Linear Regression Problems

Part A: Simple Linear Regression/Correlation

Q.A.1. A simple linear regression was fit relating number of species of arctic flora observed (Y) and July mean temperature (X, in Celsius). The results of the regression model, based on n=19 temperature stations is given below.

ANOVA						
	df	*SS*	*MS*	*F*	*gnificance F*	
Regressio	1	39858	39858	79.87	0.0000	
Residual	17	8484	499			
Total	18	48342				
Coefficients		*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-34.49	16.56	-2.08	0.0527	-69.43	0.46
JulyTemp	24.60	2.75	8.94	0.0000	18.79	30.41
SS_XX	X-bar					
65.85	5.7					

p.1.a. What proportion of the variation in number of species is "explained" by mean July temperature?

p.1.b. Compute a 95% Confidence Interval for the population mean number of species, with mean July temperature of 6 degrees.

p.1.c. Compute a 95% Prediction Interval for the number of species, at a single station with mean July temperature of 6 degrees.

p.1.d. Test H_0: $\rho = 0$ vs H_A: $\rho \neq 0$

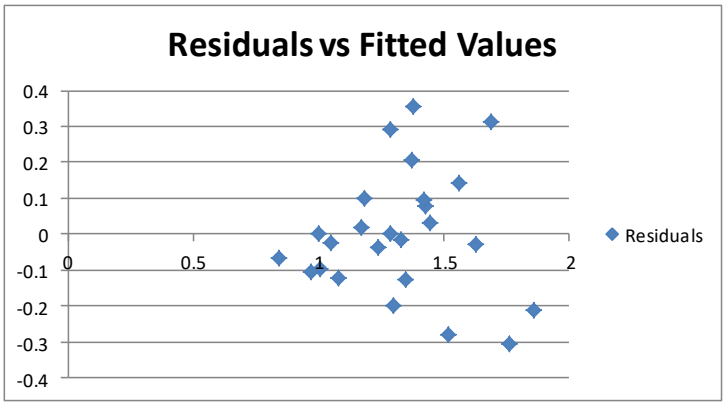
p.1.e. Obtain an approximate 95% Confidence Interval for ρ

Q.A.2. A linear regression was (inappropriately) run, relating success in throwing a frisbee through a hula-hoop (Y=1, if good, 0 if not) to children's eye-color (X = 1 if dark, 0 if light). Note that this should be conducted as a chi-square test, or logistic regression (it was a very old paper). The authors report that $r^2 = .030$ (well, that's what it should be) from a sample of n = 136 children. Based on the regression : $E(Y) = \beta_0 + \beta_1 X$, test H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$ based on the F-test.

Part B: Model Diagnostics (Simple and Multiple Regression)

Q.B.1. An experiment was conducted, relating the penetration depth of missiles (Y) to its impact factor (X). The results from the regression, and the residual versus fitted plot are given below (n=25).

ANOVA		
	df	*SS*
Regressio	1	1.585884
Residual	23	0.713406
Total	24	2.29929
	Coefficients	*tandard Err*
Intercept	0.633253	0.103076
impact	0.06	0.008391



p.1.a. Test H_0: $\beta_1 = 0$ (Penetration depth is not associated with impact factor) based on the t-test.

Test Statistic: _____ Rejection Region: _____

p.1.b. Test H_0: $\beta_1 = 0$ (Penetration depth is not associated with impact factor) based on the F-test.

Test Statistic: _____ Rejection Region: _____

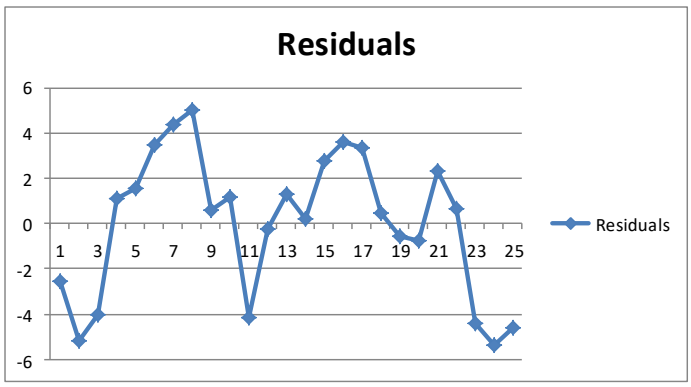
p.1.c. The residual plot appears to display non-constant error variance. A regression of the squared residuals on the impact factors (X) is fit, and the ANOVA is given below. Conduct the Breusch-Pagan test to test whether the errors are related to X. Do you reject the null hypothesis of constant variance? **Yes** or **No**

ANOVA		
	df	*SS*
Regressio	1	0.007959
Residual	23	0.025824
Total	24	0.033783

Test Statistic: _____ Rejection Region: _____

Q.B.2. A regression model was fit, relating the share of big 3 television network prime-time market share (Y, %) to household penetration of cable/satellite dish providers (X = MVPD) for the years 1980-2004 (n=25). The regression results and residual versus time plot are given below.

ANOVA				
	df	*SS*	*MS*	*F*
Regressio	1	7073.7	7073.7	685.7
Residual	23	237.3	10.3	
Total	24	7311.0		
	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	112.029	2.090	53.61	0.0000
mvpd	-0.863	0.033	-26.19	0.0000



p.2.a. Compute the correlation between big 3 market share and MVPD.

p.2.b. The residual plot appears to display serial autocorrelation over time. Conduct the Durbin-Watson test, with null hypothesis that residuals are not autocorrelated.

$$\sum_{t=2}^{25}\left(e_t - e_{t-1}\right)^2 = 161.4 \qquad d_L\left(\alpha = 0.05, n = 25, p = 1\right) = 1.29 \quad d_U\left(\alpha = 0.05, n = 25, p = 1\right) = 1.45$$

Test Statistic: _____ Reject H_0? **Yes** or **No**

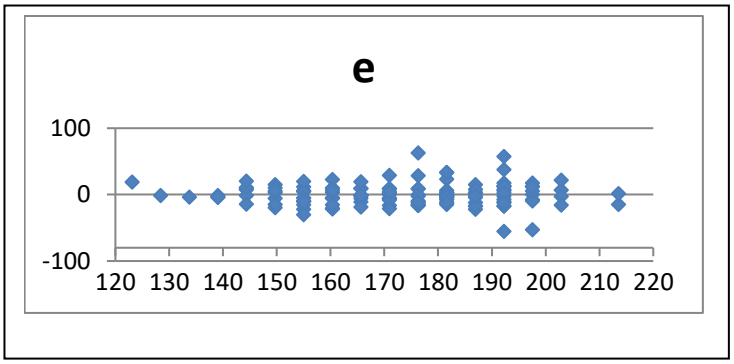
p.2.c. Data were transformed to conduct estimated generalized least squares (EGLS), to account for the auto-correlation. The parameter estimates and standard errors are given below. Obtain 95% confidence intervals for β_1, based on Ordinary Least Squares (OLS) and EGLS. Note that the error degrees' of freedom are 23 for OLS and 22 for EGLS (estimated the autocorrelation coefficient).

beta-egls	SE(b-egls)
110.577	3.469
-0.845	0.055

OLS 95% CI: _____ EGLS 95% CI: _____

Q.B.3. A regression model is fit relating weight (Y, in lbs) to height (X, in inches) among n=139 WNBA players (treating this as a random sample of all female athletes from sports such as basketball and volleyball). The results from the regression, and the residual versus fitted plot are given below.

