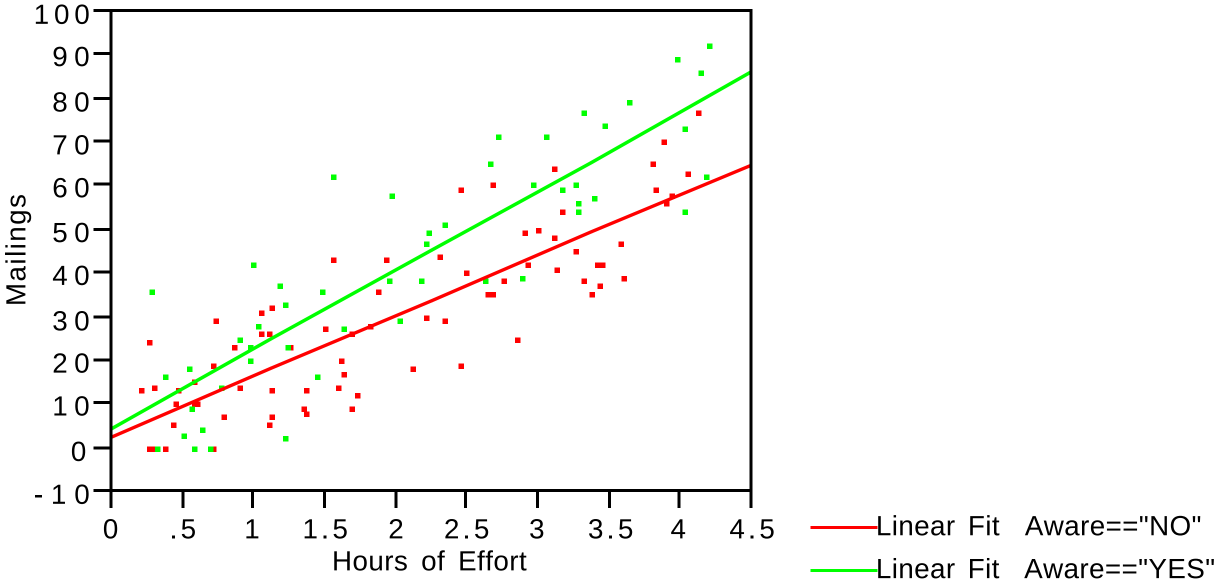
Regression with a Single, Two-Level Categorical Predictor

FedEx is promoting a type of service called "courier packs". Was the sales effort more effective for those customers who were previously aware of the product?

The response of interest is the number of courier packages (packages holding up to 2 pounds) used by a customer during a month. FedEx sales people spent time with each of the customers in this data. The goal of this effort was to increase the customers' use of courier packages. The time was spent visiting the customer, preparing special materials, and coordinating other services for the customer. While some customers were already aware that FedEx offered courier packs for larger mailings, it was a new product to many others. Given the expense of these contacts, FedEx could ask two questions (and probably many more):

> How many package shipments were generated by a typical contact hour?
>
> Was the promotion more effective for the customers who were already aware of FedEx's business, or was it more effective to those who had not known of the product before?

The following plot shows the mailings per month plotted on the hours of effort for 125 customers (75 not aware, 50 aware). The lines appear to have similar intercepts, but different slopes. "What do the slopes represent?", and "Are they different?" are our two questions. The aware customers give the higher fitted line.



Here are the fits for the two groups

Aware=="NO"	Mailings = 2.45 + 13.83 Hours of Effort
	RMSE 10.2 SE=
Aware=="YES"	Mailings = 4.16+ 18.13 Hours of Effort
	RMSE 12.4

The slope in the fit in the unaware group (“NO”) implies that about 14 packages per month were mailed per hour of contact. In the aware group, each hour of contact led to slightly more than 18 per hour, about four more per hour. The intercepts represent the number of packages per month with no direct contact of this type and are similar in the two cases.

To judge the significance of this comparison of the effects of this promotion in the two groups, we need to use a multiple regression with *Hours of Effort* and *Aware* as the two predictors. Since we are particularly interested in the difference of the slopes (and from the difference in the slopes seen in the plot), we need their interaction as well. Here is the resulting fit.

Summary of Fit

RSquare	0.767
Root Mean Square Error	11.2
Observations	125

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.31	2.00	1.66	0.1000
Hours of Effort	15.98	0.84	18.99	<.0001
Aware[NO]	-5.20	1.02	-5.09	<.0001
Aware[YES]	5.20	1.02	5.09	<.0001
(Hours of Effort-2.02)*Aware[NO]	-2.15	0.84	-2.56	0.0117
(Hours of Effort-2.02)*Aware[YES]	2.15	0.84	2.56	0.0117

Since the interaction is significant (p-value = 0.0117), we conclude that the slopes are significantly different. How many more packages are used on average per month per hour of contact? To answer this, compare the slopes. We find a difference of twice either interaction term. The aware group (Yes) has a slope of 15.98 + 2.15 = 18.13 packages shipped monthly/hour whereas the unaware group has a slope of 15.98 – 2.15 = 13.83. The difference is then just twice the interaction slope, or 2(2.15) = 4.3 packs/hour.

A confidence interval for the difference is then obtained as follows. First get a confidence interval for the interaction slope, then multiply the endpoints by 2 since what we want is an interval for twice the slope, or the rather wide range

$2[2.15 \pm 2(0.84)] = 2[2.15 \pm 1.68] = 2[0.47, 3.83] = [0.94, 7.66]$ packs/hour

Clearly, the promotion time was more effective for those already aware of the product. Whether the time spent was cost effective in either group in anther matter…