Stat 102

Introduction to Business Statistics

Class 17

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Working out the line for each group in a categorical variable regression
- The Least Squares Means table
- Multi-level categoricals
- Testing for multi-level categoricals
- Summary
- Next time

Last time

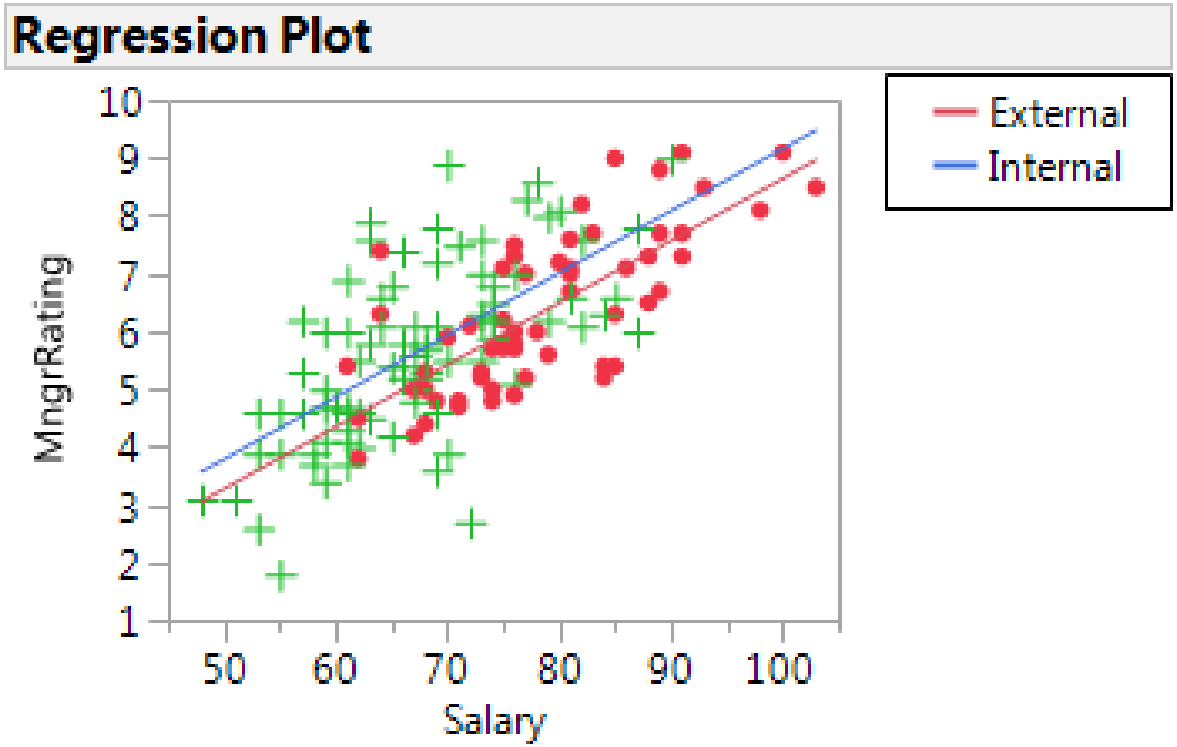
- Introduction to categorical predictor variables (two-levels)
- Three models:
  1. One line fits all
  2. Parallel lines
  3. Interaction
- Testing between the models uses the t-test on the appropriate regression coefficient

Analysis of Covariance

- Analysis of Covariance (ANCOVA) is a term sometimes used to describe regressions with categorical predictor variables.
- I don't really use it.

Back to the Manager data set

We will use the `Manager` data set and the *parallel lines model*:



Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-1.585493	0.655479	147.00	-2.42	0.0168*
Origin[External]	-0.514966	0.209029	147.00	-2.46	0.0149*
Salary	0.1074781	0.009649	147.00	11.14	<.0001*

Write down the line for each group

- We use the [1]**Indicator Function Parameterization**
- Recall, that the group that comes **last** in the alphabet is the *baseline*. Here the last group in the alphabet is `Internal`.
- The line for `Externals`:

$$\begin{aligned} \hat{y} &= -1.585 + 0.107\, Salary + -0.515 \\ &= -2.100 + 0.107\, Salary \end{aligned}$$

