Prediction!

Contents

Add in plots of data using 1-d simulated data

Ok, so I think I start off a bit loose to get people comfortable and bring out some playing cards.

- I'll point to someone and ask them to guess the suit of the next card I show.
- Reshuffle and repeat giving them five cards/guesses.
- Then I'll repeat with another three-four students.
- # correctly guessed will be noted somewhere

What's the point? How can I best predict the number of card suits the next person will get right?

Prediction

Goal: Predict a new value of a variable

- Ex: Another student will be guessing. Define $Y = \#$ of card suits guessed correctly from the five. What should we guess/predict for the next value of Y?

- Chat with them.
- Lead them to talk about a sample.
- Simulate values of Y using an app, placing them on a histogram or dot plot.
- Lead them to ideas of using something like the sample mean or median as the predicted value.
- Why something like the sample mean or sample median? What are we really trying to do? Find a value that is 'close' to the most values, i.e. something in the center being the most logical thing to do.

Loss function

Let's assume we have a sample of n people that each guessed five cards. Call these values $y_1, y_2, \ldots, y_n$.

Need: A way to quantify how well our prediction is doing... Suppose there is some best prediction, call it c. How do we measure the quality of c?

- Using the idea that we want something 'close' to all points, we find a way to compare each point to our prediction.
- Think about things like:

$$y_1 - c, (y_1 - c)^2, |y_1 - c|$$

$$\sum_{i=1}^{n}(y_i - c), \sum_{i=1}^{n}(y_i - c)^2, \sum_{i=1}^{n}|y_i - c|$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - c), \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2, \frac{1}{n}\sum_{i=1}^{n}|y_i - c|$$

- Quick app to look at how the measures work on the data set already simulated - give the ability to pick a c and at least these metrics, but maybe allow any metric to be typed in...
- In the end an objective function (mean squared error here) must be created to minimize that uses a 'Loss function', and we'll talk about why we'll use the common squared error loss:

$$g(y_1, ..., y_n) = \frac{1}{n}\sum_{i=1}^{n}L(y_i, c) = \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2$$

Can we choose an 'optimal' value for c to minimize this function? Calculus to the rescue!

Steps to minimize a function with respect to c:

1. Take the derivative with respect to c
2. Set the derivative equal to 0
3. Solve for c to obtain the potential maximum or minimum
4. Check to see if you have a maximum or minimum (or neither)

Answer comes out to be $\bar{y}$ as the minimizer.

Big wrap: This means that the sample mean is the best prediction when using squared error loss (root mean square error).

Using a Population Distribution

Rather than using sample data, suppose we think about the theoretical distribution for $Y = \#$ of card suits guessed correctly from the five. What might we use here? What assumptions do we need to make this distribution reasonable?

- $Y \sim Bin(5, 0.25)$ assuming we have independent and identical trials
- This gives
$$p_Y(y) = P(Y = y) = \binom{n}{y}p^y(1-p)^{n-y} = \binom{5}{y}0.25^y 0.75^{5-y}$$
for $y = 0, 1, 2, ..., n$ or $y = 0, 1, 2, 3, 4, 5$

Is there an optimal value c for the **expected value** of the loss function?

That is, can we minimize (as a function of c) $E\left[(Y - c)^2\right]$?

- Maybe visualize this in the same app as above (visualize the quadratic function weighted by the distribution and allow for c to be chosen)
- I think I'll start them out on this one as theory leads to more difficult ideas and I'm not sure if you did general expected values.
- In the end though we end up with np or 1.25.
- I'll show/discuss that this works generally for any distribution $p_Y(y)$ that has a mean
- Discuss the relationship with this and a sample (1/n weight for each point vs $p_Y(y)$ weight for each.)
- **Big idea:** This implies that μ is the best predictor to use if you are considering minimizing the expected squared error loss.

This would be the 2nd (short day) material.

- Recap the big idea of prediction:
  - Need to quantify how well we are doing (squared error loss and MSE)
  - Sample mean is optimal if we have a sample
  - Given a theoretical distribution, expected squared error loss is optimized at the mean of the distribution
- Introduce next material with minesweeper, because it is now browser based, nostalgia on my part, and it seems somewhat fun https://minesweeper.online/game/938135731
  - Each student will be assigned a certain number of mines for the board (15x40).
  - They'll click on the first square just below the smiley face. They'll continue to click down one block at a time until they hit a bomb.
  - They'll record in a shared spreadsheet their number of blocks down the first bomb appears.
  - Each person should play 10 games and put their data in.
- Now we'll discuss how we could predict the number of blocks until the first bomb as a function of the number of bombs.
- We'll read in the data to R and do some plotting (I'm not sure what the relationship will be exactly but I'd guess not super linear).

Relating Explanatory Variables in Prediction

Y is a random variable and we'll consider the x values fixed (we'll denote this as $Y|x$). We hope to learn about the relationship between Y and x.

