

## **DSC 330 - Business Statistics**

### **Preparation for Final Exam**

The two hour exam will consist of two parts. The first part is similar to the mid-term but will consist of just 20 questions: 18 multiple choice and 2 true/false. The second part will have a "short-answer" format and will consist of 10 questions. Each question in part 1 will be worth 10 points, while each question in part 2 will be worth 20 points, giving a total of 400 points.

This will be a closed-book exam: you cannot use the course packet or your class-notes. You may only use any notes you make on the space provided on this "Preparation for Final Exam." You may find a (regular) calculator useful; a statistical calculator may be used (although it will provide no advantage over a regular calculator).

Most of the questions will be based on the following report. You may want to read this over ahead of time so that you are familiar with it. The report was written as part of an analysis of data on annual salary and other characteristics of managers in a medium-sized business. The data were collected as evidence in a class-action case to judge a claim of salary discrimination against women. Unfortunately, the report was burnt in a fire and parts of it are now illegible (denoted by "#####" below). You have been asked to answer questions that the prosecuting attorney has about the original report. Since you have just taken a statistics class, you should be able to figure out most of the information you need to answer those questions from the damaged report.

You should bring the report to the final exam since you will need to refer to it to answer the questions. Following the report are two blank pages for you to write notes on.

By the way, the report also provides a nice example of the type of *layout* (not content) of the report expected for the "Real-world Case" Project. (Just don't drop your project report in a fire!)

# An Analysis of Annual Salary Data for ACME Corporation

## A Report by XYZ Statistical Consultants

### Introduction

This report describes data on annual salary and other characteristics of managers at ACME Corporation. The data consist of 60 managers with the following variables measured:

Name	Information
logY	Natural logarithm of annual salary in log-dollars
X1	Number of years in current rank
X2	Working years since highest degree was earned
D3	Highest degree: 0=Bachelors, 1=MBA
D4	Gender: 0=Male, 1=Female
RANK	Type of manager: 1=low-level, 2=mid-level, 3=senior
D5	1=mid-level manager, 0=low-level or senior manager
D6	1=senior manager, 0=low-level or mid-level manager

The main goal of the analysis is to determine whether there is compelling evidence that women are being paid less than men, all other things being equal. Since a multiple linear regression analysis is able to isolate the effects of individual predictors adjusting for other predictors in the model (i.e. holding them fixed), we can build such a model to answer this question.

### Analysis

Salaries ranged from \$31,802 to \$86,853, with the 30 male managers earning \$52,425 on average, and the 30 female managers earning \$45,568 on average. Salary information was analyzed using natural logarithms for a number of reasons. First, the distribution of salaries in this firm is highly **skewed**, whereas regression models often work better when the response variable is more **normal** (the natural logarithm transformation can make **skewed** data look more **normal**). Second, multiple regression inference requires that the random errors in the model satisfy certain assumptions: **independence**, **normality**, **constant variance**, and **linearity**. These assumptions appear to be met with this example when fitting regression models that use the natural logarithm of salary rather than salary itself. Third, using natural logarithm of salary allows regression parameter estimates to be interpreted in terms of percentage changes in salary (rather than absolute changes in salary). Since salary increases are usually expressed in terms of percentages, this interpretation is more applicable in real-life.

To give a broad overview of the data, consider the following summary statistics for logY:

	N	Minimum	Maximum	Mean	Std. Deviation
logY	60	10.37	11.37	10.767	.251

If this dataset could be considered a random sample from a population of similar managers, our best guess for the population mean of logY (ignoring other factors such as experience, gender, rank, etc.) is **10.767**.

Since the 97.5th percentile of the t-distribution with 59 degrees of freedom is 2.001, a 95% confidence interval for the population mean of logY is equal to:

#####

Similarly, consider these summary statistics for X1 = years in current rank and X2 = work years since degree:

	N	Minimum	Maximum	Mean	Std. Deviation
X1	60	1	21	8.93	4.878
X2	60	3	27	15.28	5.770

