Multiple Linear Regression

Prof. Maria Tackett

02.05.20

STA 210

[Click for PDF of slides]

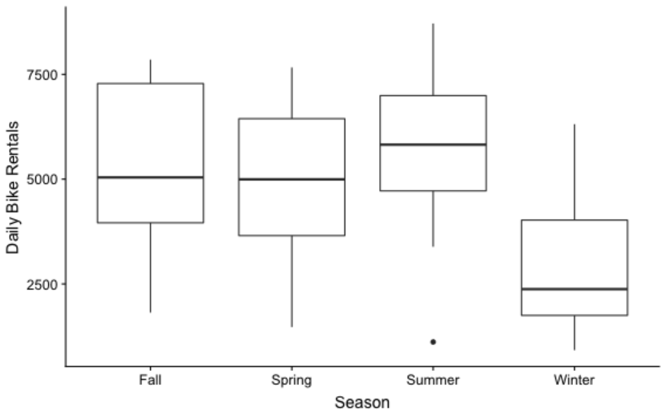
Announcements

- HW 02 due Wed, Feb 12 at 11:59p
- Reading for today
- Reading for Monday

Today's Agenda

- ANOVA
- Introducing multiple linear regression

ANOVA




```
## # A tibble: 4 x 4
##   season      n  mean    sd
##   <chr>  <int> <dbl> <dbl>
## 1 Fall      25 5180. 1848.
## 2 Spring    23 4924. 1889.
## 3 Summer    27 5739. 1662.
## 4 Winter    25 2779. 1465.
```

STA 210

ANOVA for Capital Bike Share

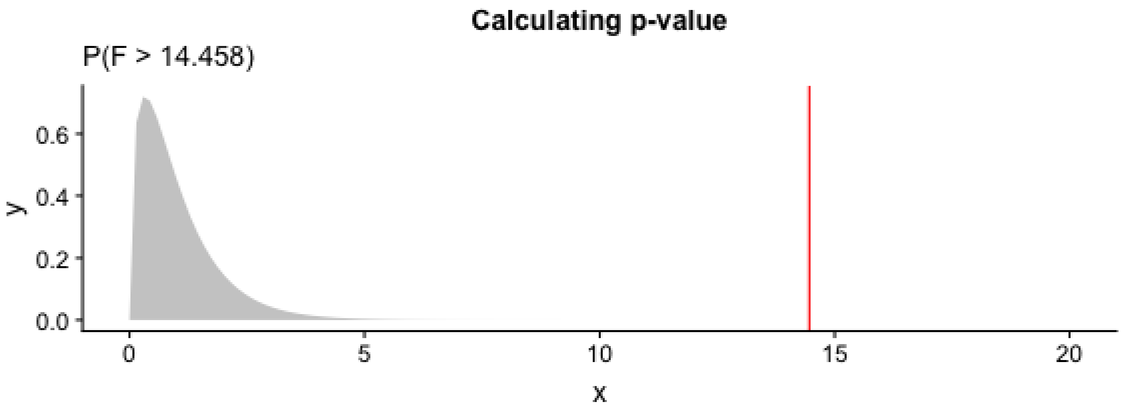
$$H_0 : \mu_W = \mu_{Sp} = \mu_{Su} = \mu_F$$
$$H_a : \text{at least 1 } \mu_i \text{ is not equal to the others}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	3	128202929	42734310	14.458	0
Residuals	96	283747246	2955700	NA	NA

STA 210

Calculate p-value

- Calculate the p-value using an F distribution with $K - 1$ and $n - K$ degrees of freedom.
  - We use F distribution since the test statistic is the ratio of two variances.
- In the Capital Bike Share example, the p-value is calculated from the F distribution with 3 and 96 degrees of freedom.



