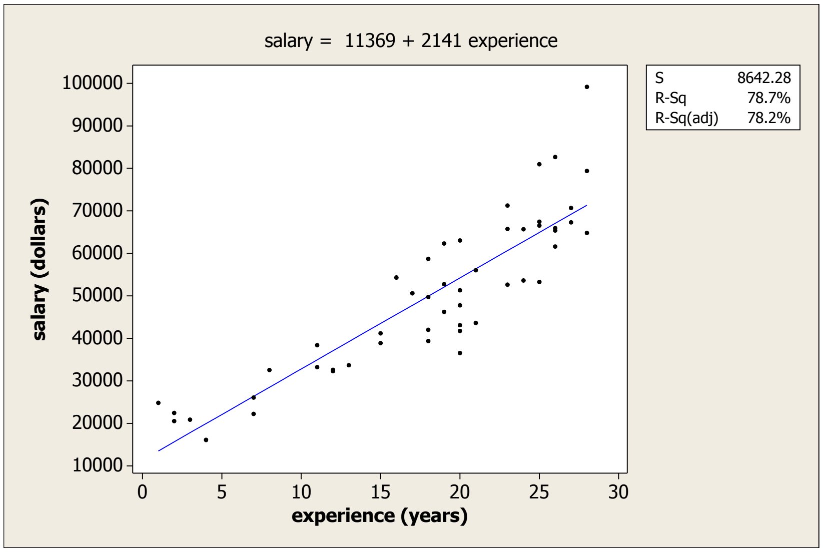
Lecture 21

Weighted Regression

If you have heterogeneity of variance, you can sometimes deal with it by using weighted regression. The weights should be proportional to the reciprocal of the variances. A typical model is that the variance is a function of the predictor variable(s) which deals with typical increasing/decreasing heterogeneity of variances cases. We discussed the various common choices of weights in class $w_i \propto 1/x_i$, $w_i \propto 1/\sqrt{x_i}$, $w_i \propto 1/x_i^2$, $w_i \propto x_i$, etc... To get accurate prediction intervals after a weighted regression analysis in Minitab, then there is a 10-step process to go through. These are given in Minitab's help menu and below.

Identification of Unequal Variances and Choice of Weights

Since most X values may not be repeated in a dataset, it is reasonable to sort the data by increasing X-value and pool data points with similar X-values into groups. For instance, with the social work salary data, it seems from the regression that we have heterogeneity of variance:



Formal tests for heterogeneity can be applied. In this case, if we break the data up into too many groups and test for heterogeneity of variance, we lose the statistical power needed to see it. If we break the data up in half, we have some power to see that the residuals on either side probably do have different variances.

After observing heterogeneity of variance, we might create categories of roughly the same size and range and form a table of using Minitab's Stat > Basic Statistics > Store Descriptive Statistics... menu storing the group variances and means.

