Week 6: Multivariate Regression

Marc Meredith*

*Statistical Methods for Data Science

June 20, 2019

```
name =   q24
label =  Preferrence dog/cat
record = 1
column = 54
width =  1
md1 =    0
md2 =    0
labels =
            1 Dogs
            2 Cats
            9 DK/NA
text =
      Which do you prefer -- dogs or cats?
```

```
> library(Hmisc)
> setwd("~/Box Sync/Teaching/Data201/Pet/")
> mydata <- spss.get("cbs201103c.por", use.value.labels=TRUE)
There were 12 warnings (use warnings() to see them)
> table(mydata$Q24)

 Dogs  Cats DK/NA
  683   201   137
```

```
name =   q1
label =  Obama Job Approval
record = 1
column = 31
width =  1
md1 =    0
md2 =    0
labels =
            1 Approve
            2 Disapprove
            9 DK/NA
text =
     Do you approve or disapprove of the way Barack Obama is handling his
     job as President?
```

How do we interpret these results?

```
> temp_table <- table(mydata$Q1, mydata$Q24)
> prop.table(temp_table, 2)

                 Dogs      Cats     DK/NA
  Approve    0.4128843 0.5273632 0.4014599
  Disapprove 0.4802343 0.3482587 0.3941606
  DK/NA      0.1068814 0.1243781 0.2043796
```

First potential pathway consistent with these data:

First potential pathway consistent with these data (continued):

```
> mydata$obamaapp <- NA
> mydata$obamaapp[mydata$Q1 == "Approve"] <- 1
> mydata$obamaapp[mydata$Q1 == "Disapprove"] <- 0
> table(mydata$obamaapp)

  0   1
452 443
>
> mydata$cats <- NA
> mydata$cats[mydata$Q24 == "Cats"] <- 1
> mydata$cats[mydata$Q24 == "Dogs"] <- 0
> table(mydata$cats)

  0   1
683 201
```

First potential pathway consistent with these data (continued):

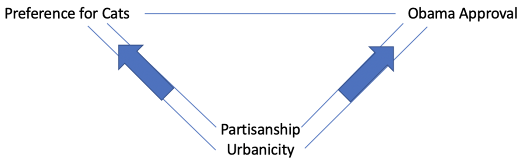
```
> reg1 <- lm(obamaapp ~ cats, data = mydata)
> summary(reg1)

Call:
lm(formula = obamaapp ~ cats, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6023 -0.4623 -0.4623  0.5377  0.5377

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46230    0.02013   22.96  < 2e-16 ***
cats         0.13998    0.04254    3.29  0.00104 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4972 on 784 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.01362,Adjusted R-squared:  0.01236
F-statistic: 10.83 on 1 and 784 DF,  p-value: 0.001045
```

Second potential pathway consistent with these data:

Second potential pathway consistent with these data (continued):

```
> reg2 <- lm(cats ~ obamaapp, data = mydata)
> summary(reg2)

Call:
lm(formula = cats ~ obamaapp, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2732 -0.2732 -0.1759 -0.1759  0.8241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17588    0.02078   8.464  < 2e-16 ***
obamaapp     0.09732    0.02958   3.290  0.00104 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4145 on 784 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.01362,Adjusted R-squared:  0.01236
F-statistic: 10.83 on 1 and 784 DF,  p-value: 0.001045
```

Third potential pathway consistent with these data:



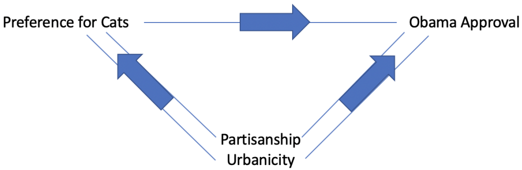
Third potential pathway consistent with these data (continued):

```
> temp_table <- table(mydata$Q24, mydata$PRTY)
> prop.table(temp_table, 2)

        Republican  Democrat Independent Don't know/No answer
  Dogs   0.7407407 0.6446281   0.6369863            0.6231884
  Cats   0.1380471 0.2369146   0.2020548            0.2173913
  DK/NA  0.1212121 0.1184573   0.1609589            0.1594203
```

Third potential pathway consistent with these data (continued):

```
> mydata$URBN[is.na(mydata$URBN)] <- 9
> mydata$urban <- factor(mydata$URBN, labels = c("Large City",
+                         "Mid City", "Suburbs", "Rural", "Unknown"))
> temp_table <- table(mydata$Q24, mydata$urban)
> prop.table(temp_table, 2)

       Large City  Mid City   Suburbs     Rural   Unknown
  Dogs  0.5818182 0.6190476 0.6508876 0.6434783 0.7649402
  Cats  0.2909091 0.2176871 0.2189349 0.2000000 0.1314741
  DK/NA 0.1272727 0.1632653 0.1301775 0.1565217 0.1035857
```

Fourth potential pathway consistent with these data:



- The topic for this week is multivariate regression
- A multivariate regression models the realization of a dependent variable as a function of two or more explanatory variables
    - Allowing us to estimate how much change we expect in the value of a dependent variable from a unit increase in an explanatory variable, while holding all other explanatory variables fixed
- Doing so can be useful for determining which of these potential pathway is most consistent with the data

Agenda for week:

- Derive the regression formula when applying the least squares criterion to a regression with two independent variables
- Explain how to interpret a regression coefficient when controlling for a variable
- Derive the generic formula for a multivariate regression with any number of independent variables
- Present the Gauss-Markov Theorem and explain layout the five conditions that are necessary for a linear regression to be the best linear unbiased estimator
- Discuss is detail the concepts of multicollinearity and functional form

Key takeaways:

- Multivariable regressions are appropriate in many, but not all circumstances, for understanding how a dependent variable varies as a function of the value of an independent variable while holding fixed the value of some other variable(s)
- Multivariate regression can estimate and test hypotheses about a variety of different empirical quantities of interest
- While it is easy to run a multivariate regression in R, structuring and interpreting the output properly is hard
- It is important to interpret regression coefficients in terms of their implications for your quantity of interest

- The real advantage of a regression is that we can look at the relationship between X and Y while holding fixed the value of some other variables
  - Unlike a difference-in-means or bivariate regression
  - Deals with concern that units with more X also systematically differ in the value of Z, which we believe also affects the value of Y
- Simplest example is $Y_i = \alpha + \beta X_i + \theta Z_i + \epsilon_i$
- Interpretation:
  - Y typically changes by β units for every unit increase in X holding constant the level of Z
  - Y typically changes by θ units for every unit increase in Z holding constant the level of X
  - Given the values of X_i and Z_i, Y_i is ϵ_i units different than we would expect it to be

- Suppose we estimate the values of α, β, and θ with $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$, respectively
- Define $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + \hat{\theta}Z_i$
  - Where $\hat{Y}_i$ is the fitted value of Y_i
- The least squares criterion means that we solve for that values of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$ by finding those that make $L(\hat{\alpha}, \hat{\beta}, \hat{\theta}) = \sum_{i=1}^{N} e_i{}^2$ as small as possible
  - Where $e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i - \hat{\theta}Z_i$

- Just like when we solved for the bivariate regression coefficients, we solve for our parameter estimates by
  1. Taking the derivatives of the loss function w.r.t the parameters that we wish to minimize function w.r.t
  2. Solving for the set of parameter estimates that set these equations equal to zero simultaneously
- But now we have three equations, instead of two, because we have three parameters that we are maximizing the function w.r.t.
  1. $\frac{dL(\hat{\alpha},\hat{\beta},\hat{\theta})}{d\hat{\alpha}} = \sum_{i=1}^{N} -2(Y_i - \hat{\alpha} - \hat{\beta}X_i - \hat{\theta}Z_i) = 0$
  2. $\frac{dL(\hat{\alpha},\hat{\beta},\hat{\theta})}{d\hat{\beta}} = \sum_{i=1}^{N} -2X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i - \hat{\theta}Z_i) = 0$
  3. $\frac{dL(\hat{\alpha},\hat{\beta},\hat{\theta})}{d\hat{\theta}} = \sum_{i=1}^{N} -2Z_i(Y_i - \hat{\alpha} - \hat{\beta}X_i - \hat{\theta}Z_i) = 0$

- The first equation on the previous slide implies that $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} - \hat{\theta}\bar{Z}$
- Plugging into the second equation on the previous slide gives us that $\hat{\beta} = \frac{\hat{cov}(X,Y) - \hat{\theta}\hat{cov}(X,Z)}{\hat{var}(X)}$
  - See next slide for proof
- Using similar logic, $\hat{\theta} = \frac{\hat{cov}(Z,Y) - \hat{\beta}\hat{cov}(Z,X)}{\hat{var}(Z)}$

Solving for $\hat{\beta} = \frac{cov(\hat{X,Y}) - \hat{\theta} cov(\hat{X,Z})}{var(\hat{X})}$:

