Multiple Linear Regression

Special Predictors

Prof. Maria Tackett

02.10.20

STA 210

[Click for PDF of slides]

Announcements

- Lab 04 **due tomorrow at 11:59p**
  - pdf of instructions in GitHub repo
- HW 02 **due Wed, Feb 12 at 11:59p**
  - pdf of instructions in GitHub repo
- Reading for today & Wednesday

Today's agenda

- Inference for regression coefficients
- Prediction
- Quick math details
- Special predictors

R packages

House prices in Levittown (sec. 1.4)

- Public data on the sales of 85 homes in Levittown, NY from June 2010 to May 2011
- Levittown was built right after WWI and was the first planned suburban community built using mass production techniques

Questions:

- What is the relationship between the characteristics of a house in Levittown and its sale price?
- Given its characteristics, what is the expected sale price of a house in Levittown?

Data

```
glimpse(homes)
```

```
## Observations: 85
## Variables: 7
## $ bedrooms     <dbl> 4, 4, 4, 5, 5, 4, 4, 4, 4, 3, 4, 4, 3, 4, 3, 5, 4, .
## $ bathrooms    <dbl> 1.0, 2.0, 2.0, 2.0, 2.5, 2.0, 1.0, 1.0, 1.5, 2.0, 2.
## $ living_area  <dbl> 1380, 1761, 1564, 2904, 1942, 1830, 1585, 941, 1481.
## $ lot_size     <dbl> 6000, 7400, 6000, 9898, 7788, 6000, 6000, 6800, 600.
## $ year_built   <dbl> 1948, 1951, 1948, 1949, 1948, 1948, 1948, 1951, 194.
## $ property_tax <dbl> 8360, 5754, 8982, 11664, 8120, 8197, 6223, 2448, 90.
## $ sale_price   <dbl> 350000, 360000, 350000, 375000, 370000, 335000, 295.
```

Variables

Predictors

- `bedrooms`: Number of bedrooms
- `bathrooms`: Number of bathrooms
- `living_area`: Total living area of the house (in square feet)
- `lot_size`: Total area of the lot (in square feet)
- `year_built`: Year the house was built
- `property_tax`: Annual property taxes (in U.S. dollars)

