Name: ____________________________________
MAT 222
Spring 2017
Exam 2 Review

Problem 1: Mark the following statements as true or false in the blank space provided. You should be able to explain why the statement is True/False.

(a) __________: A simple linear regression is modeled simply as $\beta_0 + \beta_1 x$.

(b) __________: For linear regression models, a small p–value indicates a strong linear relationship.

(c) __________: The population models for a simple linear regression are β_0, β_1, σ.

(d) __________: The larger the expected counts, the better X^2 approximates χ^2.

(e) __________: If one rejects the null hypothesis in a chi–squared test, then there is a cause/effect relationship between the rows/columns.

(f) __________: Chi–Squared calculations apply to more than simple counts.

(g) __________: Decreasing the confidence level or a decrease in the MSE *both* result in a decrease in the width of a confidence interval.

(h) __________: Chi–Squared hypothesis testing can only be done using a two–sided test.

(i) __________: Just because a particular explanatory variable is not statistically significant does not imply that it is not important in the model.

(j) __________: At a significance level of $\alpha = 0.05$ while testing the hypothesis that $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, a p–value of 0.001 is found. This means there is a strong linear relationship between the explanatory and response variables.

(k) __________: In a multivariate linear regression with very small F–value, all the coefficients must be close to zero.

(l) __________: Individual regression coefficients and all calculations about them can only be meaningfully interpreted within the context of other predictor variables.

(m) __________: In a chi–squared test, if one rejects the null hypothesis there must be a relation between every row/column.

(n) __________: The row/column sums must always agree in a computer analysis of chi–squared tests.

(o) __________: There is no way to apply linear methods to non–linear data sets.

(p) __________: One should always plot the data before creating a linear model.

(q) __________: It is always the case that SST=SSM+SSE.

(r) __________: Error in the measured x's for the model can have a great impact on the validity of a Simple Linear Regression.

(s) __________: Stating that a coefficient is statistically significant is the same as saying its effect is important.

(t) __________: Given any table of counts, one can use a chi–squared analysis.

(u) __________: A F–test and t–test can both be used to construct confidence intervals.

(v) __________: If you test $H_0 : \beta_1 = 0$ and reject the null hypothesis, then the only possibilities are that there is a linear relationship between the variables or there is a relationship between the variables–just not necessarily linear.

Problem 2: A cancer research group is investigating whether gene BRCA1 has any relation to how one responds to a new cancer treatment. The researchers examined 119 individuals undergoing cancer treatments, recording whether they possessed the gene and whether they responded to the treatment. The results are summarized below:

Table 1: BRCA1 Response Table

	Gene	No Gene
Yes	39	31
No	32	17

(a) What percent of individuals on the study had the gene?

(b) What percent of individuals in the study did not respond to the treatment?

(c) Given that a patient responded to the treatment, what is the probability that they contained the BRCA1 gene?

(d) Create a table of expected counts.

(e) Crate a table of 'residuals'.

(f) What is the degrees of freedom for this table?

(g) State the null and alternative hypothesis.

(h) What is X^2? What is the corresponding p–value? State the conclusion of the hypothesis test at $\alpha = 0.10$.

Problem 3: On a long road trip to discover yourselves, you and your friend decide to place bets on coin flips. Your friend provides the coin. At the end of the day, you have lost a lot of money to your friend. When you arrive at the hotel for the night while your friend heads to bed, you go down to the lobby and test the coin. Flipping the coin 10,000 times, you find a total of 5,103 heads.

(a) What proportion of heads should one expect?

(b) What proportion of heads was obtained?

(c) What is the percent residual from the previous parts?

(d) Determine whether or not it is probable that the coin was fair.

Problem 4: A university education group is testing whether the prestige of a university has any effect on the post–graduation employment. They break universities across the country into four categories: average (A), somewhat elite (SE), elite (E), and very elite (VE). The group surveys 385 students from the various groups and determines if they have found post–graduation employment.

Table 2: Post–Graduation Employment Counts

	A	SE	E	VE
Yes	35	50	61	52
No	60	47	36	44

Table 3: Post–Graduation Employment Expected Counts

	A	SE	E	VE
Yes		49.89	49.89	
No	46.14	47.11		46.63

Table 4: Post–Graduation Employment Residuals

	A	SE	E	VE
Yes		0.114		2.629
No	13.857	-0.114	-11.114	

(a) What percent of very elite university students find post–graduation employment?

(b) Of those students whom did not find post–graduation employment, what percent graduated from 'average' universities?

(c) What percent of the study individuals were from 'somewhat elite' universities?

(d) Complete the table of expected counts above.

(e) Complete the count of residuals above.