- $\sum_{i=1}^{N} -X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i - \hat{\theta}Z_i) = 0 \implies$
  $\sum_{i=1}^{N} -X_i(Y_i - (\bar{Y} - \hat{\beta}\bar{X} - \hat{\theta}\bar{Z}) - \hat{\beta}X_i - \hat{\theta}Z_i) = 0 \implies$
  $\sum_{i=1}^{N} -X_i\hat{\beta}\bar{X} + \hat{\beta}{X_i}^2 =$
  $\sum_{i=1}^{N} X_iY_i - X_i\bar{Y} + \hat{\theta}X_i\bar{Z} - \hat{\theta}X_iZ_i \implies$
  $\hat{\beta}\sum_{i=1}^{N} {X_i}^2 - X_i\bar{X} =$
  $\sum_{i=1}^{N} X_iY_i - X_i\bar{Y} - \hat{\theta}\sum_{i=1}^{N} X_iZ_i - X_i\bar{Z} \implies$
  $\hat{\beta} = \frac{\sum_{i=1}^{N} X_iY_i - X_i\bar{Y} - \hat{\theta}\sum_{i=1}^{N} X_iZ_i - X_i\bar{Z}}{\sum_{i=1}^{N} {X_i}^2 - X_i\bar{X}} \implies$
  $\hat{\beta} = \frac{\frac{1}{n-1}\sum_{i=1}^{N} X_iY_i - X_i\bar{Y} - \hat{\theta}\frac{1}{n-1}\sum_{i=1}^{N} X_iZ_i - X_i\bar{Z}}{\frac{1}{n-1}\sum_{i=1}^{N} {X_i}^2 - X_i\bar{X}} \implies$
  $\hat{\beta} = \frac{cov(\hat{X,Y}) - \hat{\theta} cov(\hat{X,Z})}{var(\hat{X})}$

- Comparing the formulas for $\hat{\beta}$ when we do and do not control for Z is useful for understanding why regression coefficients change when we control for different variables
  - $\hat{\beta} = \frac{cov(X,Y)}{var(X)}$ when running a regression of Y on X
  - $\hat{\beta} = \frac{cov(\hat{X},Y) - \hat{\theta} cov(\hat{X},Z)}{var(\hat{X})}$ when running a regression of Y on X and Z
- So controlling for Z causes a change in $\hat{\beta}$ relative to the bivariate regression when:
  1. $\hat{\theta}$ which, speaking a bit loosely, means changes in Z affect Y AND
  2. $cov(X,Z) \neq 0$, which again speaking a bit loosely, means X and Y are not independent of each other

- Comparing the formulas for $\hat{\beta}$ when we do and do not control for Z also helps us make predictions about the direction of change when the conditions for change on the previous slide are met
- $\frac{cov(\hat{X},Y)-\hat{\theta}cov(\hat{X},Z)}{var\hat{(}X)} - \frac{cov(X,Y)}{var(X)} = -\frac{\hat{\theta}cov(\hat{X},Z)}{var\hat{(}X)}$
- So we expect that controlling for Z will cause $\hat{\beta}$ to:
  - Get larger when the signs of θ and $cov(X,Z)$ are different
    - E.g., an increase in Z generally causes Y to get smaller and associates with more X
  - Get smaller when the signs of θ and $cov(X,Z)$ are the same
    - E.g., an increase in Z generally causes Y to get bigger and associates with more X

- Lets apply the logic from the previous slide to think about how the association ($\hat{\beta}$) between liking cats (X) and supporting Obama (Y) will differ depending on whether we control for a measure of Democratic partisanship (Z)
- Based on what we have seen, our expectation is that
  - $cov(X, Z) > 0$ (i.e., Democrats are more likely to prefer cats)
  - $\theta > 0$ (i.e., Democrats are more likely to approve of Obama)
- Because the signs of θ and $cov(X, Z)$ are the same, we expect $\hat{\beta}$ to decrease when we control for Democratic partisanship

```
> reg1 <- lm(obamaapp ~ cats, data = mydata)
> summary(reg1)

Call:
lm(formula = obamaapp ~ cats, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6023 -0.4623 -0.4623  0.5377  0.5377

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46230    0.02013   22.96  < 2e-16 ***
cats         0.13998    0.04254    3.29  0.00104 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4972 on 784 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.01362,Adjusted R-squared:  0.01236
F-statistic: 10.83 on 1 and 784 DF,  p-value: 0.001045
```

Interpretation of this regression:

- 46.2 percent of respondents who prefer dogs to cats approved of Obama
  - The constant reports the expected value of the dependent variable when all expiatory variables are set to zero
  - Because the dependent variable (d.v.) is binary, an expected value of .462 implies a 46.2 percent chance that the d.v. takes on a value of one and a 53.8 percent chance that the d.v. takes on the value of zero

Interpretation of this regression (continued):

- $46.2 + 14.0 = 60.2$ percent of respondents who prefer cats to dog approved of Obama
  - The coefficient on "cats" reports the increase in the expected value of the dependent variable from a unit increase in "cats"
    - Where a unit increase in cats represents a shift from a respondent who prefers dogs to cat to a respondent who prefers cats to dogs
  - Again we can interpret an expected value of .602 as a 60.2 percent chance that the d.v. takes on a value of one and a 39.8 percent chance that the d.v. takes on the value of zero

```
> mydata$partisanship <- 0
> mydata$partisanship[mydata$PRTY == "Republican"] <- -1
> mydata$partisanship[mydata$PRTY == "Democrat"] <- 1
>
> reg3 <- lm(obamaapp ~ cats + partisanship, data = mydata)
> summary(reg3)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8698 -0.2019 -0.1228  0.2092  0.8772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45682    0.01685   27.11   <2e-16 ***
cats          0.07904    0.03576    2.21   0.0274 *
partisanship  0.33398    0.01822   18.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4161 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,Adjusted R-squared:  0.3082
F-statistic: 175.8 on 2 and 783 DF,  p-value: < 2.2e-16
```

Interpretation of this regression:

- Consistent with expectations, the association between liking cats and Obama support declines once we controlled for partisanship
    - Respondents who prefer cats were about 14 percentage points more likely to support Obama than respondents who prefer dogs when we didn't control for partisanship
    - Respondents who prefer cats were about 8 percentage points more likely to support Obama than than respondents who prefer dogs when we controlled for partisanship

Interpretation of this regression (continued):

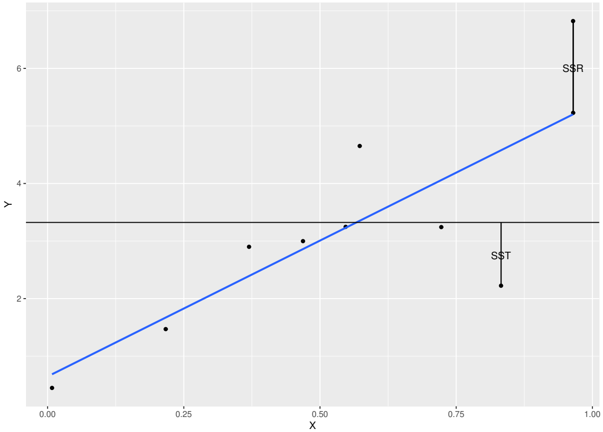
- Consistent with expectations, there was a strong association between a respondent's partisanship and Obama support controlling for someone's dog/cat preferences
    - A unit increase in partisanship associates with a 33.4 percentage point increase in Obama approval
    - Because switching from Republican to Democrat is a two-unit increase in partisanship, this means that Democrats were 66.8 percentage points more likely to support Obama than Republicans controlling for dog/cat preferences

- Near the bottom of the regression output two slides ago you see "Multiple R-squared: 0.3099"
- R-squared (or R^2) is a measure of percentage of the variation in our dependent variable that can be explain by our regression output
- So we interpret an R^2 of 0.3099 as saying that about 31 percent of the variation in Obama support is explained by partisanship and dog/cat preferences
    - Implying that about 69 percent of variation in Obama support is not explained by partisanship and dog/cat preferences

How is R^2 calculated?