ANOVA		
	df	*SS*
Regressio	1	45933.13
Residual	137	35074.41
Total	138	81007.54
	Coefficients	*andard Err*
Intercept	-212.05	28.78
Height	5.32	0.40



p.3.a. Test H_0: $\beta_1 = 0$ (Weight is not associated with Height) based on the t-test.

Test Statistic: _____ Rejection Region: _____ Reject H_0? **Yes** or **No**

p.3.b. Test H_0: $\beta_1 = 0$ (Weight is not associated with Height) based on the F-test.

Test Statistic: _____ Rejection Region: _____ Reject H_0? **Yes** or **No**

p.3.c. The residual plot appears to display non-constant error variance. A regression of the squared residuals on height (X) is fit, and the ANOVA is given below. Conduct the Breusch-Pagan test to test whether the errors are related to X. Do you reject the null hypothesis of constant variance? **Yes** or **No**

ANOVA		
	df	*SS*
Regressio	1	1070500
Residual	137	44788405
Total	138	45858905

Test Statistic: _____ Rejection Region: _____ Reject H_0? **Yes** or **No**

Q.B.4. A linear regression model was fit, relating the electric potential (Y) to the current density (X) for a series of n=18 consecutively observed pairs of X and Y.

Model: $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \qquad \nu_t \sim N\left(0, \sigma^2\right)$

p.4.a. The estimated regression coefficients and standard errors for ordinary least squares are given below. Obtain a 95% Confidence Interval for the slope coefficient for current density (β_1).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 891.460256   4.290707  207.76   <2e-16 ***
x            -0.431633   0.009523  -45.33   <2e-16 ***
```

Lower Bound: _____ Upper Bound: _____

p.4.b. The Durbin-Watson test is used to test H_0: $\rho = 0$. The test statistic and P-value are given below. Do you conclude that the errors are serially correlated (not independent)? **Yes** or **No**

```
lag Autocorrelation D-W Statistic p-value
   1       0.4923645      0.668015       0
```

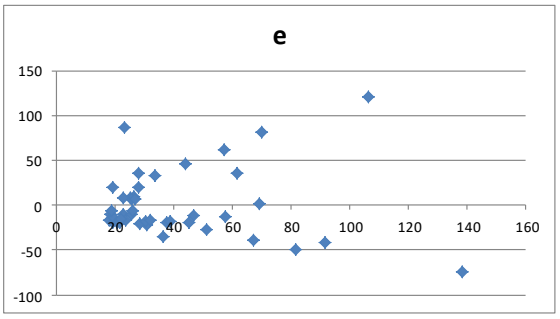
p.4.c. The estimated regression coefficients and standard errors for estimated generalized least squares are given below. Obtain a 95% Confidence Interval for the slope coefficient for current density (β_1)