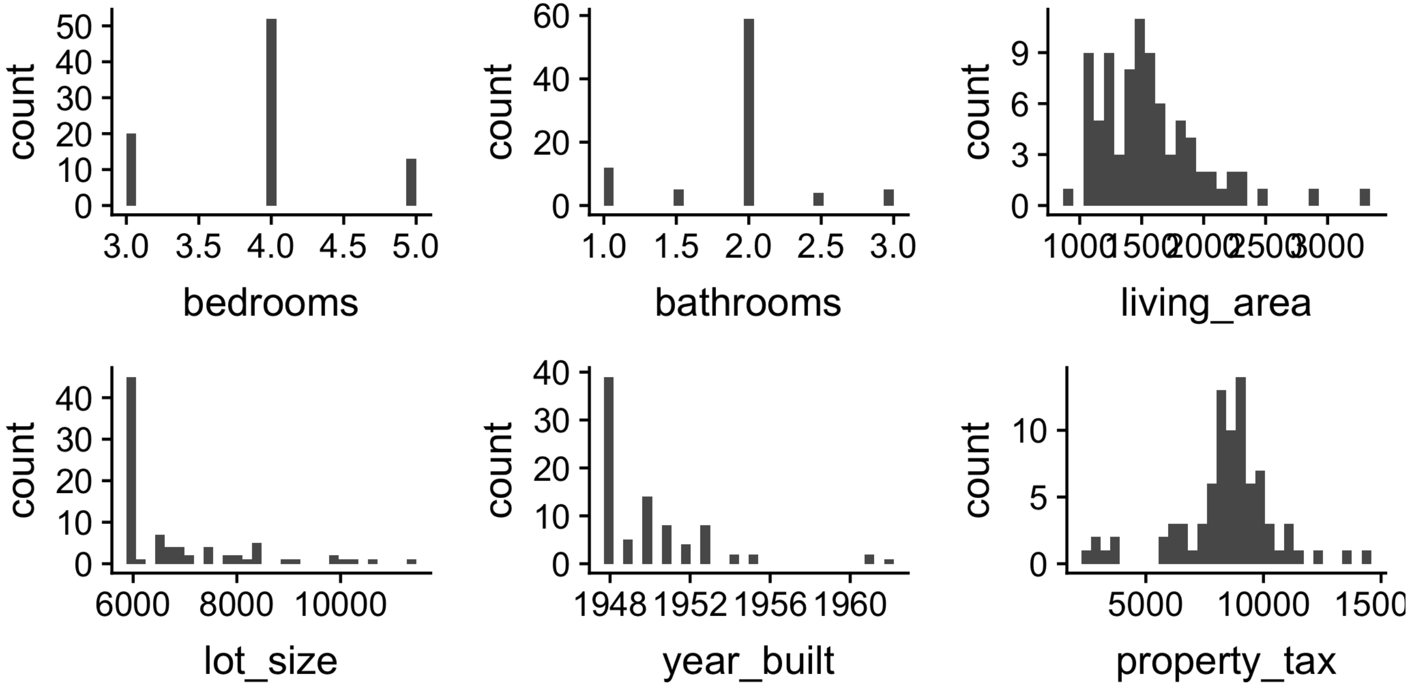
Response

- `sale_price`: Sales price (in U.S. dollars)

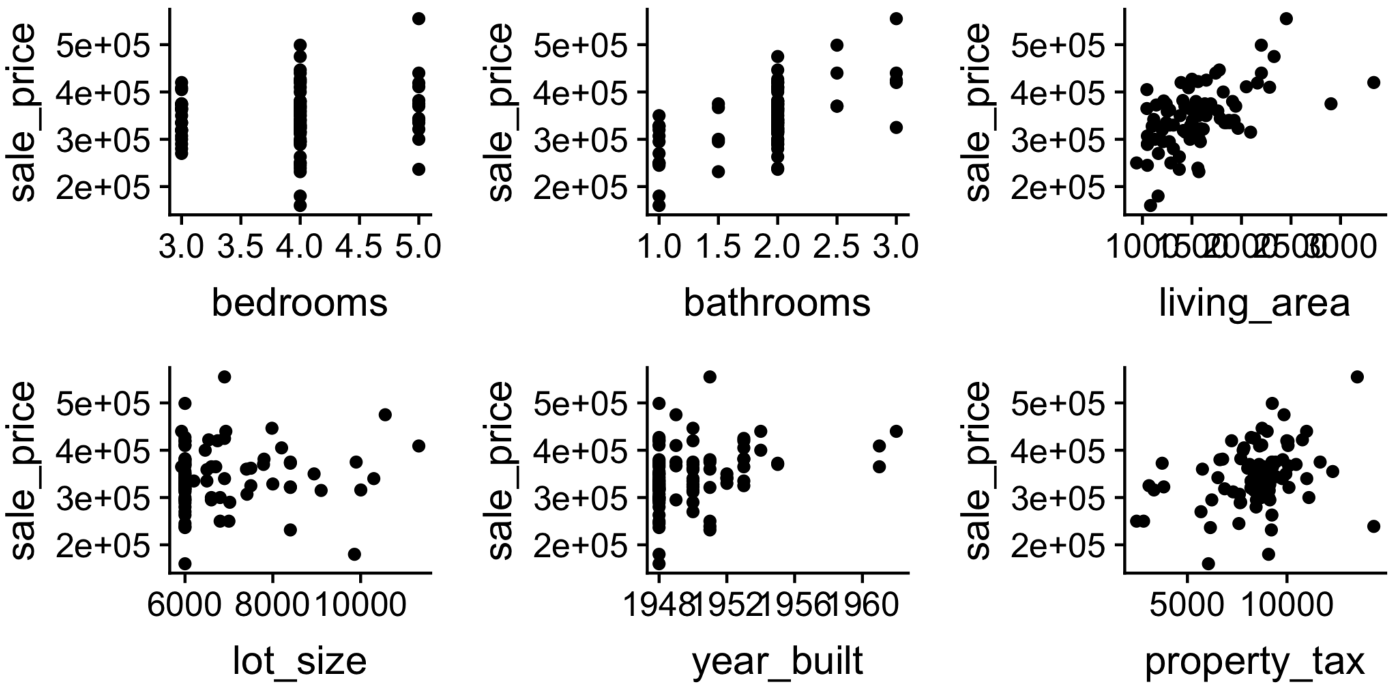
STA 210

EDA: Response variable

EDA: Predictor variables



EDA: Response vs. Predictors



Regression Output

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-7148818.957	3820093.694	-1.871	0.065	-14754041.291	456403.376
bedrooms	-12291.011	9346.727	-1.315	0.192	-30898.915	6316.893
bathrooms	51699.236	13094.170	3.948	0.000	25630.746	77767.726
living_area	65.903	15.979	4.124	0.000	34.091	97.715
lot_size	-0.897	4.194	-0.214	0.831	-9.247	7.453
year_built	3760.898	1962.504	1.916	0.059	-146.148	7667.944
property_tax	1.476	2.832	0.521	0.604	-4.163	7.115

Interpreting $\hat{\beta}_j$

- An estimated coefficient $\hat{\beta}_j$ is the expected change in y to change when x_j increases by one unit **holding the values of all other predictor variables constant**.