- $R^2 = \frac{SSR}{SST}$, where
  - $SST = \sum_{i=1}^{n} SST_i = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, which measures the total variation in a dependent variable
  - It can be shown that $SST = SSR + SSE$
    - $SSR = \sum_{i=1}^{n} SSR_i = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, which measures the variation in a dependent variable that can be explain by a regression
    - $SSE = \sum_{i=1}^{n} SSE_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, which measures the variation in a dependent variable remains unexplained by a regression

Visualizing R^2:



Decomposing SST

- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 =$
  $\sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 =$
  $\sum_{i=1}^{n}((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2 =$
  $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2 =$
  $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 =$
    - See next slide

  $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 =$
  $SSR + SSE$

Proof that $\sum_{i=1}^{n} 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$

- $\sum_{i=1}^{n} 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) =$
  $2\sum_{i=1}^{n} e_i(X_i\hat{\beta} - \bar{X}\hat{\beta}) =$
  $2\sum_{i=1}^{n} e_i(X_i - \bar{X})\hat{\beta}$
- But we proved back on an implication slide that for all k, $\sum_{i=1}^{n} x_{ik}e_i = 0$

Facts about R^2:

- Always lies between 0 and 1
  - R^2 equals 0 when our X's have no explanatory power over Y
  - R^2 equals 1 when our X's completely explain Y
- Comparing R^2 across models is generally not a good way to judge which model is better
  - Only assesses which model explains more variation
- In part because adding an additional variable to the model will make the R^2 no worse
  - Can achieve an R^2 of 1 by including a "dummy variable" for each data point
  - R also reports an "Adj R-squared" includes a penalty that increases in the number of explanatory variables

- We now will generalizes the logic from a regression model with two independent variables to a regression model with k independent variables:
  $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$
  - Interpretation is that we expect Y to change by β_1 units for every unit increase in X_1 holding constant the level of $X_2, X_3, \ldots X_k$
  - Interpretation is that we expect Y to change by β_2 units for every unit increase in X_2 holding constant the level of $X_1, X_3, \ldots X_k$
  - $\ldots$
  - Interpretation is that. we expect Y to change by β_k units for every unit increase in X_k holding constant the level of $X_2, X_3, \ldots X_{k-1}$

- While in theory we could solve for $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ using the same techniques that we used in the previous section, the algebra quickly become impossible to manage
- Thus, we turn to using matrices to represent a collection of data as a way to make the notation much simpler
  - For example, we can use the matrix $\hat{\beta}$ to represent the collection of $k + 1$ parameter estimates $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$
- The next section shows how we derive and understand the formula $\hat{\beta} = (X^T X)^{-1} X^T Y$ that we can use to estimate regression coefficients for a regression for any number of explanatory variables
  - Although there will be limits on the nature and number of explanatory variables that can be included in a regression based on characteristics of the data being analyzed

- Understanding $\hat{\beta} = (X^TX)^{-1}X^TY$ first requires us to develop an understanding of matrices
- A matrix is a rectangular array of numbers, which we refer to as elements
- The number in the *ith* row and the *jth* column of a matrix is called the i, jth element and is written in lower case
- We write the matrix A as:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix}$$

- So $b_{21} = 3$ when

$$B = \begin{pmatrix} 2 & 1 & 6 \\ 3 & 1 & 2 \end{pmatrix}$$

- The size of a matrix is indicated by the number of rows and columns.
- A matrix with n rows and k columns is said to be size n by k (or nXk)
- Thus, B is size $2X3$ when

$$B = \begin{pmatrix} 2 & 1 & 6 \\ 3 & 1 & 2 \end{pmatrix}$$

- A square matrix is a matrix such that $k = n$

Some types of matrices that are useful to know about:

- A is a diagonal matrix iff $k = n$ and $a_{ij} = 0$ if $i \neq j$

$$\textit{Example: } A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

- A is a symmetric matrix iff $k = n$ and $a_{ij} = a_{ji}$

$$\textit{Example: } A = \begin{pmatrix} 1 & 2 & 6 \\ 2 & 5 & -1 \\ 6 & -1 & 3 \end{pmatrix}$$

- A is the identity matrix iff $k = n$, $a_{ij} = 1$ if $i = j$, and $a_{ij} = 0$ if $i \neq j$

$$\textit{Example: } A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Addition:

- Let A be a $n_1 X k_1$ matrix
- Let B be a $n_2 X k_2$ matrix
- If $n_1 = n_2$ and $k_1 = k_2$ then

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1k} + b_{1k} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2k} + b_{2k} \\ \dots & \dots & \dots & \dots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \dots & a_{nk} + b_{nk} \end{pmatrix}$$

$$\textit{Example:} \begin{pmatrix} 1 & 2 \\ 5 & 6 \end{pmatrix} + \begin{pmatrix} 2 & 4 \\ 5 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 10 & 9 \end{pmatrix}$$

- Otherwise

$$A + B = \varnothing$$

Scalar multiplication:

- Let c be a real valued number
- Then

$$cA = \begin{pmatrix} ca_{11} & ca_{12} & \dots & ca_{1k} \\ ca_{21} & ca_{22} & \dots & ca_{2k} \\ \dots & \dots & \dots & \dots \\ ca_{n1} & ca_{n2} & \dots & ca_{nk} \end{pmatrix}$$

$$Example:\ 2\begin{pmatrix} 1 & 2 \\ 5 & 6 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 10 & 12 \end{pmatrix}$$

Matrix multiplication:

- Let A be a $n_1 X k_1$ matrix
- Let B be a $n_2 X k_2$ matrix
- If $k_1 = n_2$ then AB is a $n_1 X k_2$ matrix
- Otherwise $AB = \varnothing$

Matrix multiplication:

- To obtain the value of ab_{ij} we multiply the *ith* row of A by the *jth* column of B
- Specifically, $ab_{ij} = \sum_{w=1}^{n=k_1} a_{iw} b_{wj}$

$$Example: \; A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$ab_{11} = 1 * 5 + 2 * 7 = 19$$

$$ab_{12} = 1 * 6 + 2 * 8 = 22$$

$$ab_{21} = 3 * 5 + 4 * 7 = 43$$

$$ab_{22} = 3 * 6 + 4 * 8 = 50$$

$$So, \; AB = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

Matrix multiplication:

- Note that order matters
  - Unlike with addition or scalar multiplication
- Before we saw that:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

- In contrast:

$$\begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 23 & 34 \\ 31 & 36 \end{pmatrix}$$

Inverse matrix:

- There is no such thing as matrix division
- But the inverse something somewhat similar
  - For some square matrices
- Let inverse of matrix A, A^{-1}, is the matrix st $AA^{-1} = A^{-1}A = I$

$$Example:\ A = \begin{pmatrix} 2 & 4 \\ -3 & 1 \end{pmatrix}, A^{-1} = \begin{pmatrix} \frac{1}{14} & \frac{-2}{7} \\ \frac{3}{14} & \frac{1}{7} \end{pmatrix}$$

$$aa^{-1}{}_{11} = 2 * \frac{1}{14} + 4 * \frac{3}{14} = 1$$

$$aa^{-1}{}_{12} = 2 * \frac{-2}{7} + 4 * \frac{1}{7} = 0$$

$$aa^{-1}{}_{21} = -3 * \frac{1}{14} + 1\frac{3}{14} = 0$$

$$aa^{-1}{}_{22} = -3 * \frac{-2}{7} + 1\frac{1}{7} = 1$$

Matrix Rank:

- The rank of the matrix is the number of linearly independent columns of the matrix
- A column is linearly independent if it cannot be constructed by adding other columns in the matrix

$$\textit{Example: rank}\left(\begin{pmatrix} 2 & 1 & 0 \\ 2 & 1 & 0 \\ 3 & 1 & 1 \end{pmatrix}\right) = 2$$

$$Col.1 = 2 * Col.2 + Col.3$$

- A square matrix of size nXn can only be inverted if it has a rank of n
  - This is the math behind the concept of multicolinearity

Transpose:

- We use the notation A^T or A' to denote the transpose of matrix A
- The transpose operation interchanges the rows and columns of a matrix (i.e., $a_{ij} = a^T_{ji}$)

$$A = \begin{pmatrix} 2 & 1 & 6 \\ 3 & 1 & 2 \end{pmatrix} \Longrightarrow A^T = \begin{pmatrix} 2 & 3 \\ 1 & 1 \\ 6 & 2 \end{pmatrix}$$

- We will apply the following property of the transpose: $(AB)^T = B^T A^T$

- Suppose we summarize all of our data in matrix form:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}$$

- Using this notation $Y = X\beta + \epsilon$

- Suppose we have an estimate $\hat{\beta}$ of the vector β
- We can use this estimate to construct $\hat{Y} = X\hat{\beta}$ so

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \dots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{pmatrix}$$

- Which we can use to construct $e = Y - \hat{Y}$

$$\begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} - \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \dots \\ \hat{Y}_n \end{pmatrix}$$

- Remember that our least squares criterion is to select the $\hat{\beta}$ that minimizes $s(\hat{\beta}) = \sum_{i=1}^{N} {e_i}^2$
- $s(\hat{\beta}) = \sum_{i=1}^{N} {e_i}^2 =$

$$\begin{pmatrix} e_1 & e_2 & \dots & e_n \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = e^T e \implies$$

- $s(\hat{\beta}) = e^T e =$
$(Y - X\hat{\beta})^T (Y - X\hat{\beta}) =$
$Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}$
- We solve for $\hat{\beta}$ by taking the derivative of $s(\hat{\beta})$ wrt to $\hat{\beta}$ and solving for the $\hat{\beta}$ that sets this derivative equal to zero

- $s(\hat{\beta}) = Y^T Y - Y^T X \hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta} \implies$
  $\frac{ds(\hat{\beta})}{d\hat{\beta}^T} = -X^T Y - X^T Y + X^T X \hat{\beta} + X^T X \hat{\beta} = 0 \implies$
  $X^T X \hat{\beta} = X^T Y \implies$
  $(X^T X)^{-1} X^T X \hat{\beta} = (X^T X)^{-1} X^T Y \implies$
  $I \hat{\beta} = (X^T X)^{-1} X^T Y \implies$
  $\hat{\beta} = (X^T X)^{-1} X^T Y$

