Agenda

1. Multiple Linear Regression
   - Formula/notation
   - Parallel slopes
   - Parallel planes

Multiple Regression Multiple regression is a natural extension of simple linear regression.

- SLR: one response variable, one explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon)$$

- MLR: one response variable, *more than one* explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon)$$

- Estimated coefficients (e.g. $\hat{\beta}_i$'s) now are interpreted in relation to (or "conditional on") the other variables
- β_i reflects the *predicted* change in Y associated with a one unit increase in X_i, conditional upon the rest of the X_i's.
- R^2 has the same interpretation (proportion of variability explained by the model)

Regression with a Categorical Variable Consider first the case where X_1 is an *indicator* variable that can only be 0 or 1 (e.g. $isFemale$). Then,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1$$

So then,

$$\begin{aligned} \text{For men,} \qquad & \hat{Y}|_{X_1=0} = \hat{\beta}_0 \qquad \qquad \text{[what value does } \hat{\beta}_0 \text{ take on?]} \\ \text{For women,} \qquad & \hat{Y}|_{X_1=1} = \hat{\beta}_0 + \hat{\beta}_1 \end{aligned}$$

Multiple Regression with a Categorical Variable More generally, consider the case where X_1 is quantitative, but X_2 is an indicator variable. Then,

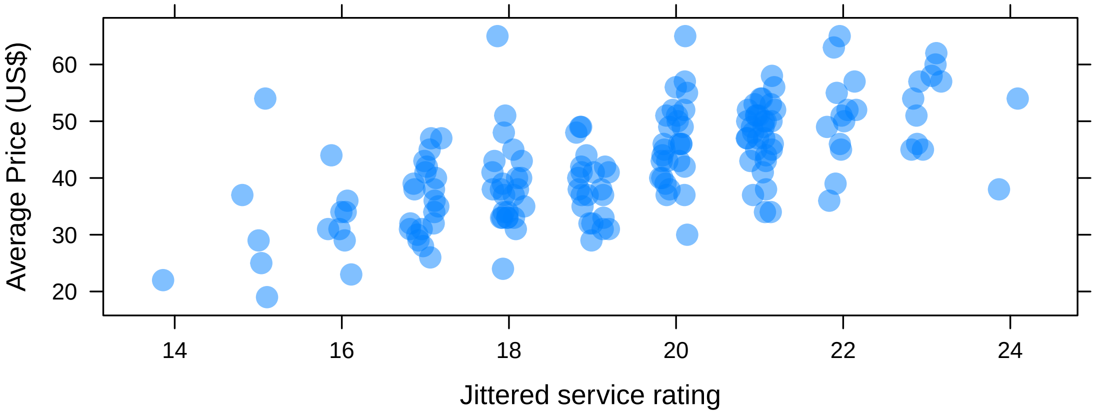
$$\begin{aligned} \hat{Y}|_{X_1, X_2=0} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 \\ \hat{Y}|_{X_1, X_2=1} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot 1 \\ &= \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \hat{\beta}_1 \cdot X_1 \end{aligned}$$

This is called a *parallel slopes* model. [Why?]

Example: Italian Restaurants The Zagat guide contains restaurant ratings and reviews for many major world cities. We want to understand variation in the average *Price* of a dinner in Italian restaurants in New York City. Specifically, we want to know how customer ratings (measured on a scale of 0 to 30) of the *Food*, *Decor*, and *Service*, as well as whether the restaurant is located to the *East* or west of 5th Avenue, are associated with the average *Price* of a meal. The data contains ratings and prices for 168 Italian restaurants in 2001.

```
require(mosaic)
NYC <- read.csv("http://www.math.smith.edu/~bbaumer/mth241/nyc.csv")
xyplot(Price ~ jitter(Service), alpha=0.5,
  xlab="Jittered service rating", ylab="Average Price (US$)", pch=19,
  cex=1.5, data=NYC)
m1 <- lm(Price ~ Service, data=NYC)
coef(m1)

## (Intercept)     Service
##  -11.977811    2.818433
```



In-Class Activity

1. Use `lm()` to build a SLR model for *Price* as a function of *Food*. Interpret the coefficients of this model. How is the quality of the food at these restaurants associated with its price?

2. Build a parallel slopes model by conditioning on the *East* variable. Interpret the coefficients of this model. What is the value of being on the East Side of Fifth Avenue?

3. Calculate the expected *Price* of a restaurant in the East Village with a *Food* rating of 23.

4. Add both regression lines to your scatterplot! [You can use plotModel to do this]

   ```
   m2 <- lm(Price ~ Food + East, data=NYC)
   plotModel(m2)
   ```

Multiple Regression with a Second Quantitative Variable If X_2 is a quantitative variable, then we have

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2$$

Notice that our model is no longer a line, rather it is a *plane* that lives in three dimensions!

Example: Italian Restaurants (continued) Now suppose that we want to improve our model by considering not only the quality of the *Food*, but also the quality of the *Service*. We can do this in `R` by simply adding another variable to our regression model.

```
m3 <- lm(Price ~ Food + Service, data=NYC)
coef(m3)

## (Intercept)        Food     Service
##  -21.158582    1.495369    1.704101
```

1. Interpret the coefficients of this model. What does the coefficient of *Food* mean in the real-world context of the problem? *Service*?

2. How important is *Service* relative to *Food*? Is it fair to compare the two coefficients?

3. Find the expected *Price* of a restaurant with a *Food* rating of 21 but a *Service* rating of 28. [Use `predict()` or calculate by hand]

4. Calculate the residual for San Pietro. Is it overpriced?

   ```
   NYC %>% filter(Restaurant == "San Pietro")

   ##   Case Restaurant Price Food Decor Service East
   ## 1   44 San Pietro    58   24    21       23    1
   ```

5. What if we added all three explanatory variables to the model? What geometric shape would we have then?