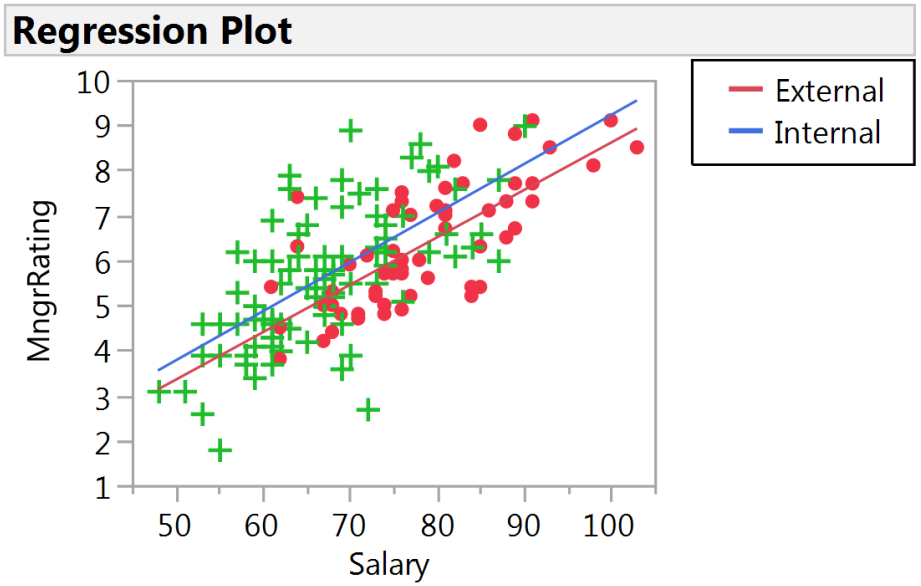
- The line for `Internals`:

$$\hat{y} = -1.585 + 0.107\, Salary$$

[1]You get this output by right-clicking in the title bar of the regression output, then choosing :Estimates" and then Indicator Parameterization Estimates

Write down the line for each group

Now for the [2] *non-parallel lines* (interaction) model:



Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-1.693524	0.870072	146.00	-1.95	0.0535
Origin[External]	-0.243417	1.44723	146.00	-0.17	0.8667
Salary	0.1090929	0.012893	146.00	8.46	<.0001*
Salary*Origin[External]	-0.003702	0.01952	146.00	-0.19	0.8499

[2] Remember to uncheck the "Center polynomials" option under `Model specification` in the fit model dialog.

Write down the line for each group

- Recall, that the group that comes **last** in the alphabet is the *baseline*. Here the last group in the alphabet is `Internal`.
- The line for `Externals`:

$$\begin{aligned} \hat{y} &= -1.694 + 0.109\, Salary + -0.243 + -0.0037 Salary \\ &= -1.937 + 0.103\, Salary \end{aligned}$$

