Regression Models

Ryan Miller

Introduction

- Recently, we introduced ANOVA as a statistical method for simultaneously comparing the means of many groups
  - More specifically, the method we discussed is known as **one-way ANOVA**, it uses *one categorical variable* to model a *numerical outcome*

Introduction

- Recently, we introduced ANOVA as a statistical method for simultaneously comparing the means of many groups
  - More specifically, the method we discussed is known as **one-way ANOVA**, it uses *one categorical variable* to model a *numerical outcome*
- One-way ANOVA is actually a special type of **regression modeling**, a general approach where a numerical outcome is modeled by a linear combination of explanatory variables
  - This presentation will focus on regression modeling, focusing primarily on **simple linear regression**, or models with a single numeric explanatory variable

Simple Linear Regression

- As mentioned previously, statistical models are often expressed in the form:

$$Y_i = f(X_i) + \epsilon_i$$

  - In words, this model states that the observed outcome for the i^{th} case equals some function of the explanatory variables for that case, plus random error (ϵ_i)

Simple Linear Regression

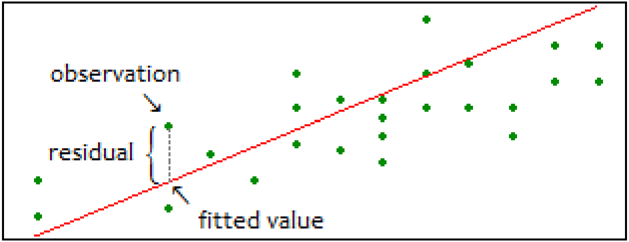
- As mentioned previously, statistical models are often expressed in the form:

$$Y_i = f(X_i) + \epsilon_i$$

  - In words, this model states that the observed outcome for the i^{th} case equals some function of the explanatory variables for that case, plus random error (ϵ_i)
- In a *linear regression model*, $f(X_i)$ is a linear combination of explanatory variables (belonging to the i^{th} subject)
  - In simple linear regression, only a single numeric explanatory variable is used
  - In this case, $f(X_i) = \beta_0 + \beta_1 X_{1i}$, notice this model is akin to a straight line with error (ie: $Y = mX + b + \epsilon$)

Simple Linear Regression

- To utilize a regression model, we must estimate the **coefficients** (β_0 and β_1) involved in the linear combination
- Without getting into the mathematical details, this is done using *least squares estimation*, a method which *minimizes* the squared residuals:

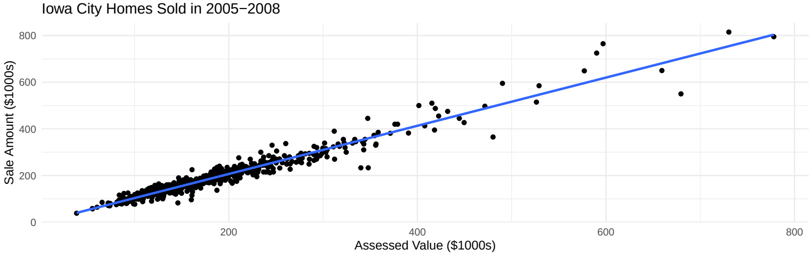


Simple Linear Regression (example)

Below is an estimated regression model that uses a home's assessed value to predict its sale price (Iowa City home sales in 2005-2008)

```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
lm(sale.amount ~ assessed, data = IC)
```

```
##
## Call:
## lm(formula = sale.amount ~ assessed, data = IC)
##
## Coefficients:
## (Intercept)     assessed
##      -1.523        1.033
```



Simple Linear Regression (Notation and Inference)

- We use the notation b_0 and b_1 to denote our *estimates* of the *model parameters* $\{\beta_0, \beta_1\}$
  - These estimates (b_0 and b_1) describe how the x and y variables are related *in our data*

Simple Linear Regression (Notation and Inference)

- We use the notation b_0 and b_1 to denote our *estimates* of the *model parameters* $\{\beta_0, \beta_1\}$
  - These estimates (b_0 and b_1) describe how the x and y variables are related *in our data*
- Like any estimate, the regression estimates, b_0, b_1, won't *exactly* match the population parameters, β_0, β_1

Simple Linear Regression (Notation and Inference)

