Multiple Regression

November 30, 2022

Today

- Extension of Linear Regression to multiple IVs
  - Why include other IVs?
  - Prediction with multiple IVs
  - Interpreting the effects of individual coefficients
  - Interpreting statistical significance
  - Evaluating Model fit
    - R^2
    - F-test
- *Note*: You will spend most of next semester (722) on multiple linear regression

Linear Regression--Review

- Linear regression gives the equation for the "best fit" line
  - Only appropriate for linear relationships
  - Line minimizes the squared residuals
- Gives regression equation $Y = \beta_0 + \beta_1 \mathrm{X}$
  - β_0="constant" or "intercept"
  - β_1="slope"
- Can use regression equation to generate predicted values for y (i. e. , $\hat{y}$)
  - $\hat{y} = \beta_0 + \beta_1 X_i$
- Can compare predicted values to observed values
  - $Y_i = \beta_0 + \beta_1 X_i + e_i$
- Residual (e_i) is the difference between the observed value and the predicted value

Regression and Means--Review

- Regression equation predicts the mean value of Y in sub-populations
  - Sub-populations are those with different values of X
- Individual observations for Y vary around the mean of Y in sub-populations
- Linear relationship assumes that these means fall along a line
  - A one-unit change in X has the same effect on the mean of Y across the entire range of X

Regression and Inference

- In the population, there is a regression equation
- The regression equation that we generate in our sample is an estimate of the population regression equation
  - $\hat{\beta}_0$ is an estimate of the "true" β_0 in the population
  - $\hat{\beta}_1$ is an estimate of the "true" β_1 in the population
- When we test hypotheses – we are interested in whether X has an effect on Y in the population
- Test the **null hypothesis** that X has "no effect" on Y
  - If true, then $\beta_1 = 0$
- Need to determine how likely it is that we would observe $\hat{\beta}_1$ of the size that we find in the sample if $\beta_1 = 0$

Rejecting the null hypothesis

- Determining this requires knowing the **standard error**
  - You know how to compute this using the model RMSE—also given by statistical software
- With the coefficient and s.e., we can find a t-statistic
  - $t = \frac{\beta_1}{SE_{\beta_1}}$
- Then, **use the t-distribution chart** to determine the p-value
  - Remember, degrees of freedom = *n*-*k*-1
  - Decide whether you are looking at one or two tails
- P-value tells you the probability you would get a coefficient of the observed magnitude if the null hypothesis were true
- Compare to the level of statistical significance chosen
  - Ex: if $p \leq .05$, statistically significant at the .05 level; can reject null hypothesis with 95% confidence

Confidence Intervals—Review

- Statistical significance is a test of how confident we are that, in the population, β_1 does not equal 0
- Also interested in estimating a range within which we think β_1 lies
- To generate a confidence interval, find the *t* for a specific level of confidence
  - From the t-distribution with degrees of freedom *n*-*k*-1
- Multiply this by the standard error, add and subtract from coefficient
- A level C confidence interval for β_1 is:
  - $\beta_1 +/-(t)(SE_{\beta_1})$

Regression Example

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Model Fit

- $Y_i = \hat{Y}_i + \hat{e}_i$
- Total Sum of Squares (TTS): sum of difference between observed Y and mean of Y, squared
- Explained Sum of Squares (ESS): sum of difference between predicted Y and mean of Y, squared
- Residual sum of squares (RSS): sum of the squared residuals
- TSS=ESS + RSS
- Can judge goodness-of-fit using R^2
- $R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$
- R^2 indicates the fraction of the variation in Y that is explained by X