```
Coefficients:
              Value Std.Error   t-value p-value
(Intercept) 895.8873  9.161308  97.79033       0
x            -0.4488  0.018576 -24.16305       0
```

Lower Bound: _____ Upper Bound: _____

Q.B.5. A study compared n = 43 island nations with respect to various demographic and transportation measures. The authors fit a linear regression relating Vehicles/road length (Y, cars/km) to GDP (X, $1000s/capita). The regression output and residual versus predicted plot are given below.

	df	*SS*
Regressio	1	29066
Residual	41	57786
Total	42	86852
	Coefficients	*Standard Error*
Intercept	17.41	7.66
GDP_K	4.33	0.95



p.5.a. Test H_0: $\beta_1 = 0$ (Car density is not related to GDP) based on the t-test.

Test Statistic: _____ Rejection Region: _____

p.5.b. Test H_0: $\beta_1 = 0$ (Car density is not related to GDP) based on the F-test.

Test Statistic: _____ Rejection Region: _____

p.5.c. The residual plot appears to potentially display non-constant error variance. A regression of the squared residuals on GDP (X) is fit, and the ANOVA is given below. Conduct the Breusch-Pagan test to test whether the errors are related to X. Do you reject the null hypothesis of constant variance? **Yes** or **No**

	df	*SS*
Regressio	1	111098168
Residual	41	194285510
Total	42	305383678

Test Statistic: _____ Rejection Region: _____

Q.B.6. A regression model was fit, relating points scored in 2014 WNBA games by Skylar Diggins (Y) to whether the game was a Home game (X_1 = 1 if Home, 0 if Away) and the number of minutes she played (X_2) over a season of n = 34 games. The regression results are given below for the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

ANOVA						
	df	***SS***	***MS***	***F***	***F(0.05)***	***R^2***
Regression	**2**	**404.6**	**202.3**			
Residual	**31**	**958.1**	**30.9**	**#N/A**	**#N/A**	**#N/A**
Total	**33**	**1362.7**	**#N/A**	**#N/A**	**#N/A**	**#N/A**

p.6.a. Complete the ANOVA table. Do you conclude that Skylar's average point total is associated with the game being at home, and/or the number of minutes she played? **Yes** **No**

p.6.b. Conduct the Durbin-Watson test, with null hypothesis that residuals are autocorrelated.

$$\sum_{t=2}^{25}\left(e_t - e_{t-1}\right)^2 = 1794.5 \qquad d_L\left(\alpha = 0.05, n = 34, p = 2\right) = 1.33 \quad d_U\left(\alpha = 0.05, n = 34, p = 2\right) = 1.58$$

Test Statistic: _____ Reject H_0? **Yes** or **No**

p.6.c. Data were transformed to conduct estimated generalized least squares (EGLS), to account for potential auto-correlation. The parameter estimates and standard errors are given below. Obtain 95% confidence intervals for β_2, based on Ordinary Least Squares (OLS) and EGLS. Note that the error degrees' of freedom are 34-3=31 for OLS and 30 for EGLS (estimated the autocorrelation coefficient).

	OLS	**OLS**	**EGLS**	**EGLS**
Parameter	**Estimate**	**StdError**	**Estimate**	**StdError**
Intercept	**-15.6**	**10.09**	**-15.38**	**10.19**
Home	**-2.13**	**1.95**	**-1.99**	**1.95**
Minutes	**1.05**	**0.29**	**1.04**	**0.29**

OLS 95% CI: _____ EGLS 95% CI: _____

Q.B.7. A simple linear regression model was fit relating Weight (Y, in pounds) to Height (X, in inches) for a random sample of n=52 National Hockey League players. The total sum of squares, TSS = 9237, and $R^2 = 0.262$.

p.7.a. Complete the following ANOVA table.

ANOVA					
Source	*df*	*SS*	*MS*	*F*	*F(0.05)*
Regression					
Residual				#N/A	#N/A
Total		9236.7	#N/A	#N/A	#N/A

p.7.b. Complete the following table and use it to conduct the F-test for Lack-of-Fit (there are c = 9 distinct heights).

$H_0 : \mu_j = \beta_0 + \beta_1 X_j \quad j = 1,...,c \qquad H_A : \mu_j \neq \beta_0 + \beta_1 X_j$

$SSLF = \sum_{j=1}^{c} n_j \left(\overline{Y}_j - \hat{Y}_j \right)^2 \qquad df_{LF} = c - 2 \qquad SSPE = \sum_{j=1}^{c} \left(n_j - 1 \right) S_j^2 \qquad df_{PE} = n - c$

Height	n	Y-bar	Y-hat	n*(YB-YH)	(n-1)S^2
70	2	193.00	190.31	14.42	128.00
71	6	198.50	194.12	114.95	785.50
72	7	197.86	197.93	0.04	330.86
73	13	199.15	201.74	86.87	1357.69
74	11	202.27	205.55	117.91	1782.18
75	6	211.50	209.35	27.61	433.50
76	4	219.50	213.16	160.65	1451.00
77	2	216.50	216.97	0.44	24.50
78	1	222.00	220.78		0.00
Sum	52	#N/A	#N/A		

Test Statistic: _____ Reject H_0 if Test Stat falls in Range: _____

Q.B.8. A simple linear regression model is fit, relating Orlando June Total Precipitation (Y) to Mean Temperature (X) over an n = 45 year period. The following table gives the results.

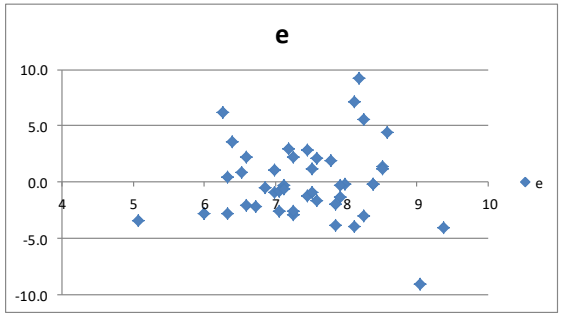
ANOVA		
	df	*SS*
Regression	1	32.59982
Residual	43	489.2068
Total	44	521.8067
	Coefficients	*andard Err*
Intercept	61.3280	31.8335
meanTemp	-0.6626	0.3915

p.8.a. Use the t-test to test H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$ at $\alpha = 0.10$ significance level

Test Statistic: _____

Reject H_0 if Test Stat falls in range: _____

A plot of the residuals displays a possible case of unequal variances. The regression of the squared residuals on X is given below:



	df	*SS*
Regression	**1**	**1326.60**
Residual	**43**	**14436.63**
Total	**44**	**15763.24**

p.8.b. Conduct the Breusch-Pagan test to test H_0: Equal Variances H_A: Variance is related to X. (Use $\alpha = 0.05$):

Test Statistic: _____ Reject H_0 if Test Statistic falls in the Range: _____

Q.B.9. Monthly mean temperatures for Boston (Y, in Fahrenheit) for the years 1920-2014 are fit using a linear regression model to Year (X_1=Year-1920) and 11 monthly dummy variables (X_2 = 1 if January, 0 otherwise,…, X_{12} = 1 if November, 0 otherwise, Note that December is the reference month). The ANOVA table and regression coefficient estimates are given below.

ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regression	12	270975.961	22581.330	2842.345	0.000
