CS 134 Data Visualization: Week 1

Joshua Goldberg

Edmonds College

Thank you to Allison Obourn for parts of these slides

Recap

Understanding data and different data types

Distributions, PDF, CDF

Sampling and descriptive statistics

Hypothesis testing to evaluate a single parameter

Bivariate linear model

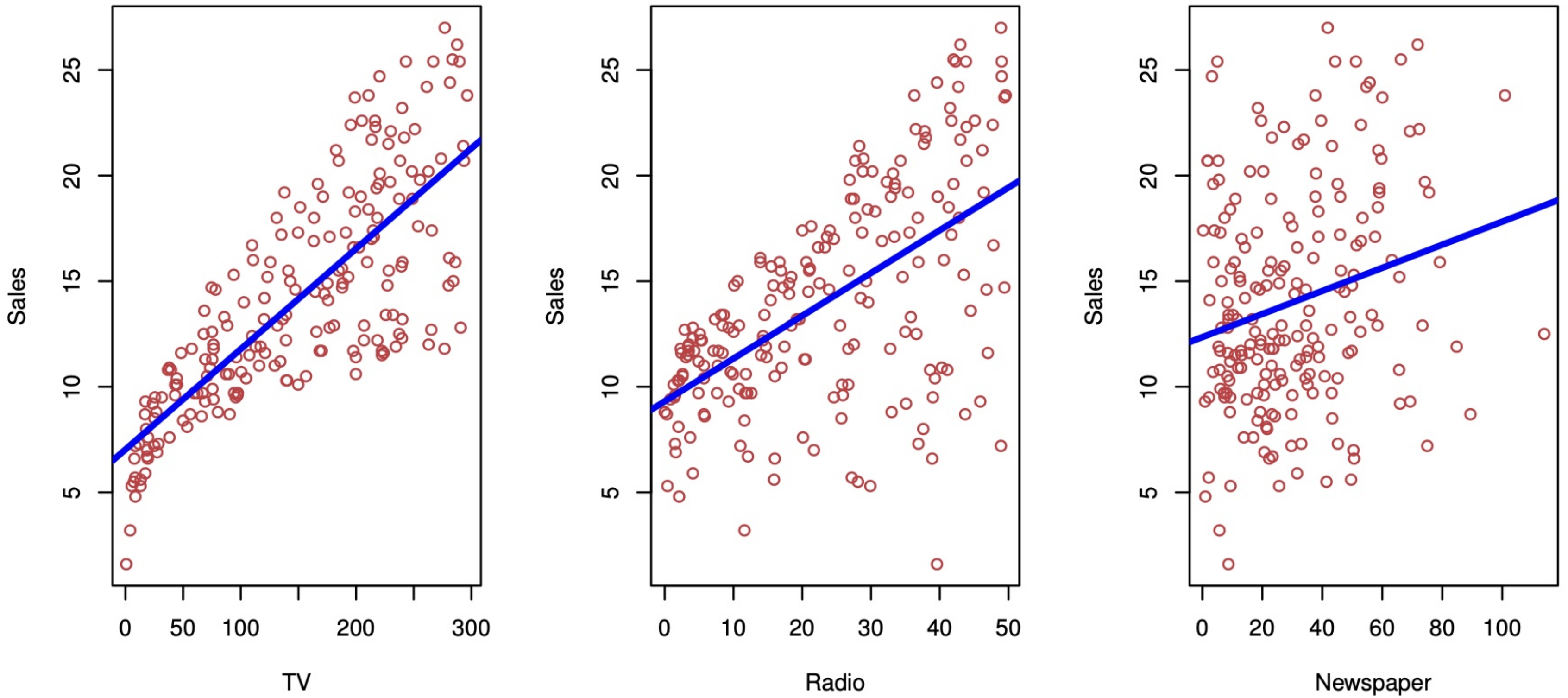
Correlation vs. Causation

Agenda

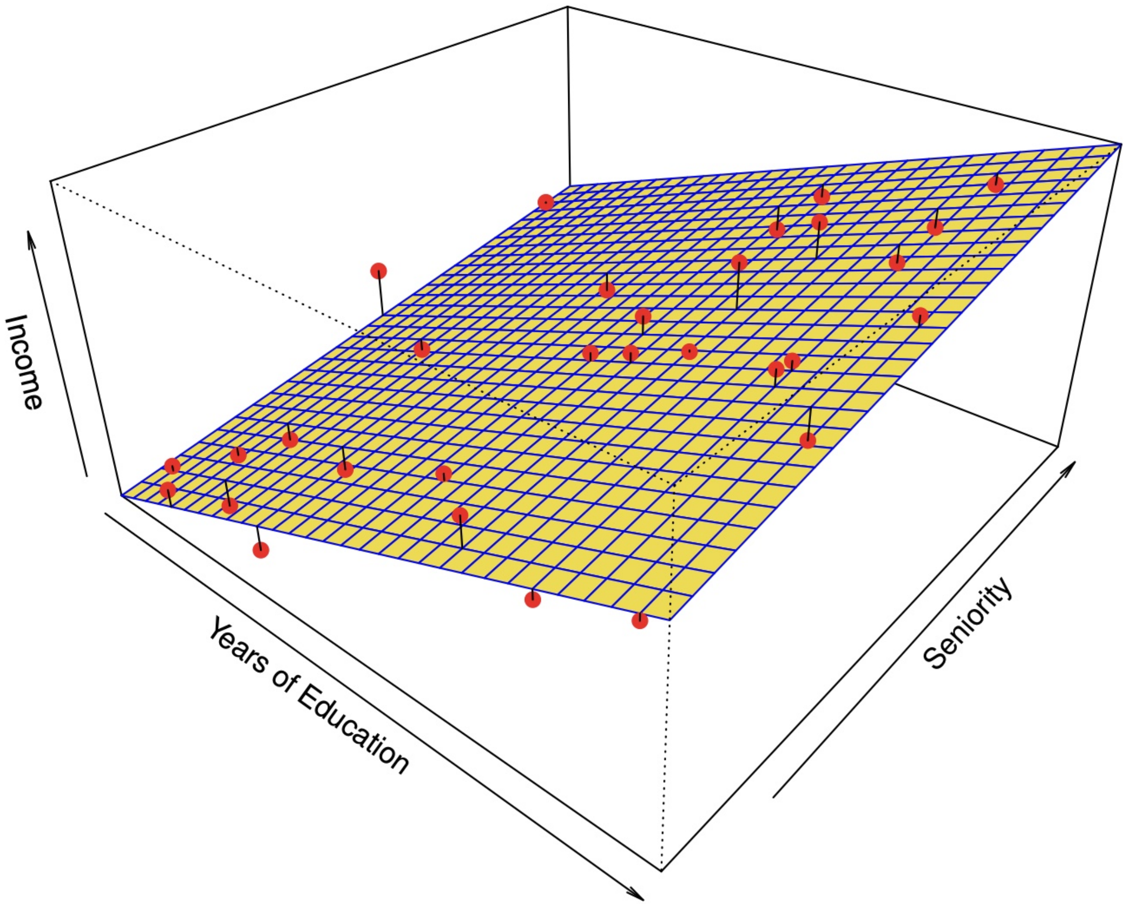
- Multivariate linear regression
  - Model evaluation
  - Omitted variable bias
  - Multicollinearity – correlated independent variable
- Hypothesis testing
  - Testing multiple parameters – T test vs. F test
- Variable transformations – interpreting results
  - Affine
  - Polynomial
  - Logarithmic
  - Dummy variables

Multivariate Regression

Simple Regression



Multivariate Regression



Multivariate Regression

$y = \beta_1 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

How do we interpret β_1, β_2?