- We use the notation b_0 and b_1 to denote our *estimates* of the *model parameters* $\{\beta_0, \beta_1\}$
  - These estimates (b_0 and b_1) describe how the x and y variables are related *in our data*
- Like any estimate, the regression estimates, b_0, b_1, won't *exactly* match the population parameters, β_0, β_1
- We won't go too far into the details, but `R` can be used to produce confidence interval estimates for the population parameters using the t-distribution (with $df = n - 2$)
  - We can also perform hypothesis testing using the t-distribution (by default, software will test $H_0 : \beta = 0$)

Simple Linear Regression - Example

1) Interpret the hypothesis test results (ie: the *p*-value) for the slope coefficient
2) Can you use this output to come up with a 95% *t*-distribution CI for the population's slope coefficient (β_1)?

```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(sale.amount ~ assessed, data = IC)
summary(model)
```

```
##
## Call:
## lm(formula = sale.amount ~ assessed, data = IC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152050   -7137    -347    7496  148286
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.523e+00  1.712e+03  -0.001    0.999
## assessed     1.033e+00  8.819e-03 117.142   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20970 on 775 degrees of freedom
## Multiple R-squared:  0.9465, Adjusted R-squared:  0.9465
## F-statistic: 1.372e+04 on 1 and 775 DF,  p-value: < 2.2e-16
```

Simple Linear Regression - Solution

1) There is overwhelming evidence ($p < 0.001$) of an association between assessed value and sale price in Iowa City homes
2) Shown below:

```
## Notice df = 775
t_star <- qt(.975, df = 775)

## The point estimate of the slope
point_est <- model$coefficients[2]

## Standard error
se <- 8.819e-03

## 95% CI
c(point_est - t_star*se, point_est + t_star*se)
```

```
## assessed assessed
## 1.015815 1.050439
```

```
## Using confint
confint(model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -3361.652060 3358.60640
## assessed        1.015815    1.05044
```

Testing Hypotheses Other Than $\beta = 0$

- In the Iowa City home sales example, we might want to test $H_0 : \beta_1 = 1$, which would imply that differences between assessed and sale prices remain consistent across homes with different values
  - How might you test this hypothesis using the output shown below?

```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(sale.amount ~ assessed, data = IC)
summary(model)
```

```
##
## Call:
## lm(formula = sale.amount ~ assessed, data = IC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152050   -7137    -347    7496  148286
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.523e+00  1.712e+03  -0.001    0.999
## assessed     1.033e+00  8.819e-03 117.142   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20970 on 775 degrees of freedom
## Multiple R-squared:  0.9465, Adjusted R-squared:  0.9465
## F-statistic: 1.372e+04 on 1 and 775 DF,  p-value: < 2.2e-16
```

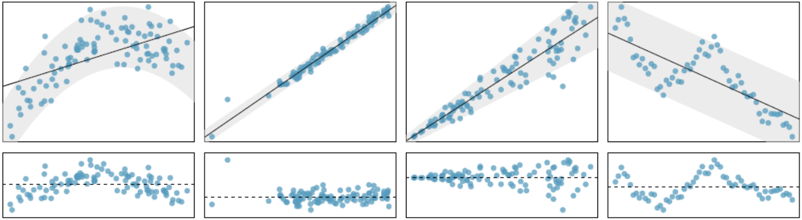
Testing Other Hypotheses

1) For testing $H_0 : \beta_1 = 1$, we can use:
   $T = \frac{b_1 - 1}{SE(b_1)} = \frac{1.033 - 1}{8.819e-03} = 3.74$
2) Then, using a t-distribution with $df = n - 2 = 755$, the two-sided p-value is 9.88e-05 (nearly zero)
3) Thus, we conclude that the deviation between assessed and sale amount is not constant across differently priced homes (ie: $\beta_1 \neq 1$)

Simple Linear Regression - Assumptions

A simple linear regression model can be estimated using any data, but *statistical inference* involving that model is only valid when four conditions are met:

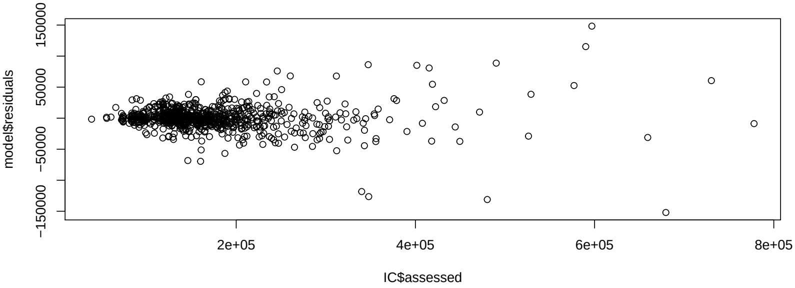
1) Linearity
2) Normally distributed residuals
3) Constant variance
4) Independent observations



Example - Residuals in R

Do these assumptions appear to be met for our Iowa City homes model?