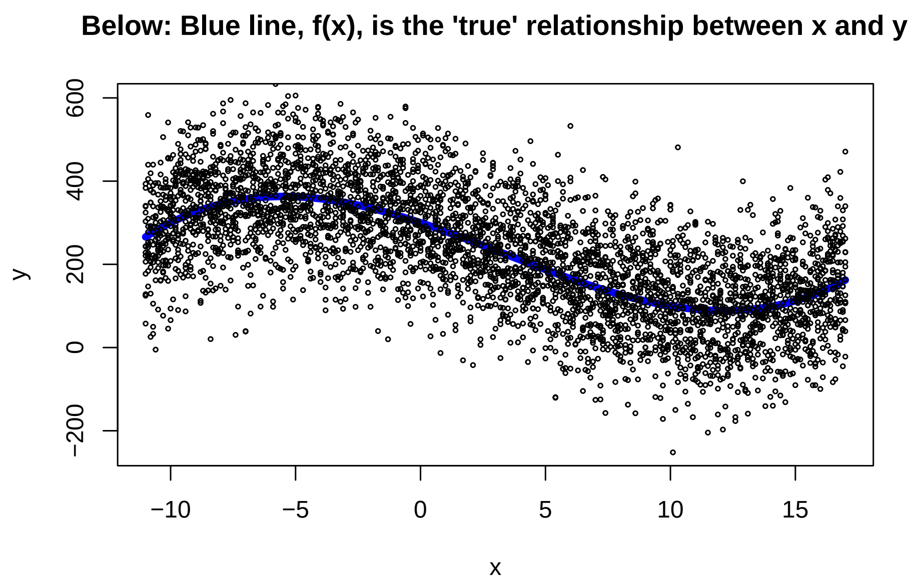
When we considered just Y by itself and used squared error loss, we know that $E(Y) = \mu$ minimizes

$$E\left[(Y - c)^2\right]$$

as a function of c. Given data, we used $\hat{\mu} = \bar{y}$ as our prediction.

** Insert Plot **

Harder (and more interesting) problem is to consider predicting a (response) variable Y as a function of an explanatory variable x.



Now that we have an x, $E(Y|x)$ will minimize

$$E\left[(Y-c)^2|x\right]$$

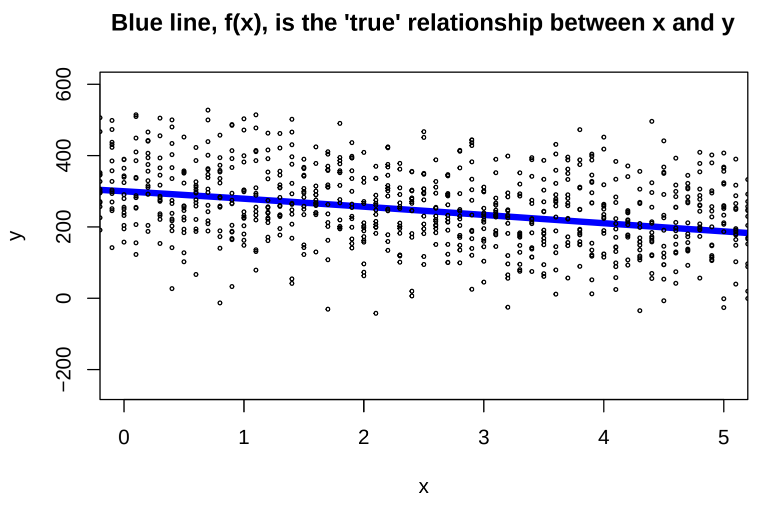
We can call this true unknown value $E(Y|x) = f(x)$. That is, the average value of Y will now be considered as a function of x.

Given observed Y's and x's, we can estimate this function as $\hat{f}(x)$ (think $\bar{y}$ from before). This $\hat{f}(x)$ will minimize

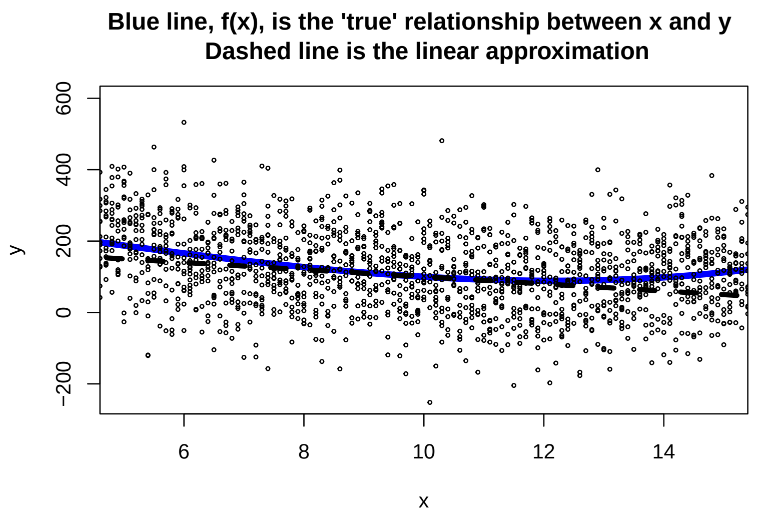
$$g(y_1,...,y_n|x_1,...,x_n) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Approximating $f(x)$

Although the true relationship is most certainly nonlinear, we may be ok approximating the relationship linearly. For example, consider the same plot as above but between 0 and 5 only:



That's pretty linear. Consider plot between 5 and 15:



Line still does a reasonable job and is often used as a basic approximation.