- $y = 10 + 3x_1 + 4x_2, x_1 = 5, x_2 = 3$

- $y = 10 + 18 + 20 = 48$

- 1 unit increase in x_1 led to a β_1 increase in y (just like bivariate regression)
- But what about x_2? It did not change. So this change is only true holding x_2 constant
- We can hold x_2 constant to see how y changes as x_1 changes at that level of x_2

Evaluating the Model: Adjusted R^2

- Recall we can use $\mathrm{R}^2 = 1 - \mathrm{SSR}/\mathrm{TSS}$
- When we add a new independent variable, TSS does not change. $\mathrm{TSS} = (u - \mathrm{mean}(y))^2$
- However, the new variable will always cause SSR, $(y - \hat{y})^2$ to decrease. Therefore, R^2 will always decrease, which makes adding more variables ostensibly better
- Adjusted R^2 adds a disincentive (penalty) for adding new variables:

$$\mathrm{Adj\ R}^2 = 1 - \frac{(n-1)}{n-k-1} \frac{\mathrm{SSR}}{TSS}$$

Omitted variable bias

- If we do not use multiple regression, we may get biased estimate of the variable we do include
- "The bias results in the model attributing the effect of the missing variables to the estimated effects of the included variable."
- In other words, there are two variables that determine y, but our model only knows about one.
- The model we estimate with one variable accounts for the full effect of y, when we know the effect should be split between the two variables

Omitted variable bias

- When will there be no omitted variable bias effect?
  1. The second variable has no effect on y. Therefore, there is no extra effect to go into the first variable
  2. x_1 and x_2 are completely unrelated. Even though x_2 has an effect on y, x_1 lacks that information

$$\hat{\beta}_1 = \frac{\hat{\text{Cov}}(X, Y)}{\hat{Var}(X)}$$

$$\begin{aligned}
\hat{\text{Cov}}(\text{educ}, \text{wages}) &= \hat{\text{Cov}}(\text{educ}, \beta_1 \text{educ} + \beta_2 \text{exp} + \epsilon) \\
&= \beta_1 \hat{\text{Var}}(\text{educ}) + \beta_2 \hat{\text{Cov}}(\text{educ}, \text{exp}) + \hat{\text{Cov}}(\text{educ}, \epsilon) \\
&= \beta_1 \hat{\text{Var}}(\text{educ}) + \beta_2 \hat{\text{Cov}}(\text{educ}, \text{exp})
\end{aligned}$$

$$\text{Omitted variable bias: } \hat{\beta}_1 = \beta_1 + \beta_2 \frac{\hat{\text{Cov}(\text{educ}, \text{exp})}}{\hat{\text{Var}}(\text{educ})}$$

Calculating the bias effect

1. Population model (true relationship): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \nu$
2. Our model: $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \upsilon$
3. Auxiliary model: $x_2 = \delta_0 + \delta_1 x_1 + \epsilon$

- In the simple case of one regression and one omitted variable, estimating equation (2) by OLS will yield:

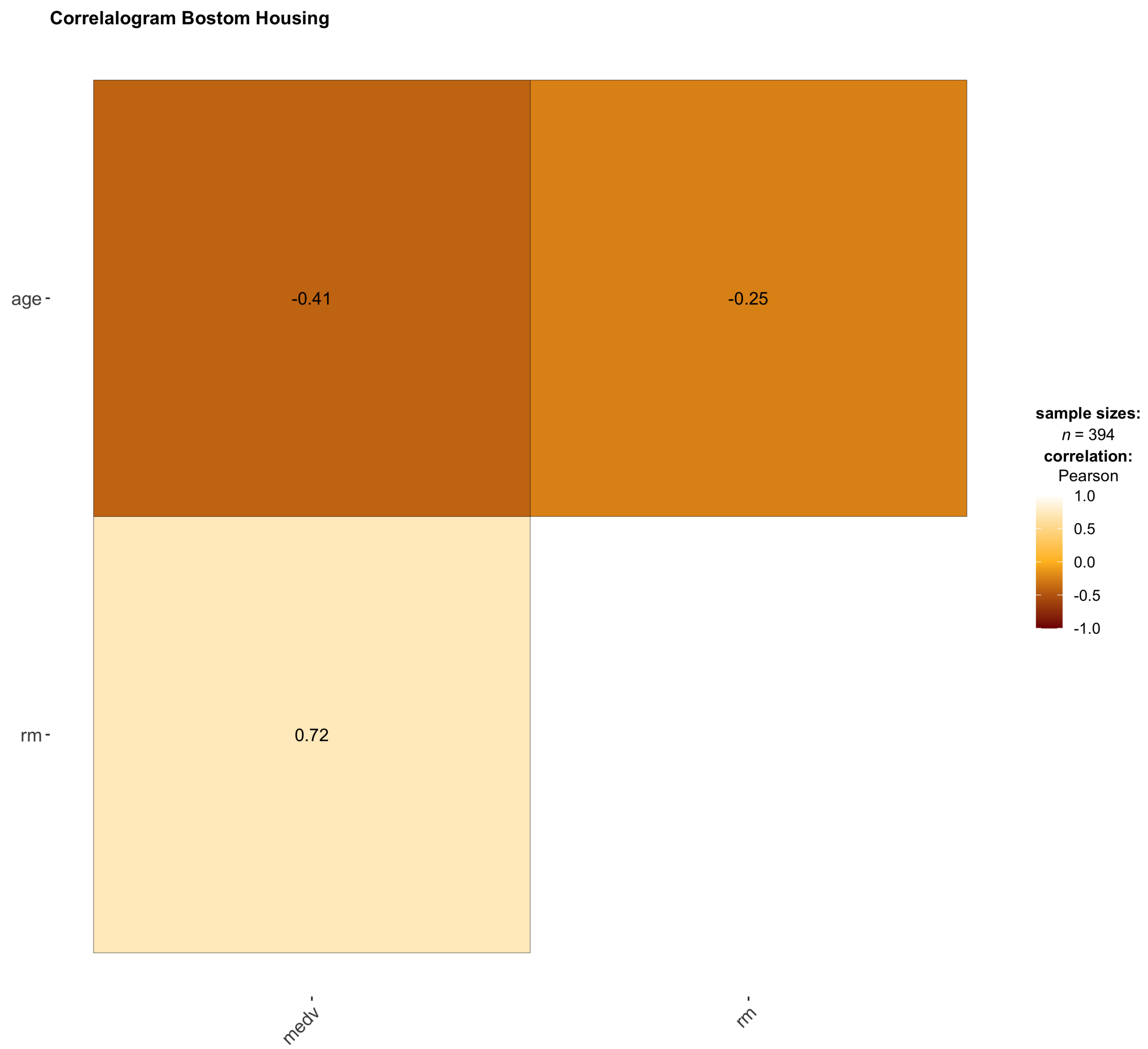
$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \beta_2 \delta$$

Equivalently, the bias is: $\mathrm{E}(\hat{\beta}_1) - \beta_1 = \beta_2 \delta$

	A and B are positively correlated	A and B are negatively correlated
B is positively correlated with y	Positive bias	Negative bias
B is negatively correlated with y	Negative bias	Positive bias

Example: Bostom Housing Data

variable	description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NO	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full value property tax rate per $10,000
PTRATIO	pupil teacher ratio by town
B	1000(Bk 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's


Correlalogram Bostom Housing
age
rm
-0.41
-0.25
0.72
medv
rm
sample sizes:
n = 394
correlation:
Pearson
1.0
0.5
0.0
-0.5
-1.0

