# STAT 100: Week 11

Ricky’s Section

# Introductions and Attendance

**Introduction**: Name

**Question of the Week**: Thanksgiving is coming up! What's something you're grateful for?

# Important Reminders

## Anonymous Feedback

https://docs.google.com/forms/d/e/1FAIpQLSfKv_FGvsooqm-IvtxKx3Vf6bBzSJE2jamK1gklAzL6NkXE8w/viewform

# Upcoming Events...

- **Last section for STAT 100**: Next week 🙁
- **Last day of class for STAT 100**: Wednesday, 12/4
- **Final exam for STAT 100**: Wednesday, 12/11
- **ggparty**: Thursday, 12/5 from 11:30 AM to 1:30 PM in Science Center 316
  - RSVP: https://forms.gle/yKw6Wziiy6Lj5guu6

# Workshop (Review Session)

- **Saturday, 11/16 from 4-5 PM in Science Center 309!**
- We'll be practicing and reviewing linear regression
- We recognize flipped classroom can be difficult, so these are here to help you
  - Along with OH, 1-on-1 OH, Slack, etc.

# Teaching Fellow (TF) Applications!

- https://forms.gle/PzgYrGFNasLejap37
  - The deadline is Monday, 11/18
- Let me know if you have any questions!
- I definitely encourage you all to apply! 🙂

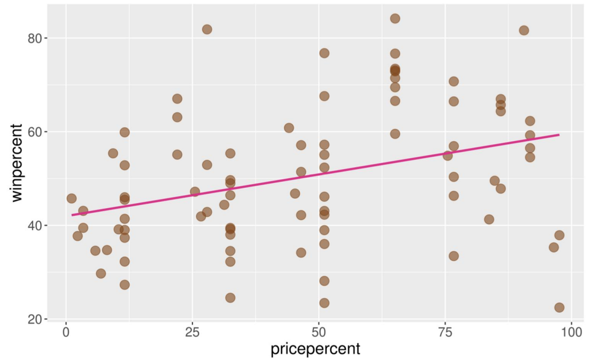
# Content Review: Week 10

# Let's (Quickly) Recap Linear Regression

- **Linear regression**: Models the **linear** relationship between **numerical response variable (y)** and **explanatory variables (x)**, which can be either **numerical** or **categorical**
  - For now, we'll focus on **simple linear regression**, which only has one **explanatory variable**
- The form of this model is $\hat{\mathbf{y}} = \hat{\mathbf{B}}_0 + \hat{\mathbf{B}}_1\mathbf{x}$
  - Note: $\hat{B}$ is supposed to represent beta hat (β + ˆ)
- The **coefficients** ($\hat{\mathbf{B}}_0$ **and** $\hat{\mathbf{B}}_1$) have different interpretations depending on whether x is **numerical** or **categorical**

# Explanatory Variable: Numerical

- When x is **numerical...**
  - The model represents a “line of best fit”
  - $\hat{B}_0$ is the **y-intercept**
    - *When price percentage equals 0%, the average win percentage is 42%*
  - $\hat{B}_1$ is the **slope**
    - *As price percentage increases by 1%, the win percentage increases by 0.178%, on average*
  - **Least-squares regression** finds the optimal values of $\hat{B}_0$ **and** $\hat{B}_1$ by minimizing **residuals** (errors)




\# A tibble: 2 × 7

| | | term | estimate | std_error | statistic | p_value | lower_ci | upper_ci |
|---|---|---|---|---|---|---|---|---|
| | | <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| $\hat{\beta}_0$ | 1 | intercept | 42.0 | 2.91 | 14.4 | 0 | 36.2 | 47.8 |
| $\hat{\beta}_1$ | 2 | pricepercent | 0.178 | 0.053 | 3.35 | 0.001 | 0.072 | 0.283 |

- Interpretation of the coefficients?

$\hat{\beta}_0 = 42$

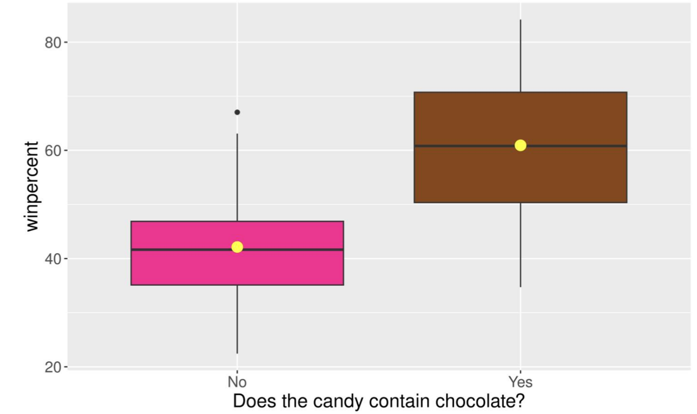
$\hat{\beta}_1 = 0.178$

intercept: when candy has $% of 0, we expect its win % to be 42

slope: as we increase $ by 1%, we expect win % to increase by 0.178% points.

