do we assess the overall fit of the whole model?

R^2

$$\hat{y} = b_0 + b_1 x$$



▶ **Interpreting R^2: "percentage of variability in *y* explained by the model"**

Higher R→ Better *overall* model fit!

▶ **Interpreting R^2: "percentage of variability in *y* explained by the model"**

Higher R→ Better *overall* model fit!

▶ **Calculating R^2 for a simple linear regression model** *(ie. 1 expl. var)*

$$R^2 = (correlation\ coeff)^2 = (R)^2$$

```
murder %>%
    summarise(r_sq = cor(annual_murders_per_mil, perc_pov)^2)
```

```
      r_sq
1 0.7052275
```

(1) R^2 assesses model fit -- higher the better

▶ **Interpreting R^2: "percentage of variability in *y* explained by the model"**

Higher R→ Better *overall* model fit!

▶ **Calculating R^2 for a simple linear regression model** *(ie. 1 expl. var)*

$$R^2 = (correlation\ coeff)^2 = (R)^2$$

▶ **Calculating R^2 for any linear regression model** *(ie. could have >1 expl. var – Unit 7)*

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{SS_{tot} - SS_{residuals}}{SS_{tot}} = 1 - \frac{SS_{residuals}}{SS_{tot}}$$

(1) R^2 assesses model fit -- higher the better

ANOVA Output for Simple Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable	1	$SS_{Reg}=SS_{Tot}-SS_{Res}$	$MS_{Reg}=SS_{Reg}/Df_{Reg}$	MS_{Reg}/MS_{Res}	$P(F>MS_{Reg}/MS_{Res})$
Residuals	n-2	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{SS_{tot} - SS_{residuals}}{SS_{tot}} = 1 - \frac{SS_{residuals}}{SS_{tot}}$$

(1) R^2 assesses model fit -- higher the better

ANOVA Output for Simple Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable	1	$SS_{Reg}=SS_{Tot}-SS_{Res}$	$MS_{Reg}=SS_{Reg}/Df_{Reg}$	MS_{Reg}/MS_{Res}	$P(F>MS_{Reg}/MS_{Res})$
Residuals	n-2	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

(1) R^2 assesses model fit -- higher the better

ANOVA Output for Simple Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable	1	$SS_{Reg}=SS_{Tot}-SS_{Res}$	$MS_{Reg}=SS_{Reg}/Df_{Reg}$	MS_{Reg}/MS_{Res}	$P(F>MS_{Reg}/MS_{Res})$
Residuals	n-2	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

```
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
anova(m_mur_pov)
```

```
Analysis of Variance Table

Response: annual_murders_per_mil
           Df  Sum Sq Mean Sq F value    Pr(>F)
perc_pov    1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

ANOVA Table for a Regression
- Use to find R^2

(1) R^2 assesses model fit -- higher the better

ANOVA Output for Simple Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable	1	$SS_{Reg}=SS_{Tot}-SS_{Res}$	$MS_{Reg}=SS_{Reg}/Df_{Reg}$	MS_{Reg}/MS_{Res}	$P(F>MS_{Reg}/MS_{Res})$
Residuals	n-2	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

```
anova(m_mur_pov)
Analysis of Variance Table

Response: annual_murders_per_mil
           Df  Sum Sq Mean Sq F value    Pr(>F)
perc_pov    1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

ANOVA Table for a Regression
- Use to find R^2

$$R^2 = \frac{explained\ variabilty}{total\ variability} = \frac{SS_{reg}}{SS_{tot}} = \frac{1308.34}{1308.34 + 546.86} = \frac{1308.34}{1855.2} \approx 0.71$$

(1) R^2 assesses model fit -- higher the better

ANOVA Output for Simple Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable	1	$SS_{Reg}=SS_{Tot}-SS_{Res}$	$MS_{Reg}=SS_{Reg}/Df_{Reg}$	MS_{Reg}/MS_{Res}	$P(F>MS_{Reg}/MS_{Res})$
Residuals	n-2	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

```
anova(m_mur_pov)
Analysis of Variance Table

Response: annual_murders_per_mil
           Df  Sum Sq Mean Sq F value    Pr(>F)
perc_pov    1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

ANOVA Table for a Regression
- Use to find R^2

Total 1855.2 $SS_{tot} = SS_{reg} + SS_{residuals}$

$$R^2 = \frac{explained\ variabilty}{total\ variability} = \frac{SS_{reg}}{SS_{tot}} = \frac{1308.34}{1308.34 + 546.86} = \frac{1308.34}{1855.2} \approx 0.71$$

(1) R^2 assesses model fit -- higher the better

ANOVA Output for Simple Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable	1	$SS_{Reg}=SS_{Tot}-SS_{Res}$	$MS_{Reg}=SS_{Reg}/Df_{Reg}$	MS_{Reg}/MS_{Res}	$P(F>MS_{Reg}/MS_{Res})$
Residuals	n-2	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

