APPLIED REGRESSION | Sample Test-3 | SPRING

Name: ______________________ | Instructor: D. Kushary

1. The data (sheet=Logistic) represents the test firing results for 25 surface-to-air anti-aircraft missiles at targets of varying speed (in knots). The results of each test is either a hit (Y=1) or a miss (Y=0).

(a) Fit a logistic regression model to the response variable y and write down the estimated model. Use a simple linear regression model as the structure of the linear predictor.

R- output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.0709	2.1090	2.88	0.0040
x	-0.0177	0.0061	-2.91	0.0036

Solution: The estimated model is $\hat{\pi} = \frac{1}{1+e^{-(6.0709-0.0177\times X)}}$

(b) Find a 98% confidence interval for the odds ratio when speed goes up by 50 (knots)

Solution: The odds ratio when speed goes up by 50 is $e^{50\times\hat{\beta}_1} = 0.4126172$ and the 98% CI is (0.2035339, 0.8364847)

(c) Does the model deviance indicate that the above model is adequate at 10% level?

	Df	Deviance	Resid. Df	Resid. Dev
NULL			24	34.62
x	1	14.25	23	20.36

Model deviance is 20.36 with df=23 which implies the p-value = 0.6199. Hence the model is adequate at 10% level.

(d) Expand the linear predicted to include a quadratic term in the Target speed. Is there any evidence that this quadratic term is required in the model. Use 5% significance level.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.1927	9.0299	0.69	0.4928
x	-0.0185	0.0551	-0.33	0.7377
x2	0.0000	0.0001	0.01	0.9889

To test the significance of the quadratic term we set the hypothesis as $H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$. Now from the above table we see that the p-value for the test is 0.9889. Hence the addition of the quadratic term is insignificant at 5% level.

2. Do average automobile insurance costs differ for different insurance companies? One of the other variables which determines the cost is location. To test the theory, estimates (in dollars) are taken for a fixed type of drivers from 3 insurance companies (1-State Farm 2-Allstate 3-AAA). Each of the three companies provided estimates for four different cities(A-Riverside, B-San Bernadino, C-Hollywood and D-Long Beach). Data was analyzed using two-way ANOVA model without interaction and one observation per cell. The following summary statistic are available. SST=86, SS(Location)=42 and F-Stat(Location) = 7. Total for each of the companies are $T_1 = 100.T_2 = 324, T_3 = 196$.

   (a) Write down the model, factor and levels for the problem.

   Model: $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ A=Insurance Company(3) B=Location(5)

(b) Complete the ANOVA table and test if there is sufficient evidence to indicate that average insurance premiums differ from company to company at 5% level?

Source	Sum of Squares	Deg of Freedom	Mean Squares	F - Stat	P -value
Insurance Company	32	2	16	8	0.0203
Location	42	3	14	7	0.0219
Error	12	6	2		
Total	86	11			

Step - 1 DFs are (k-1) = 3 -1 =2, b -1 = 4 -1 =3 and for error (k-1)(b-1) = 6.

Step - 2 - MSB = SSB/(b-1) = 42/3 = 14

Step - 3 - $\frac{MSB}{MSE} = F_B - Stat \quad \Rightarrow \quad MSE = \frac{MSB}{F_B - Stat} = \frac{14}{7} = 2$

Step - 4 - $SSE = 2 \times 6 = 12 \quad \Rightarrow \quad SSA = 86 - 12 - 42 = 32$

Step - 5 $MSA = 32/2 = 16 \quad \Rightarrow \quad F_A - stat = MSA/MSE = 16/2 = 8$

To test the carriers we set
$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \ vs \ H_1$: Not all of them 0.

As $F_A - Stat = 8 > F_{0.05,2,6} = 5.14$ we reject the null hypothesis.

Conclusion: The data provides sufficient evidence to support the fact that insurance companies are different in average cost at 5% level.

3. Over the past 20 years, inventory carrying costs for a large tire manufacturing facility have been as shown in the data file (sheet=Timeseries). Data are in thousands of dollars.

(a) Write down the time series model and plot the data and fit a simple linear regression using time (t) as the independent variable and test for positive autocorrelation at 5% level. (must write hypothesis and p-values)

The time series model is:

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t \quad \text{where} \quad \epsilon_t = \rho\, \epsilon_{t-1} + a_t \quad \text{and} \quad a_t \sim N(0, \sigma^2)$$

The estimated SLR model is: $\hat{y}_t = 4206.10 - 88.3789 \times t$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4206.1032	145.8094	28.85	0.0000
t	-88.3789	12.1719	-7.26	0.0000

To the test the auto-correlation we set up the hypothesis as $H_0 : \rho \leq 0 \; vs \; H_1 : \rho > 0$.

p-value ≈ 0. Hence we reject the null hypothesis.

lag	Autocorrelation	D-W Statistic	p-value
1	0.681758	0.308722	0

(b) If positive autocorrelation is present, use Cochran-Orcutt method and write down the updated estimated model and then predict the carrying cost for the next year.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4229.8451	167.5960	25.24	0.0000
xprime	-247.2012	36.9878	-6.68	0.0000