Final project

- After today's presentation exercise, the rest of the tutorial is for the final project.
- Forming your group for the final project by the end of this tutorial. Groups will consist of 3-4 students.

Linear Regression

- What types of learning does linear regression belong to?
- Supervised learning.
- Simple linear regression (SLR) model: with only one independent/predictor variable.

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

- Y_i: response variable (or dependent variable, target variable, ...) for i^{th} observation
- x_i: independent variable (or predictor, covariate, feature, input, ...) for i^{th} observation
- β_0: intercept parameter
- β_1: slope parameter
- ϵ_i: random error term for i^{th} observation

Learning

- Use Least square method to learn the coefficient.

Sum of squared errors (Residual sum of squares)

- This is our objective function that we are trying to minimize.

Find the line (i.e. find β_0 and β_1) which minimizes the sum of squared errors for the n observations: $\sum_{i=1}^{n} \epsilon_i^2$

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Goal: Find the values of the intercept β_0 and slope β_1 that minimize the function $f(\beta_0, \beta_1)$. We treat the data $(x_1, y_1), \ldots, (x_n, y_n)$ as constants to do this.

- Taking the **partial derivative** with respect to β_0 and β_1 on sum of squared errors *f* for the *n* observations.

$$\frac{\partial f}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial f}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

- By doing so, we are finding the values of β_0 and β_1 which minimize the on sum of squared errors.

Partial derivative

- If you have a multivariable function, e.g. the sum of square error, which is a function on β_0 and β_1:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- The partial derivative on one variable is just a regular derivative while treating the other variables as a constant.

Result

- By solving the two equations of partial derivatives equals to zero (two unknow and two equations), we get:

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

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the **least squares estimators** of β_0 and β_1

- Difference between β and $\hat{\beta}$?
- β theoretical value (you never know the true value), $\hat{\beta}$ the estimate

- Difference between y_i and $\hat{y}_i$?
- y_i: the true value of the data point: x_i; $\hat{y}_i$ prediction.

- Difference between ϵ_i and e_i ?
- ϵ_i random error of you model, and
- e_i:

The difference between the observed and predicted value of y for the i^{th} observation is called the **residual** $e_i = y_i - \hat{y}_i$

Interpreting the results

- The **slope** $\hat{\beta}_1$ is the average change in y for a 1-unit change in x
- The **intercept** $\hat{\beta}_0$ is the average of y when $x_i = 0$ (often this doesn't make sense, but tells us the height of the line).

Use the `lm()` function to fit a linear regression model

Same syntax as we used to fit classification trees to specify the response and predictors

Response variable *Predictor variable*

```
mod_height <- lm(height ~ footLength, data = heights)
summary(mod_height)$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 64.125614 11.4850534 5.583397 2.122303e-06
## footLength   4.291253  0.4459507 9.622708 9.833229e-12
```

`(Intercept)` is the **estimate of** β_0 (i.e. $\hat{\beta}_0$)

`footLength` is the **estimate of** β_1 (i.e. $\hat{\beta}_1$)

- What is your final model based on the R output?
- height = 64.125614 + 4.291253*footLength

