STATS 210P

Lecture 6

Sevan Koko Gulesserian

University of California, Irvine

Simson's Paradox

Simson's Paradox.

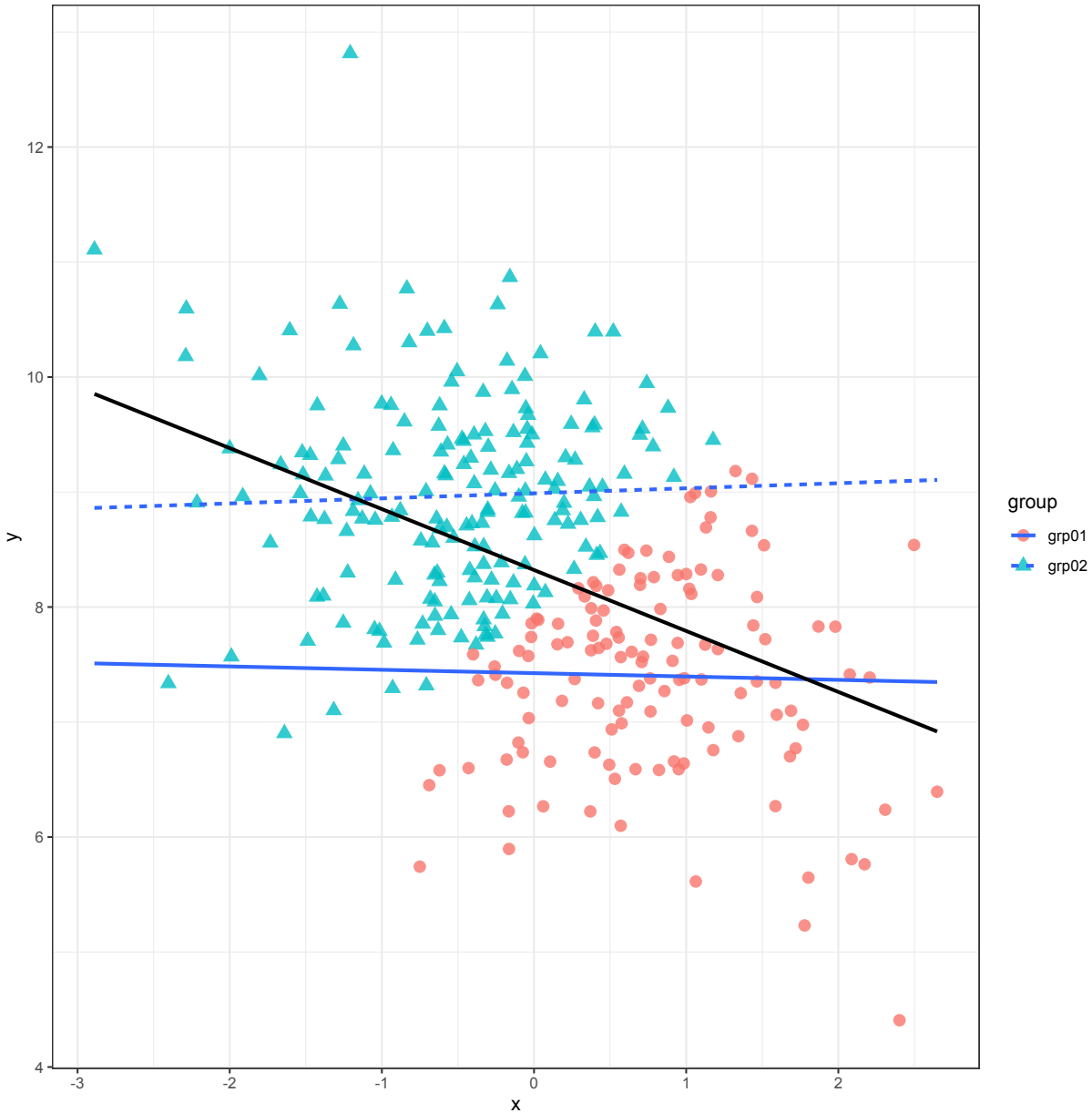
- Simpson's paradox is a phenomenon in which a trend appears in several different groups of the data but disappears or reverses when these groups are combined (or vice-versa, a trend appears when the groups are combined but reverses or disappears when the groups are accounted for).
- This highlights the importance of having the categorical covariate (that defines the groups) in our model and also on including an interaction term.

Simson's Paradox

Simson's Paradox.

- Say we have groups 1 and 2 in our data.
- The idea is that when we ignore these groups (that is to say just lump the data together and fit a model of $Y \sim X$) we can have a certain relationship between X and Y, but when we do account for the groups (by including the categorical group variable in the model) we can have another relationship between X and Y that contradicts the ungrouped relationship.
- Will take a look at an example on the next slide.