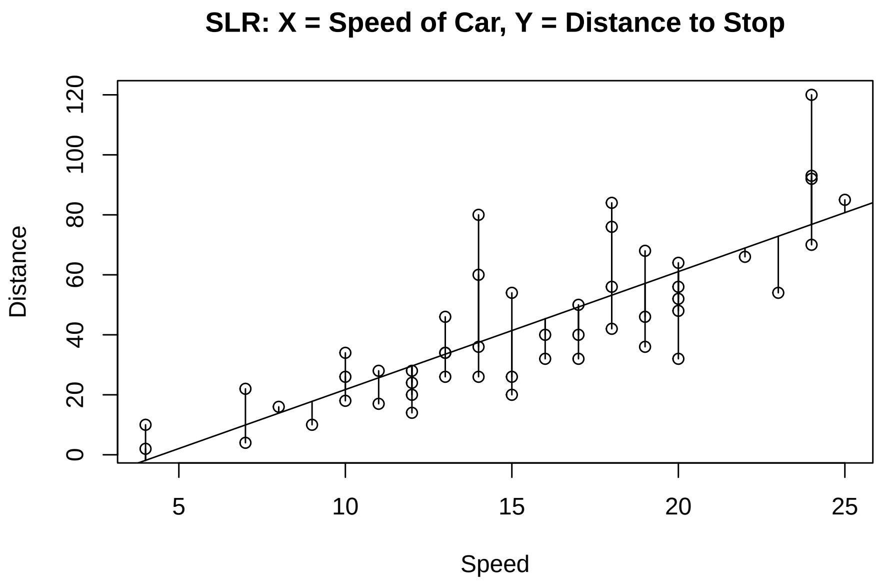
Linear Regression Model

The (fitted) linear regression model uses $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. This means we want to find the optimal values of $\hat{\beta}_0$ and $\hat{\beta}_1$ from:

$$g(y_1, ..., y_n | x_1, ..., x_n) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

This equation is often called the 'sum of squared errors (or residuals)' or the 'residual sum of squares'.

Update this with the data from minesweeper but show a plot like this or perhaps just over a portion that looks like it could be modeled linearly.



Calculus allows us to find the 'least squares' estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$ in a nice closed-form!

Do they know partial derivatives? I'm not sure. I think we'll be running low on time here anyway, so maybe I'll just talk about the idea of how to get them, set up the equations and then just give the answers.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

Ok, day 3 here. Get into R and do some model fitting and predicting. Start with a quick recap here

Fitting a Linear Regression Model in R

Recap: Our goal is to predict a value of Y while including an explanatory variable x. We are assuming we have a sample of (x_i, y_i) pairs, $i = 1, ..., n$.

The Simple Linear Regression (SLR) model can be used:

$$\hat{f}(x_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where

- y_i is our response for the i^{th} observation
- x_i is the value of our explanatory variable for the i^{th} observation
- β_0 is the y intercept
- β_1 is the slope

The best model to use if we consider squared error loss has

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

called the 'least squares estimates'.

Data Intro

This dataset contains information about used motorcycles and their cost.

This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning. The columns in the given dataset are as follows:

- name
- selling price -year -seller type -owner -km driven -ex showroom price

The data are available to download from this URL:
https://www4.stat.ncsu.edu/~online/datasets/bikeDetails.csv

I think the exploration and modeling would be better to do live and ask them for input as we go through it. I'll put some stuff here and then we can talk about it.

Read in Data and Explore!

```
library(tidyverse)
bikeData <- read_csv("https://www4.stat.ncsu.edu/~online/datasets/bikeDetails.csv")
bikeData <- bikeData %>% tidyr::drop_na()
bikeData %>%
  select(selling_price, year, km_driven, ex_showroom_price, name, everything())
```

```
## # A tibble: 626 x 7
##    selling_price  year km_driven ex_showroom_price name         seller_type owner
##            <dbl> <dbl>     <dbl>             <dbl> <chr>        <chr>       <chr>
##  1        150000  2018     12000            148114 Royal Enfi~ Individual  1st ~
##  2         65000  2015     23000             89643 Yamaha Faz~ Individual  1st ~
##  3         18000  2010     60000             53857 Honda CB T~ Individual  1st ~
##  4         78500  2018     17000            87719 Honda CB H~ Individual  1st ~
##  5         50000  2016     42000            60122 Bajaj Disc~ Individual  1st ~
##  6         35000  2015     32000            78712 Yamaha FZ16 Individual  1st ~
##  7         28000  2016     10000            47255 Honda Navi  Individual  2nd ~
##  8         80000  2018     21178            95955 Bajaj Aven~ Individual  1st ~
##  9        365000  2019      1127           351680 Yamaha YZF~ Individual  1st ~
## 10         25000  2012     55000            58314 Suzuki Acc~ Individual  1st ~
## # ... with 616 more rows
```

Our 'response' variable here is the `selling_price` and we could use the variable `year`, `km_driven`, or `ex_showroom_price` as the explanatory variable. Let's make some plots and summaries to explore.