Regression Example

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

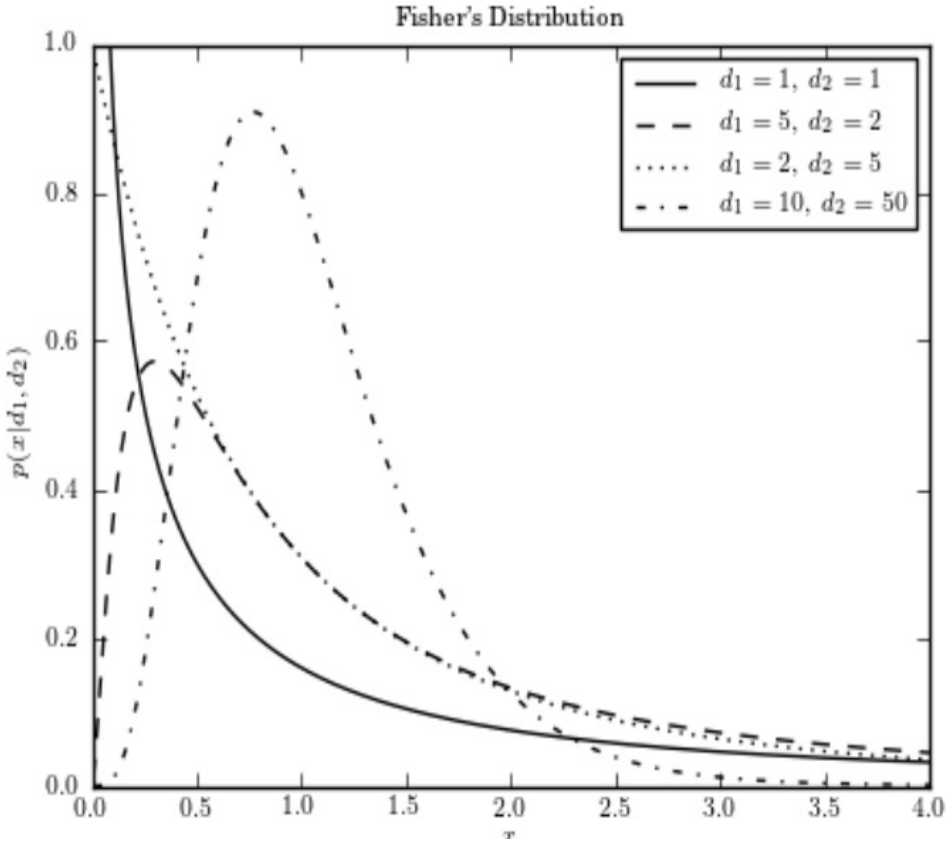
F-test

- When we have only one IV, we can use an F-test to test the null hypothesis that B_1=0
- The F-test uses the F-statistic, which is based on the "mean squares"
- Mean square (MS) = sum of squares divided by degrees of freedom
- Each part of the model has degrees of freedom
  - Model Degrees of Freedom (DFM) = # of IVs
  - Error Degrees of Freedom (DFE) = n – k - 1
  - Total Degrees of Freedom (DFT) = DFM + DFE = n-1
- So:
  - Mean Square Model (MSM)=$\frac{ESS}{DFM}$
  - Mean Square Error (MSE) =$\frac{RSS}{DFE}$
  - Mean Square Total (MST) =$\frac{TSS}{DFT}$

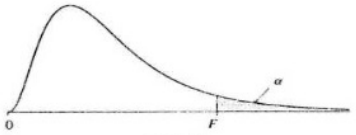
F-test—Bivariate Regression

- The F-statistic is equal to MSM divided by MSE
  - $F = \frac{MSM}{MSE}$
- When the null hypothesis is true, the F-statistic follows an F-distribution with:
  - k degrees of freedom in the numerator
  - $n - k$ - 1 degrees of freedom in the denominator
- With only one IV, the p-value is the probability that a random variable having the F(1, n-2) distribution is ≥ the calculated value of the F-statistic (if the null were true)
- When only one IV, $t^2 = F$, and the t-statistic and the F-statistic will tell us the same thing
  - The F-test is really only useful with multiple IVs

F-distribution



F-Distribution Table



$\alpha = .05$

df_2	df_1 1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Source: From Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London, 1974. (Previously published by Oliver & Boyd, Edinburgh.) Reprinted by permission of the authors and publishers.