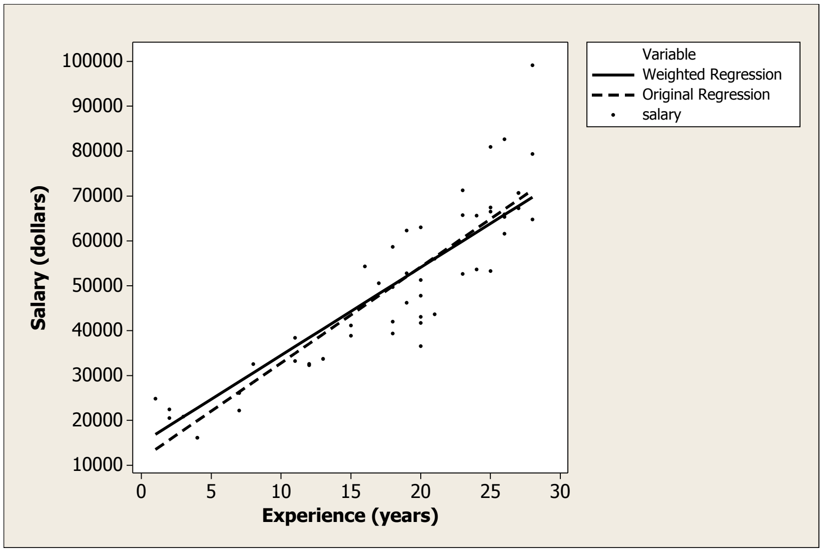
Experience range	Residual Variance	Group Mean	$s^2/\bar{x}$	$s^2/\bar{x}^2$	$s^2/\sqrt{\bar{x}}$
1-4	30849727	2.4	12854053	5355855	19913413
7-8	16873115	7.33	2300879	313756	6230811
11-13	13516439	11.8	1145461	97073	3934788
15-17	34927743	15.75	2217634	140802	8800964
18-19	64602253	18.43	3505549	190224	15048799
20	85671437	20	4283572	214179	19156716
21-23	89665859	22.2	4039003	181937	19030519
24-25	98498801	24.67	3993195	161886	19832420
26	87953041	26	3382809	130108	17249010
27-28	179946316	27.6	6519794	236224	34252196

We can then look at the variance over the mean, the variance over the mean squared, the variance over the square root of the mean, the variance times the mean, the variance times the mean squared, the variance times the square root of the mean. Since the variance increases as the mean increases in this problem, only the first three choices are relevant. (If the variance decreased as the group mean increases, the latter 3 would be the relevant options.) What we are looking for is the option that makes the variances roughly equal; that is, makes the variances on the same order of magnitude as each other. In the raw data, the variance in the residuals of the salaries for the workers with longer experiences is 10 times that of the workers with less experience. It is a bit hard to tell whether the variance is about like the mean or the square root of the mean, but the latter is probably a little better.

Running weighted regression in Minitab with weights equal to the reciprocal of the square root of experience, we see that the line shifts from

$$\text{Salary} = \$11,369 + 2141 \times \text{Experience} \;\text{ to }\; \text{Salary} = \$14,911 + 1958 \times \text{Experience}$$

The following picture shows the difference. The weighted regression puts more emphasis on points where the variance is less, so it gives a better fit for smaller salaries.



In order to get prediction bands out of Minitab when using a weighted regression, we have to follow the following 10 steps:

Preparing the data

1 Create a column of 1s, the same length as the predictor and response columns.

2 Create a column for each predictor containing the new observations. The number of predictors columns for new observations must match the number of predictor columns in your original data.

3 Create a column of weights for the new observations.

4 Using Calc > Calculator, calculate the square roots of the weights for the new observations and store in a column called SqrtWeight.

5 Using Calc > Calculator, multiply each predictor column containing the new observations by the SqrtWeight column and store in a column.

Performing weighted regression

6 Choose Stat > Regression > Regression. In Response, enter the response column. In Predictors, enter the column of 1s as your first predictor. Then enter your original predictor columns.

7 Click Options. In Weights, enter the column of weights for your original data. Then uncheck Fit intercept.

8 In Prediction intervals for new observations, enter the SqrtWeight column you calculated in Step 4. Then enter the predictor columns you created in Step 5, following the same order in which you entered the predictors in Step 6.

9 Under Storage, check Prediction limits. Click OK in each dialog box.

Transforming prediction limits

10 Using Calc > Calculator, divide the columns of stored prediction limits by the column SqrtWeight (from step 3). This transformation provides the correct prediction limits. It is important to note that if you also displayed or stored the fits, standard error of the fits, or confidence limits during this procedure, you must also divide them by the square roots of the weights to obtain the correct results.

Weights versus Transforms

In the case of these data, weighted regression is not the best approach to dealing with heterogeneity of variance. Salaries have a natural multiplicative error structure (your raise is 3% of your current salary rather than a fixed amount like $500, usually). A log transform of salaries in this data set cures both the heterogeneity of variance problem and the curvature problem in the data.

Exercises for Lecture 21

1. –

2. –