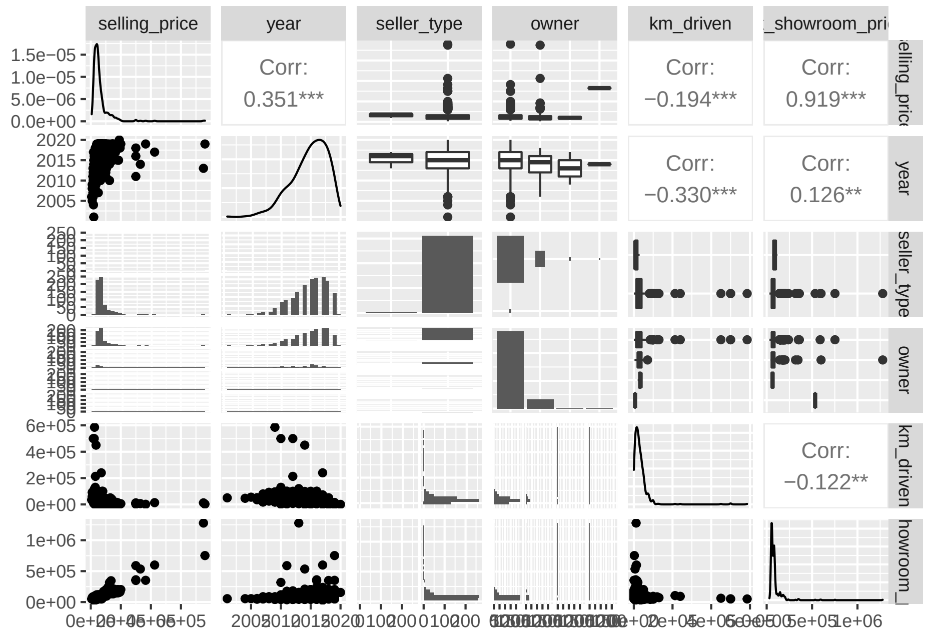
```
summary(bikeData)
```

```
##      name           selling_price         year      seller_type
##  Length:626         Min.   :  6000   Min.   :2001   Length:626
##  Class :character   1st Qu.: 30000   1st Qu.:2013   Class :character
##  Mode  :character   Median : 45000   Median :2015   Mode  :character
##                     Mean   : 59445   Mean   :2015
##                     3rd Qu.: 65000   3rd Qu.:2017
##                     Max.   :760000   Max.   :2020
##     owner             km_driven      ex_showroom_price
##  Length:626         Min.   :   380   Min.   :  30490
##  Class :character   1st Qu.: 13031   1st Qu.:  54852
##  Mode  :character   Median : 25000   Median :  72753
##                     Mean   : 32672   Mean   :  87959
##                     3rd Qu.: 40000   3rd Qu.:  87032
##                     Max.   :585659   Max.   :1278000
```

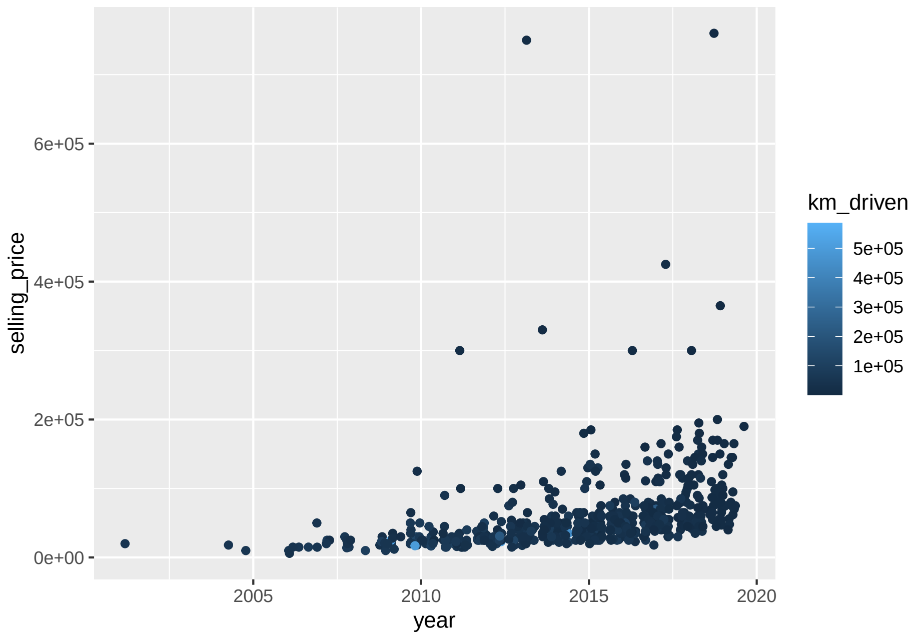
```
bikeData %>%
  group_by(owner) %>%
  summarize(mean = mean(selling_price),
            median = median(selling_price),
            sd = sd(selling_price), IQR =
              IQR(selling_price))
```

```
## # A tibble: 4 x 5
##   owner        mean median      sd   IQR
##   <chr>       <dbl>  <dbl>   <dbl> <dbl>
## 1 1st owner  58432.  45000  51125. 35000
## 2 2nd owner  64795.  35000 104861. 30750
## 3 3rd owner  39333.  40000  17010. 17000
## 4 4th owner 330000  330000     NA      0
```