Assumptions for ANOVA

- **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$
- **Constant variance:** The population distribution for each group has a common variance, σ^2
- **Independence:** The observations are independent from one another
  - This applies to observation within and between groups
- We can typically check these assumptions in the exploratory data analysis

Robustness to Assumptions

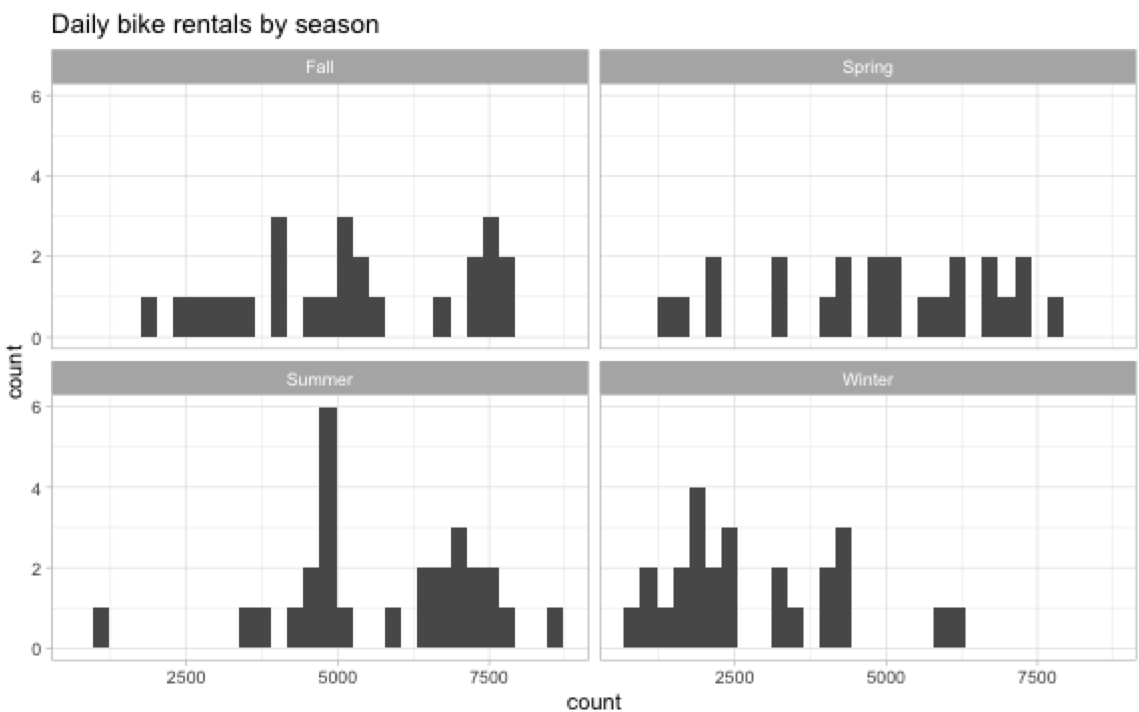
- **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$
  - ANOVA relatively robust to departures from Normality.
  - Concern when there are strongly skewed distributions with different sample sizes (especially if sample sizes are small, < 10 in each group)
- **Independence:** There is independence within and across groups
  - If this doesn't hold, should use methods that account for correlated errors

Robustness to Assumptions

- Constant variance: The population distribution for each group has a common variance, σ^2
  - Critical assumption, since the pooled (combined) variance is important for ANOVA
  - **General rule:** If the sample sizes within each group are approximately equal, the results of the F-test are valid if the largest variance is no more than 4 times the smallest variance (i.e. the largest standard deviation is no more than 2 times the smallest standard deviation)

Capital Bike Share: Normality

```r
ggplot(data = bikeshare, aes(x = count)) +
  geom_histogram() +
  facet_wrap(~season) +
  labs(title = "Daily bike rentals by season") +
  theme_light()
```



Capital Bike Share: Constant Variance

```
bikeshare %>%
  group_by(season) %>%
  summarise(sd = sd(count))
```