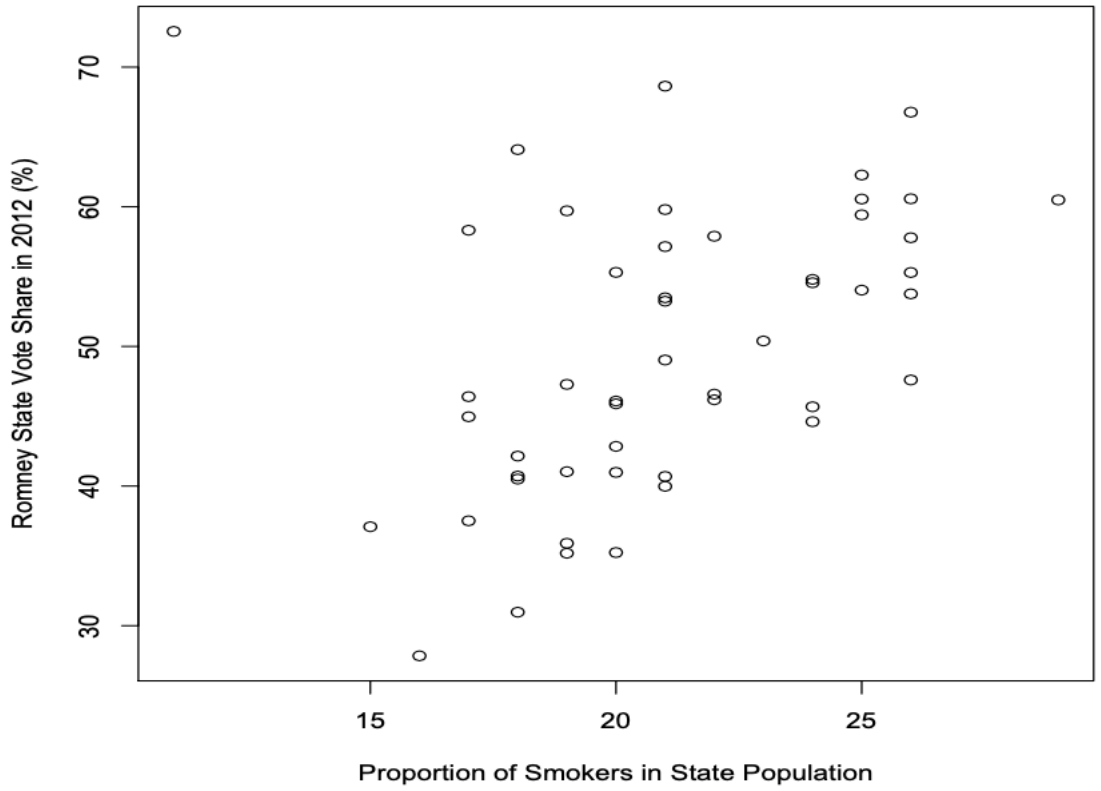
Regression Example

```
Call:
lm(formula = Obama2012 ~ reppct_m, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9484  -5.1864   0.8801   5.0077  13.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7026     5.5661  14.499  < 2e-16 ***
reppct_m     -1.0415     0.1745  -5.968 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.875 on 48 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.414
F-statistic: 35.62 on 1 and 48 DF,  p-value: 2.805e-07
```

Multiple Regression

- One important advantage of regression—easy to incorporate additional independent variables
- Why would we control for additional variables?
  - Worried about omitted variable bias
  - Problem if some underlying (unobserved) factor (X_2) is driving the relationship between X_1 and Y
  - A problem if (1) X_1 and X_2 are correlated; & (2) X_2 is correlated with Y
- Example—the proportion of a state's citizens that are smokers is predictive of state voting returns for Mitt Romney in 2012
  - Its unlikely that this is a causal relationship (and thus not spurious)
  - Rather, it's likely that states with more smokers share other features associated with voting for Romney