# Explanatory Variable: Binary Categorical

- When x is **binary categorical…**
  - The model represents means (one for each of the two group)
  - $\hat{B}_0$ is the mean of y in the **baseline group** (when x = 0)
    - *For candy without chocolate, the average win percentage is 42.1%*
  - $\hat{B}_1$ is the **difference in means** of **other group** from **baseline group** ($\bar{y}_{other}$ - $\bar{y}_{baseline}$)
    - *Candy with chocolate has a higher average win percentage than candy without chocolate by 18.8%*




```
# A tibble: 2 × 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept     42.1      1.65     25.6        0     38.9     45.4
2 chocolate     18.8      2.50      7.52       0     13.8     23.7
```

$\hat{\beta}_0$ (row 1), $\hat{\beta}_1$ (row 2)

win % of non choc

pos. slope bc. pos. relationship between choco. and win %.

# Linear Regression: Code

- **Fitting the model**: Use this to build your model
  - MODEL <- lm(Y-VAR ~ X-VAR, data = DATASET)
  - *model <- lm(winpercent ~ pricepercent, data = candy)*
- **Getting the numbers**: Use this to summarize your model
  - get_regression_table(MODEL)
  - *get_regression_table(model)*
- **Predicting**: Use this for your model to predict y-value of new instances
  - predict(MODEL, newdata = data.frame(Y-VAR = VALUE))
  - *predict(model, newdata = data.frame(pricepercent = 85))*

# Population Model vs. Estimated Model

- **Population model**: $y = B_0 + B_1x + \varepsilon$
  - $\varepsilon$ is **error/"random noise"** around the line (**population parameter** for the **residuals**)
  - $\varepsilon \sim N(0, \sigma)$
  - $B_0$ **and** $B_1$ are **population parameters**

- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1x$
  - This is what our "line of best fit" is
  - $\hat{B}_0$ **and** $\hat{B}_1$ are estimates of the **population parameters**
  - $\varepsilon$ "disappears" because the **estimated model** is a straight line

# Content Review: Week 11

# Introducing Multiple Linear Regression

- **Multiple linear regression**: Models the **linear** relationship between **numerical response variable (y)** and multiple **explanatory variables ($x_1$, $x_2$, ..., $x_p$)**, which can be either **numerical** or **categorical**
- The form of this model is $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \ldots + \hat{B}_p x_p$
  - Note: $\hat{B}$ is supposed to represent beta hat ($\beta$ + ˆ)
- $\hat{B}_k$ (coefficient of **predictor $x_k$**) is predicted mean change in **y (response variable)** corresponding to **1 unit change in $x_k$** when **all other predictors are held constant**
  - If $x_k$ is **numerical**, think of slope
  - If $x_k$ is **categorical**, think of difference in means (of group where $x_k$ = 1 from baseline group)

For houses, if I want to predict price based on living area and whether or not there’s central air, what is p (number of predictors)?

# Question:

For houses, if I want to predict price based on living area and whether or not there's central air, what is p (number of predictors)?

We'll use linear regression to model this relationship.

$\hat{y}$ = price

$x_1$ = living area (numerical)

$x_2$ = whether or not there's central air (categorical)

Thus, p = 2.

# Example: Houses

- **Variables**: price ($\hat{y}$), living area ($\mathbf{x_1}$), whether or not there's central air ($\mathbf{x_2}$)
  - $x_1$ is **numerical**, $x_2$ is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2$
  - **Line when $x_2$ = 0 (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2$ = 1 (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1$


```
## # A tibble: 3 x 2
##   term              estimate
##   <chr>                <dbl>
## 1 intercept           42595.   β̂0
## 2 livingArea            107.   β̂1
## 3 centralAir: No    -28451.   β̂2
```

when living area, $x_1$, is held constant, this is the average difference in houseprice, y

# Example: Houses

- **Variables**: price ($\hat{y}$), living area (**$x_1$**), whether or not there's central air (**$x_2$**)
  - $x_1$ is **numerical**, $x_2$ is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2$
  - **Line when $x_2$ = 0 (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2$ = 1 (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1$