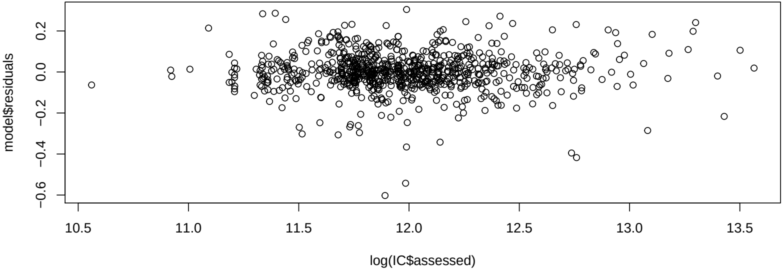
```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(sale.amount ~ assessed, data = IC)
plot(IC$assessed, model$residuals)
```



Transformations and Model Assumptions

If we apply a log-transformation to both the explanatory and response variables, these assumptions seem much more reasonable:

```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(log(sale.amount) ~ log(assessed), data = IC)
plot(log(IC$assessed), model$residuals)
```



Inference after Transformations

- After a log-transformation, interpreting the model coefficients (slope and intercept) is much trickier
- You can find a good guide to proper interpretations at this link
  - I don't plan to ask you any direct questions pertaining to log-transformed variables in the context of regression, but you might consider using this approach on your project (or some future analysis)

One-way ANOVA as Regression via Dummy Variables

- To connect regression and one-way ANOVA, we need to introduce **dummy variables**
- To create a dummy variable, we assign one category to be the **reference category**
  - The category represented by the non-reference category receives a numeric value of 1 in the dummy variable

One-way ANOVA as Regression via Dummy Variables

- To connect regression and one-way ANOVA, we need to introduce **dummy variables**
- To create a dummy variable, we assign one category to be the **reference category**
  - The category represented by the non-reference category receives a numeric value of 1 in the dummy variable
  - Below is an example of a dummy variable for a categorical variable with 2 categories

Y	group
8.5	B
11.6	A
9.0	A
9.1	A
8.0	B
9.7	A

Y	dummyB
8.5	1
11.6	0
9.0	0
9.1	0
8.0	1
9.7	0

Dummy Variables

- For a categorical predictor with k categories, $k - 1$ different dummy variables are necessary

Y	group
8.5	B
11.6	C
9.0	C
9.1	A
8.0	C
9.7	A

Y	dummyB	dummyC
8.5	1	0
11.6	0	1
9.0	0	1
9.1	0	0
8.0	0	1
9.7	0	0

Dummy Variables - Example

- First, find the mean following distance of each drug group in the Tailgating dataset
- Then, use the `lm()` function to fit a linear regression model that uses drug to predict distance
  - Which group did `R` use as the reference category? How do you interpret this model?

```
tail <- read.csv("https://remiller1450.github.io/data/Tailgating.csv")
```

Dummy Variables - Solution (some R code)

```
## Group means
mean(tail$D[tail$Drug == "ALC"])
```

```
## [1] 36.82831
```

```
mean(tail$D[tail$Drug == "THC"])
```

```
## [1] 42.60538
```

```
## Regression model
lm(D ~ Drug , data = tail)
```

```
##
## Call:
## lm(formula = D ~ Drug, data = tail)
##
## Coefficients:
## (Intercept)     DrugMDMA   DrugNODRUG      DrugTHC
##      36.828       -9.221       10.499        5.777
```

Dummy Variables - Solution

The *estimated* model is expressed by:

$$\hat{Y} = b_0 + b_1 X_{MDMA} + b_2 X_{NODRUG} + b_3 X_{THC}$$

- "Alcohol" was used as the reference category
  - $b_0 = 36.83$ is the sample mean of the alcohol group, this isn't a coincidence
- $b_1 = -9.2$ is the difference between the alcohol and MDMA group means
- $b_2 = 10.5$ is the difference between the alcohol and no drug group means
- $b_3 = 5.8$ is the difference between the alcohol and the THC group means

Two Approaches to One-way ANOVA

```r
## Using lm (Regression)
reg <- lm(D ~ Drug , data = tail)
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: D
##            Df Sum Sq Mean Sq F value Pr(>F)
## Drug        3   4989  1663.1  0.8496 0.4696
## Residuals 115 225127  1957.6
```

```r
## Using aov (ANOVA)
anov <- aov(D ~ Drug, data = tail)
summary(anov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Drug          3   4989    1663    0.85   0.47
## Residuals   115 225127    1958
```