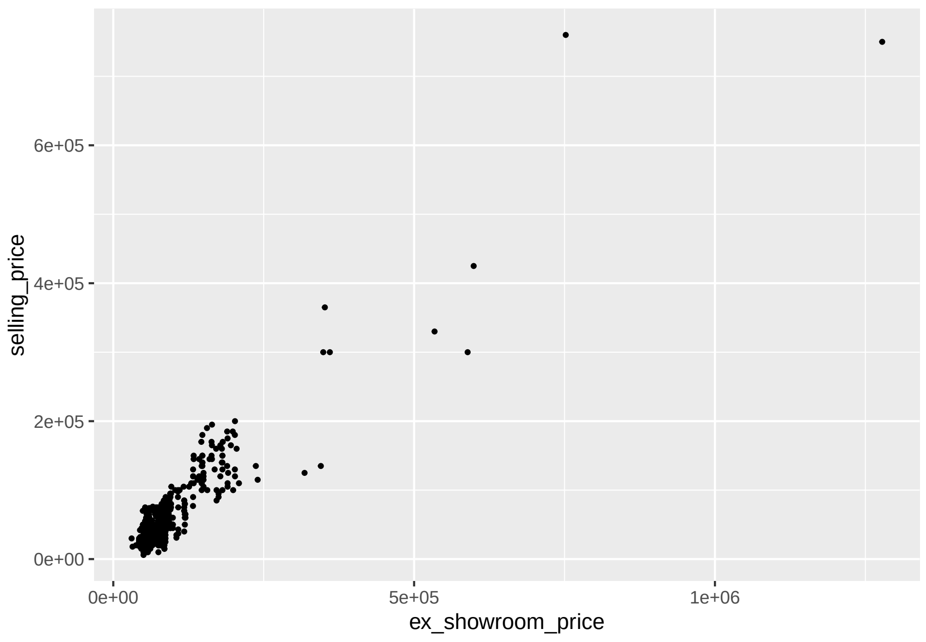
```
library(GGally)
ggpairs(bikeData %>%
          select(-name))
```



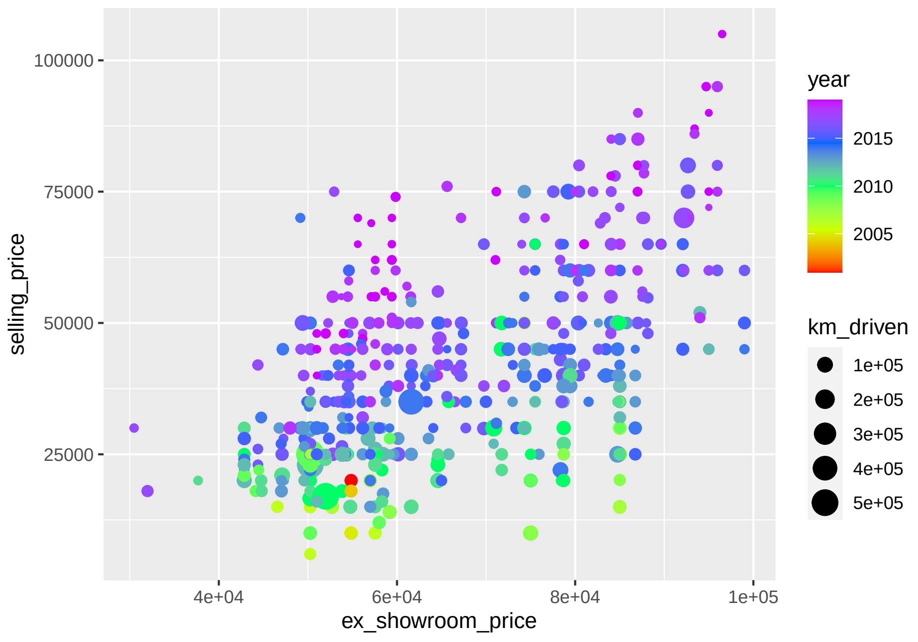
```r
g <- ggplot(data = bikeData, aes(y = selling_price))
g + geom_jitter(aes(x = year, color = km_driven))
```



```r
g + geom_point(aes(x = ex_showroom_price), size = 0.75)
```



```r
g <- ggplot(data = filter(bikeData, ex_showroom_price < 100000), aes(y = selling_price))
g +
  geom_point(aes(x = ex_showroom_price, color = year, size = km_driven)) +
  scale_color_gradientn(colours = rainbow(5))
```