Multiple Linear Regression



Simson's Paradox

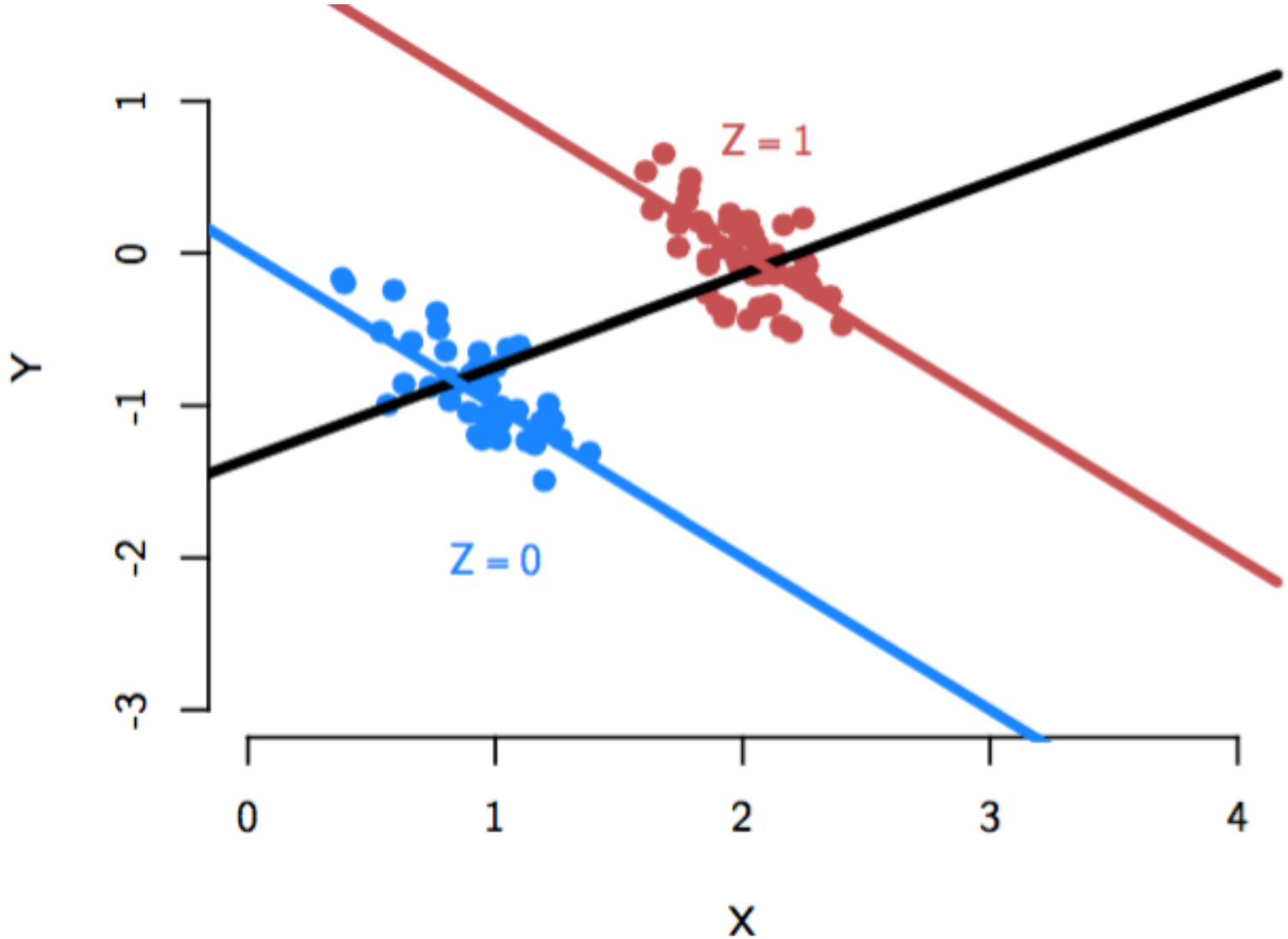
- We have a single quantitative continuous X and a single quantitative continuous Y, and we also a yes/no categorical variable of groups (group 1 and group 2).
- The blue line and the dotted blue line are the linear regression lines of X on Y for group 1 and group 2 respectively.
- The black line is the linear regression line of X on Y for the combined groups (that is to say ignoring the group labels).
- Can see that for each of the groups, there does not seem to be a steep slope (slightly positive slope for group 2 and slightly negative slope for group 1), but when the groups are combined, we see a steep negative slope.

Simson's Paradox

- If we ignored the groups, we would conclude that as X increases, Y tend to decrease substantially (as X increases one unit, the estimated expected value of Y will decrease by 0.530).
- But when we account for the groups, and go even further by including the interaction term between X and group (1 or 2), we are able to obtain the inference that is told by the blue lines.
- When accounting for groups and having an interaction term, we have that as as X increases one unit, the estimated expected value of Y will decrease by 0.029 for group 1, but will increase by 0.044 for group 2.
- All needed code to obtain plot and models is in the code for the class.

Multiple Linear Regression

Another example of Simpson's paradox.



Interaction term.

- The interaction term signified that the effect of X_1 on the response variable will depend on what the level of X_2 is (and vice versa that the effect of X_2 on the response will depend on the level of X_1.
- Example with binary covariate. Let Y be a wellness score, X_1 be treatment dosage amount (say in milligrams) and X_2 be an indicator of the subject being diabetic.
- If there is an interaction term, what we are stating is that the effect of the treatment in milligrams on the response depends on if the subject is diabetic or not.
    - Could be that as treatment dosage increases, the diabetic patients tend to have a higher positive effect on the wellness score than non-diabetic patients.

Multiple Linear Regression

Let our population model be:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon$.

- The coefficient of β_1 is called the *main effect* of X_{1i} (and β_2 is the main effect of X_{2i}).
- The coefficient β_3 signifies the interaction between X_1 and X_2.
- It is the added effect on Y when X_1 interacts with X_2 (and vice-versa).
- Focusing on X_1, without the interaction term the effect of X_1 on the response Y is captured by the β_1 term.
  - But now with the interaction term, the effect of X_1 will be captured by β_1 and also by the added term of β_3.

Multiple Linear Regression

Linear regression with several explanatory variables.

- We are still considering a single quantitative continuous response variable Y.
- We will now generalize to a general scenario with several quantitative and categorical yes/no predictors/explanatory variables.
- Also will learn how to test several of the slopes simultaneously (such as $\beta_1 = \beta_2 = 0$).
- Will generalize to categorical explanatory variables with several categories later on as well.

Multiple Linear Regression

- The multiple linear regression population model is now:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i$$

- There are now p many explanatory variables (where some could possible be interactions, i.e. $X_{3i} = X_{1i} X_{2i}$).
- The error term ε is the same as before, independent and identically distributed according to a Normal distribution with mean 0 and variance σ_ε^2.
- In total, we have $p + 1$ many β coefficients to estimate.

Multiple Linear Regression

Matrix formulation becomes a useful tool when needing to derive the estimates of β and the properties of their distribution (such as the standard error or deviation and what approximate distribution the estimates follow).

Also helps write out models in concise forms.

This is when we had only a single X predictor for each sampled unit.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

Multiple Linear Regression

Now with p many covariates per each sampled unit, the matrix X becomes.

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix}$$

and so the vector β becomes:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Multiple Linear Regression

- Estimation of the parameters $\hat{\beta}$ is done the same as before, minimizing the sum of square errors.
- By writing in matrix notation, it is quicker to get the derivatives needed and solve the system of equation.
- The estimated equation is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_p X_{pi}$$

Multiple Linear Regression

The estimates of β_0, β_1,...,β_p minimizes the following function (called the objective function):

$$f(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_p X_{pi}))^2$$

Will need to take partial derivatives with respect to each β_j and set it equal to 0 and solve.