- Since we have **multiple variables**, be careful interpreting the **coefficients**
  - **$\hat{B}_0$**: For **houses with central air ($x_2$ = 0)**, when **living area ($x_1$)** equals 0, the **price ($\hat{y}$)** is **\$42,595 ($\hat{B}_0$)**, on average
  - **$\hat{B}_1$**: Controlling for **central air ($x_2$)**, as **living area ($x_1$)** increases by 1 unit, **price ($\hat{y}$)** increases by **\$107 ($\hat{B}_1$)**, on average
  - **$\hat{B}_2$**: Controlling for **living area ($x_1$)**, **houses without central air ($x_2$ = 0)** cost **\$28,451 ($\hat{B}_2$)** less than **houses with central air ($x_2$ = 1)**, on average

# The General "Formulas" for Equal-Slopes (When $x_2$ Is Categorical)

- $\hat{B}_0$ is **y-intercept of line when $x_2$ = 0**
  - *Ex: For houses with central air ($x_2$ = 0), when living area ($x_1$) equals 0, the price ($\hat{y}$) is $42,595 ($\hat{B}_0$), on average*
- Since this is equal-slopes, $\hat{B}_1$ is **slope of both lines** (a.k.a. increase in $\hat{y}$ after 1-unit increase in $x_1$, **controlling for $x_2$**)
  - *Ex: Controlling for central air ($x_2$), as living area ($x_1$) increases by 1 unit, price ($\hat{y}$) increases by $107 ($\hat{B}_1$), on average*
- $\hat{B}_0 + \hat{B}_2$ is y-intercept of line $x_2$ = 1, so $\hat{B}_2$ is **difference in $\hat{y}$ between both lines ($\hat{y}_{other} - \hat{y}_{baseline}$), controlling for $x_1$**
  - *Ex: Controlling for living area ($x_1$), houses without central air ($x_2$ = 0) cost $28,451 ($\hat{B}_2$) less than houses with central air ($x_2$ = 1), on average*

Looking at the tibble, how can we tell what's the baseline group?

# Question:

Looking at the tibble, how can we tell what's the baseline group?

Remember the **baseline group** is when $x_k = 0$ for some categorical predictor $x_k$.

Things are relative to the **baseline group**, so the tibble presents the "change" with the **categorical predictor** (to $x_k = 1$ from $x_k = 0$).

Thus, the **baseline group** is the OPPOSITE of the group shown.

# Baseline Group

```
## # A tibble: 3 x 2
##   term            estimate
##   <chr>              <dbl>
## 1 intercept         42595.
## 2 livingArea          107.
## 3 centralAir: No   -28451.
```

$\hat{\beta}_0$

$\hat{\beta}_1$

$\hat{\beta}_2$

when living area, $x_1$, is held constant, this is the average difference in house price, $y$

The output tells us "centralAir: No" has an estimate of -28,451. Thus, "centralAir: Yes" (a.k.a. houses WITH central air) is our baseline group.

# Categorical Variables with 2+ Categories

- **Linear regression** can accommodate **categorical variables** with 2+ categories
  - *Ex: We can predict RFFT score with the categorical variable of education, which can be "Lower Secondary," "Higher Secondary," or "University"*
- When **x** is a **categorical variable** with k + 1 categories…
  - $\hat{\mathbf{B}}_0$ represents the **mean of y** in the **baseline group** (one of those k + 1 categories)
  - $\hat{\mathbf{B}}_k$ represents the **difference in means**—specifically, going from x = 0 (**baseline group**) to x = k (one of the other groups)
  - Thus, $\hat{\mathbf{B}}_k = \bar{\mathbf{y}}_{\textbf{group k}} - \bar{\mathbf{y}}_{\textbf{baseline}}$
- We can confirm our answers with some data wrangling
- Let's look at an example…

# Interpreting a categorical predictor with several levels

$$\widehat{RFFT} = 40.9 + 14.8(Edu_{LS}) + 32.1(Edu_{HS}) + 45.0(Edu_{Univ})$$

- When $x$ is a categorical variable with $k+1$ levels...
  - $\hat{\beta}_0$ represents the mean of $y$ in the baseline group
  - $\hat{\beta}_k$ represents the difference in means; specifically, going from $x = 0$ to $x = k$
- Mean RFFT score is 40.9 points among those with at most a Primary education.
- The mean RFFT score among those with at most a University education is 45 points higher than those with at most a Primary education: $40.9 + 45 = 85.9$ points.
- The `Edu_new: Univ` coefficient equals $\bar{y}_{Univ} - \bar{y}_{Primary} = 45$

```
prevend.samp %>%
  group_by(Edu_new) %>%
  summarize(mean_RFFT = mean(RFFT))
```

```
## # A tibble: 4 x 2
##   Edu_new      mean_RFFT
##   <fct>            <dbl>
## 1 Primary           40.9
## 2 Lower Sec         55.7
## 3 Higher Sec        73.1
## 4 Univ              85.9
```