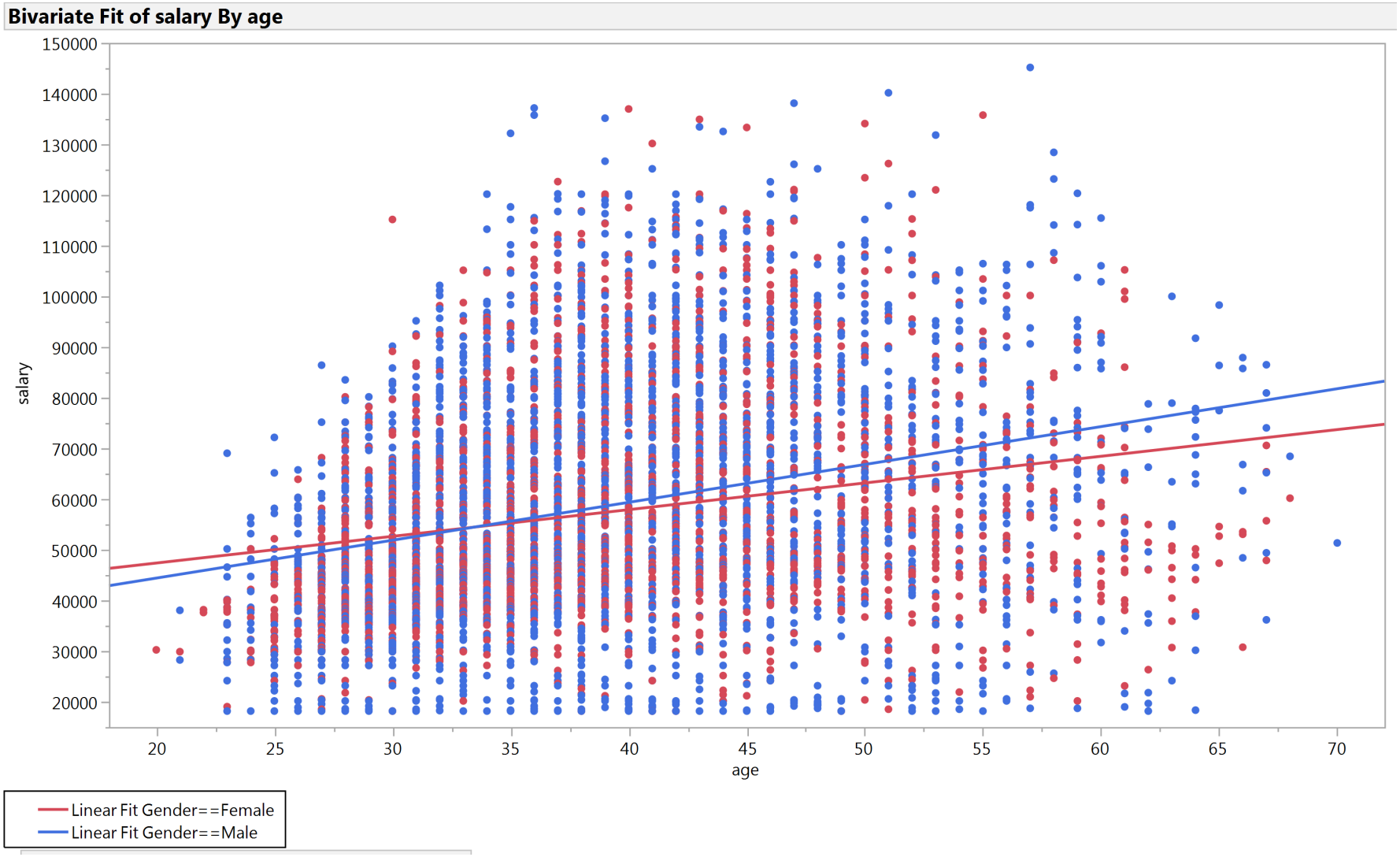
- The line for `Internals`:

$$\hat{y} = -1.694 + 0.109\, Salary$$

Looking for a gender effect in the HR data

- Goal: Examine differences in `Salary` by `Gender`, adjusting for `Age`.
- Start the analysis by exploring for an interaction using the *by* command (don't forget to color by the levels of the categorical variable).

Looking for a gender effect in the HR data



Looking for a gender effect in the HR data

- Interpret the interaction in the context of the question: "a differential impact of aging on salary, according to gender".
- Given the suggestion of interaction, we will fit the interaction model directly.
- The `Male` category will be the baseline as it is the last in the alphabet.

Output from the multiple regression with interaction

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	29452.826	1698.393	5643.0	17.34	<.0001*
age	747.33187	42.39554	5643.0	17.63	<.0001*
Gender[Female]	7381.9763	2550.197	5643.0	2.89	0.0038*
age*Gender[Female]	-220.6519	63.80512	5643.0	-3.46	0.0005*

- Note the highly significant interaction.
- The age slope for Men is 747.
- For women it is (747 - 221) = 526.
- As women get older, their pay rises at a lower rate than men ($221 less per year).

Write down the line for each group

- The line for `Females`:

$$\begin{aligned} \hat{y} &= 29453 + 747\,Age + 7382 + -221\,Age \\ &= 36835 + 526\,Age \end{aligned}$$

- The line for `Males`:

$$\hat{y} = 29453 + 747\,Age$$

- Given that the RMSE for this regression is 21,274 provide a 95% prediction interval for the salary a 40 year old Female:

$$\hat{y} \pm 2\,RMSE.$$

$$36835 + 526 \times 40 \pm 2 \times 21274 = 57875 \pm 42548 = (15327, 100423).$$

- Not a highly predictive model!

The Least Squares Means Table

- Under the leverage plot for the categorical variable you will find the Least Squares Means table:

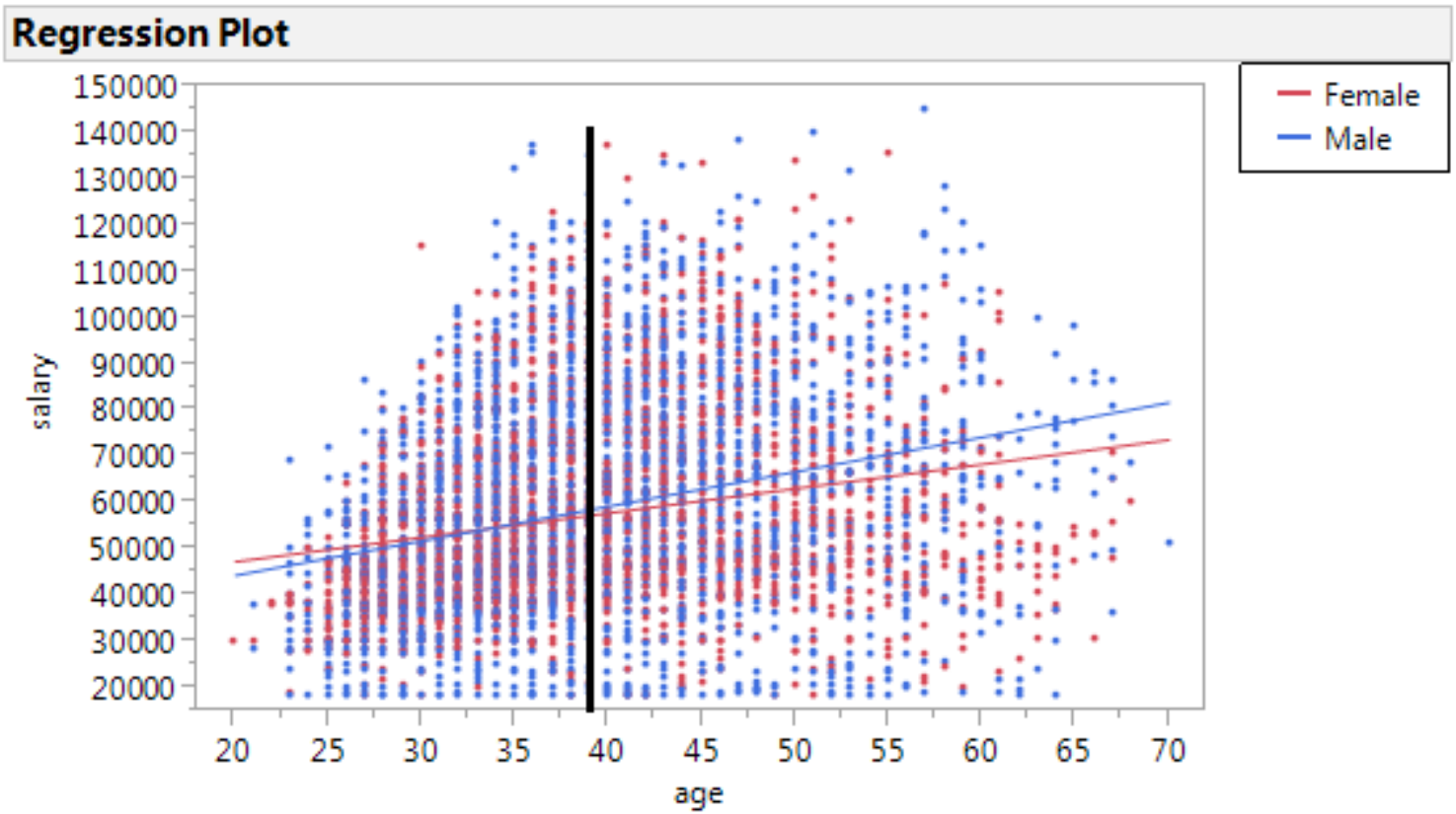
Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
Female	57362.729	424.21985	57317.9
Male	58580.903	380.13850	58631.9

- It allows for a comparison on the expected value of y, over the levels of x_1, controlling for the value of x_2.
- The least squares means are sometimes called *model based means* because they are calculated directly from the multiple regression *model*.
- Informally, they allow you to do an *apples to apples* comparison across the levels of the categorical variable.

The Least Squares Means Table

- The values under the *mean* column are the raw means for `Salary` in the data set, for Men and Women (marginal means).
- The values under Least Sq Mean are calculated by finding the average age in the data set (38.976) and calculating from the regression model, the expected salaries for men and women at this age value.



Multi-level categorical predictors

- Conceptually, we are still deciding between three models:
  1. One line fits all
  2. Parallel lines
  3. Interaction

  but now, the pictures can have more than two lines.
- The coding gets a bit more involved, but JMP wil take care of it.
- If you have a *k* level categorical, then you will need *k - 1* columns to code it up. Very messy, for example, if you had `State` in your model.
- The key difference between two-level categoricals and multi-level categoricals, is in the hypothesis testing. Now we **must** use the Partial-F test, prior to claiming interesting effects for the individual level coefficients.
- You **don't** have to do these Partial-F's via the Custom Test dialog. They will be automatically calculated and are in the **Effects Test** table.