Or can do this in more concise steps by writing out the objective function in matrix notation.

Multiple Linear Regression

Properties of our estimates $\hat{\beta}$.

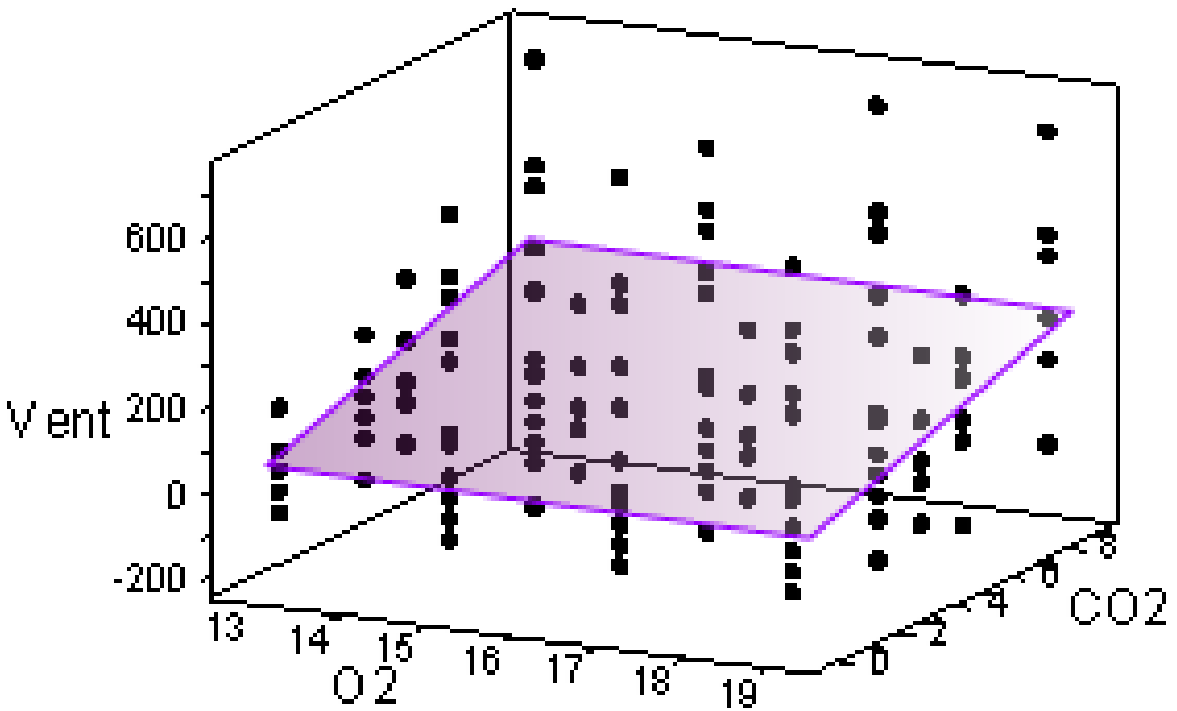
- The estimates of β_j for j=0,1,2,...,p are unbiased for the true β_j.
- That is to say the expectation of $\hat{\beta}_j$ is equal to the true β_j, $E(\hat{\beta}_j) = \beta_j$.
- The estimated standard error (or standard deviation) of $\hat{\beta}_j$, $\widehat{SE}(\hat{\beta}_j)$, will be a function of $\hat{\sigma}_\varepsilon$ and the X covariates.
- In large samples, we have that $\frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)}$ follows an approximate Normal distribution with expectation/mean 0 and variance 1.
- In small samples it will follow an approximate t distribution with n-(p+1) degrees of freedom.

Example: A study was conducted to see if the swallow bird, which burrows underground to nest, will alter its breathing behavior based on how much oxygen and carbon dioxide levels are below the surface.

The response variable Y is the precent change in the total volume of air breathed in a minute. X_{1i} is the percentage of oxygen in the air in the burrow and X_{2i} is the percentage of carbon dioxide in the air in the burrow.

Multiple Linear Regression

A 3d scatterplot with the plane of the regression equation.



Multiple Linear Regression

Creating a 3d plot is an involved process.

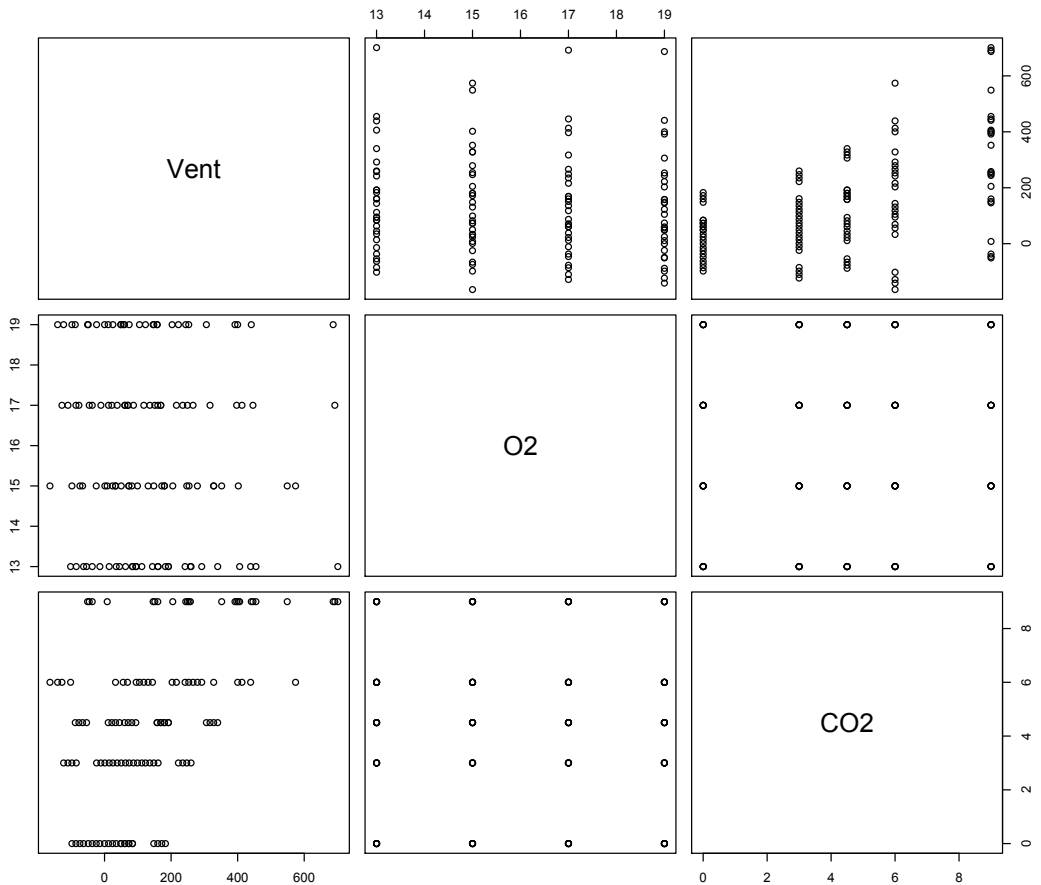
Additionally when we have more than 3 variables, a single plot is not feasible.