- Now that we have a formula for $\hat{\beta}$, we need to think about how to interpret what we estimate using it
- What we will establish over the remainder of the course is that our interpretation depends heavily on the properties of the data being analyzed
  - And particularly the properties of the determinants of the dependent variable that are not modeled, ϵ
- We'll begin by thinking about what we learn from regression coefficients when five assumptions about our data hold
- And then see how interpretations change when these assumption fail to hold

- Five assumptions about our data
  1. X has full rank
  2. The true model that generates our data is $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$
  3. $E[\epsilon_i^2 \mid X] = \sigma^2$ (homoscadasticity)
  4. $E[\epsilon_i \epsilon_j \mid X] = 0$ if $i \neq j$ (no autocorrelation)
  5. $E[\epsilon_i \mid X] = 0$

- When the five assumptions on the previous slide hold, then

1. $E[(X^TX)^{-1}X^TY] = E[\hat{\beta}] = \beta$
   - On average, we estimate the true value of β
2. $var(\hat{\beta}) = S^2(X^TX)^{-1}$
   - Where $S^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2$
3. We can interpret $\hat{\beta}$ as our best estimate of the effect of X on Y

- Point number three from the previous slide is a consequence of the Gauss-Markov Theorem
  - Tells us that under assumptions 1 through 5, the least squares estimator is the minimum variance linear unbiased estimate of β
- Sketch of proof:
  - Define $\hat{\beta}* = CY$, where C is a matrix (like $(X^TX)^{-1}X^T$)
  - We can express $C = (X^TX)^{-1}X^T + D$
  - Next slide shows that unbiased implies that $DX = \bar{0}_k$
  - Following slide shows that $var(CY) = \sigma^2((X^TX)^{-1} + D^TD)$
  - Because D^TD is a positive semidefinite matrix, this will be smallest when $D^T = \bar{0}_k$
    - Intuition is that the smallest squared real number is 0
- Bottom line: We cannot do any better by weighting observations in any other way, so we should be using least squares

Proof that $DX = \bar{0}_k$

- Unbiased means that $E[CY] = E[((X^TX)^{-1}X^T + D)(X\beta + \epsilon)] = \beta \implies$
- $E[(X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon + DX\beta + D\epsilon] = \beta \implies$
- $E[I_k\beta + (X^TX)^{-1}X^T\epsilon + DX\beta + D\epsilon] = \beta \implies$
- $E[I_k\beta] + E[(X^TX)^{-1}X^T\epsilon] + E[DX\beta] + E[D\epsilon] = \beta \implies$
- $\beta + E[(X^TX)^{-1}X^TE[\epsilon \mid X]] + DX\beta + E[DE[\epsilon \mid X]] = \beta \implies$
- $E[(X^TX)^{-1}X^T\bar{0}_n] + DX\beta + E[D\bar{0}_n] = \bar{0}_k \implies$
- $\bar{0}_k + DX\beta + \bar{0}_k = \bar{0}_k \implies$
- $DX\beta = \bar{0}_k \implies$
- $DX = \bar{0}_k\bar{0}_k^T$

Proof that $var(CY) = \sigma^2((X^TX)^{-1} + D^TD)$

- $var(CY) = E[(\hat{\beta}* - E[\hat{\beta}*])(\hat{\beta}* - E[\hat{\beta}*])^T] =$
- $E[(\hat{\beta}* - \beta)(\hat{\beta}* - \beta)^T] =$
- $E[(((X^TX)^{-1}X^T + D)(X\beta + \epsilon) - \beta)$
  $(((X^TX)^{-1}X^T + D)(X\beta + \epsilon) - \beta)^T] =$
- $E[((X^TX)^{-1}X^TX\beta + DX\beta + (X^TX)^{-1}X^T\epsilon + D\epsilon - \beta)$
  $((X^TX)^{-1}X^TX\beta + DX\beta + (X^TX)^{-1}X^T\epsilon + D\epsilon - \beta)^T] =$
- $E[(\beta + \bar{0}_k\bar{0}_k^T\beta + (X^TX)^{-1}X^T\epsilon + D\epsilon - \beta)$
  $(\beta + \bar{0}_k\bar{0}_k^T\beta + (X^TX)^{-1}X^T\epsilon + D\epsilon - \beta)^T] =$
- $E[((X^TX)^{-1}X^T\epsilon + D\epsilon)((X^TX)^{-1}X^T\epsilon + D\epsilon)^T]$
- $E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1} + (X^TX)^{-1}X^T\epsilon\epsilon^TD + D\epsilon\epsilon^TX(X^TX)^{-1} + D\epsilon\epsilon^TD^T]$

Proof that $var(CY) = \sigma^2((X^TX)^{-1} + D^TD)$ (continued)

- $E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1} + (X^TX)^{-1}X^T\epsilon\epsilon^TD + D\epsilon\epsilon^TX(X^TX)^{-1} + D\epsilon\epsilon^TD^T] =$
- $\sigma^2((X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1} + D^TD) =$
    - Applying assumption 4 and assumption 5
- $\sigma^2((X^TX)^{-1} + (X^TX)^{-1}(DX)^T + DX(X^TX)^{-1} + D^TD) =$
    - Applying $(AB)^T = B^TA^T$
- $\sigma^2((X^TX)^{-1} + (X^TX)^{-1}(\bar{0}_k\bar{0}_k^T)^T + \bar{0}_k\bar{0}_k^T(X^TX)^{-1} + D^TD) =$
- $\sigma^2((X^TX)^{-1} + D^TD)$

Earlier in class we discussed the following output:

```
> reg3 <- lm(obamaapp ~ cats + partisanship, data = mydata)
> summary(reg3)

Call:
lm(formula = obamaapp ~ cats + partisanship, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8698 -0.2019 -0.1228  0.2092  0.8772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45682    0.01685   27.11   <2e-16 ***
cats          0.07904    0.03576    2.21   0.0274 *
partisanship  0.33398    0.01822   18.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4161 on 783 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,Adjusted R-squared:  0.3082
F-statistic: 175.8 on 2 and 783 DF,  p-value: < 2.2e-16
```

- When the five Gauss-Markov assumptions are true then our best guess that:
  - Liking cats makes you 7.9 percentage points more likely to support Obama
    - With a standard error (e.g. measure of uncertainty) on this estimate of 3.6 percentage points
  - Increasing partisanship by one-unit makes you 33.4 percentage points more likely to support Obama
    - With a standard error (e.g., measure of uncertainty) on this estimate of 1.8 percentage. points
- Next week we'll talk about how we can use R to construct confidence intervals for β based on these estimates and standard errors

- The Gauss-Markov Theorem explains why least squares regressions are so ubiquitous
  - We have a straightforward analytic formula for both estimated effects and uncertainty
  - That is best formula we could use to fit a line to data
- Unfortunately, the assumptions underlying the Gauss-Markov Theorem often do not hold in practice
- So we'll spend the remainder of this class thinking about how we can adjust our approach to still get meaningful information when these assumptions do not hold

- Focus of the remaining classes:
  - Rest of week 6 focuses on assumptions one and two
  - Week 7 focuses on assumptions three and four
  - Week 8 focuses on assumptions five
- The broad goals are:
  - Recognize when the regressions you want to run are likely to violate one or more of these assumptions
  - Identify strategies to deal with these violations in order to reduce the likelihood that you reach erroneous conclusions about the relationships between an explanatory variable and a dependent variable on the basis of a regression results
- Because the ability to run regressions in R without an ability to properly structure or interpret them is a dangerous situation

- To calculate our regression coefficients, we need to be able to calculate $(X^T X)^{-1}$
- This will not be possible if X doesn't have full rank
  - Occurs when at least one column in X is a linear combination of one or more other column(s) in X
- There are two common reasons why this will happen
  1. There is no variation in a variable that you are including in your regression
     - Making it a linear combination of the constant
  2. A series of variables partition the set of possible outcomes
     - This is called multicolinearity

Assumption #1: X has full rank

- $X = \begin{pmatrix} 1 & 1 & 0 & \dots & X_{1k} \\ 1 & 0 & 1 & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \\ 1 & 0 & 1 & \dots & X_{nk} \end{pmatrix}$
- Example:
  - Suppose X_{i1} is an indicator (or dummy variable) for whether respondent i is male
  - Suppose X_{i2} is an indicator (or dummy variable) for whether respondent i is female
  - Then Col. 3 = Col. 1 - Col. 2
    - Meaning that X is not full rank

- An implication is that you always need to have an excluded group when using dummy variables
  - And the coefficient on a dummy variable is interpreted relative to that excluded group
- To illustrate the concept of an excluded group, lets return to our exploration of the association between the liking cats and supporting Obama and also control for a respondent's sex