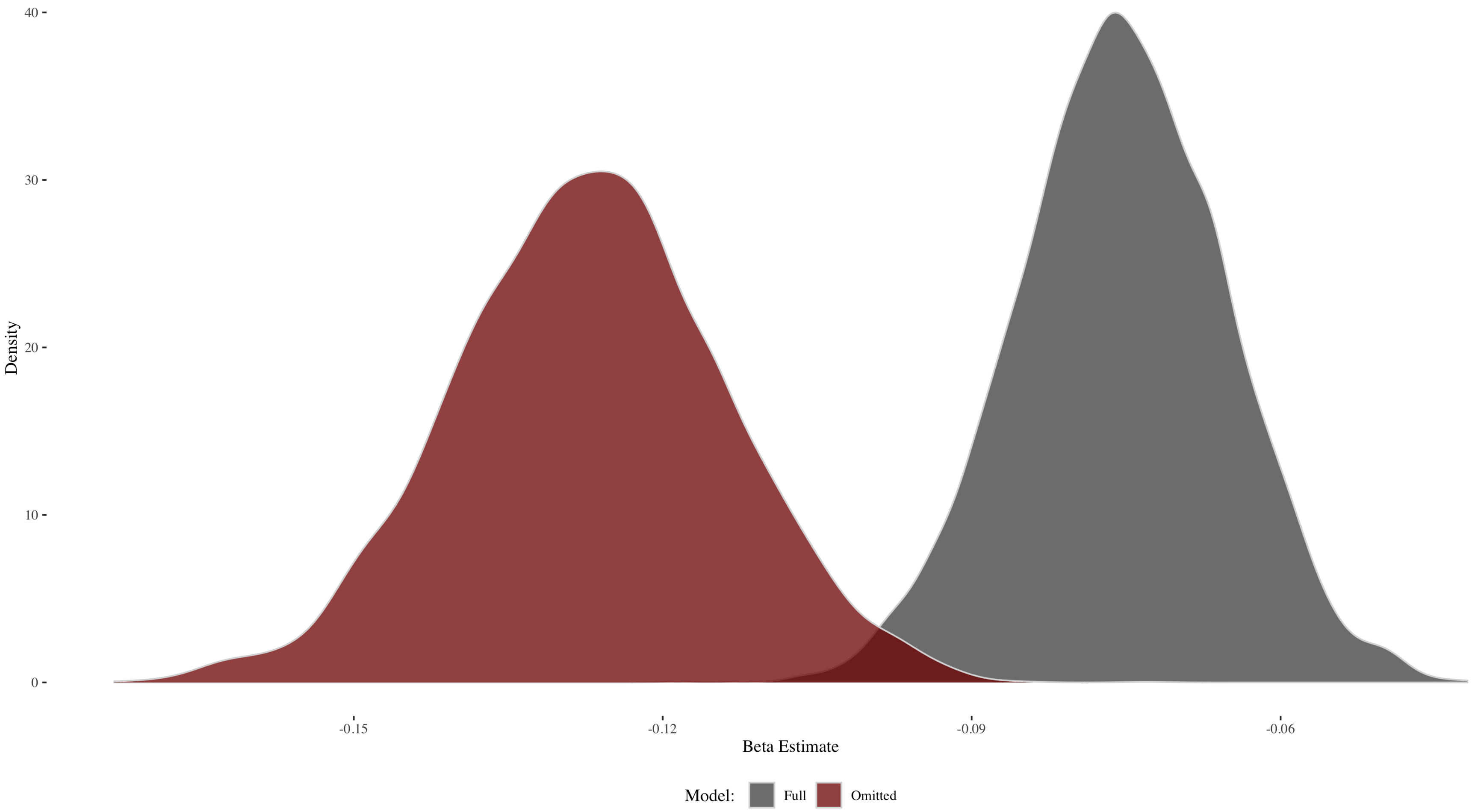
2,000 Regressions

- Take a random sample of 90% people out of the 506 that are in the Boston Housing data set
- Our model will be $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $\beta_1 = \mathrm{age}$ and $\beta_2 = \mathrm{rm}$
- Estimate β_1 using OLS (NOT controlling for rm) with the sample
- Estimate β_1 using OLS, controlling for rm with the same sample
- Repeat 2,000 times

Our data:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	NA	36.2
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

β_1 is underestimated when β_2 is ommitted



Multicollinearity

Multicollinearity

- Multivariate linear models cannot handle perfect multicollinearity
- Example: we have two variables: x_1 and $x_2 = 3 \times x_1$
- Fit model to predict y with x_1 and x_2:
    - $y = \beta_0 + \beta_1 x_1 + \mathrm{NA}$, where NA stands for not a value

- We can think of this as β_1 containing the entire effect for both x_1 and x_2. After all, these variables are the same.
- Including highly correlated variables in our model will not produce biased estimates, but it will harm our precision.