Instead, we will plot all variables against each other to assess if there is a linear relationship between the explanatory variables and the response variable.

In R, this is done using the plot(dataset) function, where you put in the name of the dataset.

Multiple Linear Regression

plot(babybirds)



Scatterplot matrix.

- The plot on the previous slide is called a scatterplot matrix.
- The row you are in represent the Y-axis and the column you are in represent the X-axis.
- So the top right box will have Vent on the Y-axis and CO2 on the X-axis.
- The bottom left box will have CO2 on the Y-axis and Vent on the X-axis.

Multiple Linear Regression

```
lm(formula = Vent ~ O2 + CO2, data = babybirds)

Residuals:
    Min      1Q  Median      3Q     Max
-356.57  -96.50    8.72   84.68  422.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   85.901    106.006   0.810    0.419
O2            -5.330      6.425  -0.830    0.408
CO2           31.103      4.789   6.495  2.1e-09 ***
---
Residual standard error: 157.4 on 117 degrees of freedom
Multiple R-squared:  0.2682,Adjusted R-squared:  0.2557
F-statistic: 21.44 on 2 and 117 DF,  p-value: 1.169e-08
```

Multiple Linear Regression

Fitting a model with 2 quantitative explanatory variables.

- Our model is: $Y_i = \beta_0 + \beta_1 O2_i + \beta_2 CO2_i + \varepsilon_i$.
- Interpretation of the intercept is the same as before, when O2 and CO2 are equal to 0, the predicted value of Vent is 85.90.
- The interpretation of the slope is also the same, **but now must mention that we are fixing the values of the other predictors.**
- For example to interpret the slope of O2, we say that given a fixed level of CO2 (no need to specify the level, just as long as it is fixed), as O2 increases by 1 unit, we predict Vent to change by -5.33 units.
- Similarly for the slope of CO2, we say that given a fixed level of O2, as CO2 increases by 1 unit, the predicted Vent will change by 31.10 units.

Multiple Linear Regression

Fitting a model with 2 quantitative explanatory variables.

- We can also interpret the slope coefficients by saying we are comparing two birds.
- For example to interpret the slope of O2, we say that comparing two birds with the same CO2 levels but differ in O2 by one unit, the bird with the higher O2 level will have an estimated expected Vent level that is -5.33 units lower than the other bird.
- For example to interpret the slope of CO2, we say that comparing two birds with the same O2 levels but differ in CO2 by one unit, the bird with the higher CO2 level will have an estimated expected Vent level that is 31.10 units higher than the other bird.

Multiple Linear Regression

- R-squared interpretation is the same as before, but now accounts for 2 predictors.
- 26.82% of the variation in Vent is explained by O2 and CO2.
- As of now, we only know how to test a single slope given that the other predictors are in the model already.
- For example we can test if CO2 is associated with Vent given that we have accounted for O2 (that is to say that is CO2 a significant predictor/explanatory variable given that O2 is already in the model).
- This would test if $\beta_2 = 0$ vs $\beta_2 \neq 0$.
- Similarly can test if O2 is associated with Vent given that we have already accounted for CO2.
- This would be the test of $\beta_1 = 0$ vs $\beta_1 \neq 0$.

Multiple Linear Regression

Adjusted R-squared and multiple R-squared.

- A new concept is called the *adjusted R-squared*, label this adjusted R-squared as $\tilde{R}^2$.
- The adjusted R-squared takes the usual R-squared (called multiple R-squared in the output of R) and adjusts it for the fact that there are numerous explanatory variables in the model now.
- Adjusted R-squared $= \tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$.
- Remember that $R^2 = 1 - \frac{SSE}{SSTO}$.
- Now $\tilde{R}^2 = 1 - \frac{SSE/(n-p-1)}{SSTO/(n-1)}$.
- Where p is the number of **slope** coefficients in the model $(\beta_1, \beta_2, ..., \beta_p)$.

Multiple Linear Regression

- The idea is that R^2 will naturally increase the more predictors are added.
- That is because even if we add a predictor/explanatory variable that has nothing to do with the response (correlation with the response is ≈ 0), it cannot lower the R^2 (worst case scenario it does not increase R^2 at all, but it cannot decrease it).
- The adjusted R-squared is an attempt to take account this issue of R^2 automatically and spuriously increasing when extra explanatory variables are added to the model.
- The adjusted R-squared will only increase when the increase in R^2 (from including a new predictor) is more than what would be expected to be seen by chance.
- Will see later on how adjusted R-squared can be used to build models (adjusted R-squared will penalize a model for adding poor predictors).

Multiple Linear Regression

Returning to bird example, run a model with only CO2 predictor.

```
lm(formula = Vent ~ CO2, data = babybirds)

Residuals:
    Min      1Q  Median      3Q     Max
-351.24  -92.41    8.23   85.48  420.45

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6208    25.8643   0.024    0.981
CO2          31.1028     4.7823   6.504 1.97e-09 ***
---
Residual standard error: 157.2 on 118 degrees of freedom
Multiple R-squared:  0.2639,Adjusted R-squared:  0.2576
F-statistic:  42.3 on 1 and 118 DF,  p-value: 1.973e-09
```

Multiple Linear Regression

Bird breathing data.