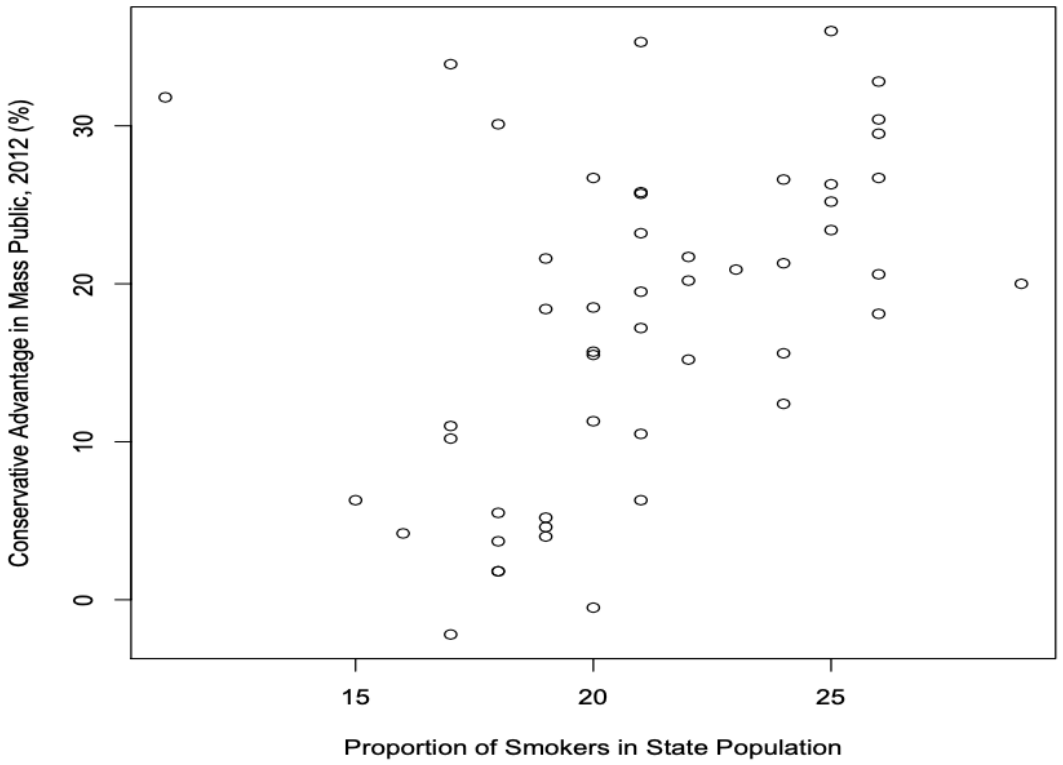
Multiple Regression Example



Multiple Regression Example

```
Call:
lm(formula = Romney2012 ~ Smokers12, data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-16.682  -6.561  -0.654   5.598  33.249

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.8456     8.2603   3.371  0.00149 **
Smokers12     1.0423     0.3869   2.694  0.00970 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.609 on 48 degrees of freedom
Multiple R-squared:  0.1313,      Adjusted R-squared:  0.1132
F-statistic: 7.258 on 1 and 48 DF,  p-value: 0.009698
```

Multiple Regression Example



Multiple Regression Example

```
Call:
lm(formula = Romney2012 ~ Smokers12 + Conserv_advantage, data =
states)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0544  -3.0022  -0.2395   2.8293   9.3041

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.18002    3.82681   9.193  4.5e-12 ***
Smokers12         -0.06831    0.19565  -0.349    0.729
Conserv_advantage  0.90652    0.06733  13.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.406 on 47 degrees of freedom
Multiple R-squared:  0.8211,      Adjusted R-squared:  0.8135
F-statistic: 107.9 on 2 and 47 DF,  p-value: < 2.2e-16
```

Multiple Regression

- When we have multiple IVs, we have a regression equation with multiple predictors:
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k$
- The multiple regression is still the linear equation that minimizes the sum of squared residuals
- Multiple regression is still a prediction about the mean of Y in sub-populations
- But, sub-populations are those who share the same value on all X's (X_1, X_2, …X_k)
- A one-unit change in each individual IV has the same effect on the mean of Y, holding all other IVs constant
  - A one-unit change in X_1 has a b_1-unit effect on Y, on average and holding all else constant

The Multiple Regression Equation

- How do we determine the coefficient estimates?
  - Don't worry about the math this semester (…yeah!)
  - Will touch on this in 722 (….doh!)