- *Example:* The estimated coefficient for `living_area` is 65.90. This means for each additional square foot of living area, we expect the sale price of a house in Levittown, NY to increase by $65.90, on average, holding all other predictor variables constant.

Hypothesis Tests for $\hat{\beta}_j$

- We want to test whether a particular coefficient has a value of 0 in the population, given all other variables in the model:

$$
\begin{aligned}
H_0 &: \beta_j = 0 \\
H_a &: \beta_j \neq 0
\end{aligned}
$$

- The test statistic reported in R is the following:

$$\text{test statistic} = t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

- Calculate the p-value using the t distribution with $n - p - 1$ degrees of freedom, where p is the number of terms in the model (not including the intercept).

STA 210

Confidence Interval for β_j

The C confidence interval for β_j

$$\hat{\beta}_j \pm t^* SE(\hat{\beta}_j)$$

where t^* follows a t distribution with with $(n - p - 1)$ degrees of freedom

- **General Interpretation**: We are C confident that the interval LB to UB contains the population coefficient of x_j. Therefore, for every one unit increase in x_j, we expect y to change by LB to UB units, holding all else constant.

Confidence interval for `living_area`

Interpret the 95% confidence interval for the coefficient of `living_area`.

Caution: Large sample sizes

If the sample size is large enough, the test will likely result in rejecting $H_0 : \beta_j = 0$ even x_j has a very small effect on y

- Consider the practical significance of the result not just the statistical significance
- Use the confidence interval to draw conclusions instead of p-values

Caution: Small sample sizes

If the sample size is small, there may not be enough evidence to reject $H_0 : \beta_j = 0$

- When you fail to reject the null hypothesis, **DON'T** immediately conclude that the variable has no association with the response.
- There may be a linear association that is just not strong enough to detect given your data, or there may be a non-linear association.

STA 210

Prediction

- We calculate predictions the same as with simple linear regression
- **Example:** What is the predicted sale price for a house in Levittown, NY with 3 bedrooms, 1 bathroom, 1050 square feet of living area, 6000 square foot lot size, built in 1948 with $6306 in property taxes?

```
-7148818.957 - 12291.011 * 3 + 51699.236 * 1 +
  65.903 * 1050 - 0.897 * 6000 + 3760.898 * 1948 + 1.476 * 6306
```

```
## [1] 265360.4
```

The predicted sale price for a house in Levittown, NY with 3 bedrooms, 1 bathroom, 1050 square feet of living area, 6000 square foot lot size, built in 1948 with $6306 in property taxes is **$265,360**.

STA 210

Intervals for predictions

- Just like with simple linear regression, we can use the **`predict`** function in R to calculate the appropriate intervals for our predicted values

```
x0 <- data.frame(bedrooms = 3, bathrooms = 1, living_area = 1050,
                 lot_size = 6000, year_built = 1948,
                 property_tax = 6306)
predict(price_model, x0, interval = "prediction")
```

- Go to http://bit.ly/sta210-sp20-pred and use the model to answer the questions
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss with 1 - 2 people around you, but **each person** response.

03:00

STA 210

Cautions

- **Do not extrapolate!** Because there are multiple explanatory variables, you can extrapolation in many ways
- The multiple regression model only shows **association, not causality**
  - To show causality, you must have a carefully designed experiment or carefully account for confounding variables in an observational study

Math details

Regression Model

- The multiple linear regression model assumes

$$y|x_1, x_2, \dots, x_p \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \sigma^2)$$

- For a given observation $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)$, we can rewrite the previous statement as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

Estimating σ^2

- For a given observation $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ the residual is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip})$$

- The estimated value of the regression variance , σ^2, is

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\sum_{i=1}^n e_i^2}{n-p-1}$$

STA 210

Estimating Coefficients

- One way to estimate the coefficients is by taking partial derivatives of the formula

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots + \beta_p x_{ip})]^2$$

- This produces messy formulas, so instead we can use matrix notation for multiple linear regression and estimate the coefficients using rules from linear algebra.
  - For more details, see Section 1.2 of the textbook and the supplemental notes [Matrix Notation for Multiple Linear Regression]
  - **Note:** You are <u>not</u> required to know matrix notation for MLR in this class

STA 210

Special Predictors

Interpreting the Intercept

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

- Interpret the intercept.
- Is this interpretation meaningful? Why or why not?