- Note that the adjusted R-squared will always be less than the regular R-squared (shown as Multiple R-squared in R).
- Comparing the model with both CO2 and O2 to the model with just CO2, can see that by adding O2 (already given CO2 in the model), the multiple R-squared barely increased.
- O2 is a poor predictor of the response of breathing rate (large 2-sided p-value).
- As a result, the adjusted R-squared virtually has no increase between the 2 models (in fact, the adjusted R-squared is lower for the bigger model with both O2 and CO2).
- The contribution to the explanation of the variance of the response does not increase more than what one would expect by random chance.

Bird breathing data.

- The adjusted R-squared for the model with both O2 and CO2 is 0.2557 and the adjusted R-squared for the model with just CO2 is 0.2576.
- The idea is that O2 is a very poor predictor/explanatory variable for the response Vent.
- Thus we are being penalized more for adding an extra explanatory variable than what is being contributed by adding the extra explanatory variable, which is O2.
- Note the large p-value on the O2 coefficient estimate.

Multiple Linear Regression

As before with a single quantitative and single categorical explanatory variable, we can have an interaction between the quantitative covariates in a multiple linear regression.

- The idea is similar to before, but now instead of the covariate having values 0 or 1 it has several (potentially infinitely) many values.
- An interaction will signify that the effect of one quantitative explanatory variable depends on the level of the other quantitive explanatory variable.
- The interpretation of the effect of one quantitative explanatory variable on the response variable will need to be conditional on a specific value of the other explanatory variable (before it was just a 0 or 1, now can be many numbers).

Multiple Linear Regression

Say we have a single response Y, and two quantitative explanatory variables X_1 and X_2.

- The model with an interaction term will be $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$.
- β_1 and β_2 correspond the main effects of the covariates X_1 and X_2 respectively.
- β_3 corresponds to the interaction effect of the covariates X_1 and X_2.
- All other assumptions and specifications is similar to the simple linear model.

Multiple Linear Regression

Say we have a single response Y, and two quantitative explanatory variables X_1 and X_2.

- The model with an interaction term will be $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$.
- To interpret the effect of X_1 on the response, must condition on a specific value of X_2.
- When X_2 was categorical, it could be 0 or 1. Now that it is quantitative, can be any discrete or continuous value.
- The same for interpreting the effect of X_2 on the response, must condition on a specific value of X_1.
- When there was no interaction, to interpret the effect of X_1 on the response, just had to condition on X_2 being fixed (without necessary needing a specific value).

Multiple Linear Regression

Example of interaction term between two quantitative explanatory variables.

- Let the response be Y = blood pressure of subject, X_1 = their weight in pounds, and X_2 = their height in inches.
- The interaction term between X_1 and X_2 will signify that the effect of weight on blood pressure depends on the height of the subject (and vice-versa).
- Weight increasing by 10 pounds will have different effects depending on the height of the subject.
- Gaining 10 pounds could have a bigger impact when the subject is 5 feet tall (60 inches), but less of an effect when the subject is 7 feet tall (84 inches tall).
- Thus the weight of the subject and their height interact when determining their effects on blood pressure.

Multiple Linear Regression

Example of letting the response be Y = blood pressure of subject, X_1 = their weight in pounds, and X_2 = their height in inches.

- Another way to view the interaction is as follows.
- Say the subject in question weighs 250 pounds.
- A doctor or health worker will need to know the height of the subject to determine how to predict their blood pressure.
- 250 pounds is a high weight amount for someone who is 60 inches tall, but is not a high amount for someone who is 84 inches tall.
- Thus if you are 250 pounds and 60 inches tall, you will tend to have high blood pressure, but if you are 250 pounds and 84 inches tall, you will tend to have a normal blood pressure.
- And so determining the effect of weight on blood pressure, we will need to know the height as well.

Multiple Linear Regression

Return to the Midwest house sale price data.

- This was the dataset where the sale price of a house in the Midwest was recorded, along with several explanatory variables related to the characteristics of the house.
- Let us consider two quantitative predictors, the square footage of the house and how many bedrooms the house has.
- The response variable is the sales price.

Multiple Linear Regression

Fit a model with no interaction term.

```
lm(formula = price ~ sqft + bed, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-227961  -38270   -8693   24670  381949

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -66971.563  13404.170  -4.996    8e-07 ***
sqft           165.832      5.855  28.326   <2e-16 ***
bed          -8647.511   4104.008  -2.107   0.0356 *
---

Residual standard error: 78860 on 519 degrees of freedom
Multiple R-squared:  0.6743,Adjusted R-squared:  0.6731
F-statistic: 537.3 on 2 and 519 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

- The estimated sample based model with no interaction is: $\widehat{price}_i = -66971.56 + 165.83 sqft - 8647.51 bed$.
- To interpret the slope on bed, given the sqft of a house is fixed (at some value), then increasing the number of bedroom by 1 will reduce the predicted price by 8647.51 (or the estimated expected price by 8647.51).
- To interpret the slope on sqft, given the number of bedrooms of a house is fixed, then increasing the number of sqft by 1 will increase the predicted price by 165.83.
- Predictions are done the same as before (just plug in the sqft and bed values into the estimated regression equation).

Multiple Linear Regression

- Now say we add an interaction term.
- The population model will be
  $Y_i = \beta_0 + \beta_1 sqfti + \beta_2 bed_i + \beta_3 sqft_i bed_i + \varepsilon_i$.
- Just like with categorical variables, the added term is just $sqft * bed$.
- This will now signify that the effect of sqft changes on the house price will depend on the number of bedrooms (and vice-versa, the changes in bedroom numbers will depend on the square footage of the house).

Multiple Linear Regression

```
lm(formula = price ~ sqft + bed + sqft * bed, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-196725  -41800   -6185   27098  376624

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.033e+05  3.625e+04  -5.607 3.36e-08 ***
sqft         2.281e+02  1.646e+01  13.857  < 2e-16 ***
bed          2.754e+04  9.835e+03   2.800  0.00529 **
sqft:bed    -1.576e+01  3.904e+00  -4.037 6.23e-05 ***
---
 Residual standard error: 77730 on 518 degrees of freedom
Multiple R-squared:  0.6843,Adjusted R-squared:  0.6824
F-statistic: 374.2 on 3 and 518 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

Will round roughly for the output.