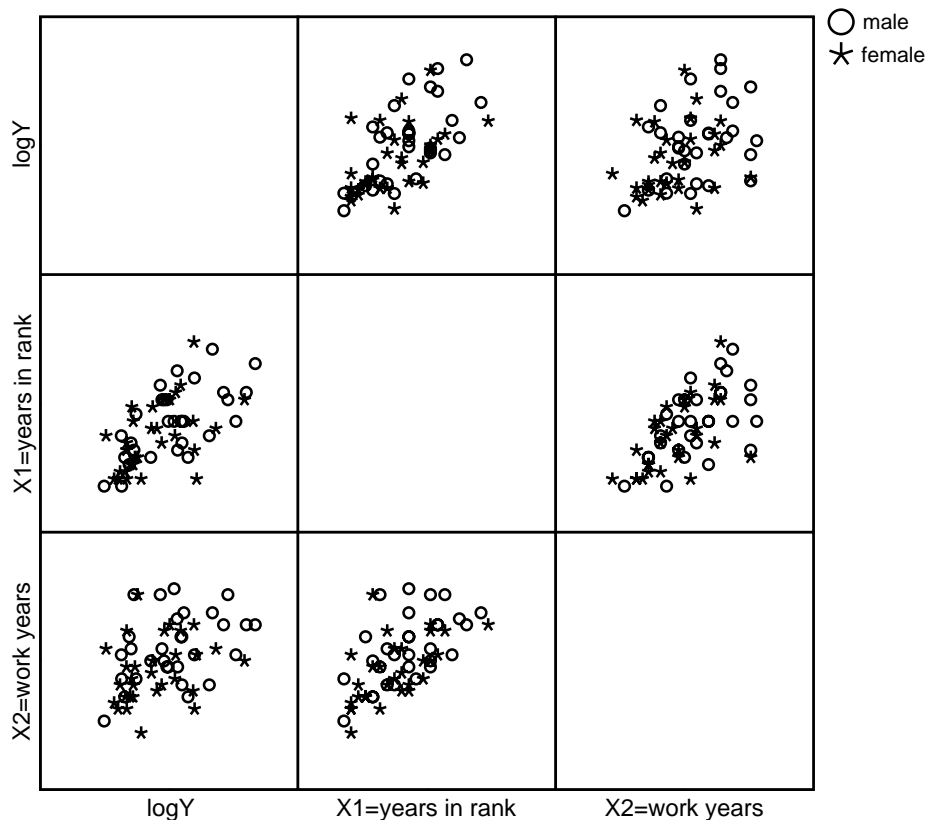
Concern was raised that the sampled managers have sufficient years in their current rank for the data to tell us about possible sex discrimination. In particular, we would like the mean number of years in current rank to be more than 7.5. Consequently, we set up the following hypothesis test:

NH: #####

AH: #####

The t-statistic for this test is #####. Using a significance level of 5%, and since the 95th percentile of the t-distribution with 59 degrees of freedom is 1.671, the p-value for this test is ##### 0.05 and so we ##### NH in favor of AH. In other words, #####.

Concern was also raised that there might be problems with ##### if we tried to use two variables in a multiple linear regression model that were highly ##### with each other. In particular, it might be problematic to use both X1 and X2 in a model together. The following scatterplot-matrix and correlation coefficients suggest that we need to keep an eye on this issue, since X1 and X2 are more correlated with each other than X2 is with logY:

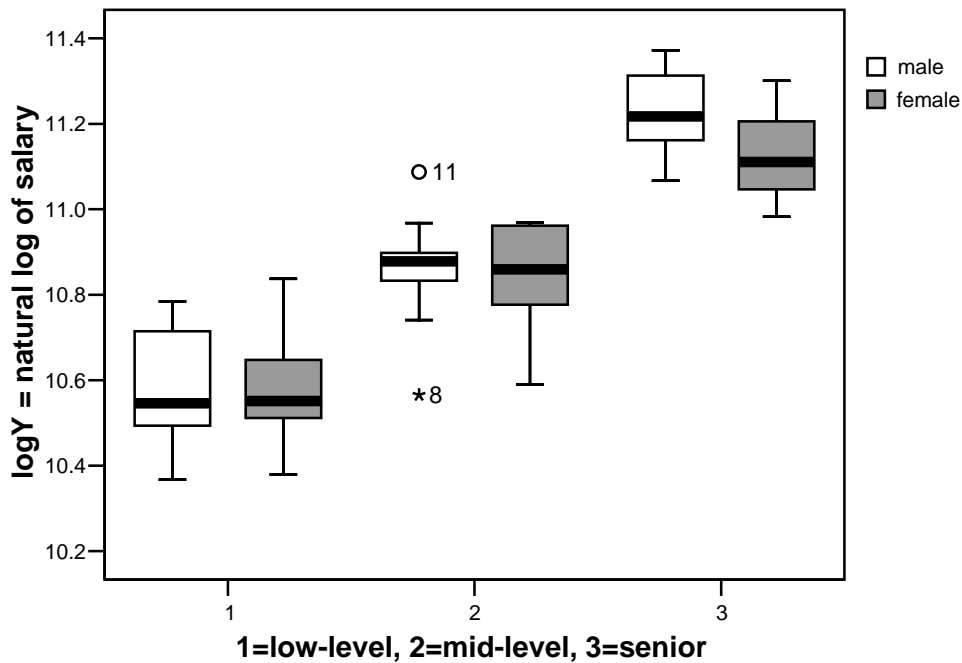
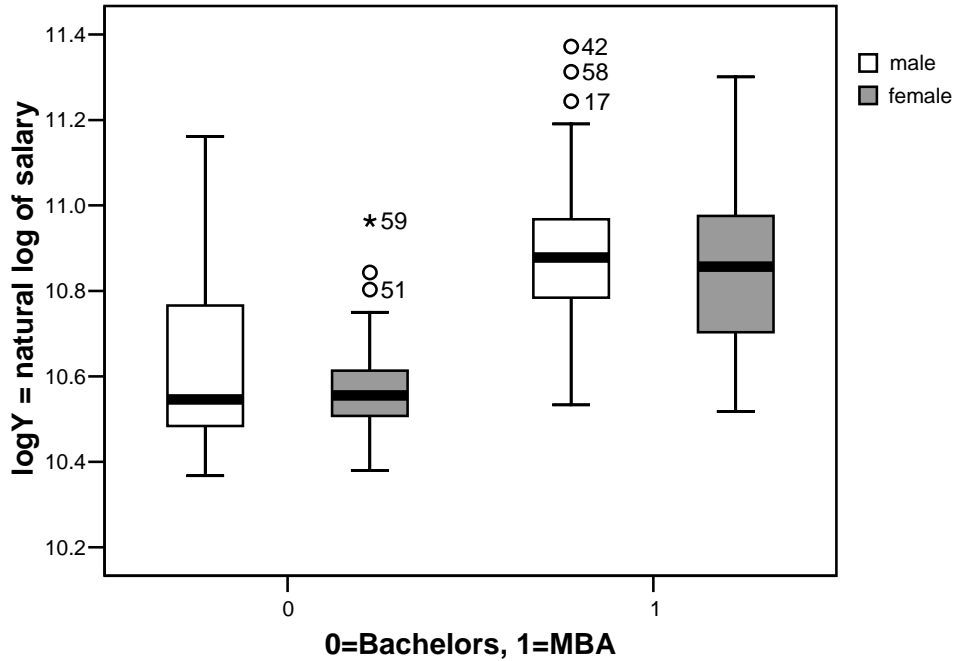