```
anova(m_mur_pov)
Analysis of Variance Table

Response: annual_murders_per_mil
           Df  Sum Sq Mean Sq F value    Pr(>F)
perc_pov    1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

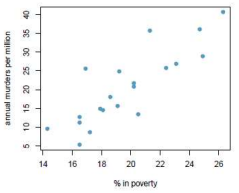
ANOVA Table for a Regression
- Use to find R^2

SS_{reg} = Sum of the Sum Squared values for all the predictors in the ANOVA table.

$$R^2 = \frac{explained\ variabilty}{total\ variability} = \frac{SS_{reg}}{SS_{tot}} = \frac{1308.34}{1308.34 + 546.86} = \frac{1308.34}{1855.2} \approx 0.71$$

Clicker question

R^2 for the regression model for predicting annual murders per million based on percentage living in poverty is roughly 71%. Which of the following is the correct interpretation of this value?



(a) 71% of the variability in percentage living in poverty is explained by the model.

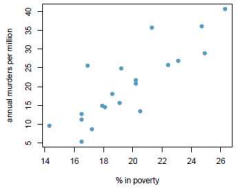
(b) 84% of the variability in the murder rates is explained by the model, i.e. percentage living in poverty.

(c) 71% of the variability in the murder rates is explained by the model, i.e. percentage living in poverty.

(d) 71% of the time percentage living in poverty predicts murder rates accurately.

Clicker question

R^2 for the regression model for predicting annual murders per million based on percentage living in poverty is roughly 71%. Which of the following is the correct interpretation of this value?



(a) 71% of the variability in percentage living in poverty is explained by the model.

(b) 84% of the variability in the murder rates is explained by the model, i.e. percentage living in poverty.

(c) 71% of the variability in the murder rates is explained by the model, i.e. percentage living in poverty.

(d) 71% of the time percentage living in poverty predicts murder rates accurately.

Outline

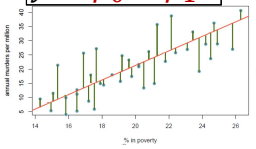
1. Housekeeping

2. Main ideas
   1. **Assessing the Fit:** ***Simple Linear Regression Model***
      1. USING the Model:
         1. Predictions: Predicted values also have uncertainty around them
      2. UNDERSTANDING Relationships in the Model:
         1. Overall Fit of Model: R^2 assesses model fit -- higher the better
         2. Individual Coefficients: Inference for regression uses the t-distribution
         3. Conditions/Diagnostic Checking
         4. Outliers: Type of outlier determines how it should be handled

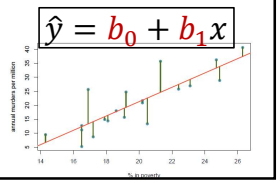
Outline

How do we test the significance of a intercept or coefficient in the population model?

$\hat{y} = \beta_0 + \beta_1 x$



$\hat{y} = b_0 + b_1 x$



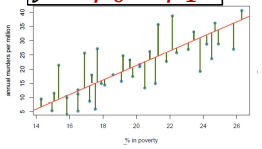
Outline

How do we test the significance of a intercept or coefficient in the population model?

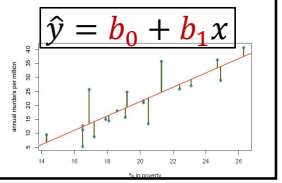
Conduct Hypothesis Testing on β_0 and β_1:

- using b_0 and b_1 as point estimates (resp).