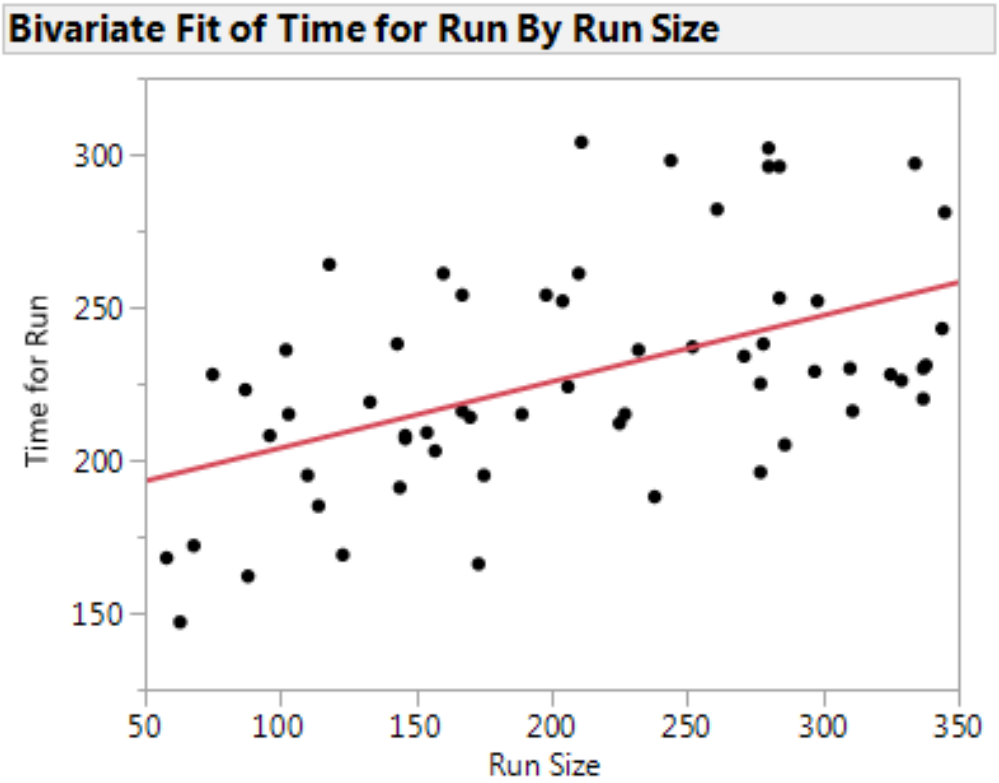
The `ProdTime` data set

This data set contains records on the amount of time it took for a production job to run. It has three variables:

1. Y: `Time for Run` measured in minutes.
2. X_1: The `Manager` doing the job. A three-level categorical.
3. X_2: The `Run Size` measuring the number of items produced.

Interpreting the coefficients

We will use the *one-line fits all* model to start off with (we do this by completely ignoring `Manager`.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	182.31371	10.96144	16.63	<.0001*
Run Size	0.2165919	0.048479	4.47	<.0001*