(f) State H_0, H_a and find the corresponding chi–squared value along with its probability. State your conclusions at $\alpha = 0.05, 0.01$.

Problem 5: Researchers try to see if there is a connection between age and the type of alcohol one regularly drinks. Their data is summarized below:

Table 5: 'Regular' alcohol usage among different age groups.

	15–22	22–29	30–35	35–40	40–60	Total
Liquor	1271	1205	607	345	401	3829
Beer	1497	2112	2115	2456	3019	
Wine	262			1772	6998	12,034
None	2063	1206	2097	2419	10,042	17,827
Total	5093		6333	6992	20460	44,889

Table 6: Expected values of 'regular' alcohol usage among different age groups.

	15–22	22–29	30–35	35–40	40–60
Liquor	434	513	540	596	1745
Beer	1271		1580	1744	5104
Wine	1365	1611	1698		5485
None		2387	2515	2777	8125

Table 7: 'Chi–Squared Residual' values of 'regular' alcohol usage among different age groups.

	15–22	22–29	30–35	35–40	40–60
Liquor	1,610.97	934.66	8.26		1,035.36
Beer	40.336	250.053		290.307	851.991
Wine	891.625		19.892	5.598	417.358
None	0.807	584.449	69.49	46.096	452.091

(a) Fill in the missing values from the table of counts.

(b) Fill in the missing values from the expected count table.

(c) Fill in the missing values from the 'chi–residual' table.

(d) What is the degree of freedom?

(e) State H_0 and H_a.

(f) What is the p–value? At $\alpha = 0.05$, what is the conclusion?

(g) Should the conclusion have been expected by looking at the original table?

Problem 6: Researchers are trying to analyze various demographics of Hispanic Americans. The factors they are interested in are influenced by one's income level. The researchers want to check that their study 'fits' the national distribution of Hispanic American income levels. On average 23.5% of Hispanic Americans are in the bottom 20% of income levels, 24.4% are in the next 20% of incomes, 22.3% are in the next income bracket, 18.3% are in the next, and 11.5% are in the top 20% of income levels. Their study consists of the following distribution of incomes among the Hispanic Americans surveyed: According to their data, do their surveyed Hispanic Americans 'fit'

Table 8: Surveyed Hispanic Americans broken down by income level.

Income Level	0–20%	20–40%	40–60%	60–80%	80–100%
Count	1215	1232	1205	921	552

the average distribution of Hispanic American income levels?

Problem 7: A company is trying to predict the average production costs for the coming fiscal year. They have a statistician create a linear model predicting the cost (in thousands) for production given the amount of items they produce (in thousands). The computer output for the model is found below.

```
Analysis of Variance

            Source       DF   Adj SS   Adj MS   F-Value   P-Value
            Regression          93657                       0.000
            Items               93657                       0.000
            Error                        472.9
            Total        16   100750

Model Summary

                    S     R-sq   R-sq (adj)   R-sq (pred)
                        92.96%       92.49%        90.47%

Coefficients

            Term          Coef   SE Coef   T-Value   P-Value    VIF
            Constant       2.0      10.1      0.20     0.846
            Items       10.101     0.718               0.000   1.00

The regression equation is

production cost = ________________________________________________
```

(a) Fill in the missing items in the tables above.

(b) What was the total amount of data points used to create the model?

(c) Predict the average production cost if the company produces 18,000 items.

(d) What is the correlation coefficient?

(e) What is the coefficient of determination?

(f) What proportion of the variation in y is explained by the variation in x for this model?

(g) What is the constant in the model? Interpret the constant in the context of the problem.

(h) Find a 95% confidence interval for β_1. Interpret the result.

(i) Conduct a hypothesis test for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. State the p–value, degrees of freedom, and interpret the result carefully. [Use $\alpha = 0.05$.]

Problem 8: A company is trying to determine if more experienced workers are more productive than 'fresh' employees. The company hires a statistician to create a linear model predicting the average number of new customers an employee processes each year based on the number of years the employee has been at the company. The model is summarized below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	727.9	727.88	8.75	
Years	1	727.9	727.88	8.75	
Error	9				
Total					

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
		43.65%	14.17%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	99.99				
Years	2.572	0.870	2.96	0.016	1.00

The regression equation is

Number Customers = ______________________________

(a) Fill in the missing entries in the tables above. [Note: $\sum(x_i - \overline{x})^2 = 110$ and the average employee used to create the model had worked at the company for 5 years.]

(b) Use the model to predict the number of customers an employee would add on average that year if the employee had worked at the company for 9 years.

(c) What is the correlation coefficient?

(d) What is the coefficient of determination?

(e) What proportion of the variation in y is explained by the variation in x for this model?

(f) What was the total number of data points used to create the model?

(g) Construct a 90% confidence interval for β_0.