$\hat{y} = \beta_0 + \beta_1 x$



$\hat{y} = b_0 + b_1 x$



Coefficient Hypothesis Testing for Simple Linear Regression

- **Hypothesis testing for a slope**:

$H_0: \beta_1 = 0;$

$H_A: \beta_1 \neq 0$

Inference for regression uses the *t*-distribution

Coefficient Hypothesis Testing for Simple Linear Regression

- **Hypothesis testing for a slope**:

$H_0: \beta_1 = 0;$

$H_A: \beta_1 \neq 0$

- **Test Statistic** $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$

Inference for regression uses the *t*-distribution

Coefficient Hypothesis Testing for Simple Linear Regression

- **Hypothesis testing for a slope**:

$H_0: \beta_1 = 0;$

$H_A: \beta_1 \neq 0$

- **Test Statistic** $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

Inference for regression uses the *t*-distribution

Coefficient Hypothesis Testing for Simple Linear Regression

- **Hypothesis testing for a slope**:

$H_0: \beta_1 = 0;$

$H_A: \beta_1 \neq 0$

- **Test Statistic** $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

- **p-value** = P(observing a slope at least as different from 0 as the one observed given there is no linear relationship between *x* and *y)*

Coefficient Hypothesis Testing for Simple Linear Regression

- **Hypothesis testing for a slope**:

$H_0: \beta_1 = 0;$

$H_A: \beta_1 \neq 0$

- **Test Statistic** $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

- **p-value** = P(observing a slope at least as different from 0 as the one observed given there is no linear relationship between *x* and *y)*

```
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
summary(m_mur_pov, newdata)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.901      7.789  -3.839   0.0012 **
perc_pov       2.559      0.390   6.562 3.64e-06 ***
```


Coefficient Hypothesis Testing for Simple Linear Regression

- **Hypothesis testing for a slope**:

$H_0: \beta_1 = 0;$

$H_A: \beta_1 \neq 0$

How to interpret:

- **Ho**: Explanatory variable is not a significant predictor of the response variable.
- **Ha**: Explanatory variable is a significant predictor of the response variable.

OR

- **Ho**: There is no linear relationship between the explanatory variable and response variable.
- **Ha**: There is a linear relationship between the explanatory variable and response variable.

OR

- **Ho**: The slope of the relationship between the explanatory and response variable is 0.
- **Ha**: The slope of the relationship between the explanatory and response variable is not 0.

Coefficient Confidence Intervals for Simple Linear Regression

- **Confidence Interval for a slope**:

$b_1 \pm t^*_{n-2} SE_{b_1}$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

Coefficient Confidence Intervals for Simple Linear Regression

- **Confidence Interval for a Slope**:

$b_1 \pm t^*_{n-2} SE_{b_1}$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

```
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
confint(m_mur_pov, level = 0.95)
```

```
                2.5 %     97.5 %
(Intercept) -46.265631 -13.536694
perc_pov      1.740003   3.378776
```

Inference for regression uses the t-distribution

Coefficient Confidence Intervals for Simple Linear Regression

▶ **Confidence Interval for a Slope:**

$$b_1 \pm t^*_{n-2} SE_{b_1}$$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

```
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
confint(m_mur_pov, level = 0.95)
```

```
                2.5 %     97.5 %
(Intercept) -46.265631 -13.536694
perc_pov      1.740003   3.378776
```

▶ **Interpreting Confidence Interval for this Slope:**

"We are 95% confident that for each additional % increase in poverty rate, we would expect the annual murder per million to increase, on average, by 1.74 to 3.38."

Inference for regression uses the t-distribution

Coefficient Confidence Intervals for Simple Linear Regression

▶ **Confidence Interval for a Slope:**

$$b_1 \pm t^*_{n-2} SE_{b_1}$$

*In general, use **df= n – k – 1**, where k = # of slopes/predictors you're estimating in the regression equation.*

```
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
confint(m_mur_pov, level = 0.95)
```

```
                2.5 %     97.5 %
(Intercept) -46.265631 -13.536694
perc_pov      1.740003   3.378776
```

Use non-causal language!

▶ **Interpreting Confidence Interval for this Slope:**

"We are 95% confident that for each additional % increase in poverty rate, we would expect the annual murder per million to increase, on average, by 1.74 to 3.38."

Important Tables Recap for Linear Regression (with one predictor)

```
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
summary(m_mur_pov, newdata)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.901      7.789  -3.839   0.0012 **
perc_pov       2.559      0.390   6.562 3.64e-06 ***
```

Table with Coefficient Test Statistics and p-values

- Use to test:
  - Ho: β_0=0; Ho: β_0≠0
  - Ho: β_1=0; Ho: β_1≠0

```
anova(m_mur_pov)
```

```
Analysis of Variance Table

Response: annual_murders_per_mil
          Df  Sum Sq Mean Sq F value    Pr(>F)
perc_pov   1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

ANOVA Table for a Regression

- Use to find R^2

Important Tables Recap for Linear Regression (with one predictor)