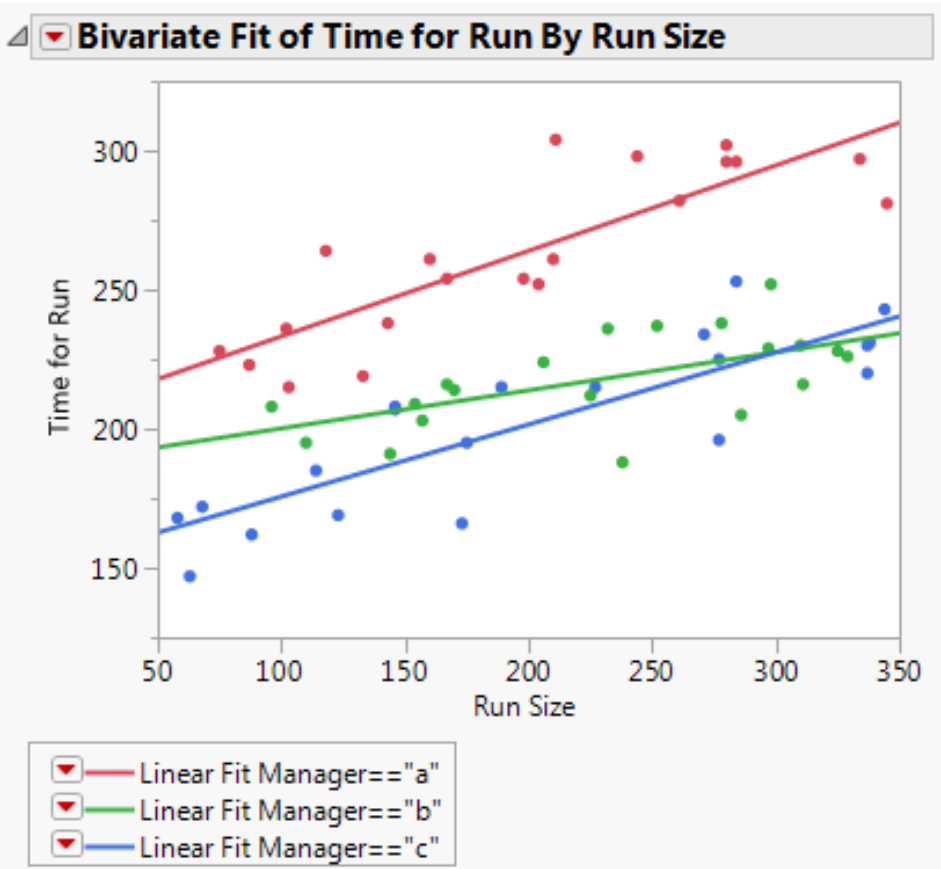
Interpret the intercept and the slope in the context of the problem.

Interpreting the coefficients

- Slope: the additional time in minutes for every extra item produced. Call it the *work-rate*. It is about 0.2, that is 1/5 of a minute or 12 seconds.
- Intercept: formally, the expected time to complete a job with 0 items, but in the context, better interpreted as the *set-up time*. It typically takes about 180 minutes to set up one of these machines, that is, there's a three hour set-up time

Interpreting the coefficients

Start with the initial exploratory analysis: create the [3]color coded scatter-plot with individual regression lines. Fit them using the *by* command from the title-bar of the scatterplot. What do you learn?



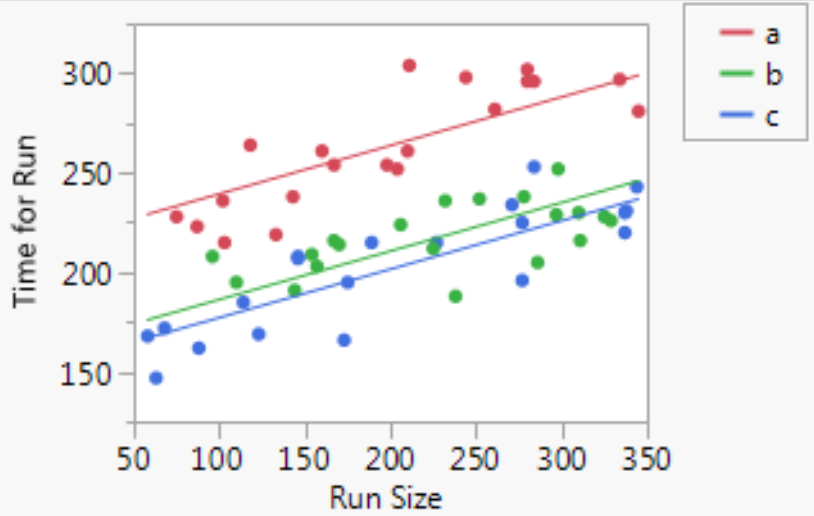
[3]Use the Rows→ color or mark by columns dialog

Parallel lines

- It looks like the interaction model may be the winner, but we'll fit parallel lines first, to build our knowledge incrementally.
- Parallel lines, allows for different intercepts, but keeps the slopes the same. That is different set-up times but the same work rate.
- We will now use the multiple regression (Fit Model) dialog to formally estimate the parameters in the model
- Don't forget to obtain the *Indicator Parameterization estimates*.

Parallel lines

Regression Plot



Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Size	1	1	25260.250	94.1906	<.0001*
Manager	2	2	44773.996	83.4768	<.0001*

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	152.95031	6.245301	56.00	24.49	<.0001*
Run Size	0.243369	0.025076	56.00	9.71	<.0001*
Manager[a]	62.168171	5.180029	56.00	12.00	<.0001*
Manager[b]	9.1073536	5.224343	56.00	1.74	0.0868

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
a	266.05966	3.6749523	263.050
b	212.99885	3.6958029	217.850
c	203.89149	3.6667574	202.050

Parallel lines

Questions to answer:

1. What is the common work-rate?
2. Which manager is the slowest at setting up their machine?
3. What is the difference between Manager a's and Manager b's setup time.
4. Is the difference in intercepts (set-up times) significant?
5. For the average job size, how long does it take Manager b to do it?
6. Write down the line for Manager a and predict how long it would take them to do a job with 300 items.