- The regression output on the next few slides highlights some general points about multicollinearity
  1. R automatically drops variables when it encounters multicolinarity
     - E.g., femaleTRUE is NA
  2. R automatically drops the dummy variable associated with the largest value of a factor variable to avoid multicolinarity
  3. While coefficient(s) change depending on the excluded group, the substantive interpretation should always remain the same
     - E.g., Men are 0.3 percentage points less likely to support Obama than women or women are 0.3 percentage points more likely to support Obama than men

Code to create a male dummy variable and a female dummy variable:

```
> table(mydata$SEX)

  Male Female
   405    616
> mydata$male <- (mydata$SEX == "Male")
> mydata$female <- (mydata$SEX == "Female")
```

```
> reg4 <- lm(obamaapp ~ cats + partisanship + male + female, data = mydata)
> summary(reg4)

Call:
lm(formula = obamaapp ~ cats + partisanship + male + female,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2020 -0.1227  0.2094  0.8773

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4569400  0.0218901  20.874   <2e-16 ***
cats          0.0789890  0.0361934   2.182   0.0294 *
partisanship  0.3339686  0.0183004  18.249   <2e-16 ***
maleTRUE     -0.0002638  0.0306818  -0.009   0.9931
femaleTRUE           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4164 on 782 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,Adjusted R-squared:  0.3073
F-statistic: 117.1 on 3 and 782 DF,  p-value: < 2.2e-16
```

```
> reg5 <- lm(obamaapp ~ cats + partisanship + SEX, data = mydata)
> summary(reg5)

Call:
lm(formula = obamaapp ~ cats + partisanship + SEX, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2020 -0.1227  0.2094  0.8773

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4566762  0.0237476  19.230   <2e-16 ***
cats         0.0789890  0.0361934   2.182   0.0294 *
partisanship 0.3339686  0.0183004  18.249   <2e-16 ***
SEXFemale    0.0002638  0.0306818   0.009   0.9931
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4164 on 782 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,Adjusted R-squared:  0.3073
F-statistic: 117.1 on 3 and 782 DF,  p-value: < 2.2e-16
```

```
> reg6 <- lm(obamaapp ~ cats + partisanship + male, data = mydata)
> summary(reg6)

Call:
lm(formula = obamaapp ~ cats + partisanship + male, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2020 -0.1227  0.2094  0.8773

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4569400  0.0218901  20.874   <2e-16 ***
cats          0.0789890  0.0361934   2.182   0.0294 *
partisanship  0.3339686  0.0183004  18.249   <2e-16 ***
maleTRUE     -0.0002638  0.0306818  -0.009   0.9931
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4164 on 782 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3099,Adjusted R-squared:  0.3073
F-statistic: 117.1 on 3 and 782 DF,  p-value: < 2.2e-16
```

- We continue to need an excluded group when we partition outcomes into more than two groups
- The code below creates dummy variables for a respondent's partisanship that partition the outcome of partisanship into four groups
- The next few slides show that the estimated difference between Democrats and Republicans doesn't depend on the excluded group

```
> table(mydata$PRTY)

         Republican            Democrat         Independent Don't know/No answer
                297                 363                 292                   69
> mydata$rep <- (mydata$PRTY == "Republican")
> mydata$dem <- (mydata$PRTY == "Democrat")
> mydata$ind <- (mydata$PRTY == "Independent")
> mydata$oth <- (mydata$PRTY == "Don't know/No answer")
```

```
 > reg7 <- lm(obamaapp ~ cats + dem + rep + oth, data = mydata)
> summary(reg7)

Call:
lm(formula = obamaapp ~ cats + dem + rep + oth, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8776 -0.2110 -0.1315  0.2020  0.8685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44784    0.02987  14.995  < 2e-16 ***
cats         0.07957    0.03580   2.223   0.0265 *
demTRUE      0.35018    0.03772   9.285  < 2e-16 ***
repTRUE     -0.31636    0.03926  -8.058  2.9e-15 ***
othTRUE     -0.04330    0.06834  -0.634   0.5266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4164 on 781 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3108,Adjusted R-squared:  0.3073
F-statistic: 88.05 on 4 and 781 DF,  p-value: < 2.2e-16
```

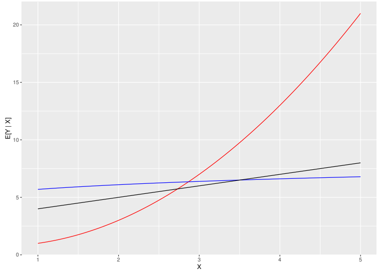
- The excluded group in previous slide are Independent respondents
- Thus, the coefficients on “demTRUE”, “repTRUE”, “othTRUE” imply that:
  - Democratic respondents were 35.0 percentage points more likely to support Obama than Independents respondents holding fixed their dog/cat preferences
  - Republican respondents were 31.6 percentage points less likely to support Obama than Independents respondents holding fixed their dog/cat preferences
  - Respondents who were not Democrats, Republicans, nor Independents were 4,3 percentage points less likely to support Obama than Independents respondents holding fixed their dog/cat preferences

- We can also use the coefficients on "demTRUE" and "repTRUE" to make comparison of Democrat and Republican respondents
- We back out that Democratic respondents were 66.6 percentage points more likely to support Obama than Republican respondents holding fixed their dog/cat preferences
  - Because Dem - Rep. = (Dem. - Ind.) - (Rep. - Ind.) = "demTRUE" - "repTRUE" = 35.0 - (-31.6) = 66.6
- Implication is that we should find a coefficient of 66.6 on "demTRUE" if Democratic respondents were the excluded group and a coefficient of -66.6 on "repTRUE" if Republican respondents were the excluded group

```
Call:
lm(formula = obamaapp ~ cats + ind + rep + oth, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8776 -0.2110 -0.1315  0.2020  0.8685 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.79802    0.02630  30.345  < 2e-16 ***
cats         0.07957    0.03580   2.223   0.0265 *  
indTRUE     -0.35018    0.03772  -9.285  < 2e-16 ***
repTRUE     -0.66655    0.03650 -18.261  < 2e-16 ***
othTRUE     -0.39348    0.06677  -5.893 5.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4164 on 781 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3108,Adjusted R-squared:  0.3073 
F-statistic: 88.05 on 4 and 781 DF,  p-value: < 2.2e-16
```

- Bottom line is that you will know if X isn't full rank, because a statistical program will not allow you to estimate that model
- The bigger concern in when one column is almost a linear combination of other column(s) in your dataset
- In such case, you will get estimates, but they can be extremely misleading or have large standard errors
  - Something called a VIF diagnostic that sometimes gets applied to help uncover the issue

- Even when all of the explanatory variable(s): $X_1, X_2, \ldots, X_k$ that determine the dependent variable, Y have been identified, a regression model must specify a functional form of the relationship between these explanatory variables and the dependent variable
- Some of the possible functional forms between X and $E[Y \mid X]$:



- Useful that linear regressions can be fit to any model in which the dependent variable is determined by a linear combination of regression coefficients and the explanatory variables
    - All of the following relationship are determined by a linear combination of regression coefficients and the explanatory variable :
      $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$
      $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 {X_{i1}}^2 + \beta_3 X_{i2} + ... + \beta_{k+1} X_{ik} + \epsilon_i$
      $Y_i = \alpha + \beta_1 ln(X_{i1}) + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$
    - But we cannot estimate the following using a linear regression:
      $Y_i = \alpha + \beta_1 X_{i1} + {\beta_1}^2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$
      $Y_i = (\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik})\epsilon_i$

- Often hard to assess whether we are likely to satisfy assumption #2, because generally easier to assess whether X affects Y than how X affects Y
- Ways to assess assumption #2
  1. Apply theory
     - Do we expect the effect of a unit change in an explanatory variable on Y to be increasing, decreasing or constant as the value of that explantory variable gets larger
     - Do we expect the effect of a unit change in an explanatory variable on Y to depend on the value of another explanatory variable
  2. Visual inspection of the data
     - Although increases the risk of overfitting model to the data

- Suppose we think that X_1, X_2, X_3, X_4 affect Y
- And we model such that:
  $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i$
- Implicit in this regression model is:
  - The expected change in Y from a unit change in X_1 is the same no matter what the value of X_1
  - The expected change in Y from a unit change in X_1 is the same for any combination of values of X_2, X_3, X_4
- Both of these facts are established mathematically by noting that $\frac{dY}{dX_1} = \beta_1$