'Fitting' the Model

- Basic *linear model* fits done with `lm()`. First argument is a `formula`:

$$\textit{response variable} \sim \textit{modeling variable(s)}$$

We specify the modeling variable(s) with `+` separating variables.

```
fit <- lm(selling_price ~ ex_showroom_price, data = bikeData)
fit
```

```
##
## Call:
## lm(formula = selling_price ~ ex_showroom_price, data = bikeData)
##
## Coefficients:
##       (Intercept)  ex_showroom_price
##        -3010.6984             0.7101
```

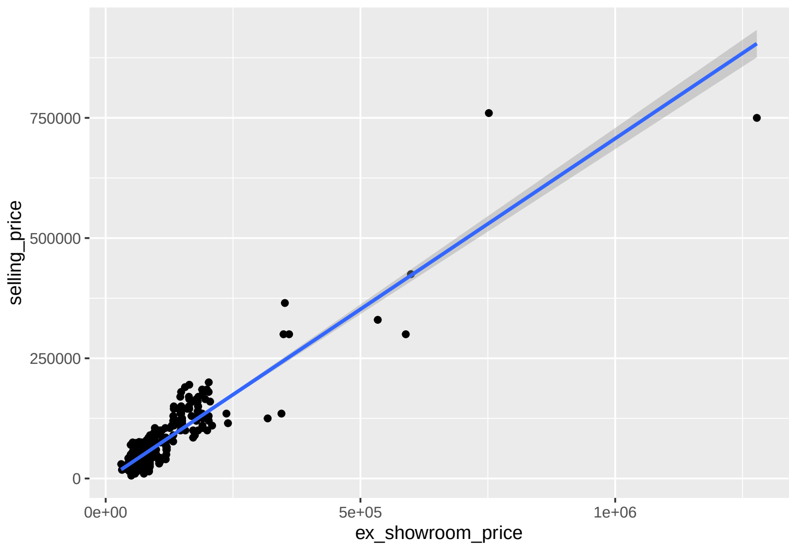
- We can easily pull off things like the coefficients.

```
coefficients(fit) #helper function
```

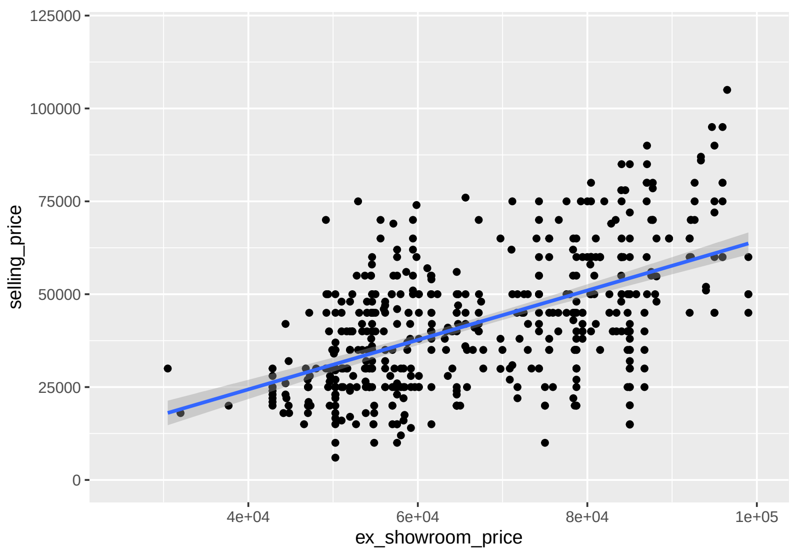
```
##       (Intercept) ex_showroom_price
##     -3010.6984021         0.7100588
```

We can also look at the fit of the line on the graph.

```
ggplot(bikeData, aes(x = ex_showroom_price, y = selling_price)) +
  geom_point() +
  geom_smooth(method = "lm")
```



```
ggplot(bikeData, aes(x = ex_showroom_price, y = selling_price)) +
  geom_point() +
  geom_smooth(method = "lm") +
  scale_x_continuous(limits = c(25000, 100000)) +
  scale_y_continuous(limits = c(0, 120000))
```



Predicting!