### Correlations

	logY	X1	X2
logY	1	.614	.387
X1	.614	1	.595
X2	.387	.595	1

The scatterplots also show that there are no obvious points well-separated from the rest that might unduly influence the analysis.

An illustration of the other two predictor variables is available in the following boxplots:



For managers with MBA degrees, females appear to have a ##### median salary than males, and for mid-level and senior managers, females appear to have a ##### median salary than males. This provides some guidance on which variables are likely to be important in determining salary. However, it can be misleading to draw conclusions from analyses that ##### the possible effects of other variables. We would like to isolate the effects of each variable controlling (or adjusting) for the effects of all other relevant variables. We can use multiple linear regression to do this. Thus we next build a regression model to provide the best prediction of logY while at the same time keeping the model as simple as possible.

We first try a model with the ##### predictors, X1 and X2, the two-level ##### variables, D3 and D4, and the three-level ##### variable, rank, coded using D5 (=1 for mid-level managers, 0 otherwise) and D6 (=1 for senior managers, 0 otherwise). Model results are:

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	10.526	.033	319.964	.000
	X1	.021	.002	8.621	.000
	X2	-.010	.002	-4.661	.000
	D3	.122	.021	5.783	.000
	D4	-.005	.020	-.272	.787
	D5	.187	.024	7.710	.000
	D6	.523	.030	17.251	.000

a. Dependent Variable: logY

It looks like gender may have ##### effect on logY after adjusting for the effects of the other predictors, since ## has a relatively ##### p-value of #####. However, it is possible that gender may have a significant effect on logY via interactions with other predictors. So, we decided to next consider the eight two-way interactions of X1 and X2 with D3, D4, D5 and D6 in the following model:

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
2	(Constant)	10.426	.070	149.026	.000
	X1	.019	.007	2.819	.007
	X2	-.001	.007	-.164	.870
	D3	.179	.073	2.442	.019
	D4	.054	.067	.815	.419
	D5	.287	.085	3.371	.002
	D6	.463	.136	3.411	.001
	D3X1	.002	.007	.295	.769
	D4X1	.000	.007	-.058	.954
	D5X1	-.001	.005	-.141	.889
	D6X1	.010	.012	.801	.427
	D3X2	-.006	.007	-.906	.370
	D4X2	-.004	.006	-.629	.533
	D5X2	-.006	.005	-1.079	.286
	D6X2	-.004	.009	-.430	.669

a. Dependent Variable: logY

At first sight, none of the interactions seem to be useful, but with so many terms in the model, important effects can become masked. So, we next see if we can remove a subset of the interactions by considering an F-test for comparing nested models. The complete model in this case is model 2 above, while the reduced model (model 3 below) excludes D3X1, D4X1, D5X1, and D6X2. Fitting the two models produces the following output:

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
3	.968 <sup>a</sup>	.938	.925	.069					
2	.969 <sup>b</sup>	.938	.919	.071	.001	.117	4	45	.976

a. Predictors: (Constant), X1, X2, D3, D4, D5, D6, D6X1, D3X2, D4X2, D5X2

b. Predictors: (Constant), X1, X2, D3, D4, D5, D6, D6X1, D3X2, D4X2, D5X2, D3X1, D4X1, D5X1, D6X2

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
3	(Constant)	10.441	.063	166.394	.000
	X1	.019	.002	7.847	.000
	X2	-.003	.005	-.568	.573
	D3	.190	.068	2.773	.008
	D4	.042	.062	.677	.502
	D5	.266	.073	3.641	.001
	D6	.410	.102	4.003	.000
	D6X1	.009	.009	1.016	.315
	D3X2	-.005	.004	-1.187	.241
	D4X2	-.003	.004	-.841	.404
	D5X2	-.005	.004	-1.135	.262

a. Dependent Variable: logY

The nested model F-test has a p-value of #####, so there is ##### evidence to ##### the null hypothesis that the regression parameters for the 4 additional terms in the complete model are all zero. In other words, the additional terms #####. Note also that the regression standard error, *s*, decreases from 0.071 to 0.069 when we take these 4 terms out, indicating that the additional terms are more harmful than beneficial. Note how the p-values for the remaining interactions have decreased, but still a couple of them don't seem to be that useful. A further nested model F-test confirms this:

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
4	.967 <sup>a</sup>	.934	.924	.069					
3	.968 <sup>b</sup>	.938	.925	.069	.003	1.243	2	49	.298

a. Predictors: (Constant), X1, X2, D3, D4, D5, D6, D3X2, D5X2

b. Predictors: (Constant), X1, X2, D3, D4, D5, D6, D3X2, D5X2, D6X1, D4X2

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
4	(Constant)	10.470	.041	252.420	.000
	X1	.020	.002	8.392	.000
	X2	-.005	.003	-1.499	.140
	D3	.159	.064	2.478	.017
	D4	-.009	.020	-.426	.672
	D5	.279	.072	3.862	.000
	D6	.508	.031	16.269	.000
	D3X2	-.003	.004	-.741	.462
	D5X2	-.006	.004	-1.422	.161

a. Dependent Variable: logY

Finally, an individual t-test allows the removal of one more term:

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics
		B	Std. Error			VIF
5	(Constant)	10.478	.040	263.797	.000	
	X1	.020	.002	#####	.000	1.685
	X2	-.006	.003	#####	.065	3.772
	D3	.114	.021	5.460	.000	
	D4	-.004	.019	-.209	.835	
	D5	.305	.063	4.818	.000	
	D6	.505	.031	16.376	.000	
	D5X2	-.007	.004	-2.003	.050	

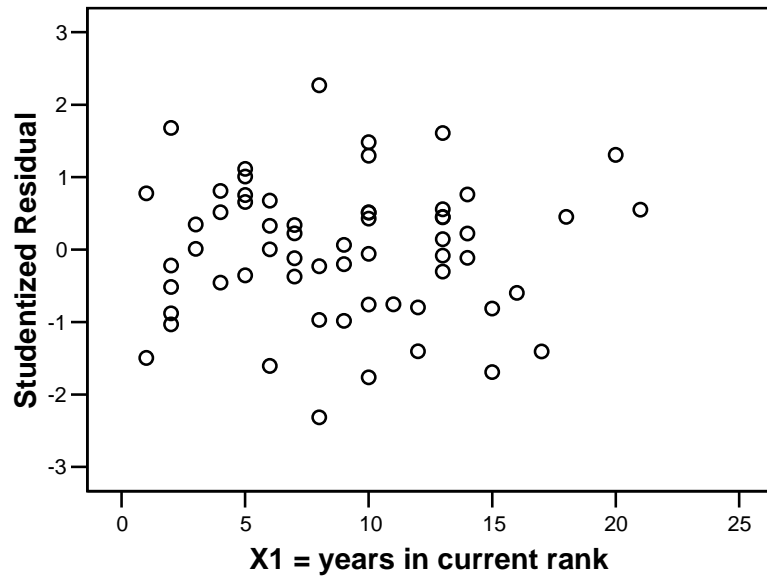
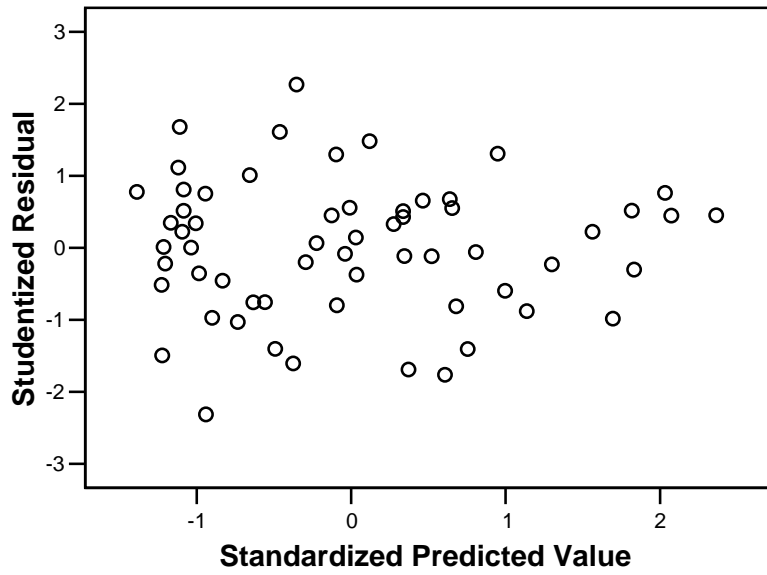
a. Dependent Variable: logY