```
confint(m_mur_pov, level = 0.95)
```

```
                2.5 %     97.5 %
(Intercept) -46.265631 -13.536694
perc_pov      1.740003   3.378776
```

Table with Confidence Intervals for β_0 and β_1

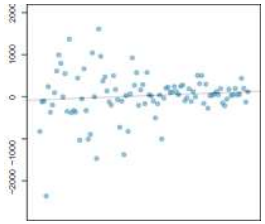
Outline

1. Housekeeping

2. Main ideas
   1. **Assessing the Fit:** ***Simple Linear Regression Model***
      1. USING the Model:
         1. Predictions: Predicted values also have uncertainty around them
      2. UNDERSTANDING Relationships in the Model:
         1. Overall Fit of Model: R^2 assesses model fit -- higher the better
         2. Individual Coefficients: Inference for regression uses the t-distribution
         3. Conditions/Diagnostic Checking
         4. Outliers: Type of outlier determines how it should be handled

Outline

What conditions/diagnostics should we also check to see if our linear regression model is appropriate for the data?



Conditions for regression

Important regardless of doing inference

- **Linearity** → randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference (or check the scatter plot for linear relationship)

Conditions for regression

Important regardless of doing inference

- **Linearity** → randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference (or check the scatter plot for linear relationship)

Important for inference

- **Nearly normally distributed residuals** → histogram or normal probability plot of residuals

Conditions for regression

Important regardless of doing inference

- **Linearity** → randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference (or check the scatter plot for linear relationship)

Important for inference

- **Nearly normally distributed residuals** → histogram or normal probability plot of residuals
- **Constant variability of residuals** (*homoscedasticity*) → no fan shape in the residuals plot (can also check in a scatter plot)

Conditions for regression

Important regardless of doing inference

- **Linearity** → randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference (or check the scatter plot for linear relationship)

Important for inference

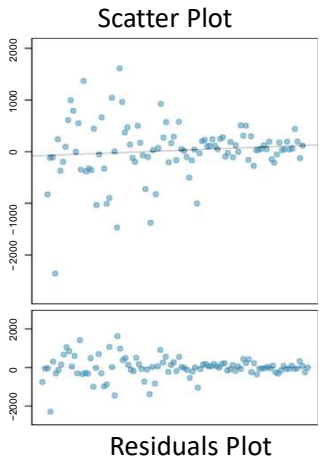
- **Nearly normally distributed residuals** → histogram or normal probability plot of residuals
- **Constant variability of residuals** (*homoscedasticity*) → no fan shape in the residuals plot