Residual	1127	8953.578	7.945		
Total	1139	279929.539			
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	
Intercept	33.343	0.323	103.344	0.000	
year	0.013	0.003	4.406	0.000	
month1	-4.563	0.409	-11.158	0.000	
month2	-3.285	0.409	-8.033	0.000	
month3	4.312	0.409	10.543	0.000	
month4	14.171	0.409	34.649	0.000	
month5	24.313	0.409	59.449	0.000	
month6	33.787	0.409	82.616	0.000	
month7	39.412	0.409	96.368	0.000	
month8	37.906	0.409	92.688	0.000	
month9	30.806	0.409	75.327	0.000	
month10	20.758	0.409	50.757	0.000	
month11	10.793	0.409	26.390	0.000	

p.9.a. Give the predicted temperatures for December 1920, June (Month 6) 1920, December 2010, and June 2010.

	1920	2010
December		
June		

p.9.b. Compute a 95% Confidence Interval for the change in annual mean temperature, controlling for month.

Lower Bound: _____ Upper Bound: _____

p.9.c. Compute the Durbin-Watson statistic. $\sum_{t=2}^{1140}(e_t - e_{t-1})^2 = 14094.8$

DW = _____

p.9.d. What proportion of the variation in temperature is explained by the model?

Q.B.10. A regression model was fit, relating Price (Y, in $1000s) to acceleration rate (X_1) and Miles per gallon (X_2) for a sample of n = 25 models of hybrid compact cars. The fitted equation and summary model statistics are given below.

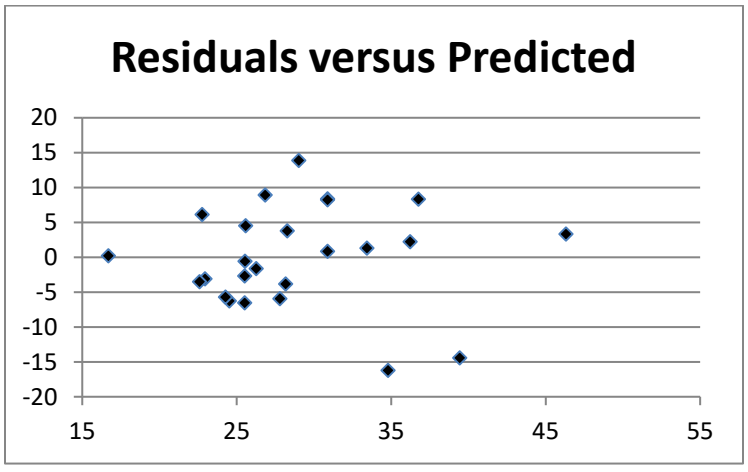
$\hat{Y} = -26.35 + 4.50X_1 + 0.20X_2 \quad SSR = 957 \quad SSE = 1239$

p.10.a. Test whether price is related to either acceleration rate and/or Miles per gallon. H_0: $\beta_1 = \beta_2 = 0$.

Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

p.10.b. A plot of the residuals versus predicted values suggests a possible non-constant error variance. A regression of the squared residuals on X_1 and X_2 yields SS(Reg*) = 3786261. Test:

H_0: Equal Variance Among Errors $\sigma^2\{\varepsilon_i\} = \sigma^2 \ \forall i$ H_A: Unequal Variance Among Errors $\sigma_i^2 = \sigma^2 h(\gamma_1 X_{i1} + \gamma_2 X_{i2})$



Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

Q.B.11. Monthly mean temperatures for Houston (Y, in Fahrenheit) for the years 1947-2014 are fit using a linear regression model to Year (X_1=Year-1947) and 11 monthly dummy variables (X_2 = 1 if January, 0 otherwise,…, X_{12} = 1 if

November, 0 otherwise, Note that December is the reference month). The ANOVA table and regression coefficient estimates are given below.

ANOVA				
	df	*SS*	*MS*	*F*
Regressio	12	98045.96	8170.497	1164.116
Residual	803	5635.958	7.018627	
Total	815	103681.9		
	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	54.7357	0.3581	152.8291	0
Year1947	0.0277	0.0047	5.860017	6.75E-09
Month01	-2.0941	0.4543	-4.60908	4.7E-06
Month02	0.8632	0.4543	1.899952	0.057797
Month03	7.0544	0.4543	15.52652	9.75E-48
Month04	14.0868	0.4543	31.00449	2.1E-139
Month05	20.7985	0.4543	45.77686	5E-226
Month06	26.2000	0.4543	57.66531	1E-287
Month07	28.1074	0.4543	61.86333	1E-307
Month08	28.2221	0.4543	62.1158	0
Month09	24.2809	0.4543	53.4414	1.1E-266
Month10	16.0529	0.4543	35.33198	1E-165
Month11	6.4794	0.4543	14.26097	2.69E-41

p.11.a. Give the predicted temperatures for December 1947, June (Month 6) 1947, December 2007, and June 2007.

	1947	2007
December		
June		

p.11.b. Compute a 95% Confidence Interval for the change in annual mean temperature, controlling for month.

Lower Bound: _____ Upper Bound: _____

p.11.c. Compute the Durbin-Watson statistic. $\sum_{t=2}^{816}\left(e_t - e_{t-1}\right)^2 = 8476.684$

DW = _____

Q.B.12. A regression model was fit, relating Total Points (Y = 3*Wins + Draws) to Total Salary (X, in millions of pounds) for the n = 20 English Premier Football League 1998/1999 season. The fitted equation and summary model statistics are given below.

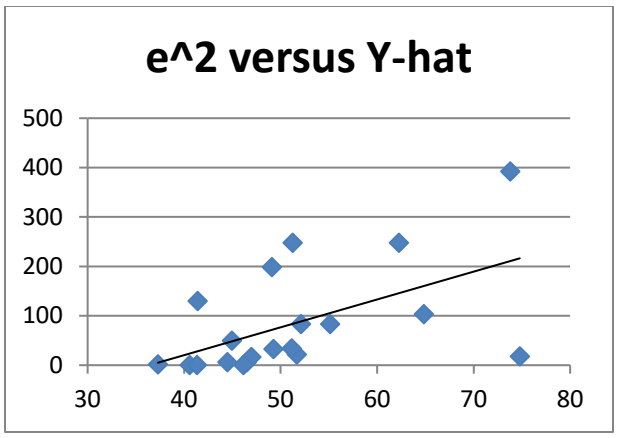
$\hat{Y} = 29.19 + 4.43X \quad SSR = 2050 \quad SSE = 1674$

p.12.a. Test whether Total Points is related to Total Salary. H_0: $\beta_1 = 0$.

Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

p.12.b. A plot of the squared residuals versus predicted values suggests a possible non-constant error variance. A regression of the squared residuals on X yields SS(Reg*) =65105. Test:

H_0: Equal Variance Among Errors $\sigma^2\{\varepsilon_i\} = \sigma^2 \ \forall \ i \quad H_A$: Unequal Variance Among Errors $\sigma_i^2 = \sigma^2 h(\gamma_1 X_i)$



Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

Q.B.13. An experiment was conducted relating strength of ice cream cones (Y, in newtons) to the temperature (X_1, in C) and relative humidity (X_2, in %) on $n = 20$ consecutive work days. The following model was fit, with the results of calculations given below.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \varepsilon \sim N\left(0, \sigma^2\right) \text{ independent}$$

$$\hat{Y} = 42.766 + 0.146 X_1 - 0.390 X_2 \quad \sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2 = 813.0 \quad \sum_{t=1}^{n} e_t^2 = 650.5 \quad \sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2 = 1168.6$$

p.13.a Compute the coefficient of determination, R^2.

p.13.b. Test $H_0 : \beta_1 = \beta_2 = 0$ by giving the Test Statistic, Rejection Region, and P-value relative to .05.

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

p.13.c. The critical values for the Durbin-Watson statistic for $n = 20$ and $p = 2$ are $d_L = 1.10$ and $d_U = 1.54$. Compute the Durbin-Watson statistic for testing H_0: the errors are not autocorrelated and circle the best conclusion.

D-W Statistic: _____ Conclude: Reject H_0 Accept H_0 Inconclusive

Q.B.14. An experiment was conducted relating springiness in berries (Y, in mm) to sugar equivalent (X, in g/L). There were $c = 6$ distinct sugar equivalent groups, with $n_j = 5$ berries per group. The lack-of-fit test for a linear relation is:

$$H_0 : E\left\{Y_{ij}\right\} = \mu_j = \beta_0 + \beta_1 X_j \quad i = 1,\ldots,5;\ j = 1,\ldots,6 \quad H_A : E\left\{Y_{ij}\right\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

ANOVA							X	n_j	Yhat_j	Ybar_j	s_j
	df	*SS*	*MS*	*F*	*gnificance F*		176.5	5	2.1249	2.0618	0.2132
Regressio	1	2.2089	2.2089	46.3966	0.0000		192.6	5	1.9703	2.0250	0.1540
Residual	28	1.3330	0.0476				209.3	5	1.8100	1.7818	0.1996
Total	29	3.5419					225	5	1.6593	1.7536	0.4052
							242.1	5	1.4951	1.4804	0.1501
	Coefficients	*andard Err*	*t Stat*	*P-value*			259.5	5		1.2852	0.1188
Intercept	3.8193	0.3091	12.3555	0.0000							
sugCont	-0.0096	0.0014	-6.8115	0.0000							

p.14.a. Give the fitted value for the linear regression for the 6th group ($X_6 = 259.5$).

p.14.b. Compute the Pure Error Sum of Squares, degrees of freedom and Mean Square.

SS_{PE} = _____ df_{PE} = _____ MS_{PE} = _____

p.14.c. Compute the Lack-of-Fit Sum of Squares, degrees of freedom and Mean Square.

SS_{LF} = _____ df_{LF} = _____ MS_{LF} = _____

p.14.d. Give the Test Statistic, Rejection Region, and P-value relative to .05 for the Lack-of-Fit test.

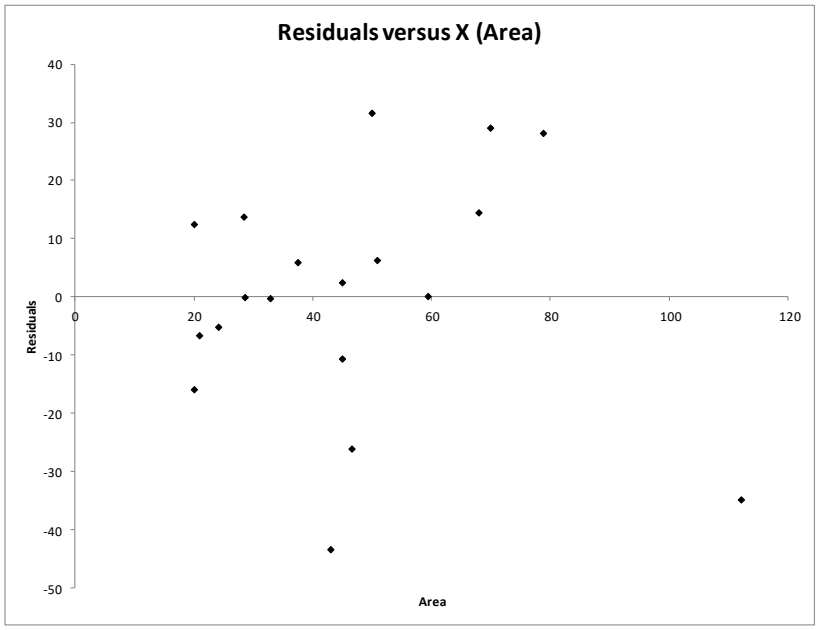
Test Statistic: _____ Rejection Region: _____ P-value > or < .05

Q.B.15. A regression model is fit, relating energy consumption (Y) to total area (X) for a sample of *n* = 19 luxury hotels in Hainan Province, China. The Analysis of Variance for the simple linear regression model is given below.

ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regressio	1	25521.76	25521.76	57.75	0.0000
Residual	17	7512.94	441.94		
Total	18	33034.70			

p.15.a. A plot of the residuals for area is given below. It demonstrates which possible violations of assumptions (circle all that apply).

Non-normal Errors Unequal Variance Serial Correlation of Errors Non-linear Relation between Y and X



p.15.b. A second regression model is fit, relating the squared residuals (Y) to area (X). Conduct the Breusch-Pagan test to test whether the equal variance assumption is reasonable. The sums of squares are given below.

ANOVA		
	df	*SS*
Regressio	1	1239379
Residual	17	3871645
Total	18	5111024

Test Statistic: _____ Rejection Region: _____ P-value $>$ or $<$ .05

Q.B.16. A study considered the effect of NaCl on chewiness of chewiness of berries. There were 6 levels of NaCl (130 to 180 by 10). There were $n_i = 5$ replicates per treatment. Note that this data can be analyzed as a 1-Way ANOVA (as the authors did) or as a simple linear regression. Consider the following 2 models (with notation from the 1-Way ANOVA: *i*=Treatments (or distinct X levels), and *j*=replicates).

Model 1: $Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1,...,t;\ j = 1,...,n_i$ Model 2: $Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij} \quad i = 1,...,t;\ j = 1,...,n_i$

p.16.a. Give the degrees of freedom for error for each model: Model 1 _____ Model 2 _____

p.16.b. The following table gives the means and standard deviations of chewiness scores for the different treatments, and most of the fitted values based on the regression model (model 2). The fitted equation is $\hat{Y}_{ij} = 7.764 - 0.031X_i$. Complete the table.

trt(i)	X_i	ybar_i	s_i	yhat_i
1	130	3.393	0.987	3.734
2	140	3.486	0.325	3.424
3	150	3.481	0.511	3.114
4	160	2.906	1.287	2.804
5	170	2.424	0.904	
6	180	1.967	0.634	

p.16.c. Conduct the F-test for lack of fit. Hint: SSPE is the same as SSE for model 1.

$H_0 : E\{Y_{ij}\} = \beta_0 + \beta_1 X_i$ vs $H_A : E\{Y_{ij}\} = \mu_i \neq \beta_0 + \beta_1 X_i$

SSLF = _____ SSPE = _____ F_{LF} = _____ Rejection Region: _____

Q.B.17. A linear regression model is fit, relating annual mean Temperature (Y, in °F) versus Year-1945 (X) for years 1945-2014 for Los Angeles (n=70). Thus, the intercept is the predicted value for year 1945. The regression coefficients are given below.

$Y = \beta_0 + \beta_1 X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$ independent

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	61.7544	0.2909	212.2543	0.0000
Year-1945	0.0304	0.0073	4.1765	0.0001

p.17.a. Test whether the mean is changing over time.

Test Statistic _____ P-value _____ Conclude: **Increasing** **Decreasing** **No Association**

p.17.b. Conduct the Durbin-Watson test. H_0: Errors are Not Autocorrelated H_A: Positive Autocorrelation

$\sum_{t=1}^{70} e_t^2 = 102.9067$ $\sum_{t=2}^{70} (e_t - e_{t-1})^2 = 106.2682$ $d_L(p=1, \alpha=0.05) = 1.58$ $d_U(p=1, \alpha=0.05) = 1.64$

DW = _____ Conclude: **Positive Autocorrelation** **No Autocorrelation** **Cannot Determine**

p.17.c. To test whether the errors have constant variance, a second regression was fit, relating squared residuals (e^2) to X.

For this regression, the regression sum of squares was $SSR_{e^2} = 19.0943$. Conduct the Breusch-Pagan Test.

Test Statistic _____ Rejection Region _____ Conclude: Variance **Constant / Non-Constant**

Q.B.18. An experiment was conducted relating energy consumption (Y, in MJ) to fiber space velocity (X, in m/h) in a carbon fiber production process. There were $c = 4$ distinct fiber space velocity "groups", with varying n_j runs per group. The lack-of-fit test for a linear relation is:

$$H_0 : E\{Y_{ij}\} = \mu_j = \beta_0 + \beta_1 X_j \quad i = 1,\ldots,n_j;\ j = 1,\ldots,4 \quad H_A : E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

ANOVA											
	df	*SS*	*MS*	*F*	*gnificance F*		fsv	n_grp	yhat_grp	ybar_grp	s_grp
Regressio	1	47.1060	47.1060	809.1265	0.0000		20	8	7.5625	7.7913	0.1283
Residual	28	1.6301	0.0582				25	9	6.4686	6.2143	0.0784
Total	29	48.7361					30	5	5.3747	5.1922	0.0802
							35	8		4.4523	0.0735