Mean-Centered Variables

- To have a meaningful interpretation of the intercept, use **mean-centered** predictor variables in the model (quantitative predictors only)

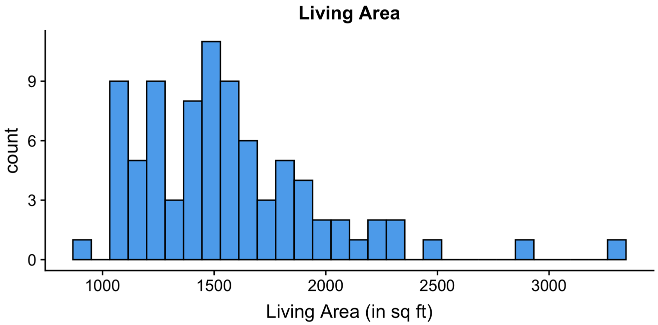
- A mean-centered variable is calculated by subtracting the mean from each value of the variable, i.e.

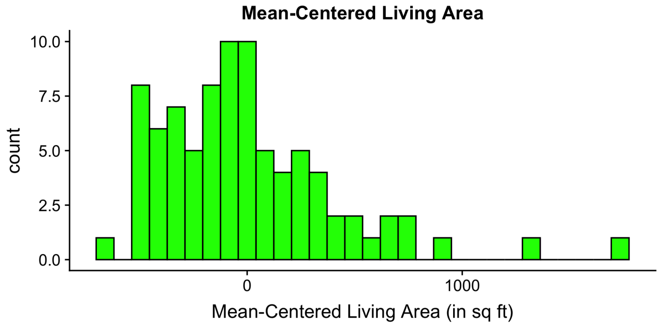
$$x_{ip} - \bar{x}_{.p}$$

- Now the intercept is interpreted as the expected value of the response at the mean value of all quantitative predictors

Salary: Mean-Centered Variables

```
homes <- homes %>%
  mutate(bedroomsCent = bedrooms - mean(bedrooms),
         bathroomsCent = bathrooms - mean(bathrooms),
         living_areaCent = living_area - mean(living_area),
         lot_sizeCent = lot_size - mean(lot_size),
         year_builtCent = year_built - mean(year_built),
         property_taxCent = property_tax - mean(property_tax))
```





STA 210

In-class exercise

Below is the original model:

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

- Go to http://bit.ly/sta210-sp20-mean-center and describe how the model would change if mean-

03:00

STA 210

How model changes with mean-centered variables

Indicator (dummy) variables

- Suppose there is a categorical variable with k levels (categories)
- Make k indicator variables (also known as dummy variables)
- Use $k-1$ of the indicator variables in the model
  - Can't uniquely estimate all k variables at once if the intercept is in the model
- Level that doesn't have a variable in the model is called the **baseline**
- Coefficients interpreted as the change in the mean of the response over the baseline