+14.8 +32.2 +45

```
model <- lm(RFFT ~ Edu_new,
            data = prevend.samp)
get_regression_table(model) %>%
  select(term, estimate)
```

```
## # A tibble: 4 x 2
##   term                   estimate
##   <chr>                     <dbl>
## 1 intercept                  40.9
## 2 Edu_new: Lower Sec         14.8
## 3 Edu_new: Higher Sec        32.1
## 4 Edu_new: Univ              45.0
```

estimate : diff in means

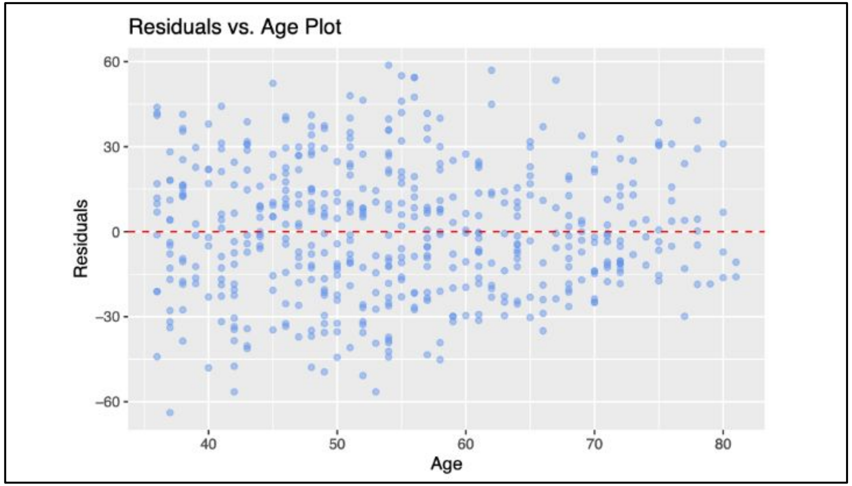
$\hat{\beta}_0 = 40.9$, $\hat{\beta}_1 = 14.8$, $\hat{\beta}_2 = 32.1$, $\hat{\beta}_3 = 45.0$

# Assumptions for (Multiple) Linear Regression

- **Linearity**: For each **predictor variable $x_k$** , the change in the **predictor** is **linearly related** to change in the **response variable** when the values of **all other predictors** are held constant
- **Constant Variability**: The **residuals** (errors) have approximately **constant variance**
- **Independence**: Each observation is **independent** (i.e., value of one observation provide no information about value of others)
- **Normality**: The **residuals** (errors) are approximately **normally distributed**

# Assumption #1: Linearity

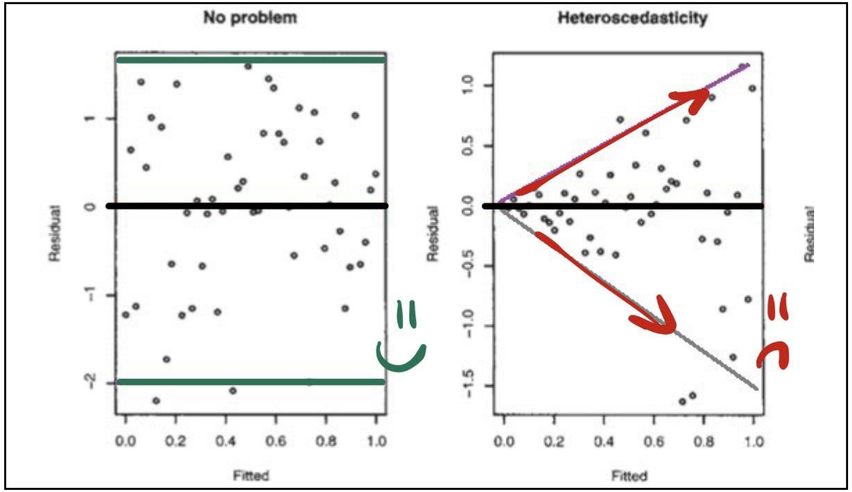
- Check via “**residual vs. predictor” plot with ggplot()**
  - For each **numerical predictor**, plot the **residuals** on the y-axis and the **predictor values** on the x-axis
- If data is linear, points should scatter from y = 0 randomly, with no pattern