Baseball example

- Use home runs, batting average, and RBI to predict salary
- Variables are defined as follows:
  - $\text{salary} = \text{homeruns} \times 10,000 + \epsilon$
  - $\text{BA} = \text{homeruns} + 270 + \epsilon$
  - $\text{RBI} = \text{homeruns} \times 3 + \epsilon$
  - Example: $\text{homeruns} = 30, \text{BA} = 300, \text{RBI} = 90, \text{salary} = 300,000$
- Fit a model for each variable individually:
  - $\text{salary} = 9,934.27 \times \text{HR}$
  - $\text{salary} = 1,002.95 \times \text{BA}$
  - $\text{salary} = 3,291.02 \times \text{RBI}$
- Fit a model with all three: $\text{salary} = 9,226.169 \times \text{HR} + 225.884 \times \text{RBI} + 2.982 \times \text{BA}$
- What is this model saying? Why not: $\text{salary} = 9,934.27 \times \text{HR} + 3,291.02 \times \text{RBI} + 1,002.95 \times \text{BA}$

Helpful resource

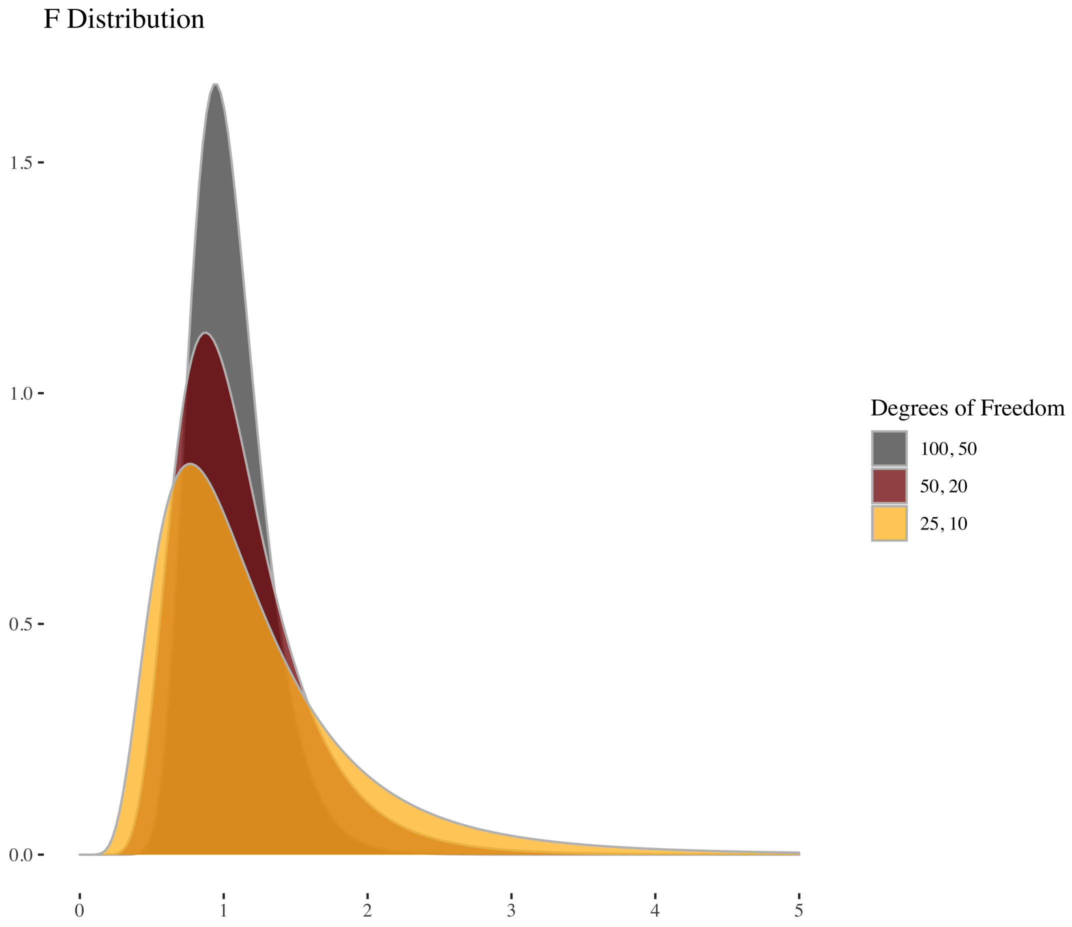
- Omitted variable bias and multicollinearity discussion: https://are.berkeley.edu/courses/EEP118/current/handouts/OVB%20versus%20Multicollinearity_eep118_sp15.pdf

Situation	Action
z is correlated with both x and y	Probably best to include z but be wary of multicollinearity
z is correlated with x but not y	Do not include z — no benefit
z is correlated with y but not x	Include z — new explanatory power
z is correlated with neither x nor y	Should not be much effect when including, but could affect hypothesis testing — no real benefit