	Coefficients	*andard Err*	*t Stat*	*P-value*							
Intercept	11.9381	0.2135	55.9060	0.0000							
fsv	-0.2188	0.0077	-28.4451	0.0000							

p.18.a. Give the fitted value for the linear regression for the 4th group ($X_4 = 35$).

p.18.b. Compute the Pure Error Sum of Squares, degrees of freedom and Mean Square.

SS_{PE} = _____ df_{PE} = _____ MS_{PE} = _____

p.18.c. Compute the Lack-of-Fit Sum of Squares, degrees of freedom and Mean Square.

SS_{LF} = _____ df_{LF} = _____ MS_{LF} = _____

p.18.d. Give the Test Statistic, Rejection Region, and P-value relative to .05 for the Lack-of-Fit test.

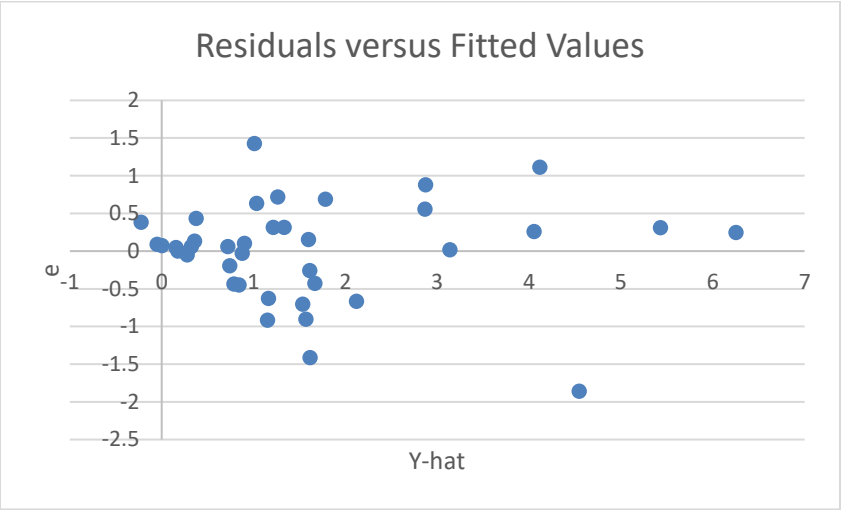
Test Statistic: _____ Rejection Region: _____ P-value > or < .05

Q.B.19. A regression model is fit, relating mobility (Y) to six predictor variables: GDP (X_1), vehicles/km of road (X_2), population density (X_3), percent urban population (X_4), land area (X_5), and population (X_6) for $n = 38$ island nations. The Analysis of Variance for the multiple linear regression model is given below.

ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regressio	6	87.57	14.60	28.83	0.0000
Residual	31	15.69	0.51		
Total	37	103.26			

p.19.a. A plot of the residuals versus predicted values is given below. It demonstrates which possible violations of assumptions (circle all that apply).

Non-normal Errors Unequal Variance Serial Correlation of Errors Non-linear Relation between Y and X



p.19.b. A second regression model is fit, relating the squared residuals (Y) to the 6 predictors (X_1,…,X_6). Conduct the Breusch-Pagan test to test whether the equal variance assumption is reasonable. The sums of squares are given below.

ANOVA		
	df	*SS*
Regressio	6	269.47
Residual	31	166.87
Total	37	436.34

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

Q.B.20. A simple linear regression model and Analysis of Variance (Completely Randomized Design) model were fit, relating stretch percentage of viscose rayon to specimen length (there were 4 lengths: 5, 10, 15, 20 inches). The results from each model are given below:

Regression Model

ANOVA				
	df	*SS*	*MS*	*F*
Regressio	1	84.53	84.53	129.63
Residual	183	119.33	0.65	
Total	184	203.87		
	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	24.77	0.14	172.63	0.0000
X Variable	-0.12	0.01	-11.39	0.0000

Completely Randomized Design (1-Way ANOVA)

ANOVA				
	df	SS	MS	F
Treatments (Lengths)	3	88.16	29.39	45.97
Error	181	115.71	0.64	
Total	184	203.87		

Below are the fitted values (regression) and sample means (1-way ANOVA), and sample sizes:

Length	Yhat(Reg)	Ybar(Grp)	n(Grp)
5	24.18	24.31	48
10	23.58	23.45	45
15	22.99	22.83	44
20	22.39	22.53	48

Conduct the Goodness-of-Fit F-test, where H_0 states that the relationship between stretch percentage and specimen length is linear. Hint: All numbers (sums of squares and degrees of freedom can be obtained (implicitly) from the ANOVA tables).

Test Statistic: _____ Rejection Region: _____

Part C: Multiple Linear Regression: t, F-tests, Extra Sums of Squares

Q.C.1. A multiple regression model is fit relating a response Y to 4 predictors: X_1, X_2, X_3, X_4. We fit 2 models (each based on a sample of n=20 cases):

$i)\quad E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad ii)\quad E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

For model i), the error sum of squares is 250. for model 2 it is 300.

Test H_0: $\beta_3=\beta_4=0$ at the α=0.05 significance level.

Q.C.2. A multiple regression model is fit relating a response Y to 6 predictors: X_1, X_2, X_3, X_4, X_5, X_6. We fit 2 models (each based on a sample of n=34 cases):

$i)\quad E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \qquad ii)\quad E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

For model i), the error sum of squares is 1500. For model 2 it is 1850.

Test H_0: $\beta_4=\beta_5=\beta_6=0$ at the α=0.05 significance level.

Q.C.3. A regression model is fit, relating breaking strength of a fiber (Y) to the amount of an additive applied to it (X_1), the amount of time it is heated (X_2), and the temperature (X_3) at which it is heated. The fitted equation and coefficient of multiple determination are given below (n=24):

$$\hat{Y} = 10.0 + 0.35 X_1 + 0.30 X_2 + 0.010 X_3 \qquad R^2 = 0.75$$

Give the predicted breaking strength of a fiber with X_1=10 units of additive, heated for X_2=15 minutes, at X_3=300 degrees

Test H_0: $\beta_1=\beta_2=\beta_3=0$ at the α=0.05 significance level.

Q.C.4. If we continue adding new predictors to a regression model, the error sum of squares will never increase, but the error mean square may increase. _____

Q.C.5. It is not possible to get a negative F-statistic when (properly) conducting a Complete versus Reduced F-test, when we are testing to determine whether one set of predictors is not associated with Y, after controlling for another set of predictors. _____

Q.C.6. A researcher states that her regression model explains 75% of the variation in her dependent variable. This means that SSE/TSS = 0.25, where SSE is the Error sum of squares and TSS is the Total sum of squares _____

Q.C.7. A regression model is fit with p=6 predictors (and an intercept), based on n=20 experimental units. How large does R^2 for the model need to be to reject H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ at $\alpha = 0.05$ significance level?

Q.C.8. Was moved to Q.D.29. Will not renumber subsequent problems in this section.

Q.C.9. A multiple regression model was fit, relating consumer's satisfaction of culinary (food) event (Y) to the following predictor variables/scales: food tasting (X_1), food/beverage prices (X_2), ease of coming/going to event (X_3), and convenient parking (X_4). The regression model was fit, based on n=277 respondents, with R^2 = 0.44. The regression coefficients are given below, for the model: $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Parameter	Estimate	StdError	t