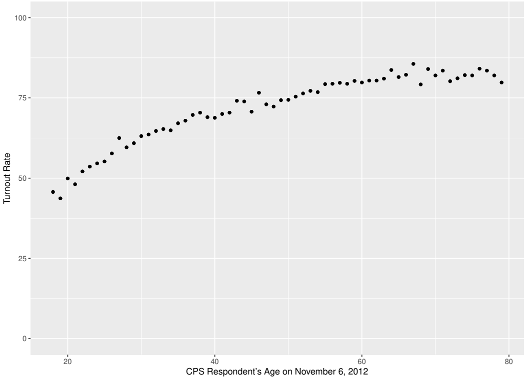
- Suppose theory said that it was not reasonable to assume that the expected change in Y from a unit change in X_1 is the same no matter what the value of X_1
- One common way to deal with this is to also include higher-order terms of X_1 (e.g., ${X_1}^2, {X_1}^3, \ldots$) as explanatory variables in the regression
- An example of such a model is: $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 {X_{i1}}^2 + \beta_3 X_{i2} + \beta_4 X_{i3} + \beta_5 X_{i4} + \epsilon_i$
- Some features of this model are:
  - $\frac{dY}{dX_1} = \beta_1 + \beta_2 X_{i1}$
  - $\frac{dY}{dX_2} = \beta_3$

Implications of the features of the model highlighted on the previous slide:

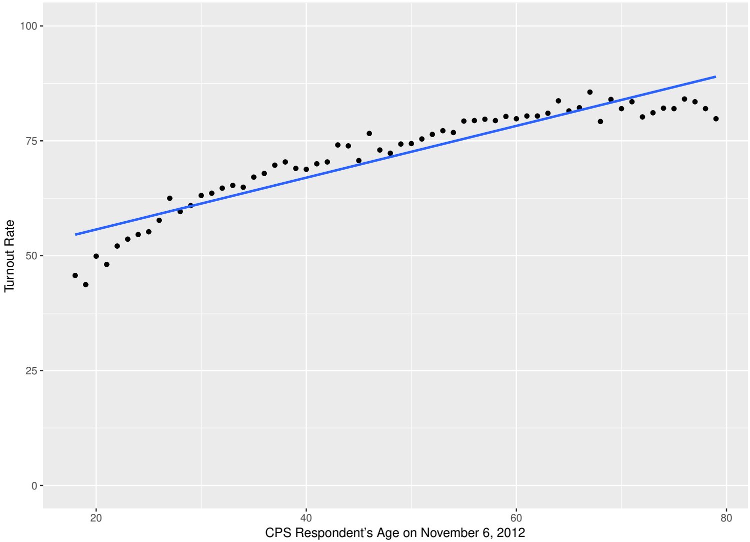
- If $\beta_2 \neq 0$, the expected change in Y from a unit change in X_1 varies based on the value of X_1
  - When $\beta_2 > 0$, then Y increases by more (or decreases by less) from a unit change in X_1 as X_1 gets larger
  - When $\beta_2 < 0$, then Y increases by less (or decreases by more) from a unit change in X_1 as X_1 gets larger
- The expected change in Y from a unit change in X_1 is the same for any combination of values of X_2, X_3, X_4
- The expected change in Y from a unit change in X_2 is the same no matter what the value of X_2

- Adding higher-order terms of X_1 to our regression model has both promise and perils in structuring our thinking about the relationship between X_1 and Y
- Promise:
  - We can describe the relationship between X_1 and Y within our sample in an increasing nuanced way as we add more higher-order terms of X_1 to our regression model
  - If the true relationship is linear, we'll coverage to estimating $\hat{\beta}$'s on these higher order terms that equal 0
- Perils:
  - We risk at overfitting our model to describe idiosyncrasies of the specific sample of data that we collect in a way that won't generalize into the broader population

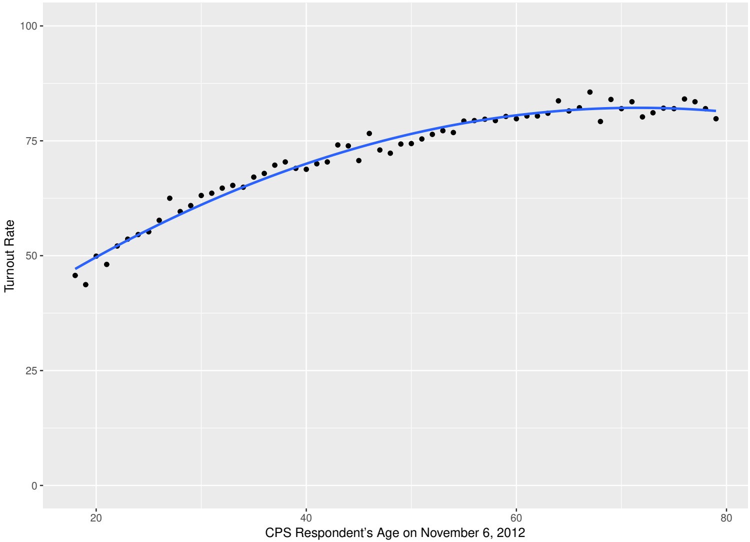
- To illustrate the points on the previous slide we are going to investigate the relationship between self-reported voter turnout and age among respondents on the 2012 Current Population Survey (CPS)



```
> summary(lm(Turnout ~ poly(AgeNum, 1, raw = TRUE), data = cps))

Call:
lm(formula = Turnout ~ poly(AgeNum, 1, raw = TRUE), data = cps)

Residuals:
    Min      1Q  Median      3Q     Max
-11.441  -3.286   1.564   2.603   6.238

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 44.42992    1.44968   30.65   <2e-16 ***
poly(AgeNum, 1, raw = TRUE)  0.56373    0.02804   20.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.951 on 60 degrees of freedom
Multiple R-squared:  0.8707,Adjusted R-squared:  0.8686
F-statistic: 404.1 on 1 and 60 DF,  p-value: < 2.2e-16
```

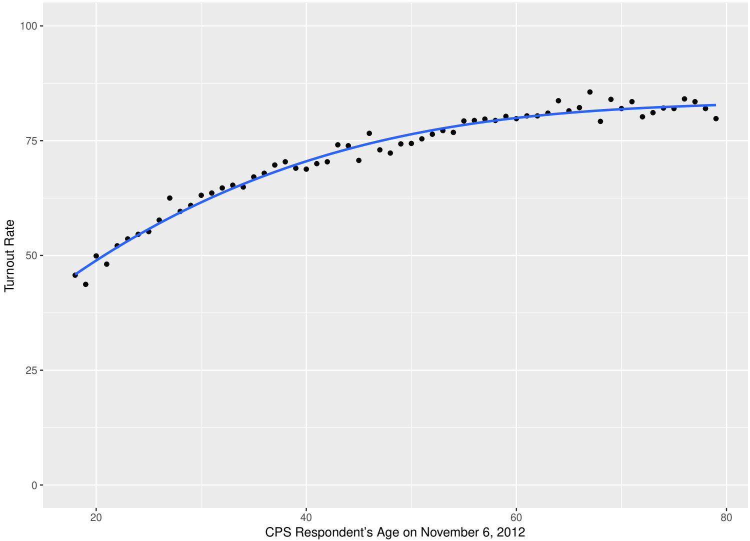

Turnout Rate
100
75
50
25
0
20
40
60
80
CPS Respondent's Age on November 6, 2012

```
> summary(lm(Turnout ~ poly(AgeNum, 2, raw = TRUE), data = cps))

Call:
lm(formula = Turnout ~ poly(AgeNum, 2, raw = TRUE), data = cps)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7118 -1.0951 -0.0225  1.2306  4.5866

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  19.569199   1.680583   11.64   <2e-16 ***
poly(AgeNum, 2, raw = TRUE)1  1.750490   0.075309   23.24   <2e-16 ***
poly(AgeNum, 2, raw = TRUE)2 -0.012235   0.000766  -15.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.727 on 59 degrees of freedom
Multiple R-squared:  0.9757,Adjusted R-squared:  0.9749
F-statistic:  1185 on 2 and 59 DF,  p-value: < 2.2e-16
```