Hypothesis testing

Hypothesis testing

- The previous example demonstrates why we must use F test to test all hypothesis simultaneously rather than a T test
- Recall the T test for $H_0 \rightarrow \hat{\beta}_1 = \theta$: $\frac{(\hat{\beta}_1 - \theta)}{\mathrm{SE}(\hat{\beta}_1)}$
- The above statistic is t-distributed under the null hypothesis, so we can see how likely it would be to get the above value from a t distribution
- If we are testing multiple hypotheses, we can apply the same logic as long as we know how that statistic is distributed. In this new test, our statistic belongs to the F distribution



Back to baseball

- To perform an F test, we compare a model with restrictions to a model without restrictions and see if there is a significant difference. Think of restrictions as features not included in the model
- $\text{salary} = \text{years} + \text{gmsYear} + \text{HR} + \text{RBI} + \text{BA}$
- If HR, RBI, BA all have no effect on salary, then the model $\text{salary} = \text{years} + \text{gmsYears}$ should perform just as well
- How do we measure *performance*? Sum of squared residuals (SSR)!
- Test statistics: $\frac{\text{SSR}_\text{r} - \text{SSR}_\text{ur}/q}{\text{SSR}_\text{ur}/(n-k-1)}$
- The above fraction is the ratio of two chi squared variables divided by their degrees of freedom, which makes this F-distributed
- Remember adding variables can only improve the model, so the F statistic will always be positive

Types of variables and transformations

Affine

- Affine transformations are transformations that do not affect the fit of the model. The most common example is scaling transformations
- Example:
  - $\text{weight}(\text{lbs}) = 5 + 2.4 \times \text{height}(\text{inches})$
  - $\text{weight}(\text{lbs}) = 5 + 0.094 \times \text{height}(\text{mm})$
- This is why scaling variables is not necessary for linear regression, but knowing the scale of your variables is important for interpretation

Polynomial

- Linear regression can still be used to fit data with a non-linear distribution
- The model is linear in parameters, not necessarily variables
- i.e. we must have β_1, β_2, β_3, but we can utilize x_1^2 or x_2/x_3
- We might leverage the above to generate a curved regression line, providing a better fit in some cases
- How do we now interpret the coefficients?

$$\hat{\text{wage}} = 3.12 + .447\text{exp} - 0.007\text{exp}^2$$

- The big difference is the effect of an increase in experience on wage now depends on the level of experience

Logarithmic

Recall that the natural logarithm is the inverse of the exponential function, so $\ln(e^x) = x$, and:

$$\ln(1) = 0$$

$$\ln(0) = -\infty$$

$$\ln(ax) = \ln(a) + \ln(x)$$

$$\ln(x^a) = a\ln(x)$$

$$\ln(\tfrac{1}{x}) = -\ln(x)$$

$$\ln(\tfrac{x}{a}) = \ln(x) - \ln(a)$$

$$\frac{d\ln(x)}{dx} = \frac{1}{x}$$

Interpreting log variables

- $\beta_0 = 5$, $\beta_1 = 0.2$
- Level-log: $y = 5 + 0.2\ln(x)$
    - 1% change in $x = \beta_1 / 100$ change in y

- Log-level: $\ln(y) = 5 + 0.2(x)$
    - 1 unit change in $x = \beta_1 \times 100\%$ change in y

- Log-log: $\ln(y) = 5 + 0.2\ln(x)$
    - 1% change in $x = \beta_1\%$ change in y

Dummy variables

- Dummy variables is how categorical variables can be mathematically represented
- They represent groups or place continuous variables into bins
- What is this regression telling us?
  - $\text{nbaSalary} = 5 \times \text{PPG} + 10.5 \times \text{guard} + 9.6 \times \text{forward} + 10.8 \times \text{center}$

- Do we need dummy variables for $guard, forward, center$?
- How would the regression change if we only used 2 out of 3?
- $\text{nbaSalary} = 10.5 + 5 \times \text{PPG} - 0.9 \times \text{forward} + 0.3 \times \text{center}$