Constant	1.805	0.156	11.571
X1	0.327	0.041	7.976
X2	0.263	0.052	
X3	0.082	0.035	2.343
X4	0.025	0.042	0.595

p.9.a. Test H_0: Customer satisfaction is not associated with any of the predictors ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$).

p.9.a.i. Test Statistic:

p.9.a.ii. Reject H_0 if the test statistic falls in the range _____

p.9.b. Test H_0: Controlling for all other predictors, customer satisfaction is not associated with food/bev prices ($\beta_2 = 0$).

p.9.b.i. Test Statistic:

p.9.b.ii. Reject H_0 if the test statistic falls in the range _____

Q.C.10. A multiple regression model is fit with 3 predictors and an intercept, based on a sample of n=25 observations. How large must $R^2/(1-R^2)$ be to reject H_0:$\beta_1 = \beta_2 = \beta_3 = 0$?

Q.C.11. It is possible to fail to reject H_{01}: $\beta_1 = 0$ and H_{02}: $\beta_2 = 0$ based on t-tests in multiple linear regression model with $p > 2$ predictors, but still reject H_{012}: $\beta_1 = \beta_2 = 0$, controlling for the remaining p-2 predictors.
True or **False**

Q.C.12. A multiple linear regression model is fit relating a dependent variable to a set of 3 numeric predictor variables, based on a sample on n=20 experimental units. How large does R^2 need to be so that the null hypothesis H_0:$\beta_1=\beta_2=\beta_3=0$ will be rejected?

Q.C.13. It is possible for a dataset to reject H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$ based on the F-test, but fail to reject H_0: $\beta_i = 0$ for i=1,…,p based on the individual t-tests. **True** or **False**

Q.C.14. A multiple regression model was fit based on a controlled experiment relating energy consumption in cars (Y, in MJ) to the following predictors: Temperature (X_1, in C), fiber space velocity (X_2, in mph), and stretch ratio (X_3, in %).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad \varepsilon \sim N\left(0, \sigma^2\right)$$

p.14.a. Complete the following ANOVA and regression coefficients tables.

ANOVA					
	df	*SS*	*MS*	*F*	*F(.05)*
Regression		47.30			
Residual				#N/A	#N/A
Total	29	48.74	#N/A	#N/A	#N/A
	Coefficients	*andard Err*	*t Stat*	*t(.025)*	
Intercept	6.0509	3.1373			
tempC	0.0253	0.0135			
fsv	-0.2186	0.0076			
stretch	0.0132	0.0378			

p.14.b. Give the point estimate of σ^2, the error variance.

p.14.c. Obtain a 95% Confidence Interval for β_3, controlling for Temperature and fiber space velocity, do you conclude that stretch ratio is associated with energy consumption?

95% Confidence Interval: _____ Conclude? **Yes** / **No**

p.14.d. For the first car in the experiment, X_1 = tempC = 227, X_2 = fsv = 20 , and X_3 = stretch = 1. Its observed energy consumption was $Y_1 = 7.637$. Compute its fitted value and residual.

$\hat{Y}_1$ = _____ e_1 = _____

Q.C.15. A multiple linear regression model is fit relating a dependent variable to a set of 2 numeric predictor variables, based on a sample on n=25 experimental units. How large does R^2 need to be so that the null hypothesis H_0:$\beta_1=\beta_2=0$ will be rejected?

Q.C.16. In a multiple regression model, with p = 3 predictors, it is possible to fail to reject the individual null hypotheses H_0:$\beta_2 = 0$ and H_0:$\beta_3 = 0$ based on t-tests and still reject H_0: $\beta_2 = \beta_3 = 0$ based on the Complete/Reduced F-test. **True** / **False**

Q.C.17. An experiment was conducted relating viscosity of flour used in baking ice cream cones (Y, in degrees MacMichael) to the conents of moisture (X_M, in %), protein (X_P, in %), and ash (X_A, in percent) for n = 39 flours obtained from different flour mills. The following models were fit, with the results for Model 3 given below. All models assume errors are independent and normally distributed.

Model 1: $Y = \beta_0 + \beta_A X_A + \varepsilon$ Model 2: $Y = \beta_0 + \beta_P X_P + \beta_A X_A + \varepsilon$

Model 3: $Y = \beta_0 + \beta_M X_M + \beta_P X_P + \beta_A X_A + \varepsilon$

ANOVA								Coefficients	Standard Err	t Stat	t(.025)
	df	SS	MS	F	F(.05)		Intercept	-115.36	63.29		
Regression		24094.91			0.0000		moisture	4.15	4.38		
Residual		10164.83					protein	19.99	2.76		
Total	38	34259.74					ash	-128.86	15.37		

p.17.a Compute the coefficient of determination, R^2 for the above model (Model 3).

p.17.b. Complete the ANOVA and regression coefficients tables and test 1) whether the Viscosity is related to any of the content variables H_0:$\beta_M = \beta_P = \beta_A = 0$ and 2) whether Viscosity is related to the individual content variables, controlling for the others H_0: $\beta_i = 0$.

p.17.c. The Regression Sums of Squares for Models 1 and 2 are $SSR_1 = 8869.33$ and $SSR_2 = 23834.15$, respectively. Give the following sequential sums of squares.

$SSR(X_A)$ = _____ $SSR(X_P \mid X_A)$ = _____ $SSR(X_M \mid X_A, X_P)$ = _____

p.17.d. Compute $R^2_{YX_P \bullet X_A}$ (the coefficient of partial determination between Y and X_P, given X_A).

Q.C.18. A model was fit, relating US annual energy consumption to the following set of predictors: X_1 = **GDP**, X_2 = price of electricity (**pElec**), X_3 = **pop**ulation, X_4 = price of natural gas (**pNatGas**), and X_5 = price of heating oil (**pHeatOil**). A second model is fit, with only **GDP** (X_1) and **pElec** (X_2). The models were fit for the years 1984-2010.

Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$ $SSE_1 = 2.860$ $SSR_1 = 100.752$ $\sum_{t=2}^{27} (e_t - e_{t-1})^2 = 5.608$

Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ $SSE_2 = 3.038$ $SSR_2 = 100.574$