- ggplot(MODEL, aes(y = .resid, x = NUM-PREDICTOR) + geom_point() + geom_hline(yintercept = 0)
- *ggplot(mod_rfft, aes(y = .resid, x = Age)) + geom_point(alpha = 0.5, col = "cornflowerblue") + geom_hline(yintercept = 0, lty = 2, col = "red") + labs(y = "Residuals", x = "Age", title = "Residuals vs. Age Plot")*

# Assumption #2: Constant Variability

- Check via **residual plot**, which plots residuals of model across domain
- Vertical spread of points should be roughly constant across domain, with no “fanning”
    - This interpretation is different from **linearity**; here, cite the upper and lower bounds (in green) to show there is no “fanning”



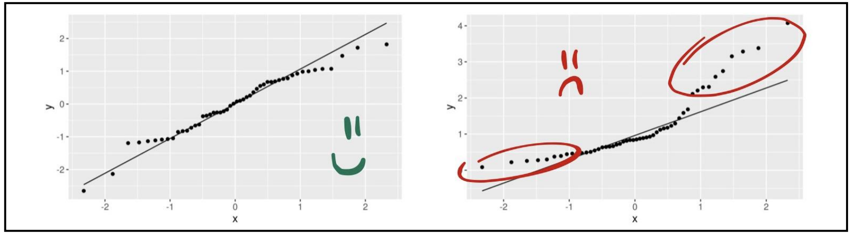

- ggplot(MODEL) + stat_fitted_resid()
- *ggplot(model) + stat_fitted_resid(alpha = 0.25)*

# Assumption #3: Independence

- Check by considering **how data was collected**
- If there's **independence**, knowing observation #1 gives no information about observation #2
  - *Ex: If data was randomly sampled, then independence can be reasonably assumed*
  - *Ex: If data was collected within a family (and we're measuring blood sugar, e.g.), then independence might not apply. Why?*

# Assumption #4: Normality

- Check via **Q-Q plot**, which plots residuals against theoretical quantiles of **normal distribution**
  - If residuals were perfectly **normally distributed**, they'd exactly follow the diagonal
  - We're not looking for perfect—just make sure it's reasonable
- Points should have a linear relationship, with no breaks at tails



- ggplot(MODEL) + stat_normal_qq()
- *ggplot(model) + stat_normal_qq(alpha = 0.25)*

# Returning to Inference: Population Model vs. Estimated Model

- **Population model**: $y = B_0 + B_1x_1 + \ldots + B_px_p + \varepsilon$
  - $\varepsilon$ is **error/"random noise"** around the line (**population parameter** for the **residuals**)
  - $\varepsilon \sim N(0, \sigma)$
  - $B_k$ is **population parameter**
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1x_1 + \ldots + \hat{B}_px_p$
  - This is what our "line of best fit" is
  - $\hat{B}_k$ is estimate of the **population parameter**
  - $\varepsilon$ "disappears" because the **estimated model** is a straight line

# Inference in (Multiple) Regression: Hypothesis Tests

- The **observed data** is assumed to have been **randomly sampled** from a population where the **explanatory variable** (X) and the **response variable** (Y) follow a **population model**
  - **Population model**: $\mathbf{Y = B_0 + B_1X_1 + \ldots + B_pX_p + \varepsilon}$
    - Like before, but we're now using capital letters to indicate **random variables**
  - **Estimated model**: $\mathbf{\hat{y} = \hat{B}_0 + \hat{B}_1x_1 + \ldots + \hat{B}_px_p}$
- Usually, we're concerned with **slope parameter** ($B_k$)
  - $\mathbf{H_0: B_k = 0}$ (i.e., there is no association between $X_k$ and Y after controlling for all other predictors in the model)
  - $\mathbf{H_A: B_k \neq 0}$ (i.e., there is an association between $X_k$ and Y after controlling for all other predictors in the model)

# Inference in (Multiple) Regression: Hypothesis Tests

- When **assumptions** are met (including 4 assumptions for multiple linear regression), then the ***t*-statistic** follows a ***t*-distribution** with **degrees of freedom n – p - 1**, where n is the number of cases and p is the number of predictors
  - $t = (\hat{B}_k - B_k^0)/SE(\hat{B}_k)$
    - Recall our null hypothesis is (often) $B_k = 0$, so the $B_k^0$ term can go away
  - $t = (\hat{B}_k)/SE(\hat{B}_k)$
- Our computers can calculate this for us!
  - get_regression_table(MODEL)
  - *get_regression_table(model)*

# Inference in (Multiple) Regression: Confidence Intervals

- **Confidence interval**: Recall the form of a confidence interval is **CI** = **sample statistic** ± **ME**
- **CI** = $\hat{B}_k \pm (t^* \times SE(\hat{B}_k))$
  - **t*** is the point on a ***t*-distribution** with **n − p - 1 degrees of freedom** and **α/2 area to the right**
  - **"With {α}% confidence, an increase in {explanatory variable} by 1 unit is associated with a change in average {response variable} between {lower bound} and {upper bound} units when holding {other explanatory variables in model} constant."**
  - *Ex: With 95% confidence, statin users have an average RFFT score that is between 4.2 points lower to 5.9 points higher than non statin users when holding age constant. Here,* $x_k$ *is categorical, so this is better interpreted as a difference in means.*
- Again, our computers can calculate this for us (use get_regression_table())!