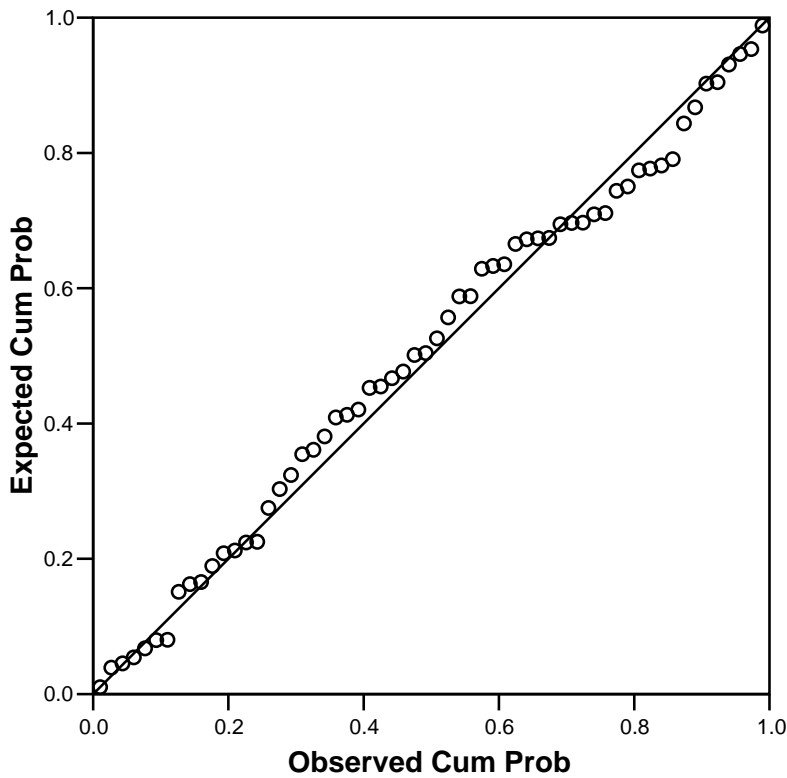
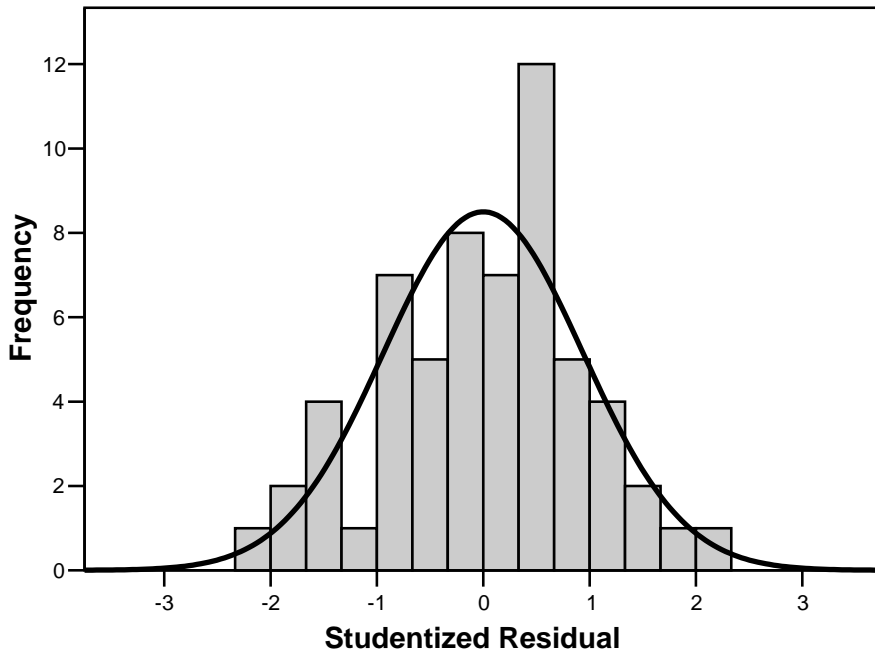
At this point we can also return to the question of whether there could be a ##### problem with this model. Since the ##### for X1 and X2 are both less than #####, we ##### concerned.