- **Independence of residuals** (and hence observations) → depends on data collection method, often violated for time-series data

Checking conditions

Clicker question

What condition is this linear model obviously and definitely violating?

(a) Linear relationship
(b) Non-normal residuals
(c) Constant variability
(d) Independence of observations

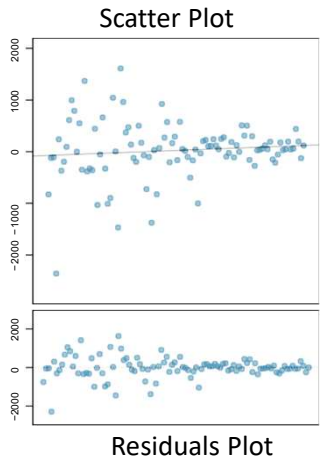


Checking conditions

Clicker question

What condition is this linear model obviously and definitely violating?

(a) Linear relationship
(b) Non-normal residuals
(c) Constant variability
(d) Independence of observations

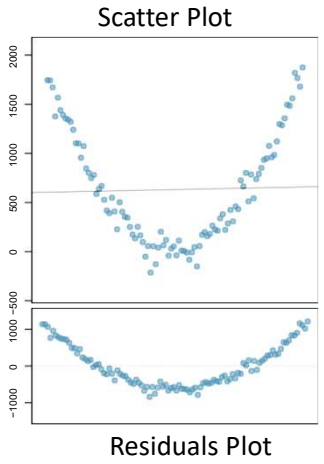


Checking conditions

Clicker question

What condition is this linear model obviously and definitely violating?

(a) Linear relationship
(b) Non-normal residuals
(c) Constant variability
(d) Independence of observations

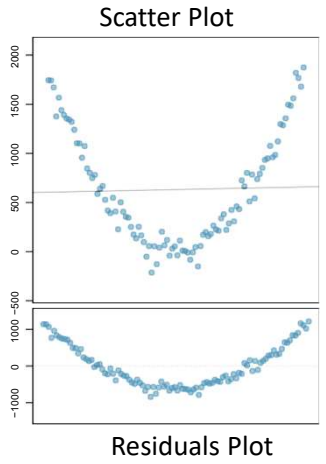


Checking conditions

Clicker question

What condition is this linear model obviously and definitely violating?

(a) Linear relationship
(b) Non-normal residuals
(c) Constant variability
(d) Independence of observations



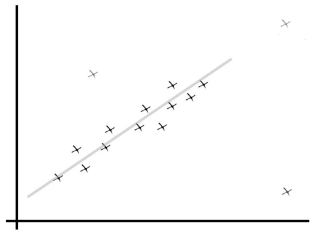
Outline

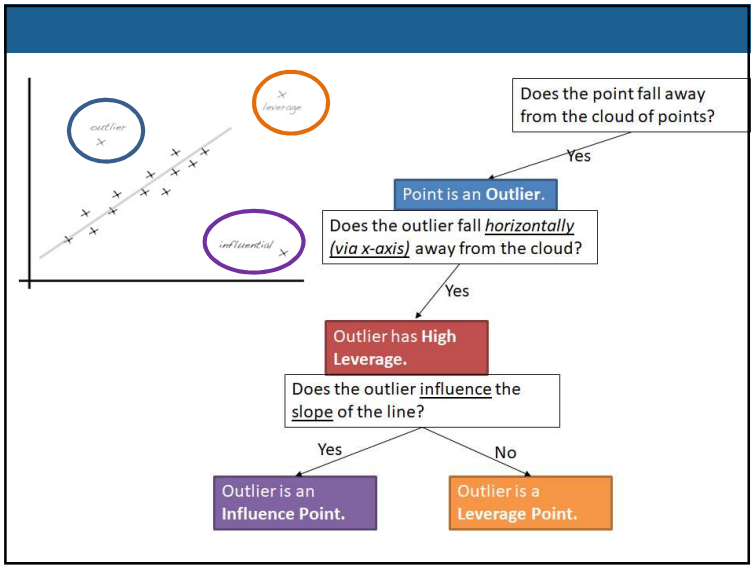
1. Housekeeping

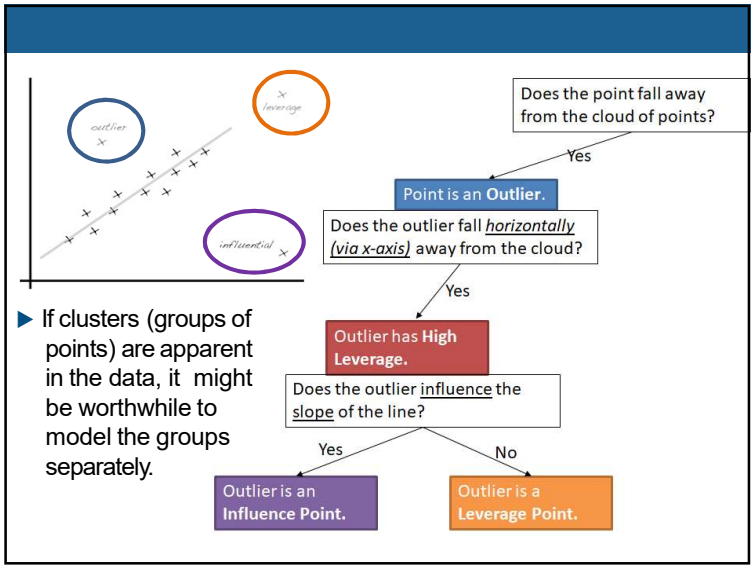
2. Main ideas
   1. **Assessing the Fit:** ***Simple Linear Regression Model***
      1. USING the Model:
         1. Predictions: Predicted values also have uncertainty around them
      2. UNDERSTANDING Relationships in the Model:
         1. Overall Fit of Model: R^2 assesses model fit -- higher the better
         2. Individual Coefficients: Inference for regression uses the t-distribution
         3. Conditions/Diagnostic Checking
         4. Outliers: Type of outlier determines how it should be handled

Outline

How can we classify outliers by affect they *might* have on the regression model?







- If clusters (groups of points) are apparent in the data, it might be worthwhile to model the groups separately.

Application exercise: 6.2 Linear regression

See course website for details

Summary of main ideas

1. Predicted values also have uncertainty around them
2. R^2 assesses model fit – higher the better
3. Inference for regression uses the t-distribution
4. Conditions for regression
5. Type of outlier determines how it should be handled