# Confidence Interval vs. Prediction Interval

- **Confidence interval for mean response**: Tries to find plausible range for **parameter**
  - Centered at $\hat{y}$, with **smaller SE**
  - *Ex: We are 95% confident that the average price of 20 year-old, 1,500 square-feet Saratoga houses with central air and 2 bathrooms is between $199,919 and $211,834*

- **Prediction interval for individual response**: Tries to find plausible range for a **single, new observation**
  - Centered at $\hat{y}$, with **larger SE**
  - *Ex: For a 20 year-old, 1,500 square-foot Saratoga house with central air and 2 bathrooms, we predict, with 95% confidence, the price will be between $73,885 and $337,869*

# Confidence Interval vs. Prediction Interval: Code

- OBSERVATION-OF-INTEREST <- data.frame(EXPL-VAR(S) = VALUE(S))
- predict(MODEL, newdata = OBSERVATION-OF-INTEREST, interval = "confidence", level = CONF-LEVEL)
    - *house_of_interest <- data.frame(livingArea = 1500, age = 20, bathrooms = 2, centralAir = "yes")*
    - *predict(model, house_of_interest, interval = "confidence", level = 0.95)*

- OBSERVATION-OF-INTEREST <- data.frame(EXPL-VAR(S) = VALUE(S))
- predict(MODEL, newdata = OBSERVATION-OF-INTEREST, interval = "prediction", level = CONF-LEVEL)
    - *house_of_interest <- data.frame(livingArea = 1500, age = 20, bathrooms = 2, centralAir = "yes")*
    - *predict(model, house_of_interest, interval = "prediction", level = 0.95)*

# Two Types of Mult. Linear Regression: Equal-Slopes, Varying-Slopes

- **Equal-Slopes**: Assumes **change in y** associated with change in **1 explanatory variable**—a.k.a. the slope—DOES NOT DEPEND on **other explanatory variable(s)** in model
  - Visually, we see equal slopes in the lines
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \ldots + \hat{B}_p x_p$
  - We see there are no terms where the x variables interact with each other
- Code: `— <- lm(— ~ — + —, data = —)`

- **Varying-slopes model**: Assumes **change in y** associated with change in **1 explanatory variable**—a.k.a. the slope—DOES DEPEND on **other explanatory variable(s)** in model, so interaction term(s) is present
  - Visually, we see different slopes in the lines
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \hat{B}_3 x_1 x_2 + \ldots + \hat{B}_p x_p$
  - We see there is an interaction term between $x_1$ and $x_2$: $\boldsymbol{\hat{B}_3 x_1 x_2}$
- Code: `— <- lm(— ~ — * —, data = —)`

For houses, if I want to predict price based on living area and whether or not there's central air—now with a varying slopes model—what is p (number of predictors)?

# Question:

For houses, if I want to predict price based on living area and whether or not there's central air—now with a varying slopes model—what is p (number of predictors)?

We'll use linear regression (with varying-slopes) to model this relationship.

$\hat{y}$ = price

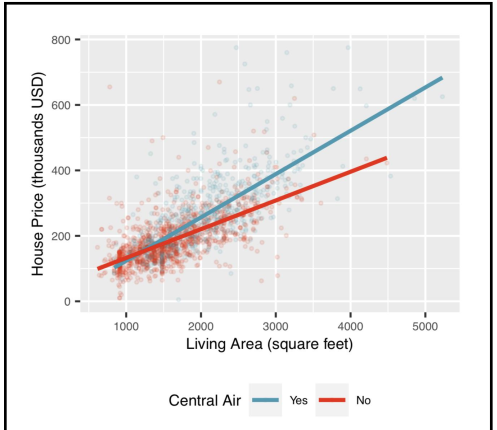
$x_1$ = living area (numerical)

$x_2$ = whether or not there's central air (categorical)

Thus, $p = 2$—like last time!