The only interaction remaining in model 5 is #####. Note that with this interaction in the model, the p-value for X2 is higher than the usual 0.05 threshold. This might suggest that we can remove X2 from the model. However, using the principal of ##### (retaining lower-order terms like D5 and X2, if higher-order terms like the D5X2 interaction are in the model) suggests retaining X2.

There are other possible interactions we have not yet considered, for example D3D4, D3D5, D3D6, D4D5, and D4D6. (Note that a D5D6 interaction makes no sense.) However, a nested model F-test (output not shown) demonstrates that including these 5 interactions would not significantly improve the model.

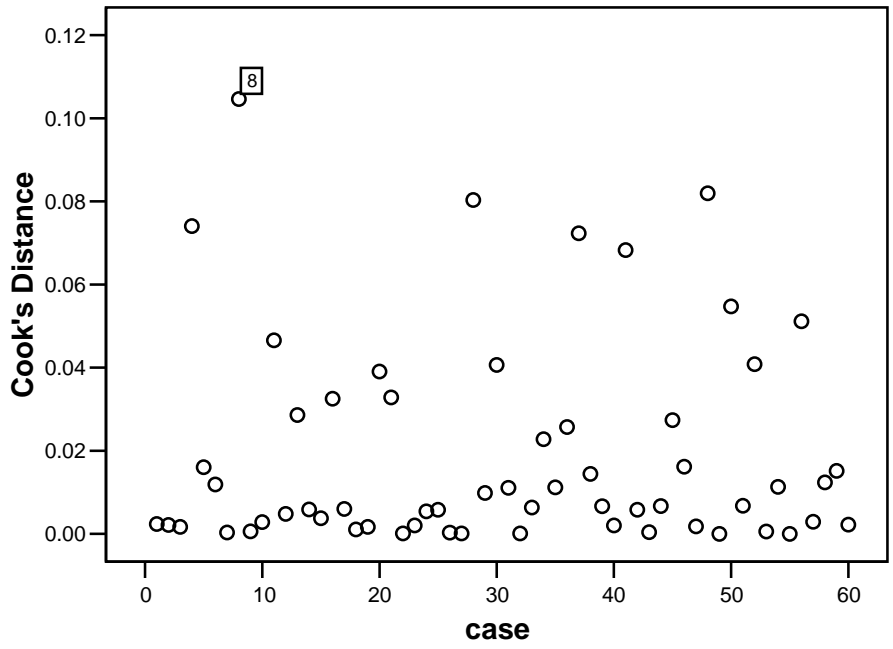
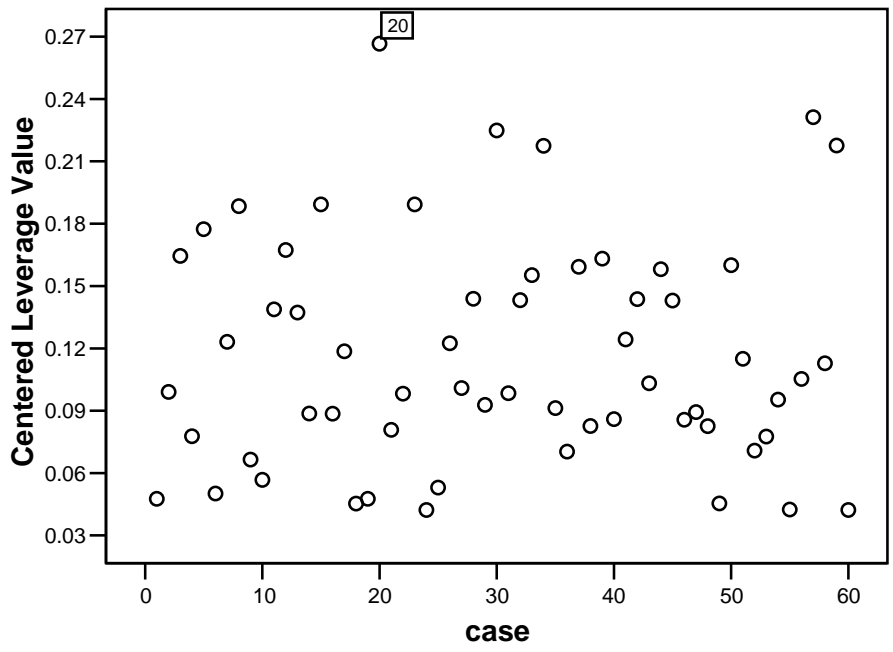
Next, we construct some residual plots for model 5 to check on the assumptions underlying multiple linear regression:





There appear to be no clear non-random patterns in the scatterplots on page 8, so the assumptions of `#####`, `#####`, and `#####` appear to be satisfied. In addition the histogram is approximately `#####` and the points in the QQ-plot `#####`, so the assumption of `#####` appears to be satisfied. Finally, there are no studentized residuals less than  $-3$  or larger than  $+3$ , so there appear to be no `#####` to worry about.