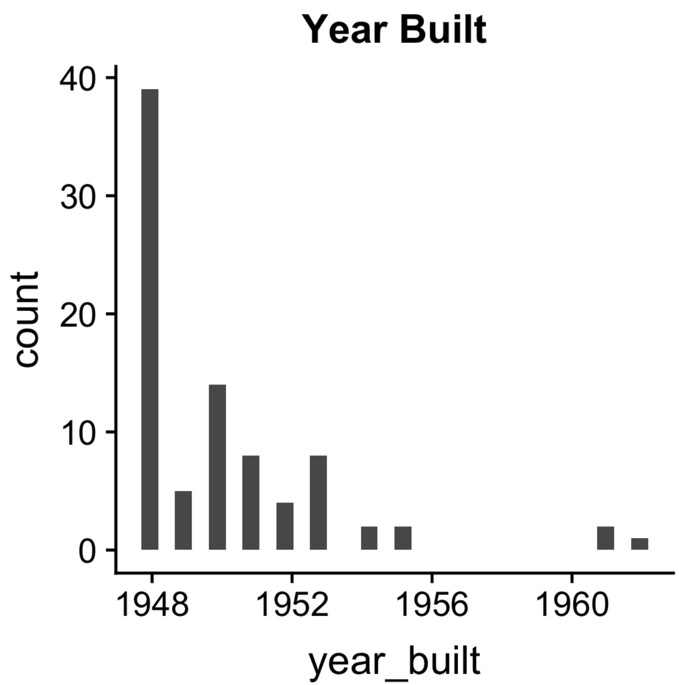
Indicator variables when $k > 2$

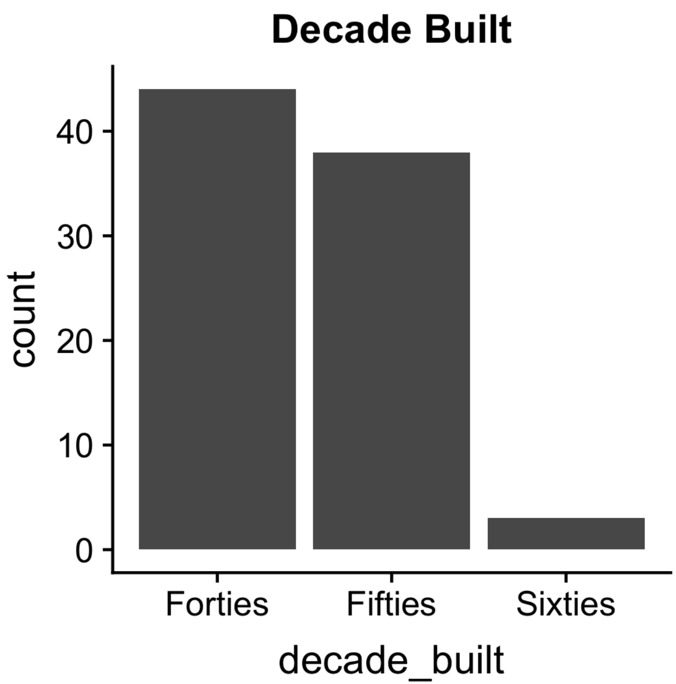
Suppose we create a new variable called **`decade_built`** that is the decade the house is built

```
homes <- homes %>%
  mutate(decade_built = case_when(
    year_built %in% c(1948, 1949) ~ "Forties",
    year_built %in% c(1961, 1962) ~ "Sixties",
    TRUE ~ "Fifties"
  ),
  decade_built = factor(decade_built,
                        levels = c("Forties", "Fifties", "Sixties'
  )
```

STA 210

`year_built` and `decade_built`





STA 210

Model with categorical predictor

- Let's fit the model with `decade_built` instead of `year_built`

term	estimate	std.error	statistic	p.value
(Intercept)	171943.215	49977.768	3.440	0.001
bedrooms	-12024.725	9611.054	-1.251	0.215
bathrooms	56179.551	12993.791	4.324	0.000
living_area	64.194	16.241	3.953	0.000
lot_size	-0.568	4.264	-0.133	0.894
property_tax	1.249	2.885	0.433	0.666
decade_builtFifties	10492.738	11325.820	0.926	0.357
decade_builtSixties	40300.518	30393.515	1.326	0.189

Interaction Terms

- **Case**: Relationship of the predictor variable with the response depends on the value of another predictor variable
  - This is an **interaction effect**
- Create a new interaction variable that is one predictor variable times the other in the interaction
- **Good Practice**: When including an interaction term, also *include the associated main effects* (each predictor variable on its own) even if their coefficients are not statistically significant

STA 210

Interaction: decase_built * bathrooms

```
price_model <- lm(sale_price ~ bedrooms + bathrooms + living_area
                   lot_size + property_tax + decade_built +
                   decade_built * bathrooms,
                 data = homes)
```

term	estimate	std.error	statistic	p.value
(Intercept)	171675.745	55663.287	3.084	0.003
bedrooms	-12580.988	9816.949	-1.282	0.204
bathrooms	56363.759	18081.988	3.117	0.003
living_area	64.319	16.478	3.903	0.000
lot_size	-0.357	4.345	-0.082	0.935
property_tax	1.312	2.929	0.448	0.655
decade_builtFifties	12931.382	46800.696	0.276	0.783
decade_builtSixties	-70531.549	267991.500	-0.263	0.793
bathrooms:decade_builtFifties	-1263.513	23798.721	-0.053	0.958
bathrooms:decade_builtSixties	50802.984	122521.380	0.415	0.680

STA 210