Turnout Rate
100
75
50
25
0
20
40
60
80
CPS Respondent's Age on November 6, 2012

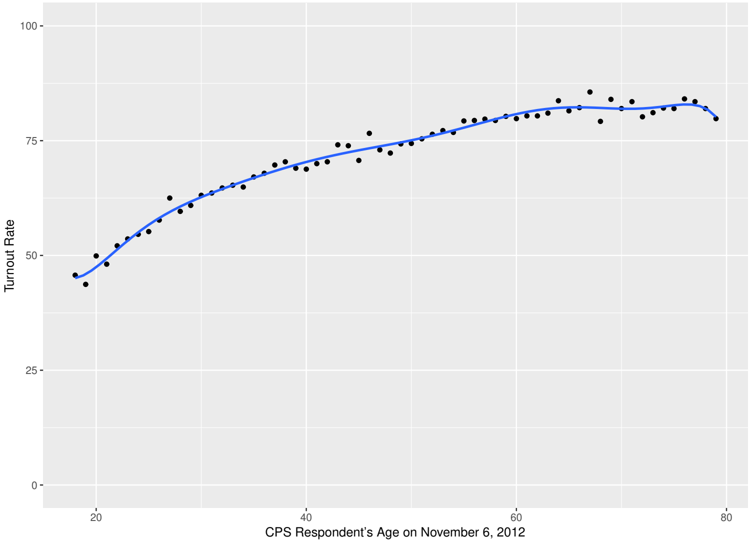
- Lets compare the change in expected rate of turnout from a unit change in age when age is 30 and age is 70 when using a quadratic to model age
- $Turnout_i = 19.569199 + 1.750490 * Age_i - 0.012235 * {Age_i}^2 \implies \frac{dTurnout}{dAge} = 1.750490 - 2 * 0.012235 * Age$
    - $\frac{dTurnout}{dAge} \approx 1.02$ for someone who is 30 years old
    - $\frac{dTurnout}{dAge} \approx 0.04$ for someone who is 70 years old

```
> summary(lm(Turnout ~ poly(AgeNum, 3, raw = TRUE), data = cps))

Call:
lm(formula = Turnout ~ poly(AgeNum, 3, raw = TRUE), data = cps)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7008 -1.0645  0.2121  0.8905  4.3009

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   9.528e+00  4.338e+00   2.196   0.0321 *
poly(AgeNum, 3, raw = TRUE)1  2.506e+00  3.115e-01   8.044 5.17e-11 ***
poly(AgeNum, 3, raw = TRUE)2 -2.920e-02  6.844e-03  -4.266 7.44e-05 ***
poly(AgeNum, 3, raw = TRUE)3  1.166e-04  4.676e-05   2.493   0.0156 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.655 on 58 degrees of freedom
Multiple R-squared:  0.9781,Adjusted R-squared:  0.9769
F-statistic: 862.1 on 3 and 58 DF,  p-value: < 2.2e-16
```


100
75
50
25
0
Turnout Rate
20
40
60
80
CPS Respondent's Age on November 6, 2012

- Lets compare the change in expected rate of turnout from a unit change in age when age is 30 and age is 70 when using a cubic to model age
- $Turnout_i =$ $9.528 + 2.506 * Age_i - 0.0292 * {Age_i}^2 + 0.0001166 * {Age_i}^3 \implies \frac{dTurnout}{dAge} = 2.506 - 2 * 0.0292 * Age + 3 * 0.0001166 * {Age_i}^2$
  - $\frac{dTurnout}{dAge} \approx 1.07$ for someone who is 30 years old
  - $\frac{dTurnout}{dAge} \approx 0.13$ for someone who is 70 years old

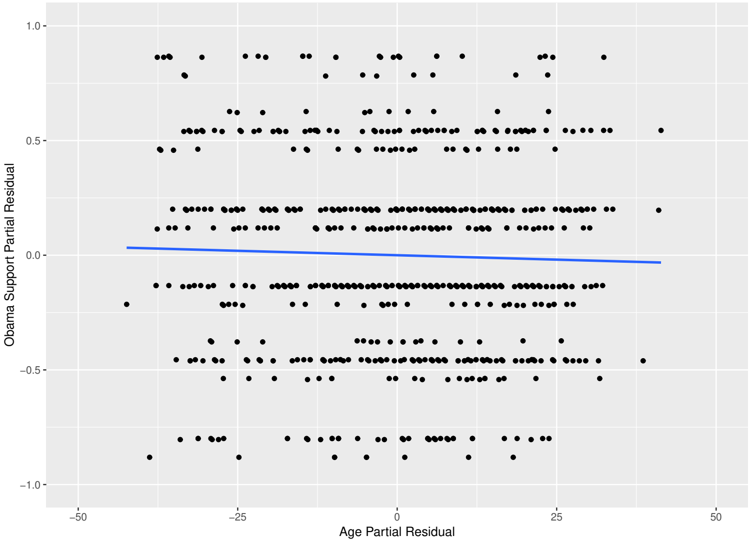
```
> summary(lm(Turnout ~ poly(AgeNum, 9, raw = TRUE), data = cps))

Call:
lm(formula = Turnout ~ poly(AgeNum, 9, raw = TRUE), data = cps)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9345 -0.9218 -0.1200  0.7009  3.3767

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.737e+03  1.710e+03   1.601   0.1155
poly(AgeNum, 9, raw = TRUE)1 -6.443e+02  3.982e+02  -1.618   0.1117
poly(AgeNum, 9, raw = TRUE)2  6.575e+01  3.985e+01   1.650   0.1050
poly(AgeNum, 9, raw = TRUE)3 -3.776e+00  2.253e+00  -1.676   0.0997 .
poly(AgeNum, 9, raw = TRUE)4  1.351e-01  7.936e-02   1.703   0.0946 .
poly(AgeNum, 9, raw = TRUE)5 -3.133e-03  1.810e-03  -1.731   0.0893 .
poly(AgeNum, 9, raw = TRUE)6  4.711e-05  2.675e-05   1.761   0.0841 .
poly(AgeNum, 9, raw = TRUE)7 -4.436e-07  2.477e-07  -1.791   0.0791 .
poly(AgeNum, 9, raw = TRUE)8  2.377e-09  1.305e-09   1.821   0.0744 .
poly(AgeNum, 9, raw = TRUE)9 -5.526e-12  2.989e-12  -1.849   0.0702 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.448 on 52 degrees of freedom
Multiple R-squared:  0.985,Adjusted R-squared:  0.9823
F-statistic: 378.2 on 9 and 52 DF,  p-value: < 2.2e-16
```


100
75
50
25
0
Turnout Rate
20
40
60
80
CPS Respondent's Age on November 6, 2012

- No easy way to tell which of these is the right way to model the relationship between turnout and age
    - Although pretty clear evidence that the right relationship is not linear
- One approach that people sometime use when making a choice like this is to separate the data into training and validation data
    - Fit the models using training data
    - Apply the models to predict the outcome in the validation data
    - Select the model in which the predictive and actual outcomes are the most similar in the validation data

- In the previous example, I was assessing the form of the relationship between X and Y without controlling for any other variables
- Partial residual plots are a good way to assess the form of this same relationship if we are controlling for other variables
- Steps to generate a partial residual plot for Y on X_1
  1. Regress $X_{i1} = \gamma_1 + \gamma_2 X_{i2} + ... + \gamma_k X_{ik} + \epsilon_i$
  2. Construct $X_{i1}{}^* = X_{i1} - (\hat{\gamma}_1 + \hat{\gamma}_2 X_{i2} + ... + \hat{\gamma}_k X_{ik})$
  3. Regress $Y_i = \lambda_1 + \lambda_2 X_{i2} + ... + \lambda_k X_{ik} + \epsilon_i$
  4. Construct $Y_i{}^* = Y_i - (\hat{\lambda}_1 + \hat{\lambda}_2 X_{i2} + ... + \hat{\lambda}_k X_{ik})$
  5. Make a scatter plot with $Y_i{}^*$ on the y-axis and $X_{i1}{}^*$ on the x-axis

```
> reg9 <- lm(AGENUM ~ cats + ind + rep + oth + male, data = fulldata)
> fulldata$ageresid <- resid(reg9)
>
> reg10 <- lm(obamaapp ~ cats + ind + rep + oth + male, data = fulldata)
> fulldata$obamaresid <- resid(reg10)
```


1.0
0.5
0.0
−0.5
−1.0
Obama Support Partial Residual
−50
−25
0
25
50
Age Partial Residual

```
> reg11 <- lm(obamaresid ~ ageresid, data = fulldata)
> summary(reg11)

Call:
lm(formula = obamaresid ~ ageresid, data = fulldata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9110 -0.2171 -0.1079  0.2091  0.8881

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.917e-17  1.505e-02   0.000    1.000
ageresid    -7.709e-04  8.639e-04  -0.892    0.372

Residual standard error: 0.4139 on 754 degrees of freedom
Multiple R-squared:  0.001055,Adjusted R-squared:  -0.0002698
F-statistic: 0.7963 on 1 and 754 DF,  p-value: 0.3725
```

```
> reg12 <- lm(obamaapp ~ cats + ind + rep + oth + male + AGENUM, data = fulldata)
> summary(reg12)

Call:
lm(formula = obamaapp ~ cats + ind + rep + oth + male + AGENUM,
    data = fulldata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9110 -0.2171 -0.1079  0.2091  0.8881

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8402809  0.0545118  15.415  < 2e-16 ***
cats         0.0846375  0.0369533   2.290   0.0223 *
indTRUE     -0.3439145  0.0386861  -8.890  < 2e-16 ***
repTRUE     -0.6641998  0.0371128 -17.897  < 2e-16 ***
othTRUE     -0.4296607  0.0719523  -5.971 3.63e-09 ***
maleTRUE     0.0037105  0.0313550   0.118   0.9058
AGENUM      -0.0007709  0.0008668  -0.889   0.3741
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4152 on 749 degrees of freedom
Multiple R-squared:  0.3167,Adjusted R-squared:  0.3112
F-statistic: 57.86 on 6 and 749 DF,  p-value: < 2.2e-16
```

- Thus far we have been assuming that the relationship between an explanatory variable and a dependent variable does not depend on the value of another explanatory variable
- When such an assumption does not accurately describe how the world works, we need to use an interactive regression model
- Our baseline interactive model is: $Y_i = \alpha + \beta X_i + \theta Z_i + \gamma X_i Z_i + \epsilon_i$
- Some features of this model are:
  - $\frac{dY_i}{dX_i} = \beta + \gamma Z_i$
  - $\frac{dY_i}{dZ_i} = \theta + \gamma X_i$