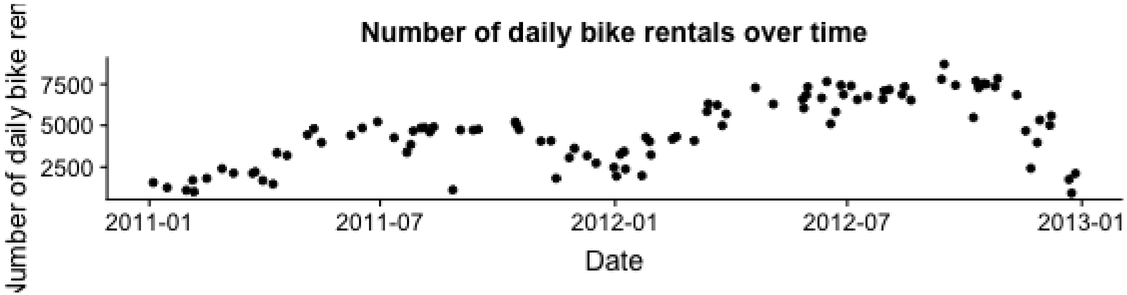
```
## # A tibble: 4 x 2
##   season    sd
##   <chr>  <dbl>
## 1 Fall   1848.
## 2 Spring 1889.
## 3 Summer 1662.
## 4 Winter 1465.
```

The largest variance 1889^2 is 1.663 times the smallest variance 1465^2, so the constant variance assumption is satisfied.

STA 210

Capital Bike Share: Independence

- Recall that the data is 100 randomly selected days in 2011 and 2012.
- Let's look at the counts in date order to see if a pattern still exists



Though the days were randomly selected, it still appears the independence assumption is violated.

- Additional methods may be required to fully examine this data.

Why not just use the model output?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5180.200	343.843	15.066	0.000	4497.677	5862.723
seasonSpring	-256.591	496.726	-0.517	0.607	-1242.585	729.402
seasonSummer	558.911	477.178	1.171	0.244	-388.279	1506.101
seasonWinter	-2400.760	486.267	-4.937	0.000	-3365.993	-1435.527

- The model coefficients and associated hypothesis test / confidence interval are interpreted in relation to the baseline level
    - The coefficients, test statistics, confidence intervals, and p-values all change if the baseline category changes (more on this later!)
- An ANOVA test gives indication if <u>any</u> category has a significantly different mean regardless of the baseline
    - The sum of squares, mean squares, test statistic, and p-value stay the same even if the baseline changes

Muddiest Point

- Go to http://bit.ly/sta210-sp20-anova-q and write one question you have about simple linear regression or ANOVA.
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

04:00

STA 210

Multiple Linear Regression

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?