Finally, we consider ##### by looking at leverages and Cook's distances:



Case 20 has the highest leverage, and it is right around the threshold of #####. However, if we exclude this case from the analysis, the ##### do not change much, and so we needn't worry that it is having an undue influence on the fit of the model. Case 8 has the highest Cook's distance, but it is not ##### and so we needn't worry that it is having an undue influence on the fit of the model.

We could perhaps improve on model 5 further, for example by considering ##### or #####. However, since the regression assumptions check out, and there are no unduly influential cases, model 5 appears to be reasonable, and we probably could not improve on this model too much more.

## Results

Our final best-fitting model is therefore model 5 with the following results:

### Model Summary<sup>b</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
5	.966 <sup>a</sup>	.934	.925	.069

a. Predictors: (Constant), X1, X2, D3, D4, D5, D6, D5X2

b. Dependent Variable: logY

### ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
5	Regression	3.474	7	.496	104.658	.000 <sup>a</sup>
	Residual	.247	52	.005		
	Total	3.721	59			

a. Predictors: (Constant), X1, X2, D3, D4, D5, D6, D5X2

b. Dependent Variable: logY

### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
5	(Constant)	10.478	.040	263.797	.000
	X1	.020	.002	#####	.000
	X2	-.006	.003	#####	.065
	D3	.114	.021	5.460	.000
	D4	-.004	.019	-.209	.835
	D5	.305	.063	4.818	.000
	D6	.505	.031	16.376	.000
	D5X2	-.007	.004	-2.003	.050

a. Dependent Variable: logY

This model is able to explain ##### of the variation in logY. Predictions using this model are likely to be accurate within approximately plus/minus #####. A test of the global usefulness of the model results in a p-value of #####, so that #####.

Individually, each predictor in the model has a p-value less than 0.05 (suggesting they are useful predictors of logY adjusting for the effects of the other predictors in the model), except X2, which as described above has been retained to preserve #####, and D4, which is the central focus of this analysis.

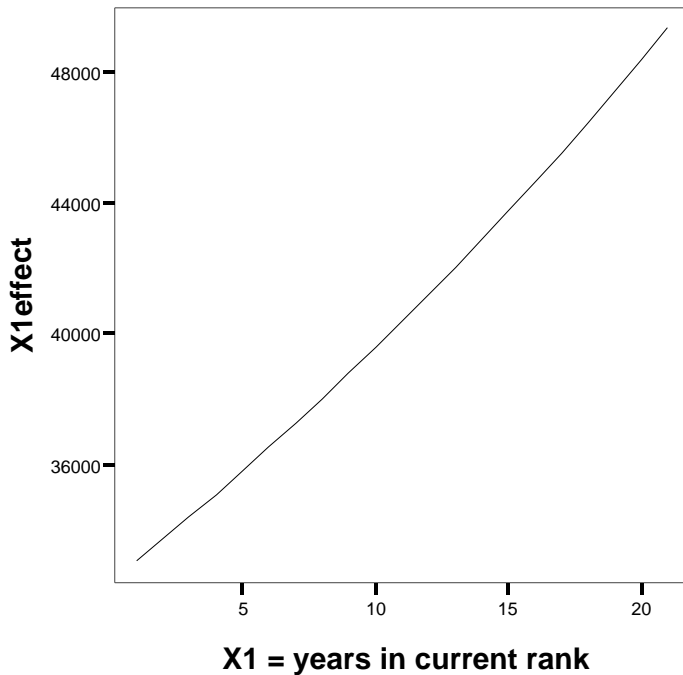
## Conclusions

The regression equation is  $E(\log Y) = \text{#####}$

The X1 effect does not vary with X2, degree, gender, or rank. To illustrate, we can calculate how X1 affects salary for managers with average work years since degree (15.28), who have a Bachelors degree (D3=0), are male (D4=0), and have the rank of low-level manager (D5=D6=0):

X1 effect =  $\exp(10.478+0.020X1-0.006(15.28)) = \text{#####}$

This can be illustrated on the following line graph:



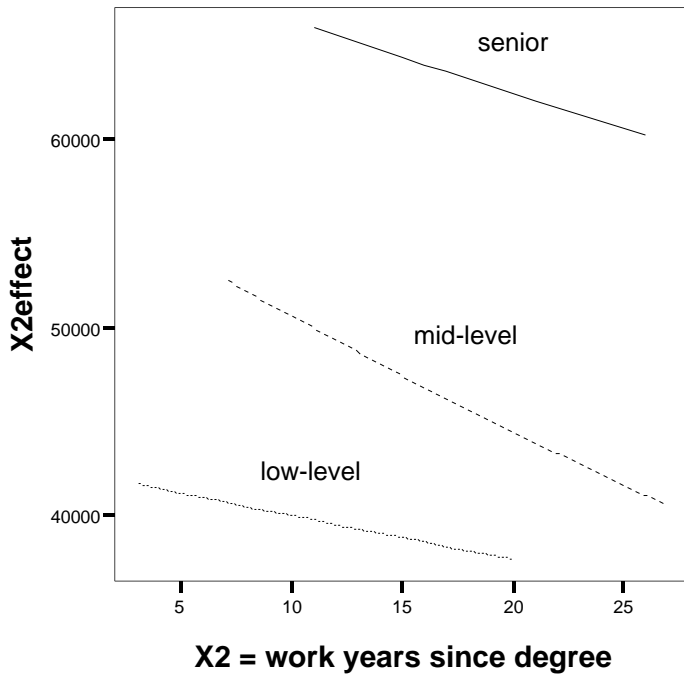
Since our response variable is the natural logarithm of salary, calculating  $\exp(b) - 1$ , where b is the corresponding parameter estimate, allows effects to be expressed as percentage changes. Since  $\exp(0.020) - 1 = 2.02\%$ , each additional year in current rank increases salary by 2.02% (holding X2, degree, gender and rank fixed).

For example, the expected salary for a male, low-level manager with 5 years in their current rank and 15.28 years since their Bachelor's degree is  $\exp(10.478+0.020(5)-0.006(15.28)) = \$35,822$ . A similar manager with 6 years in their current rank (and identical values for everything else) has an expected salary of  $\exp(10.478+0.020(6)-0.006(15.28)) = \$36,546$ , which is 2.02% higher than \$35,822. Similar calculations show that this 2.02% difference holds for any pair of managers who differ by one year in their current rank (and have identical values for everything else).

The X2 effect is complicated by the presence of an interaction between D5 and X2, where D5 is one of the indicator variables for rank. We can calculate how X2 affects salary for managers with average years in current rank (8.93), who have a Bachelors degree (D3=0), and are male (D4=0):

$$\begin{aligned} \text{X2 effect} &= \exp(10.478+0.020(8.93)-0.006X2+0.305D5+0.505D6-0.007D5X2) \\ &= \exp(10.657+0.305D5+0.505D6-(0.006+0.007D5)X2) \end{aligned}$$

In other words, the X2 effect for low-level managers (D5 = 0, D6 = 0) is  $\exp(10.657-0.006X2)$ , while for mid-level managers (D5 = 1, D6 = 0) it is #####, and for senior managers (D5 = 0, D6 = 1) it is  $\exp(11.162-0.006X2)$ . This can be illustrated on the following line graph:



Since  $\exp(-0.006) - 1 = -0.60\%$ , we can expect low-level and senior managers to have lower salaries if their degree was awarded a longer time ago, in particular 0.60% less for each additional year (holding X1, degree, and gender fixed). Similarly, since ##### = ##### %, we expect mid-level managers to #####.

Also, since  $\exp(0.305) - 1 = 35.7\%$ , mid-level managers with 0 work years since degree could expect a salary 35.7% higher than low-level managers, all other things being equal. However, with 10 work years since degree the salary premium is  $\exp(0.305 - 0.007(10)) - 1 = \exp(0.235) - 1 = 26.5\%$ , while with 20 work years since degree it has reduced to #####. Also, since ##### = ##### %, senior managers could expect a salary ##### % ##### than low-level managers, whatever the number of work years since degree, all other things being equal.

The degree effect does not vary with X1, X2, gender, or rank. The relevant regression parameter estimate is ##### with a p-value of #####, and so since  $\exp(0.114) - 1 = 12.1\%$ , we expect #####.

The parameter estimate for the gender effect is ##### with a p-value of #####, and so #####. Ordinarily, a statistically insignificant term such as this that is also not involved in interactions would be removed from a regression model; it has been retained here since it is the central focus of this analysis and its removal has very little effect on the regression parameter estimates for the other terms in the model.

To illustrate prediction and estimation using this model, we would like to know how much (in dollars) a new male low-level manager who got their Bachelors degree 2 years ago should be paid. Our best estimate is:

#####

A 95% confidence interval for the population mean salary of all such new low-level managers with these characteristics was computed to be:

$(\exp(10.397), \exp(\text{#####})) = (\$32,761, \$\text{#####})$

A 95% prediction interval for the actual salary of this individual low-level manager was calculated as:

$(\exp(10.312), \exp(\text{#####})) = (\$30,092, \$\text{#####})$

Overall, there is ##### evidence that women are being paid less than men, all other things being equal. In building a model to predict salary from X1, X2, degree, gender, and rank, neither D4 (the indicator variable for gender) nor its interactions were significant predictors of  $\log(\text{salary})$  (at the ##### significance level), after controlling for the effects of the other predictors included in the model.

**Space for Additional Notes:**

**Space for Additional Notes:**