# Example: Houses (But with Varying-Slopes)

- **Variables**: price ($\hat{y}$), living area ($\mathbf{x_1}$), whether or not there's central air ($\mathbf{x_2}$)
  - $x_1$ is **numerical**, $x_2$ is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \hat{B}_3 x_1 x_2$
  - **Line when $x_2 = 0$ (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** $= \hat{B}_0$, **slope** $= \hat{B}_1$
  - **Line when $x_2 = 1$ (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + (\hat{B}_1 + \hat{B}_3)x_1$
    - **y-intercept** $= \hat{B}_0 + \hat{B}_2$, **slope** $= \hat{B}_1 + \hat{B}_3$
    - Notice the **slopes** are different!




```
## # A tibble: 4 x 2
##   term                         estimate
##   <chr>                           <dbl>
## 1 intercept   difference in intercepts  -8248.  β̂0
## 2 livingArea                        132.  β̂1
## 3 centralAir: No                  53226.  β̂2
## 4 livingArea:centralAirNo          -44.6  β̂3
```

coefficient on interaction term

# Example: Houses (But with Varying-Slopes)

- **Variables**: price ($\hat{y}$), living area ($\mathbf{x_1}$), whether or not there's central air ($\mathbf{x_2}$)
  - $x_1$ is **numerical**, $x_2$ is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \hat{B}_3 x_1 x_2$
  - **Line when $x_2$ = 0 (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2$ = 1 (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + (\hat{B}_1 + \hat{B}_3)x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1 + \hat{B}_3$
    - Notice the **slopes** are different!

- **$\hat{B}_0$**: For **houses with central air ($x_2$ = 0)**, when **living area ($x_1$)** equals 0, the **price ($\hat{y}$)** is **-\$8,248 ($\hat{B}_0$)**, on average
- **$\hat{B}_1$**: For **houses with central air ($x_2$ = 0)**, as **living area ($x_1$)** increases by 1 unit, **price ($\hat{y}$)** increases by **\$132 ($\hat{B}_1$)**, on average
- **$\hat{B}_2$**: When **living area ($x_1$)** equals 0, **houses without central air ($x_2$ = 1)** cost **\$53,226 ($\hat{B}_2$)** more than **houses with central air ($x_2$ = 0)**, on average
- **$\hat{B}_3$**: **Houses without central air ($x_2$ = 1)** have a lower slope than houses with central air by **\$44.6/unit ($\hat{B}_3$)**. For houses without central air **($x_2$ = 1)**, as **living area ($x_1$)** increases by 1 unit, **price ($\hat{y}$)** increases by **\$87.4 ($\hat{B}_1$ - $\hat{B}_3$)**, on average

# The General "Formulas" for Varying-Slopes (When $x_2$ Is Categorical)

- **$\hat{B}_0$** is **y-intercept of line when $x_2$ = 0**
  - *Ex: For houses with central air ($x_2$ = 0), when living area ($x_1$) equals 0, the price ($\hat{y}$) is -$8,248 ($\hat{B}_0$), on average*
- **$\hat{B}_1$** is **slope of line when $x_2$ = 0**
  - *Ex: For houses with central air ($x_2$ = 0), as living area ($x_1$) increases by 1 unit, price ($\hat{y}$) increases by $132 ($\hat{B}_1$), on average*
- $\hat{B}_0 + \hat{B}_2$ is y-intercept of line when $x_2$ = 1 (houses without central air), so **$\hat{B}_2$** is **difference in y-intercepts between both lines ($b_{other}$ - $b_{baseline}$)**
  - *Ex: When living area ($x_1$) equals 0, houses without central air ($x_2$ = 1) cost $53,226 ($\hat{B}_2$) more than houses with central air ($x_2$ = 0), on average*
- $\hat{B}_1 + \hat{B}_3$ is slope of line when $x_2$ = 1 (houses without central air), so **$\hat{B}_3$** is **difference in slopes between both lines ($m_{other}$ - $m_{baseline}$)**
  - *Ex: Houses without central air ($x_2$ = 1) have a lower slope than houses with central air by $44.6/unit ($\hat{B}_3$)*

# Inference with Varying-Slopes

- Same idea as before, but now we can infer about **population interaction coefficient** ($B_3$) instead of **population slope coefficient** ($B_1$)
  - $H_0$: $B_3 = 0$ (i.e., association/slope between y and $x_1$ doesn't differ by category)
  - $H_A$: $B_3 \neq 0$ (i.e., association/slope between y and $x_1$ differs by category)
- Again, our computers give us this info with `get_regression_table()`!