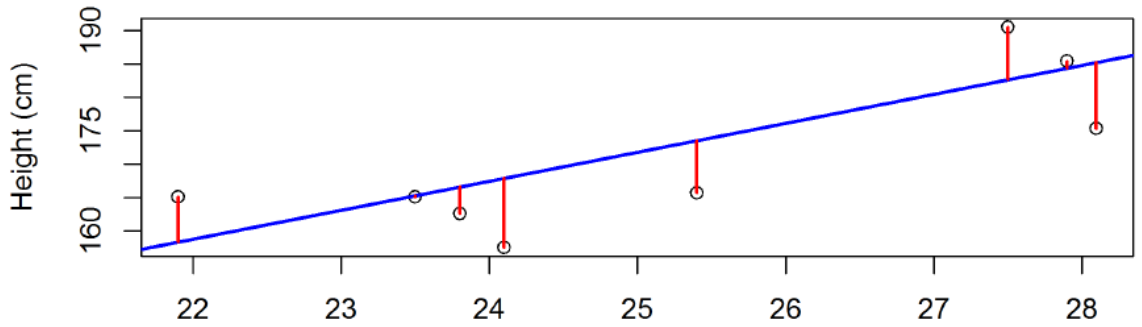
Testing

- How do we determine whether our linear regression model is good or bad, what is a good measurement?
- By the root mean squared error (RMSE)

The **root mean squared error (RMSE)** measures the prediction error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Taking a square root means that RMSE is in the same units (and scale) as y



Overfitting

- Fit to the training data too well, generalize badly to the new data, e.g. testing data or the data we haven’t seen before.

> If there is a big difference between RMSE for predictions based on training and testing data, this suggest our model "learned" the training data "too well": this is called **overfitting**.

- Example from the slides

```
# Fit model to training data
mod_height_train <- lm(height ~ footLength, data=train)
```

RMSE for predictions of individuals in the testing dataset

```
# Make predictions for testing data using training model
yhat_test <- predict(mod_height_train, newdata = test)
y_test <- test$height;                    n_test <- nrow(test)
# RMSE for predictions in testing dataset
sqrt(sum((y_test - yhat_test)^2) / n_test)
```

```
## [1] 7.003343
```

Vectorized computation, so that no for loop needed for computing RMSE.

RMSE for predictions of individuals in the training dataset

```
# Make predictions for training data using training model
yhat_train <- predict(mod_height_train, newdata = train)
y_train <- train$height;                    n_train <- nrow(train)
# RMSE for predictions in testing dataset
sqrt(sum((y_train - yhat_train)^2) / n_train)
```

```
## [1] 4.894985
```

RMSE for predictions on training data: `4.8949854`

RMSE for predictions on testing data: `7.0033435`

Here there is a relatively big difference: the RMSE is over 40% larger for predictions on the testing dataset than on the training dataset

This suggests that our predictive model may be overfitting the training data and is not very useful for to make predictions for new observations

Presentation exercise

- Pick one of the following topics regarding your homework assignment and prepare a **5** minutes presentation.
- Hand in your draft for your presentation.
- As usual, I will assign students to provided feedback to each group.

Topic 1

- Questions 1a and 1c
  - Describe your plot produced in question 1a. Make sure to note the x- and y-axis and to describe the association you observe, if any. E.g. the association linear, positive, negative, strong, weak, etc.?
  - What is the correlation between head size and brain weight? Make sure to explain how you calculated this value and what it means; i.e., provide an interpretation of the value.
  - Does this make sense based on your prior expectations? Are there any other variables you think may be important factors influencing brain weight?
- Do there appear to be many outliers? Why might this matter?

Topic 2

- Questions 1d-f
  - Provide a simple linear regression equation for the association between head size and brain weight. Explain what each part of the model means in lay terms.
  - Based on your answer to part e, report the estimated values of your model and provide an interpretation of these values.
- How well does your model fit the data? Explain what the coefficient of determination means and provide an interpretation.

Topic 3

- Question 2c
  - Present your regression model of msrp on year based on the training set.
  - What is the model equation and estimated values?
  - What is the coefficient of determination? Explain what these values mean and an interpretation in lay terms.
  - How well does your model perform?

Topic 4

- Question 2d
  - What is your predicted 2013 msrp for a 2010 model hybrid vehicle? Make sure to present your regression equation, including all coefficients.
  - Suppose the actual 2013 msrp for this 2010 model hybrid vehicle was $27,000. What is the residual?
  - Provide an interpretation in lay terms. Is this a large difference?
  - Based on previous work done in this question, why do you think this may be the case? Hint: Think about how well the model fits the data, if there may be other important factors, etc.