Parallel lines

Answers:

1. The slope is 0.24, that's about 15 seconds per incremental item.
2. Manager [a], who takes $152.95 + 62.168 = 215.118$ minutes.
3. 62.168 - 9.107 $= 53.061$ minutes (a little less than an hour).
4. Yes. Look in the effect test table for the manager effect. It has a p-value of less than 0.0001. It is highly significant. The *parallel lines* model beats *one line fits all*.
5. Straight from the Least Squares Means table: 212.999 minutes.
6. $$152.950 + 0.243\,\mathrm{RunSize} + 62.168.$$

   Letting `Run Size` equal 300, we get

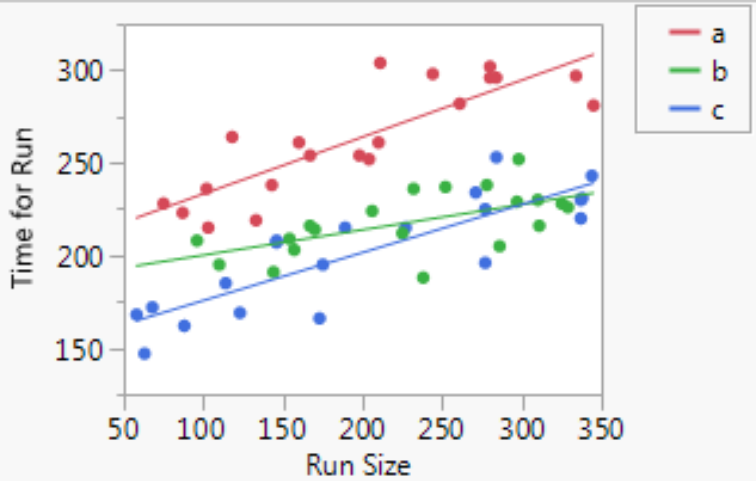
   $$152.950 + 0.243 \times 300 + 62.168 = 288.018.$$

The interaction model

- This model will allow for the potential of different work rates between the three managers.
- Make sure you uncheck *center polynomials* in the Model Specification options, in the Fit Model dialog to replicate the output I am using.
- Make sure you get the Indicator Parameterization estimates.

The interaction model

Regression Plot



Summary of Fit

RSquare	0.835267
RSquare Adj	0.820014
Root Mean Square Error	15.65812
Mean of Response	227.65
Observations (or Sum Wgts)	60

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Size	1	1	22070.614	90.0192	<.0001*
Manager	2	2	4832.335	9.8548	0.0002*
Manager*Run Size	2	2	1778.661	3.6273	0.0333*

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	149.7477	8.023586	54.00	18.66	<.0001*
Run Size	0.2592431	0.035784	54.00	7.24	<.0001*
Manager[a]	52.786451	12.25963	54.00	4.31	<.0001*
Manager[b]	36.746157	14.08201	54.00	2.61	0.0117*
Manager[a]*Run Size	0.0480219	0.056387	54.00	0.85	0.3982
Manager[b]*Run Size	-0.122466	0.059962	54.00	-2.04	0.0460*

Interaction model

Questions to answer:

1. What is Manager [c]'s work-rate?
2. Which manager has the best work-rate?
3. Does this model improve on the parallel lines model, in terms of predictive power?
4. What should I do with the Manager [a] * Run Size coefficient as it is not significant?
5. Write down the line for Manager [b].
6. Provide an approximate 95% Prediction Interval for the amount of time it is expected for Manager [b] to take for a job of 300 items.