- The estimated model with an interaction is:
  $\widehat{price}_i = -203300 + 228sqft + 27540bed - 15.7sqft * bed$.
- For a house that is 2000 sqft, the equation is
  $\hat{Y}_i = -203300 + 228 * 2000 + 27540 * bed - 15.7 * 2000 * bed = 252700 - 3860 * bed$.
- For a house that is 4000 sqft, the equation is
  $\hat{Y}_i = -203300 + 228 * 4000 + 27540 * bed - 15.7 * 4000 * bed = 708700 - 35260 * bed$.
- Now say the house is given to have 3 bedrooms, the equation for the sale price is
  $\hat{Y}_i = -203300 + 228sqft + 27540 * 3 - 15.7 * 3 * sqft = -120680 + 180.9sqft$.

Multiple Linear Regression

- The estimated sample based model with an interaction has $\widehat{price}_i = -203300 + 228sqft + 27540bed - 15.7sqft * bed$.
- To interpret this model with respect to one covariate, must condition on a specific known value of the other covariate (to interpret the effect of sqft need to condition on a specific value of bedroom, and vice-versa).
- Take for example a house with 4 bedrooms.
- To interpret the slope on sqft, given the number of bedrooms of a house is fixed at 4 bedrooms, then increasing the number of sqft by 1 will increase the predicted price (of a house with 4 bedrooms) by 228-15.7*4=165.2 dollars.
- If the house had 6 bedrooms, then a sqft increase will result in a predicted change of 228-15.7*6=133.8 dollars.

- The model with an interaction has $\hat{Y}_i = -203300 + 228sqft + 27540bed - 15.7sqft * bed$.
- Now say the house is 2500 sqft.
- To interpret the slope on bedroom, given the sqft of the house is fixed at 2500 sqft, then increasing the number of bedrooms by 1 will increase the predicted price (of a house 2500 sqft) by 27540-15.7*2500=-11709.
- Now say the house was 1500 sqft. Then increasing the number of bedrooms by 1 will increase the predicted price (of a house 1500 sqft) by 27540-15.7*1500=-3991.

How to interpret the interaction term:

- Say the house is 2500 sqft.
- The interaction term is interpreted to say that the added effect of increasing a bedroom by 1 for a 2500 sqft house is -15.76*2500 =-39400.
- Now say the house has 3 bedrooms. Then the added effect of increasing sqft by 1 for a house with 3 bedrooms is -15.76*3=-47.28.

Interpreting interaction term.

- Notice to interpret any explanatory involved in the interaction, we had to condition on the other explanatory variable in the interaction being set to a specific value.
- Thus in the Midwest house sales example, to interpret the effect of bedroom number on sales price, we had to set the sqft variable to a specific known value.
- And to interpret the effect of sqft on sales price, we had to set bedroom number to a specific known value.

Multiple Linear Regression

Multiple linear regression.

Note that we can easily extend to have more than 2 covariates.

Interpretation stays the same, but for an interaction model we must condition on all covariates that interact with the one we are interested in to be set to a specific known value.

Even without interaction, must condition on all other predictors being fixed at some level (need not be a specific value).

Multiple Linear Regression

- We can generalize the multiple linear regression model to have $p > 2$ covariates/predictors.
- The population model is $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i$.
- Each unit in the dataset has p many explanatory variables, and a single response.
- Showed how to have a single quantitative and single categorical predictors with interaction.
- Then saw how to have a model with two quantitative explanatory variables and interactions.
    - The model with 2 covariates and their interactions can be easily extended to have several covariates.

Three way interactions and more.

- Say we have covariates X_1, X_2, and X_3 (can be quantitative or categorical yes/no explanatory variables).
- We can have an interaction term that involves all three covariates: $X_1X_2X_3$.
- This would imply that the effect of X_1 on the response depends on the levels of X_2 and X_3 (also that the effect of X_2 on the response depends on the levels of X_1 and X_3, and the effect of X_3 depends on the levels of X_1 and X_2).

Multiple Linear Regression

- The model with numerous quantitative covariates can also be easily extended to include categorical variables.
- Just like with the simple model, the main effect of the categorical yes/no variable will adjust the intercept for that group.
- The interaction term with the quantitative covariate is the adjustment term to the slope estimate for the quantitative covariate.
- With numerous quantitative covariates, the addition of a categorical covariate is straightforward.
    - The extension now comes to when the categorical variable has several categories (not just yes/no).

Multiple Linear Regression: Categorical Covariates

Several categories for an explanatory variable.

- Say we have a categorical explanatory variable with k many levels (say k=3 for example).
- A yes/no categorical variable has k=2 categories.
- The model will assign one of these groups to be the base group and will compute coefficient estimates for all the other $k-1$ groups effects.
- By having a categorical variable with k levels will result in an additional $k-1$ terms being estimated for just the main effect.

Multiple Linear Regression: Categorical Covariates

- A categorical covariate with $k \geq 2$ levels will result in $k-1$ yes/no (binary indicator) covariates being generated.
- Interaction term with a single quantitative covariate and a k category categorical covariate will result in $k-1$ additional terms being created (the $k-1$ binary indicators times the quantitative).
- Having a categorical variable with k levels will result in an additional $k-1$ terms being estimated for just the main effect and an additional $k-1$ terms for any interaction.

Multiple Linear Regression: Categorical Covariates

Indicator function.

- We will denote categorical explanatory variables with more than 2 categories as indicator variables $I(.)$ (capital i).
- The indicator function is as follows: $I(condition) = 1$ if the condition is satisfied and 0 otherwise.
- If we have k categories, then we will have $k - 1$ indicator variables in the model.
- The same as converting yes/no to 1/0.
- Will have a base group and then $k - 1$ options, to make up k in total.

Multiple Linear Regression: Categorical Covariates

Example to highlight using indicators for several categories.