STA 210

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

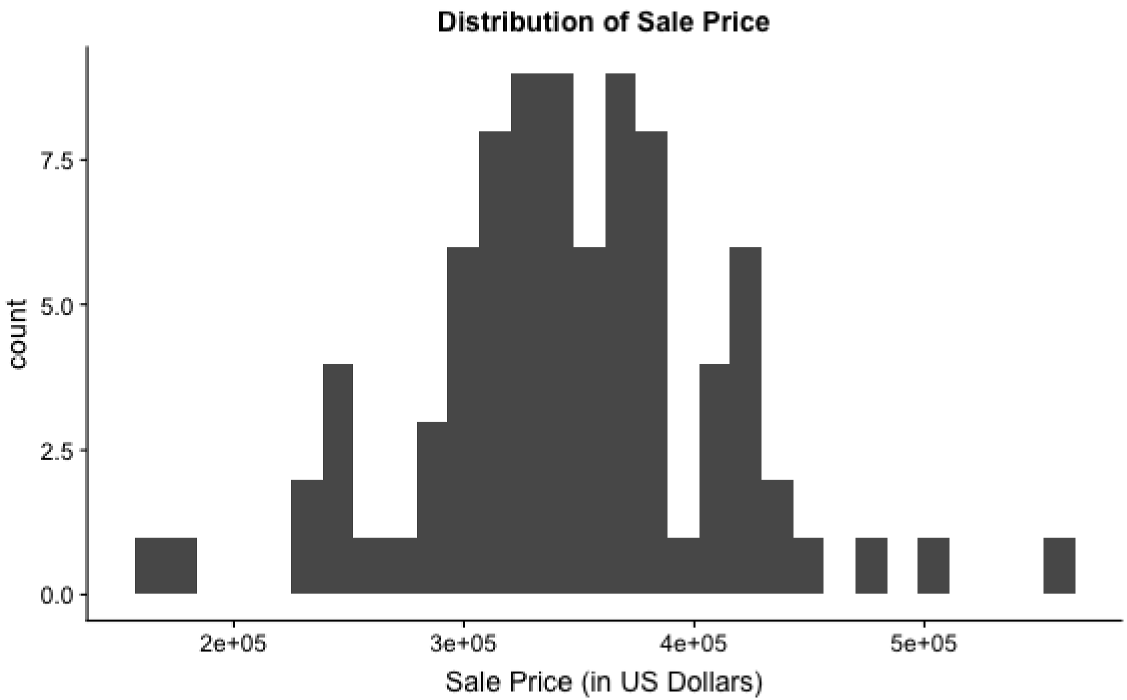
Predictors

- **`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)