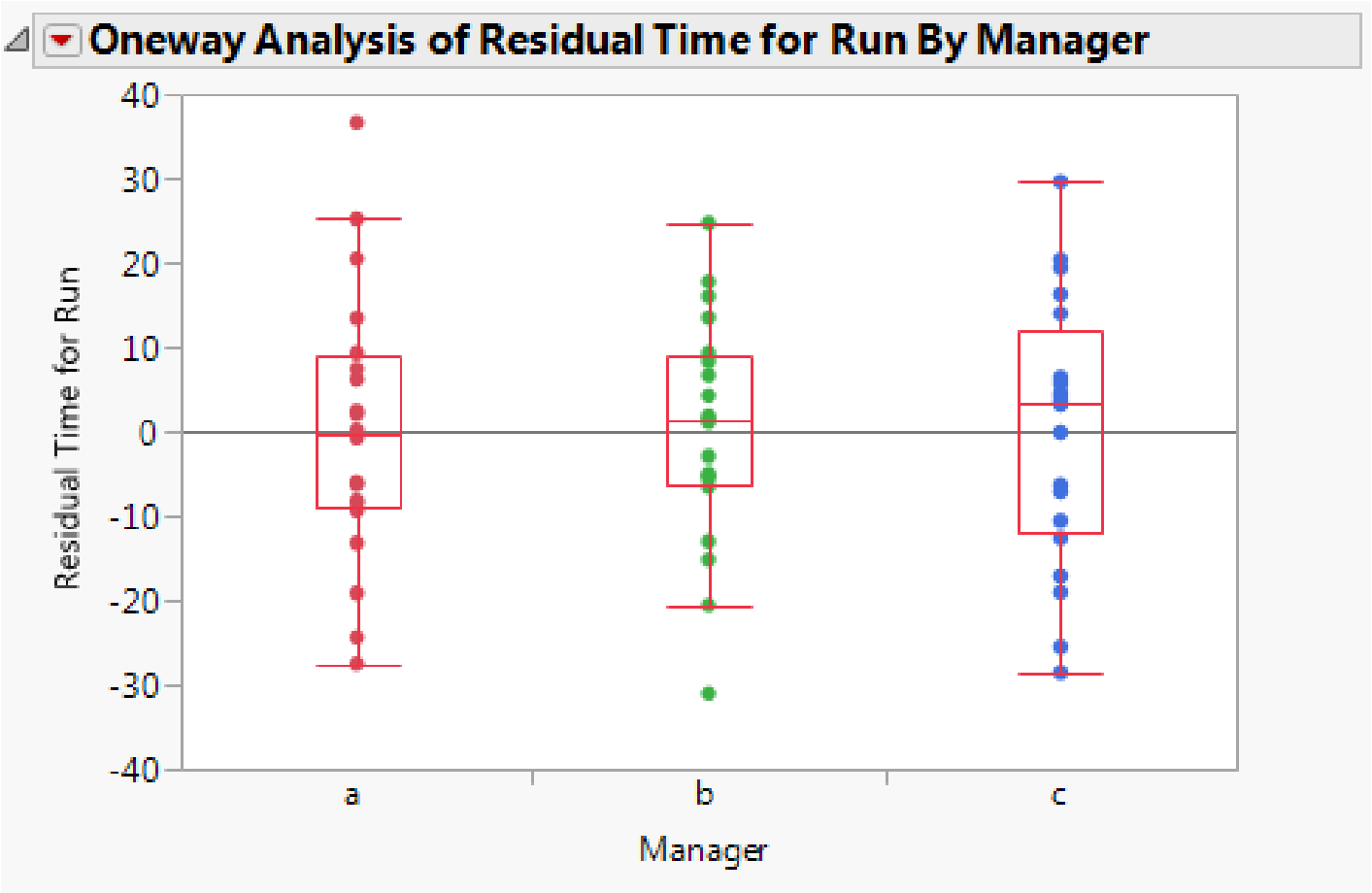
Interaction model

Answers to questions:

1. Manager [c] is the baseline, so 0.259 minutes per extra item (about 15 seconds).
2. Manager [b]'s slope is 0.122 beneath [c]'s, so is the best in terms of work rate.
3. Yes, because the interaction term for `Manager*Run Size` in the Effect Tests table is significant.
4. Nothing. Just leave it as is. If the interaction term is significant as a whole, then keep all the individual pieces (do not zero any of them out).
5. $$149.748 + 0.259\,\mathrm{Run\,Size} + 36.746 + -0.122\,\mathrm{Run\,Size}.$$
6. $$186.494 + 0.137\mathrm{Run\,Size} \pm 2RMSE = 227.594 \pm 33.316.$$

Checking residuals for categorical variables

You should plot the residuals across the levels of the categorical variable, checking for the *constant variance* assumption.



There is no evidence against equality of variances, so no heteroscedasticity.

Summary

- The Least Squares Means table
- Multi-level categorical predictor variables
- The partial-F for multi-level categorical variables

Next time

- More on multiplicity
- Introduction to step-wise regression