- Say we have a response Y and a single categorical predictor with 3 many categories (say they are A, B, and C).
- Then we will have two indicator variables, one for B and one for C (assuming A is the base group).
- The population model will be $Y = \beta_0 + \beta_1 I(\textit{group } B) + \beta_2 I(\textit{group } C) + \varepsilon$.
- For example, $I(\textit{group } B) = 1$ if the observation is in group B and 0 otherwise, and $I(\textit{group } C) = 1$ if the observation is in group C. If both are 0, then it is in the base group of A.

Multiple Linear Regression: Categorical Covariates

Example to highlight using indicators for several categories.

- Thus we need to ask two yes/no questions to determine which of the 3 categories the sampling unit is in.
- Question 1 is: Are you in group B? Yes or no.
- Question 2 is: Are you in group C? Yes or no.
- If the answer to one of the 2 previous questions is yes, then the sampling unit is in that category.
- If the answer to both questions is no, then the sampling unit is in the base category A.
- Can extend this to $k-1$ yes/no questions for a k category covariate.

Multiple Linear Regression: Categorical Covariates

Several categories and the ANOVA test.

- Return to the example with:
  $Y = \beta_0 + \beta_1 I(\textit{group } B) + \beta_2 I(\textit{group } C) + \varepsilon$.
- Thus category A will have the expected Y being β_0, category B will have $\beta_0 + \beta_1$, and category C will have $\beta_0 + \beta_2$.
- Will see in next lecture how we can test if $\beta_1 = \beta_2 = 0$.
- This is the same as testing $\mu_A = \mu_B = \mu_C$, the population means of groups A, B, and C.
- This is the ANOVA test of several population means (the extension to the two sample t-test).

Multiple Linear Regression: Categorical Covariates

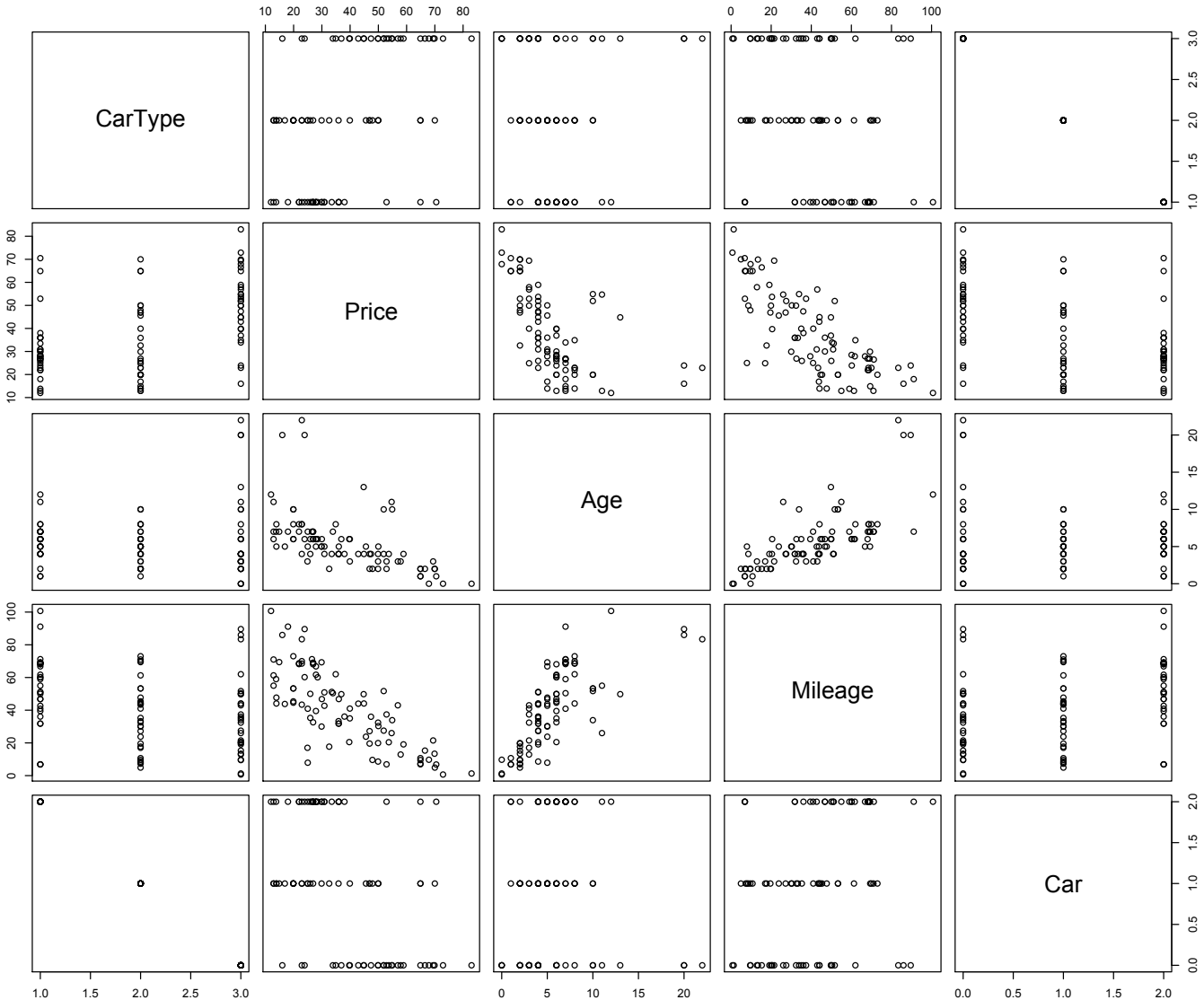
Interactions between several categorical covariates.

- Say we have two categorical covariates, one with K many categories/levels and the other with J many categories/levels.
- We will have a K-1 main effect for the first covariate and J-1 main effects for the second covariate.
- Then the number of interaction terms we will have is (K-1)(J-1).
- Thus we will have (K-1)+(J-1)+(K-1)(J-1) **slope** coefficients in the model for just these two categorical covariates and their interactions.
- The total number of beta (β) coefficients (including intercept) would then be (K-1)+(J-1)+(K-1)(J-1)+**1**.