Now can predict the `selling_price` for a given `ex_showroom_price`.

```
predict(fit, newdata = data.frame(ex_showroom_price = c(50000, 75000)))
```

```
##        1        2
## 32492.24 50243.71
```

Error Assumptions

Not sure about adding this in but it would help connect things with what they are doing in class…

We often assume that we observe our response variable Y as a function of the line plus random errors:

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

where the errors come from a Normal distribution with mean 0 and variance σ^2 ($E_i \overset{iid}{\sim} N(0, \sigma^2)$)

If we do this and use probability theory (maximum likelihood), we will get the same estimates as above! If the normality assumption is reasonable, we then also gain knowledge of the distribution of our estimators of the intercept and slope ($\hat{\beta}_0$ and $\hat{\beta}_1$).

These allow us to create confidence intervals or conduct hypothesis tests. However, these assumptions aren't needed to simply do prediction (but are useful to get error bounds on our prediction

- Get SE for prediction

```
predict(fit, newdata = data.frame(ex_showroom_price = c(50000, 75000)), se.fit = TRUE)
```

```
## $fit
##        1        2
## 32492.24 50243.71
##
## $se.fit
##         1         2
## 1054.7005  960.2046
##
## $df
## [1] 624
```

```
##
## $residual.scale
## [1] 23694.8
```

- Get confidence interval for mean response

```
predict(fit, newdata = data.frame(ex_showroom_price = c(50000, 75000)),
        se.fit = TRUE, interval = "confidence")
```

```
## $fit
##        fit      lwr      upr
## 1 32492.24 30421.05 34563.44
## 2 50243.71 48358.09 52129.34
##
## $se.fit
##         1         2
## 1054.7005  960.2046
##
## $df
## [1] 624
##
## $residual.scale
## [1] 23694.8
```

- Get prediction interval for new response

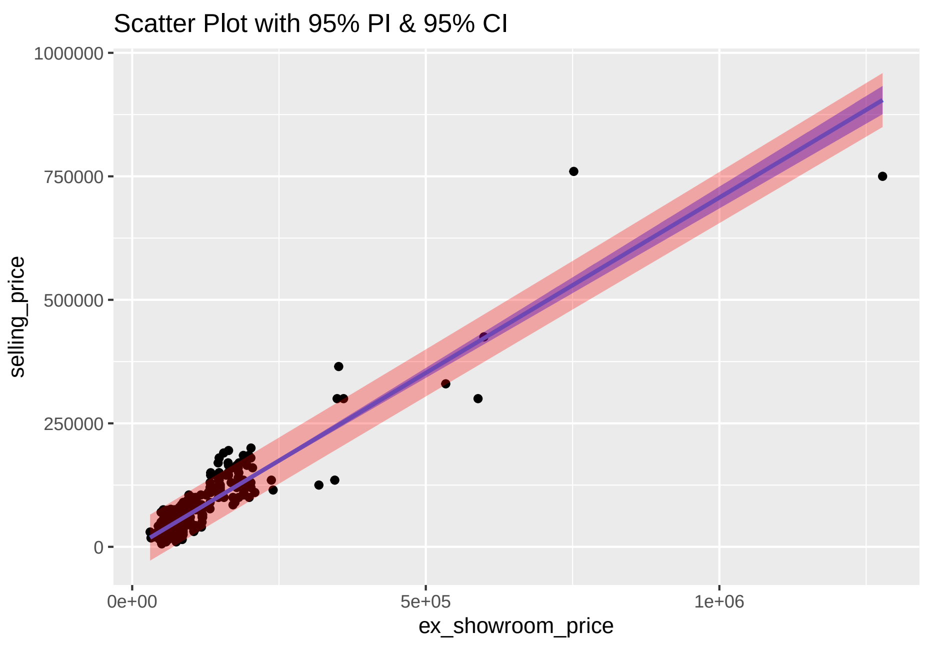
```
predict(fit, newdata = data.frame(ex_showroom_price = c(50000, 75000)),
        se.fit = TRUE, interval = "prediction")
```

```
## $fit
##        fit        lwr      upr
## 1 32492.24 -14085.045 79069.53
## 2 50243.71   3674.309 96813.12
##
## $se.fit
##         1         2
## 1054.7005  960.2046
##
## $df
## [1] 624
##
## $residual.scale
## [1] 23694.8
```

- Can see the confidence and prediction bands on the plot:

```
library(ciTools)
bikeData <- bikeData %>%
  add_pi(fit, names = c("lower", "upper"))

ggplot(bikeData, aes(x = ex_showroom_price, y = selling_price)) +
  geom_point() +
    geom_smooth(method = "lm", fill = "Blue") +
    geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.3, fill = "Red") +
    ggtitle("Scatter Plot with 95% PI & 95% CI")
```



Multiple Linear Regression

We can add in more than one explanatory variable using the `formula` for `lm()`. The ideas all follow through!

```
fit <- lm(selling_price ~ ex_showroom_price + year + km_driven, data = bikeData)
fit
```

```
##
## Call:
## lm(formula = selling_price ~ ex_showroom_price + year + km_driven,
##     data = bikeData)
##
## Coefficients:
##       (Intercept)  ex_showroom_price               year          km_driven
##        -9.429e+06          6.863e-01          4.679e+03         -1.053e-02
```

