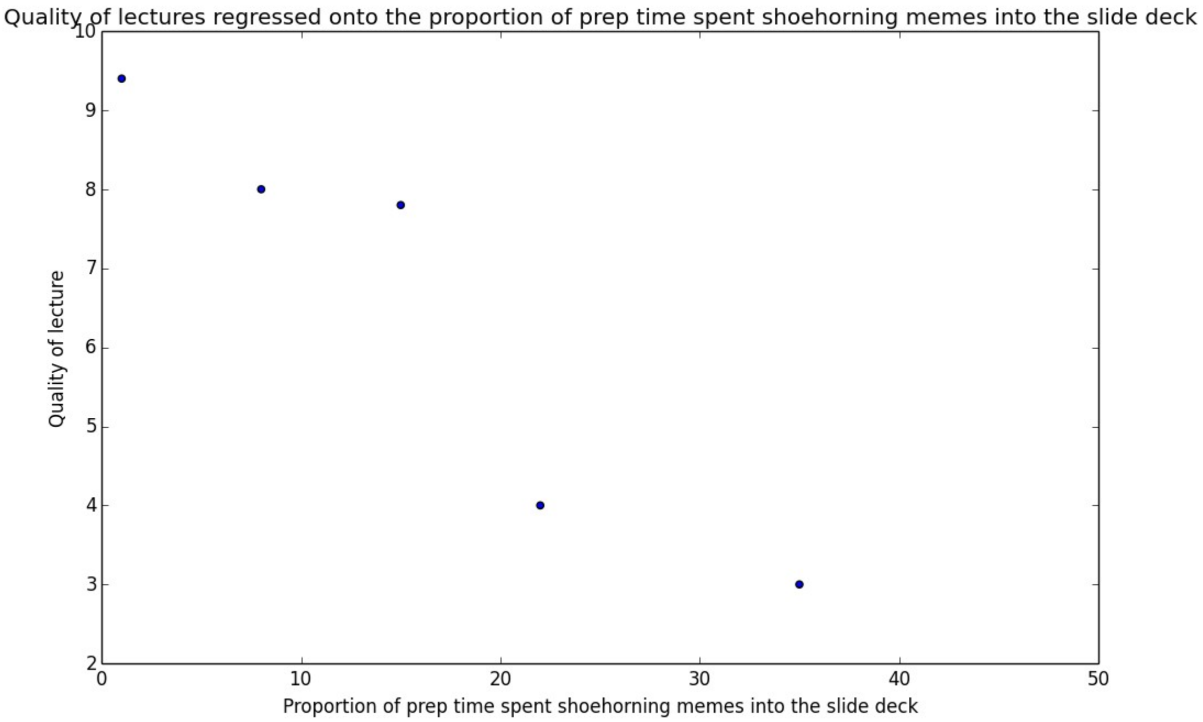
Lab 2a

In this lab we will fit a simple linear regression model. We will regress the quality of my lectures onto the proportion of prep time that I spent shoehorning memes into the lecture's slide deck.

We will use the following data set:

Training Instance	Proportion of preparation time spent shoehorning memes into the lecture's slide deck	Quality of lecture
1	8	8
2	15	7.8
3	22	4
4	1	9.4
5	35	3



Recall that the model for simple linear regression is given by the following equation:

$$y = \alpha + \beta x$$

where y is the value of the response variable, x is the value of the explanatory variable, α is the y-intercept, and β is the coefficient.

Our goal is to find the values of the model parameters that minimize the value of the residual sum of squares cost function, which is given by the following:

$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

where y_i is the value of the response variable for the ith training instance, and $f(x_i)$ is the predicted value of the response variable for the ith training instance.

First we can solve for the value of β using the following:

$$\beta = \frac{cov(x,y)}{var(x)}$$

Variance is a measure of how far a set of values are spread out. If all of the numbers in the set are equal, the variance of the set is zero. A small variance indicates that the numbers are near the mean of the set, while a set containing numbers that are far from the mean and each other will have a large variance. Variance can be calculated using the following equation:

$$var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Question 1. Calculate the variance for the explanatory variable.

$$var(x) = \frac{(8-16.2)^2 + (15-16.2)^2 + (22-16.2)^2 + (1-16.2)^2 + (35-16.2)**2)}{4} = 171.7$$

Covariance is a measure of how much two variables change together. If the variables increase together, their covariance is positive. If one variable tends to increase while the other decreases, their covariance is negative. If there is no linear relationship between the two variables, their covariance will be equal to zero; they are uncorrelated. Covariance can be calculated using the following:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Question 2. Calculate the covariance of the explanatory and response variables.

$$cov(x,y) = \frac{(8-16.2)*(8-6.44) + (15-16.2)*(7.8-6.44) + \cdots + (35-16.2)*(3-6.44)}{4}$$

Question 3. Calculate the value of β.

$$\beta = \frac{-34.56}{171.7} = -0.20128130460104837$$

Now that we have solved for beta, we can solve for \alpha using \beta and the centroid as follows:

$$\alpha = \bar{y} - \beta\bar{x}$$

Question 4. Calculate the value of \alpha.

$\alpha = 6.44 - (-0.2012 * 16.2) = 9.70075713$

Now assume that we have the following test set.

Training Instance	Proportion of preparation time spent shoehorning memes into	Quality of lecture

	the lecture's slide deck	
1	11	7.9
2	10	7.2
3	3	9.2
4	24	4

Question 5. Predict the values of the response variable for the test instances.

Training Instance	Proportion of preparation time spent shoehorning memes into the lecture's slide deck	Quality of lecture	Predicted quality of lecture
1	11	7.9	7.48666278
2	10	7.2	7.68794409
3	3	9.2	9.09691322
4	24	4	4.87000582

Recall that r-squared measures how well the observed values of the response variable are predicted by the model. More concretely, r-squared is the proportion of the variance in the response variable that is explained by the model. Values for r-squared range from zero to one. An r-squared of one indicates that the response variable can be predicted without any error using the model. An r-squared score of one half indicates that half of the variance in the response variable can be predicted from the model. R-squared is given by the following:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where

$$SS_{res} = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Question 6. Calculate r-squared for the predictions.

$$\bar{y} = (7.9 + 7.2 + 9.2 + 4)/4 = 7.075$$
$$SS_{res} = (7.9 - 7.48666278)^2 + (7.2 - 7.68794409)^2 + (9.2 - 9.09691322)^2 + (4 - 4.87000582)^2$$
$$SS_{res} = 1.17647410344789855$$
$$SS_{tot} = (7.9 - 7.075)^2 + (7.2 - 7.075)^2 + (9.2 - 7.075)^2 + (4 - 7.075)^2 = 14.667499999999997$$
$$R^2 = 1 - \frac{1.17647}{14.66749} = 0.9197906390254911$$

Question 7. In a sentence, interpret the r-squared score for our model.
Most of the variance in the response variable is explained by the model.