Multiple Linear Regression: Categorical Covariates

Example: The ThreeCars dataset on the class site has 90 observations of vehicles. Each vehicle has its make recorded (one of 3 categories: Porsche, Jaguar, and BMW), the age of it recorded (how old it is in years, 0 means new), the value of the car (in thousands of dollars), and the mileage the car has (in 1000s of miles).

Multiple Linear Regression: Categorical Covariates



Multiple Linear Regression: Categorical Covariates

- Say we have a response Y that is the value of the car and a single categorical predictor of car type (Porsche, Jaguar, or BMW).
- Say we set the base group to be BMW.
- The model will be
  $Y = \beta_0 + \beta_1 I(\textit{group Jaguar}) + \beta_2 I(\textit{group Porsche}) + \varepsilon.$

Multiple Linear Regression: Categorical Covariates

```
lm(formula = Price ~ CarType, data = ThreeCars)

Residuals:
    Min      1Q  Median      3Q     Max
-34.537 -10.242  -2.333   7.113  40.267

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      30.233      2.775  10.896  < 2e-16 ***
CarTypeJaguar     1.723      3.924   0.439    0.662
CarTypePorsche   20.303      3.924   5.174 1.45e-06 ***
---
Residual standard error: 15.2 on 87 degrees of freedom
Multiple R-squared:  0.2745,Adjusted R-squared:  0.2579
F-statistic: 16.46 on 2 and 87 DF,  p-value: 8.646e-07
```

Multiple Linear Regression: Categorical Covariates

- Can be seen that R picks BMW to be the base group.
- Since there is no quantitative predictor, only a single multi-category covariate, then this is similar to seeing if the means of all the groups are the same or not.
- The predicted price for a BMW without any other covariates given, is \$30,233. The Jaguar is a predicted 1,723 more than that and the Porsche is a predicted 20,303 above the BMW.

Multiple Linear Regression: Categorical Covariates

- Now say we add a quantitative predictor of Mileage, along with interactions with the categorical car type.
- By adding mileage, we will see if the mileage of the car is associated with the value of the car.
- By adding an interaction with car type, we will see if the association between mileage and price differs among the different brands.
- The model will be $Y_i = \beta_0 + \beta_1 Mileage_i + \beta_2 I(group_i\ Jaguar) + \beta_3 I(group_i\ Porsche) + \beta_4 Mileage_i * I(group_i\ Jaguar) + \beta_5 Mileage_i * I(group_i\ Porsche) + \varepsilon$

Multiple Linear Regression: Categorical Covariates

```
lm(formula = Price ~ CarType + Mileage + CarType * Mileage,
data = ThreeCars)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               56.29007    4.15512  13.547  < 2e-16 ***
CarTypeJaguar             -2.06261    5.23575  -0.394  0.69462
CarTypePorsche            14.80038    5.04149   2.936  0.00429 **
Mileage                   -0.48988    0.07227  -6.778 1.58e-09 ***
CarTypeJaguar:Mileage     -0.13042    0.10567  -1.234  0.22057
CarTypePorsche:Mileage    -0.09952    0.09940  -1.001  0.31962

Residual standard error: 8.638 on 84 degrees of freedom
Multiple R-squared:  0.7737,Adjusted R-squared:  0.7602
F-statistic: 57.43 on 5 and 84 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression: Categorical Covariates

- The estimated equation for a BMW is $\hat{Y}_i = 56.29 - 0.49 Mileage_i$.
- The estimated equation for a Jaguar is $\hat{Y}_i = 56.29 - 2.06 - 0.49 Mileage_i - 0.13 Mileage_i = 54.23 - 0.62 Mileage_i$.
- The estimated equation for a Porsche is $\hat{Y}_i = 56.29 + 14.80 - 0.49 Mileage_i - 0.09 Mileage_i = 71.09 - 0.58 Mileage_i$.

Multiple Linear Regression: Categorical Covariates

- The estimated model is
  $\hat{Y} = 56.29 - 0.48 Mileage_i - 2.06 I(group_i\ Jaguar) + 14.80 I(group_i\ Porsche) - 0.13 Mileage_i * I(group_i\ Jaguar) - 0.09 Mileage_i * I(group_i\ Porsche)$
- To interpret mileage, must condition on a type, say Jaguar. An increase in mileage by 1 unit (a 1000 miles) will change the predicted price by -0.48-0.13=-0.61
- And to interpret the effect of type, must condition on Mileage, say 20,000 miles (Mileage=20). Then going from BMW to Porsche will change predicted price by 14.80-0.09*20=13 (thousand).

Multiple Linear Regression: Categorical Covariates

Testing coefficients in multiple linear regression:

$$Y_i = \beta_0 + \beta_1 Mileage_i + \beta_2 I(group_i\ Jaguar) + \beta_3 I(group_i\ Porsche) + \beta_4 Mileage_i * I(group_i\ Jaguar) + \beta_5 Mileage_i * I(group_i\ Porsche) + \varepsilon$$

- Can test single coefficients just like before (R output gives us the two sided p-value).
- Can test if $\beta_1 = 0$ or if $\beta_5 > 0$ for example.
- Testing several coefficients at the same time (such as $\beta_4 = \beta_5 = 0$) will be covered in the next lecture.

Multiple Linear Regression: Categorical Covariates

Categorical explanatory variables in R.

- In R, we can use the as.factor() function to ensure that R recognizes the variable as being categorical.
- Say our categorical variable is called "group", then we would use it as as.factor(group).
- Sometimes R won't recognize the categories (for example when the categories are numbers like group 1, group 2, and so on).
- That is why you will see as.factor(group) in our code (where we replace group with the name of the variable).

Multiple Linear Regression: Categorical Covariates

Residual standard error in multiple linear regression.

- Remember that the residual standard error in R was the square root of MSE.
- In simple linear regression, $\hat{\sigma}_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$.
- Now, in multiple linear regression setting, it is:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{MSE}$$

, where p is the number of covariates we have in the model (number of slope coefficients).

- Or can define it as:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{SSE}{n-k}} = \sqrt{MSE}$$

, where k is the number of β coefficients we have in the model (number of slope coefficients plus one for the intercept coefficient).

- Thus $k = p + 1$.