# Inference with interaction

- Do our observed data suggest that the association between total cholesterol and age differs by diabetic status in the population?
- Conduct a hypothesis test for the slope of the interaction term, $H_0 : \beta_3 = 0$ vs. $H_0 : \beta_3 \neq 0$
- If the population-level association between total cholesterol and age were the same between diabetics and non-diabetics, there would only be a 0.019 probability of observing a difference in slopes of -0.032 or larger in magnitude.
- With 95% confidence, the average change in total cholesterol per 1 year increase in age for diabetics is between 0.005 to 0.06 units smaller than for non-diabetics.

```
get_regression_table(mod_chol_int) %>%
  select(term, estimate, p_value)
```

```
## # A tibble: 4 x 3
##   term              estimate p_value
##   <chr>                <dbl>   <dbl>
## 1 intercept            4.77    0
## 2 Age                  0.01    0.001
## 3 Diabetes: Yes        1.54    0.074
## 4 Age:DiabetesYes     -0.032   0.019
```

```
get_regression_table(mod_chol_int) %>%
  select(term, estimate, lower_ci, upper_ci)
```

```
## # A tibble: 4 x 4
##   term              estimate lower_ci upper_ci
##   <chr>                <dbl>    <dbl>    <dbl>
## 1 intercept            4.77     4.47     5.06
## 2 Age                  0.01     0.004    0.016
## 3 Diabetes: Yes        1.54    -0.149    3.22
## 4 Age:DiabetesYes     -0.032   -0.06    -0.005
```

# When should I use equal-slopes vs. varying-slopes?

# Question:

When should I use equal-slopes vs. varying-slopes?

Consider your goal with the model.

With varying-slopes, certain questions (like the average difference in cholesterol between diabetic groups, controlling for age) can't be answered.

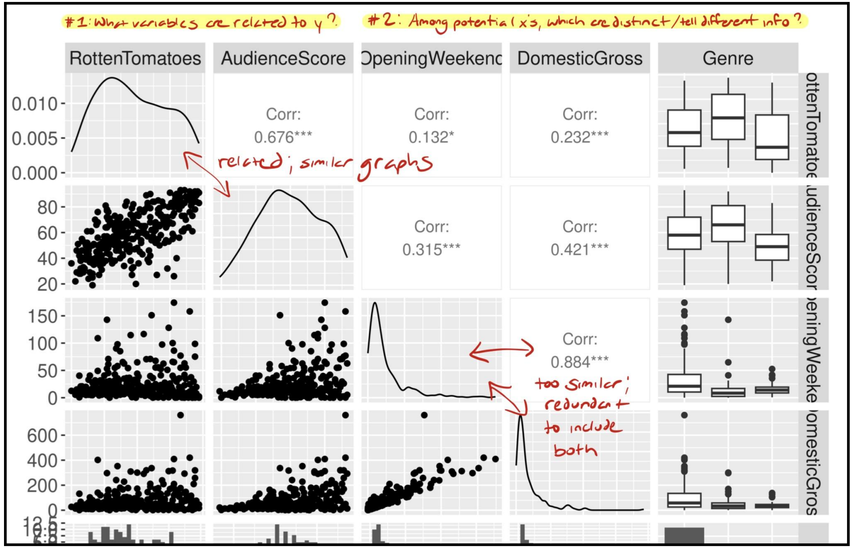
With equal-slopes, certain questions (like whether or not the relationship/slope differs between groups) can't be answered.

# $r^2$: Coefficient of Determination

- **$r^2$**: Percent of **total variation** in y (**response variable**) explained by the **model**
  - **$r^2 = (r)^2 = Var(\hat{y}_i)/Var(y_i)$**
  - If the **linear model** perfectly captured the **variability** in the observed data, then $Var(\hat{y}_i) = Var(y_i)$; thus, **$r^2$** would be 1
  - If **$r^2$** is too low, try different model; however, **$r^2$** only increases as new **predictors** are added to a model
- **adj($r^2$)**: Value of **$r^2$** adjusted for size of model (penalizes too-large models)
  - **$adj(r^2) = r^2 \times ((n - 1)/(n - p - 1))$**
  - n is sample size, p is number of predictors in model
- Basically, graph your data and pick the model with **highest adj($r^2$)**
  - glance(MODEL)
  - *glance(model)*

# Model Building Guidance

- In addition to looking at **adj($r^2$)**, consider your **explanatory variables** in the **model**
  - You want them to **explain different aspects** of the **response variable**
  - It would be redundant to have both RottenTomatoes and AudienceScore in a model, for example
- Use ggpair() to see relationship between multiple **explanatory variables**
  - If the graphs look alike, this tells you the **variables** are similar—consider removing one of them



# Questions?

# P-Set 8

# Have a great rest of your week!