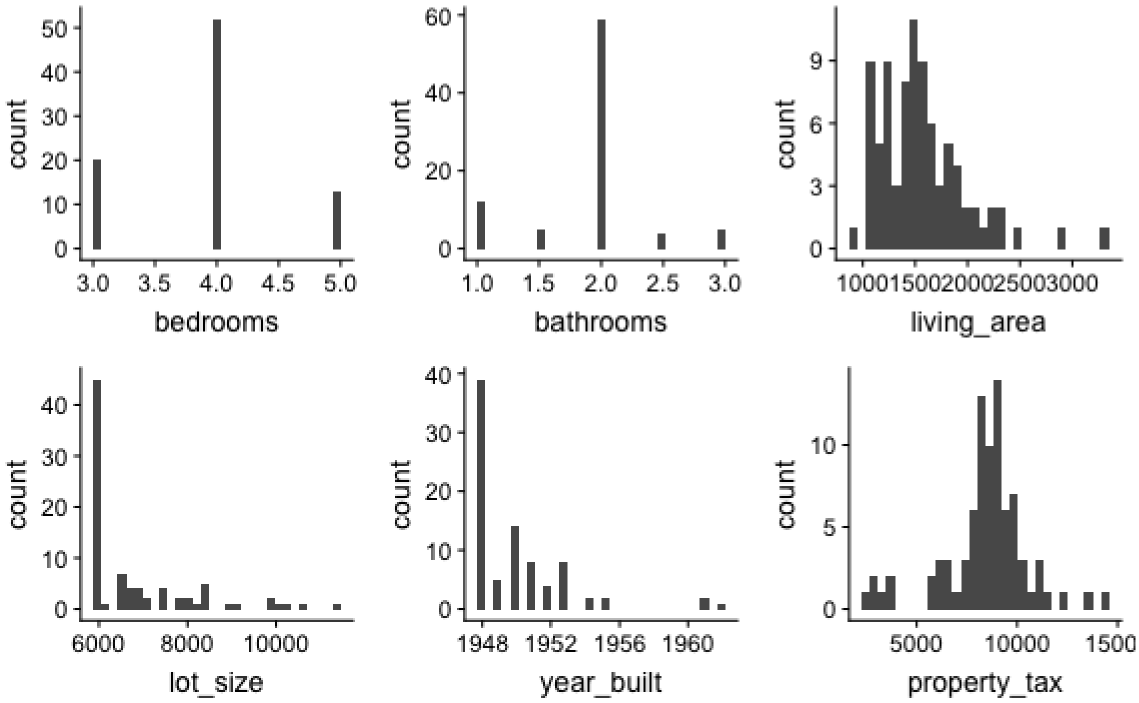
Response

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

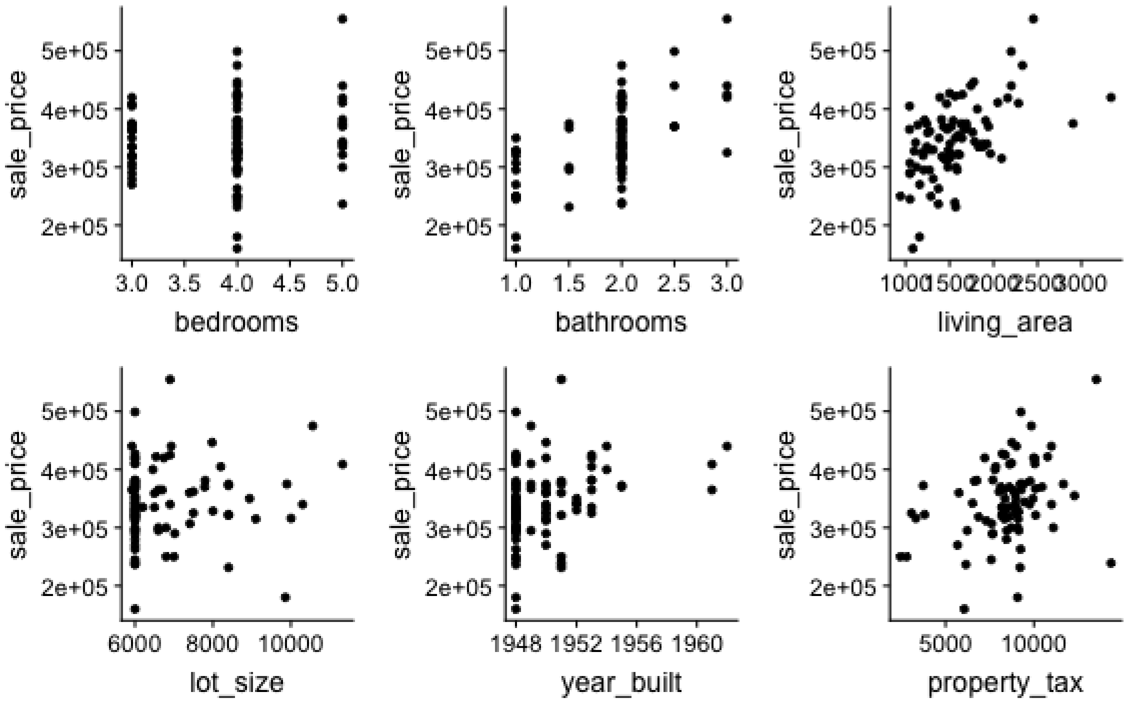
EDA: Response variable



EDA: Predictor variables



EDA: Response vs. Predictors



What is a disadvantage to fitting a separate model for each predictor variable?

Multiple Regression Model

We will calculate a multiple linear regression model with the following form:

$$
\begin{aligned}
\text{sale\_price} = \beta_0 &+ \beta_1 \text{bedrooms} + \beta_2 \text{bathrooms} + \beta_3 \text{living\_area} \\
&+ \beta_4 \text{lot\_size} + \beta_5 \text{year\_built} + \beta_6 \text{property\_tax}
\end{aligned}
$$

Similar to simple linear regression, this model assumes that at each combination of the predictor variables, the values **`sale_price`** follow a Normal distribution

STA 210

Regression Model

- Recall: The simple linear regression model assumes

$$y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

- Similarly: The multiple linear regression model assumes

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

For a given observation $(x_{i1}, x_{i2} \ldots, x_{iP}, y_i)$

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

STA 210

Regression Model

- At any combination of $x's$, the true mean value of y is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- We will use multiple linear regression to estimate the mean y for any combination of $x's$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Regression Output

```
price_model <- lm(sale_price ~ bedrooms + bathrooms + living_area
                  data = homes)

tidy(price_model, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

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

STA 210

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.

STA 210

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).

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.

STA 210

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

STA 210

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.

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**.

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

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