- What about the standard errors?
  - Likewise…

- R will generate coefficients & standard errors for us
  - Just focus on interpretation this semester

Multiple Regression Example

```
Call:
lm(formula = Romney2012 ~ Smokers12 + Conserv_advantage, data =
states)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0544  -3.0022  -0.2395   2.8293   9.3041

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.18002    3.82681   9.193  4.5e-12 ***
Smokers12         -0.06831    0.19565  -0.349    0.729
Conserv_advantage  0.90652    0.06733  13.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.406 on 47 degrees of freedom
Multiple R-squared:  0.8211,      Adjusted R-squared:  0.8135
F-statistic: 107.9 on 2 and 47 DF,  p-value: < 2.2e-16
```

Multiple Regression and Prediction

- We can generate predicted values for individual observations
- Now, we have to substitute values for **all** of the IVs into the regression equation
- Example—(1) Massachusetts has a -2.2 for Conservative Advantage in Mass Public in 2012 and 17% smokers; and (2) Virginia has a 18.4 for Conservative Advantage and 19% smokers
  - What are the predicted values?
- Can compare predicted values to observed values
  - Romney received 37.51% of the vote in Massachusetts and 47.28% in Virginia
- Residual is the difference between the predicted value and observed value
  - What are the residual values?

Interpreting Coefficients

- Coefficient (β_1) is the effect of a one-unit change in X_1 on Y, holding all other X's at constant values
  - Example: the effect of an additional one-percentage-point increase in a conservative public advantage at constant values of the state's proportion of smokers
- This is important if, for example, X_1 and X_2 are correlated
  - If states with more smokers also have more conservative citizens, then states at higher values of X_1 will generally be more clustered at the higher values of X_2
- If X_1 and X_2 are completely uncorrelated, then adding X_2 to the model will have no effect on the coefficient for X_1

Confidence Intervals

- Each individual regression coefficient in the model has a corresponding standard error
- We can use the standard error and the t-statistic to compute confidence intervals
  - Degrees of freedom equal to $n - k$ - 1
- For β_1, a level-C confidence interval is:
  - $\beta_1 \pm (\text{t})(SE_{b_1})$
- Remember, this tell us our estimate of the effect of X_1 on Y in the population, holding all other IVs constant

Statistical Significance

- Interpret the statistical significance of variables in similar way
  - $t = \frac{\beta_1}{SE_{\beta_1}}$
- Statistical significance tells us whether we can reject the null hypothesis that $\beta_1 = 0$ at constant values of other IVs in the model
- If the relationship between X_1 and Y is entirely driven by lurking variable Z, and we control for Z, we should not find statistical significance for β_1

Multiple Regression Example

```
Call:
lm(formula = Romney2012 ~ Smokers12 + Conserv_advantage, data =
states)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0544  -3.0022  -0.2395   2.8293   9.3041

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.18002    3.82681   9.193  4.5e-12 ***
Smokers12         -0.06831    0.19565  -0.349    0.729
Conserv_advantage  0.90652    0.06733  13.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.406 on 47 degrees of freedom
Multiple R-squared:  0.8211,      Adjusted R-squared:  0.8135
F-statistic: 107.9 on 2 and 47 DF,  p-value: < 2.2e-16
```

Model Fit

- Can compute sums of squares with multiple IVs
- Same basic calculation:
  - Total Sum of Squares (TSS): Sum of difference between observed y and mean of y, squared
  - Model Sum of Squares (ESS): Sum of difference between predicted y and mean y, squared
  - Residual sum of squares (RSS): squared residuals
  - TSS=ESS + RSS
- One difference, again, is the number of degrees of freedom:
  - DFM = # of IVs in model (i.e., k)
  - DFE = n - # of IVs in model - 1 (i.e., $n - k - 1$)

R-squared

- R-squared is calculated the same way
  - $R^2 = \frac{ESS}{TSS}$
- Now, however, R-squared tells us the total amount of variation in *Y* explained by the model (i.e., all *X*'s rather than by an individual *X*)