To predict we now need to specify values for all the explanatory variables.

```
data.frame(ex_showroom_price = c(50000, 75000),
                                  year = c(2010, 2011),
                                  km_driven = c(15000, 10000))
```

```
##   ex_showroom_price year km_driven
## 1             50000 2010     15000
## 2             75000 2011     10000
```

```
predict(fit, newdata = data.frame(ex_showroom_price = c(50000, 75000),
                                  year = c(2010, 2011),
                                  km_driven = c(15000, 10000)),
        se.fit = TRUE, interval = "confidence")
```

```
## $fit
##        fit       lwr      upr
## 1 11118.83  7914.815 14322.85
## 2 33007.56 30202.482 35812.63
##
```

```
## $se.fit
##        1        2 
## 1631.552 1428.402 
## 
## $df
## [1] 622
## 
## $residual.scale
## [1] 19011.31
```

Difficult to visualize the model fit though!

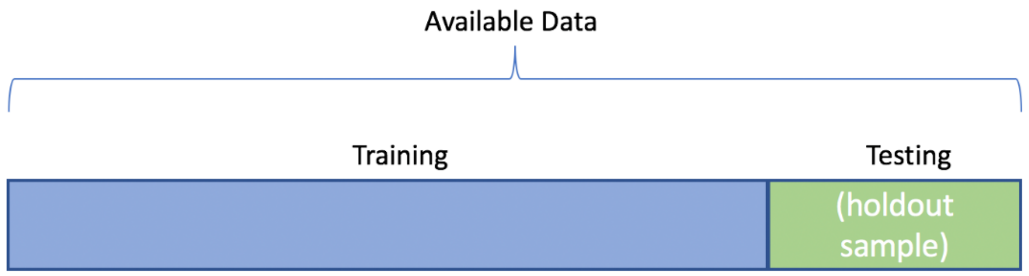
Not sure if we'll have time for this or if these would go in day 4. Will depend a bit on the inclusion/exclusion of the confidence/prediction bands.

Evaluating Model Accuracy

Which model is better? Ideally we want a model that can predict **new** data better, not the data we've already seen. We need a **test** set to predict on. We also need to quantify what me mean by better!

Training and Test Sets

We can split the data into a **training set** and **test set**.



- On the training set we can fit (or train) our models. The data from the test set isn't used at all in this process.
- We can then predict for the test set observations (for the combinations of explanatory variables seen in the test set). Can then compare the predicted values to the actual observed responses from the test set.

Split data randomly:

```r
set.seed(1)
numObs <- nrow(bikeData)
index <- sample(1:numObs, size = 0.7*numObs, replace = FALSE)
train <- bikeData[index, ]
test <- bikeData[-index, ]
```

Fit the models on the training data only.

```r
fitSLR <- lm(selling_price ~ ex_showroom_price , data = train)
fitMLR <- lm(selling_price ~ ex_showroom_price + year + km_driven, data = train)
```

Predict on the test set.

```r
predSLR <- predict(fitSLR, newdata = test)
predMLR <- predict(fitMLR, newdata = test)
tibble(predSLR, predMLR, test$selling_price)
```

```
## # A tibble: 188 x 3
##    predSLR predMLR `test$selling_price`
##      <dbl>   <dbl>                <dbl>
##  1  58966.  73988.                78500
##  2 244701. 258749.               365000
##  3  80221.  58088.                40000
##  4 101463. 114943.               150000
##  5  90603. 103928.               120000
##  6  28477.  39802.                42000
##  7  37743.  53770.                60000
##  8  40588.  56071.                45000
##  9  32173.  34042.                28000
## 10 101463. 114974.               140000
## # ... with 178 more rows
```

Root Mean Square Error

Which is better?? Can use squared error loss to evaluate! (Square root of the mean squared error loss is often reported instead and is called RMSE or Root Mean Square Error.)

```
sqrt(mean((predSLR - test$selling_price)^2))
```

```
## [1] 23026.97
```

```
sqrt(mean((predMLR - test$selling_price)^2))
```

```
## [1] 17439.81
```

MLR fit does much better at predicting!

Another Modeling Approach

Day 4:

I was going to either talk about kNN or a basic regression tree here. Only issue is that neither really uses stuff from your class explicitly. Thoughts? I'd like to do some fun-ish thing to introduce it. Talk a little about the math. Fit it in R and then predict with it.

Kaggle Competition

Last day would be discussion the Kaggle competition here: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

Perhaps split them into teams of two or three. Have them fit some models and submit to see how they do.

I'd have some basic code they could use to read the data in, split it into training and test so they can select their model, create predictions, and output them to a csv for submission (evaluation is log prediction minus log of the actual value).