Implications of the features of the model highlighted on the previous slide:

- If $\gamma \neq 0$, the expected change in Y from a unit change in X varies based on the value of Z
  - When $\gamma > 0$, then Y increases by more (or decreases by less) from a unit change in X as Z gets larger
  - When $\gamma < 0$, then Y increases by less (or decreases by more) from a unit change in X as Z gets larger
- If $\gamma \neq 0$, the expected change in Y from a unit change in Z varies based on the value of X
  - When $\gamma > 0$, then Y increases by more (or decreases by less) from a unit change in Z as X gets larger
  - When $\gamma < 0$, then Y increases by less (or decreases by more) from a unit change in Z as X gets larger

Additional implications of the baseline interaction model when $Z_i \in \{0, 1\}$:

- β is the expected change in Y from a unit increase in X if $Z = 0$
- $\beta + \gamma$ is the expected change in Y from a unit increase in X if $Z = 1$
- γ is the difference in the expected change in Y from a unit increase in X when $Z = 1$ relative to when $Z = 0$

- When both $X_i, Z_i \in \{0, 1\}$, we can use regression coefficients from our baseline interaction model to make a 2 X 2 table that represent $E[Y_i \mid X_i, Z_i]$

		X_i:	
		0	1
Z_i:	0	α	$\alpha + \beta$
	1	$\alpha + \theta$	$\alpha + \beta + \theta + \gamma$

```
> reg12 <- lm(obamaapp ~ cats*havepet, data = mydata)
> summary(reg12)

Call:
lm(formula = obamaapp ~ cats * havepet, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6183 -0.4384 -0.4384  0.5020  0.5616

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.49796    0.03175  15.684   <2e-16 ***
cats          0.05760    0.08060   0.715    0.475
havepet      -0.05960    0.04105  -1.452    0.147
cats:havepet  0.12237    0.09518   1.286    0.199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.497 on 782 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.01694,Adjusted R-squared:  0.01317
F-statistic: 4.493 on 3 and 782 DF,  p-value: 0.003896
```

- The previous slide shows a regression of Obama support on whether someone prefers cats to dogs, whether someone has a pet, and the interaction of these two
- Here is how we can combine these coefficients to get expected Obama support among every possible combination:

		$Cats_i$: 0	$Cats_i$: 1
$Pets_i$:	0	0.498	$0.498 + 0.058 = 0.556$
$Pets_i$:	1	$0.498 - 0.060 = 0.438$	$0.498 + 0.058 - 0.060 + 0.122 = 0.618$

- A common conditional hypothesis is that X increases Y when $Z = 1$, but not when $Z = 0$
- Three null hypotheses generated by this model are:
  1. $\beta = 0$
     - No relationship between X and Y when $Z = 0$
     - We can test the null that $\beta = 0$ using the p-value reported in baseline R output
  2. $\gamma > 0$
     - Greater relationship between X and Y when $Z = 1$ than when $Z = 0$
     - We can test the null that $\gamma = 0$ using the p-value reported in baseline R output
  3. $\beta + \gamma > 0$
     - Relationship between X and Y when $Z = 1$
     - Not contained in baseline R output, so need to use the linearHypothesis() function

Testing $\beta + \gamma > 0$:

```
> linearHypothesis(reg12, c("cats + cats:havepet = 0"))
Linear hypothesis test

Hypothesis:
cats  + cats:havepet = 0

Model 1: restricted model
Model 2: obamaapp ~ cats * havepet

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    783 196.26
2    782 193.14  1    3.1222 12.641  4e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Interpreting coefficients from interactive regression is challenging
- One of the challenges is that including interaction terms changes the interpretation of non-interaction terms
- To illustrate, consider the difference in the interpretation of β in these two regressions:
  1. $Y_i = \alpha + \beta X_i + \theta Z_i + \epsilon_i$
     - β represents the expected change in Y from a unit change in X
  2. $Y_i = \alpha + \beta X_i + \theta Z_i + \gamma X_i Z_i + \epsilon_i$
     - β represents the expected change in Y from a unit change in X conditional on Z equalling zero
- Implication: no coefficient summarizes the unconditional expected change in Y from a unit change in X when X is interacted with another variable

- Assessing significance is also more challenging in regressions with interaction terms
- The next three slides show that
  - Can be. hard to assess the statistical significance of the expected change in Y from a unit change in X when looking at the output of a regression in which X is interacted with another variable Z
  - Not necessary to have statistically significant coefficients for there to be a statistically significant interactive relationship between X and Z
- Implication: Only include interaction terms if you primarily care about the heterogeneity in the relationship between X and Y or isolating the relationship between X and Y when certain conditions are present

```
> reg13 <- lm(obamaapp ~ cats + havepet + PRTY + urban, data = mydata)
> summary(reg13)

Call:
lm(formula = obamaapp ~ cats + havepet + PRTY + urban, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03599 -0.28342 -0.03106  0.21979  0.96894

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.28683    0.06942   4.132 4.00e-05 ***
cats                      0.08253    0.03609   2.287  0.02248 *
havepet                  -0.04859    0.03096  -1.569  0.11695
PRTYDemocrat              0.66663    0.03634  18.342  < 2e-16 ***
PRTYIndependent           0.32507    0.03910   8.313 4.15e-16 ***
PRTYDon't know/No answer  0.28040    0.06727   4.168 3.42e-05 ***
urbanMid City            -0.16265    0.07411  -2.195  0.02849 *
urbanSuburbs             -0.09615    0.06754  -1.424  0.15499
urbanRural               -0.20718    0.07070  -2.930  0.00349 **
urbanUnknown             -0.11671    0.06930  -1.684  0.09254 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4136 on 776 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3242,Adjusted R-squared:  0.3164
F-statistic: 41.36 on 9 and 776 DF,  p-value: < 2.2e-16
```

```
> reg14 <- lm(obamaapp ~ cats*havepet + PRTY + urban, data = mydata)
> summary(reg14)

Call:
lm(formula = obamaapp ~ cats * havepet + PRTY + urban, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02473 -0.28229 -0.03007  0.21617  0.96993

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.28881    0.07003   4.124 4.12e-05 ***
cats                      0.06986    0.06717   1.040  0.29863
havepet                  -0.05192    0.03437  -1.511  0.13129
PRTYDemocrat              0.66606    0.03645  18.271  < 2e-16 ***
PRTYIndependent           0.32452    0.03920   8.278 5.46e-16 ***
PRTYDon't know/No answer  0.27956    0.06742   4.147 3.75e-05 ***
urbanMid City            -0.16235    0.07417  -2.189  0.02890 *
urbanSuburbs             -0.09594    0.06759  -1.419  0.15616
urbanRural               -0.20683    0.07077  -2.923  0.00357 **
urbanUnknown             -0.11593    0.06943  -1.670  0.09534 .
cats:havepet              0.01784    0.07976   0.224  0.82307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4139 on 775 degrees of freedom
  (235 observations deleted due to missingness)
Multiple R-squared:  0.3242,Adjusted R-squared:  0.3155
F-statistic: 37.19 on 10 and 775 DF,  p-value: < 2.2e-16
```

Testing $\beta + \gamma > 0$:

```
> linearHypothesis(reg14, c("cats + cats:havepet = 0"))
Linear hypothesis test

Hypothesis:
cats  + cats:havepet = 0

Model 1: restricted model
Model 2: obamaapp ~ cats * havepet + PRTY + urban

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    776 133.48
2    775 132.76  1   0.71649 4.1824 0.04118 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- A common mistake when estimating interactive models is omitting Z_i as an explanatory variable and estimating: $Y_i = \alpha + \beta X_i + \gamma X_i Z_i + \epsilon_i$
  - E.g., including the product of having a higher than normal class size and below average income as an explanatory variable, but not below average income by itself
- This model can incorrectly attribute an effect of Z on Y as an interaction
  - If Z has an independent effect on Y then it is an omitted variable that is positively associated with XZ
  - E.g., attribute the omitted direct effect of below average income to the coefficient on the interaction
- Also important to remember that controlling for W does not control for XW
  - And so if our primary coefficient of interest is γ, want to think about what other interactions that we want to control for

Key takeaways:

- We frequently want to learn how a dependent variable varies as a function an independent variable while holding fixed some other independent variable(s)
- Multivariate regression can estimate and test hypotheses about a variety of such quantities of interest
    - Despite being called linear regression, not limited to estimating a linear relationship between X and Y
    - Interaction terms allow for exploration of relationship between X and Y in particular cases of interest
- While it is easy to run a multivariate regression in R, structuring and interpreting the output properly is hard
    - What is the excluded group?
    - Are there interdependencies between my variables?
- It is important to interpret regression coefficients in terms of their implications for your quantity of interest
    - Both in terms of statistical and substantive significance