Multiple Regression Example

```
Call:
lm(formula = Romney2012 ~ Smokers12 + Conserv_advantage, data =
states)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0544  -3.0022  -0.2395   2.8293   9.3041

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.18002    3.82681   9.193  4.5e-12 ***
Smokers12         -0.06831    0.19565  -0.349    0.729
Conserv_advantage  0.90652    0.06733  13.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.406 on 47 degrees of freedom
Multiple R-squared:  0.8211,      Adjusted R-squared:  0.8135
F-statistic: 107.9 on 2 and 47 DF,  p-value: < 2.2e-16
```

F-test

- In simple linear regression (with one IV), an F-test was not useful
- But, it is more useful in multiple regression
- Can be used to test whether all of the coefficients in the population are equal to zero
  - H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots \beta_k = 0$
- The F statistic is again $F = \frac{MSM}{MSE}$
  - Based on the F(k, $n - k - 1$) distribution
- Thus, we have an omnibus test of whether the whole model fits well:
  - H_0: model does not fit at the 1-α level
  - H_1: model fits at the 1-α level

Inference on Multiple (Partial) Regression Coefficients

- Suppose you want to test a hypothesis about more than one coefficient at once.
- Estimate a *restricted* model as if the null hypothesis is true and an *unrestricted* model that does not impose that assumption.
- For example, consider the unrestricted model:
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_1$
- You state the null hypothesis: $H_0: \beta_1 = \beta_2 = 0$
- If the null hypothesis is true, then the restricted model is:
  - $Y_i = \beta_0 + \beta_3 X_{3i} + \mu_1$
- The unrestricted model will explain more variance in Y_i. Does it have a *significantly* better fit?

General *F*-Testing

- Provided the restricted model is *nested within* the unrestricted model, we can use an *F*-ratio to test whether the null hypothesis is true.
- Record the RSS for the unrestricted model (RSS_{UR}) and for the restricted model (RSS_R) and use them to calculate the test statistic:
  - $F = \frac{\frac{RSS_R - RSS_{UR}}{m}}{\frac{RSS_{UR}}{n-k-1}}$
  - In this equation, *n* is the sample size, *k* is the number of parameters in the unrestricted model, and *m* is the number of parameters (predictors) omitted in the restricted model.
- This test statistic follows a *F*-distribution with $\frac{m}{n-k-1}$ degrees of freedom.

Multiple Regression Example

```
Call:
lm(formula = Romney2012 ~ Smokers12 + Conserv_advantage, data =
states)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0544  -3.0022  -0.2395   2.8293   9.3041

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.18002    3.82681   9.193  4.5e-12 ***
Smokers12         -0.06831    0.19565  -0.349    0.729
Conserv_advantage  0.90652    0.06733  13.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.406 on 47 degrees of freedom
Multiple R-squared:  0.8211,      Adjusted R-squared:  0.8135
F-statistic: 107.9 on 2 and 47 DF,  p-value: < 2.2e-16
```

Substantive Significance—Very Critical

- In multiple regression, we can compare the magnitude of the effects of different variables
- But, remember that coefficient is the predicted impact of a one-unit change, and a *one-unit change means different things depending on how variable is measured/scaled*
- Can make similar comparisons to evaluate substantive significance:
  - Effect of a one-standard deviation change—a more "common" variation in the data
  - Min-to-max changes—interprets the full (observed) range of predicted effects, but may not be a reasonable (i.e., common) change
    - Generally a good idea to report both

Additive & Interactive Effects

- Multiple regression model assumes that the effect of IVs are additive
  - X_1 has the same effect (magnitude) on Y across all different values of X_2
- But, that may not be true—especially theoretically
- If the effect of X_1 on Y changes across different values of X_2, then there is an interactive effect of X_1 and X_2
  - The effect of X_1 on Y is conditional on X_2
- We can model interactive effects by including a multiplicative term:
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

Next Steps

- Finished with the substantive material for the final exam
- Next time, we will (briefly) discuss logistic regression)
  - Not covered in this course or the exam
  - But, you will likely encounter it—will discuss the basic logic
- Final homework due next week
- Final exam distributed through elms