Regression with Multiple Variables

- Dummy variables express a single categorical predictor using a set of binary variables
  - This illustrates how a regression model can involve more than one explanatory variable

Regression with Multiple Variables

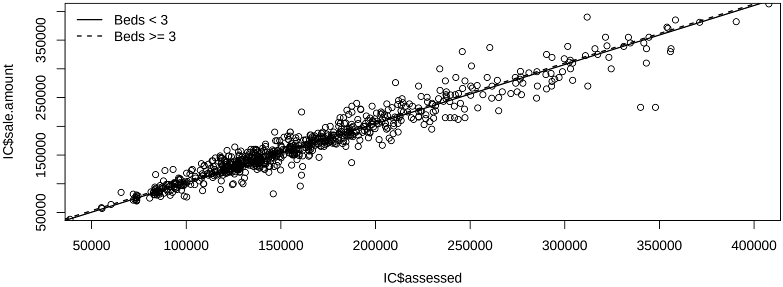
- Dummy variables express a single categorical predictor using a set of binary variables
  - This illustrates how a regression model can involve more than one explanatory variable
- **Multiple regression** models quantitative outcome using a *linear combination* of many variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

Regression with Multiple Variables

A relatively simple illustration of this framework is a model that includes a single categorical and a single numeric explanatory variable

```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(sale.amount ~ assessed + (bedrooms > 2), data = IC)
```



Regression with Multiple Variables

- Statistical inference now comes with the caveat that we've *adjusted for the other variables* in the model
  - For homes with the *same assessed value*, those with 3+ bedrooms are expected to sell for $2,626 more than homes with 1 or 2 bedrooms

```
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(sale.amount ~ assessed + (bedrooms > 2), data = IC)
summary(model)
```

```
##
## Call:
## lm(formula = sale.amount ~ assessed + (bedrooms > 2), data = IC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -150101   -7440    -211    7049  149776
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.914e+02  1.788e+03  -0.443    0.658
## assessed           1.028e+00  9.548e-03 107.617   <2e-16 ***
## bedrooms > 2TRUE   2.626e+03  1.733e+03   1.515    0.130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20960 on 774 degrees of freedom
## Multiple R-squared:  0.9467, Adjusted R-squared:  0.9466
## F-statistic:  6874 on 2 and 774 DF,  p-value: < 2.2e-16
```

Regression with Multiple Variables

- Without adjusting for assessed value, homes with 3+ bedrooms are expected to sell for $74,440 more than homes with 1 or 2 bedrooms

```r
IC <- read.csv('https://remiller1450.github.io/data/IowaCityHomeSales.csv')
model <- lm(sale.amount ~ (bedrooms > 2), data = IC)
summary(model)
```

```
##
## Call:
## lm(formula = sale.amount ~ (bedrooms > 2), data = IC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122124  -47624  -14724   19876  610376
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        130184       5230   24.89   <2e-16 ***
## bedrooms > 2TRUE    74440       6387   11.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83680 on 775 degrees of freedom
## Multiple R-squared:  0.1492, Adjusted R-squared:  0.1481
## F-statistic: 135.9 on 1 and 775 DF,  p-value: < 2.2e-16
```

Comments on Multiple Regression

- Multiple regression is a very powerful modeling framework as it allows us adjust for correlations between variables
  - In this class, I simply would like you to be aware that multiple regression exists and have some basic knowledge of when it might be used
  - I encourage you to take MATH-257 Data Modeling if this is a topic that interests you

Conclusion

- Regression is a flexible modeling approach that can be used in a variety of situations
  - *Simple linear regression* uses a single *numeric explanatory variable* to predict a *numeric response*
  - *One-way ANOVA* is a regression model that uses a single *categorical explanatory variable* to predict a *numeric response*

Conclusion

- Regression is a flexible modeling approach that can be used in a variety of situations
  - *Simple linear regression* uses a single *numeric explanatory variable* to predict a *numeric response*
  - *One-way ANOVA* is a regression model that uses a single *categorical explanatory variable* to predict a *numeric response*
- Regression is a *statistical model* because it is built upon an assumption of Normally distributed errors (and a few other assumptions)
  - Whenever using regression, you should check your model's residuals to ensure the assumptions for valid statistical inference are met