(h) Does β_0 have any meaning in this problem? Explain.

(i) Conduct a hypothesis test $H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 > 0$ using $\alpha = 0.10$. Interpret your results. Be sure to give the t–value, p–value, and degrees of freedom.

(j) Can we say that the data is linear?

(k) Conduct an F–test for the regression. State H_0 and H_a. How does this test differ from the hypothesis test you performed above?

Problem 9: A statistics student is preparing for an exam. They have to fill in the missing values from the model given below. However, they are nervous and want to be able to check their answers. Fill in the values so that the student will have a solution manual to which compare their answers. Be sure to indicate for the student whether this was a simple linear regression or a multiple linear regression and how many data values were used to create the model.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression		0.29477		4.00	0.067
Error	13	0.95920	0.07378		
Lack-of-Fit	11	0.90620	0.08238	3.11	0.268
Pure Error	2		0.02650		
Total					

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
		17.62%	0.19%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.185	0.126		0.166	
Regression	0.541	0.270	2.00	0.067	1.00

The regression equation is

y = ____

Problem 10: The same student returns to you for more help–you'd better start charging! Fill in the values so that the student will have a solution manual to which compare their answers. Be sure to indicate for the student whether this was a simple linear regression or a multiple linear regression and how many data values were used to create the model. Furthermore, explain to the student whether one can predict y using the input variable(s). Explain whether the model is linear or not. Then help the student construct a confidence interval (a 95% confidence interval) and hypothesis test (of $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$) on the variable(s).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression		3317.23	1658.62		0.000
Var1		2653.19		3151.70	0.000
Var2			664.04	788.81	0.000
Error		27.78	0.84		
Total	35				

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
		99.12%	99.01%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	15.081	0.352	42.90	0.000	
Var1	2.5134	0.0448	56.14	0.000	1.00
Var2	-1.2574	0.0448	-28.09	0.000	1.00

The regression equation is

y = ____

Problem 11: A finance company is investigating the effects of different variables on total average pre–tax income (measured in US dollars). The variables used in the model were education (in number of years), age (in years), residence (number of years living at the current residence), savings (in US dollars), debt (in US dollars), and the number of credit cards one has. Mathematicians at the company use a statistics program to create a multivariate linear regression. The output is given below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression		5673515090	810502156	4.19	
Education		2021974124	2021974124	10.45	0.004
Age		603358164	603358164		0.091
Residence		6336044	6336044	0.03	0.858
Employment		45377529		0.23	0.633
Savings		727084527	727084527	3.76	0.066
Debt		253851061	253851061	1.31	0.264
Credit Cards			51710387	0.27	0.610
Error		4257582430			
Total	29	9931097520			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
	57.13%	43.49%	8.85%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-90642	37172	-2.44		
Education	4827	1493	3.23	0.004	1.52
Age	2098		1.77	0.091	4.53
Residence		2313	-0.18	0.858	12.71
Employment	1291	2666	0.48	0.633	14.31
Savings	-1.015	0.524	-1.94	0.066	3.00
Debt	-0.854	0.746	-1.15	0.264	2.12
Credit Cards	1348	2607	0.52	0.610	1.60

The regression equation is

income = ____

(a) Fill in the missing values in the table above.

(b) Which variable(s) are statistically significant?

(c) Which variable(s) are *not* statistically significant?

(d) Predict the average income of an individual of 40 years with 5 years of education, having lived at their current residence for 3 years, worked at their current job for 6 years, has saved $5,000, has 4 credit cards, and is $2,300 in debt.

(e) Construct a 95% confidence interval for the coefficient for debt.

(f) Perform a hypothesis test for $H_0 : \beta_5 = 0$ versus $H_a : \beta_5 < 0$, where β_5 is the coefficient for savings. Perform the test at $\alpha = 0.10, 0.05, 0.01$.

(g) Is the model statistically significant? Explain. Is the model linear?

(h) Below one finds an alternative model used to predict income. Explain what is different about this model.

(i) Was 'age' statistically significant in the original model? In the new model? Comment on your answers.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	3579865782	1193288594	4.88	0.008
Age	1	2989704170	2989704170	12.24	0.002
Savings	1	494566869	494566869	2.02	0.167
Debt	1	865748478	865748478	3.54	0.071
Error	26	6351231738	865748478	3.54	0.071
Total	29	9931097520			

Model Summary

S	R-sq	R-sq (adj)	R-sq (pred)
15629.4	36.05%	28.67%	11.69%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-32377	23513	-1.38	0.180	
Age	2784	796	3.50	0.002	1.61
Savings	-0.666	0.468	-1.42	0.167	1.90
Debt	-1.246	0.662	-1.88	0.071	1.32