p.18.a. The critical values for the Durbin-Watson statistic for $n = 27$ and $p = 5$ are $d_L = 1.01$ and $d_U = 1.86$. Compute the Durbin-Watson statistic for testing H_0: the errors are not autocorrelated and circle the best conclusion.

D-W Statistic: _____ Conclude: Reject H_0 Accept H_0 Inconclusive

p.18.b. Compute $SSR(X_3, X_4, X_5 \mid X_1, X_2)$

p.18.c. Test H_0: $\beta_3 = \beta_4 = \beta_5 = 0$

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

Q.C.19. In multiple regression it is possible to have a significant F-test for the full model, while none of the individual t-tests for the partial coefficients are significant. **True / False**

Q.C.20. A researcher states that her regression model explains 75% of the variation in her dependent variable. This means that SSE/TSS = 0.25, where SSE is the Error sum of squares and TSS is the Total sum of squares

True / False

Part D: Models with Categorical Predictors, Interactions, and/or Polynomial Terms

Q.D.1. A linear regression model is fit, relating breaking strength of steel bars to thickness, length, and material type with 3 nominal levels (no interaction terms or polynomial terms are included in the model). The model is fit based on 50 experimental units and includes an intercept term β_0.

DF(Regression) _____ DF(Error)_____ DF(Total) _____

Q.D.2. A study was conducted to determine whether company size (X_1=#Employees) and presence/absence of an active safety program (X_2=1 if yes, 0 if no) were related with the lost work hours by employees in a 1-year period (Y). The sample consisted of 40 firms, 20 used the safety program, 20 did not. The model fit is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

p.2.a. Complete the following tables (ANOVA and Coefficients) for the multiple regression:

ANOVA					
Source	df	SS	MS	F(obs)	F(0.05)
Regression		90058			
Error				#N/A	#N/A
Total		104811	#N/A	#N/A	#N/A

Estimates					
Predictor	oefficier	SE(Coef)	t(obs)	t(.025)	
Intercept	31.67	8.56	#N/A	#N/A	
X1	0.014	0.0012			
X2	-58.22	6.32			

p.2.b. Controlling for company size, firms with the safety program are estimated on average to have _____ **more / less** lost work hours in 1 year than firms without the safety program.

p.2.c. What proportion of variation in lost work hours is "explained" by the variables X_1 and X_2

p.2.d. Give the predicted number of lost work hours in 1 year for a firm with 10,000 workers and the safety program.

Q.D.3. A study considered failure times of tools (Y, in minutes) for n=24 tools. Variables used to predict Y were: cutting Speed (X_1 feet/minute), feed rate (X_2), and Depth (X_3). The following 2 models were fit (TSS=3618):

Model 1: $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$ $SSE_1 = 874$

Model 2: $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1*X_2 + \beta_5X_1*X_3 + \beta_6X_2*X_3 + \beta_7X_1*X_2*X_3$ $SSE_2 = 483$

p.3.a. Treating Model 2 as the Complete Model with all potential predictors, compute C_p for Model 1

p.3.b. Test whether all of the interaction terms can be removed from the model, controlling for all main effects. That is, test H_0: $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ at $\alpha = 0.05$ significance level:

p.3.b.i. Test Statistic:

p.3.b.ii. Reject H_0 if the test statistic falls in the range: _____

p.3.b.iii. Based on this test, are we justified in dropping all interaction terms from the model? **Yes / No**

Q.D.4. A regression model is fit relating Peak Power Load (Y, in megawatts) to daily high temperature (X, in Degrees F) for a sample of n=10 days. The analyst believes that Peak Power Load will increase with temperature, and that the RATE of change will increase as temperature increases. The model to be fit is (along with estimates and standard errors):

Model 1: $E(Y) = \beta_0 + \beta_1X + \beta_2X^2$

Model 2: $E(Y) = \gamma_0 + \gamma_1W + \gamma_2W^2$ (where W is "centered" value of X, that is: W = X-91.5)

	Coeff	StdErr	t Stat	P-value			Coeff	StdErr	t Stat	P-value
Intercept	1784.1883	944.1230	1.8898	0.1007		Intercept	184.4443	6.0318	30.5786	0.0000
X	-42.3862	21.0008	-2.0183	0.0833		W	7.4192	0.7288	10.1800	0.0000
X^2	0.2722	0.1163	2.3394	0.0519		W^2	0.2722	0.1163	2.3394	0.0519

p.4.a. Test H_0: $\beta_2 = 0$ vs H_A: $\beta_2 > 0$ based on Model 1 with $\alpha = 0.05$:

p.4.a.i. Test Statistic: _____ p.4.a.ii. Reject H_0 if test stat falls in range: _____

p.4.b. Obtain a 95% Confidence Interval for β_1 for Model 1. Does it contain 0?

p.4.c. Obtain a 95% Confidence Interval for γ_1 for Model 2. Does it contain 0?

p.4.d. The correlation between X and X^2 for Model 1 is 0.9995 and between W and W^2 for Model 2 is -0.4116. Obtain the Variance Inflation Factors for Models 1 and 2 (Note that since there are only 2 predictors for each model, there will only be one VIF for each model). **This makes use of collinearity at end of Chapter 2**.

Model 1: $VIF_X = VIF_{X^2}$ = Model 2: $VIF_W = VIF_{W^2}$ =

p.4.e. Obtain predicted Peak Power Load when X=90 for Model 1, and when W = 90-91.5 = -1.5 for Model 2:

Model 1:

Model 2:

Q.D.5. An experiment was conducted to measure air permeability of fabric (Y) as a function of the following factors: warp density (X_1), weft density (X_2), and Mass per unit area (X_3). There were n=30 observations, and 4 models are fit:

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 \quad SSE = 72.4$

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 \quad SSE = 86.5$

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad SSE = 813.6$

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{23} X_2 X_3 \quad SSE = 122.7$

p.5.a. Use the first two models to test H_0: $\beta_{11} = \beta_{22} = \beta_{33} = 0$.

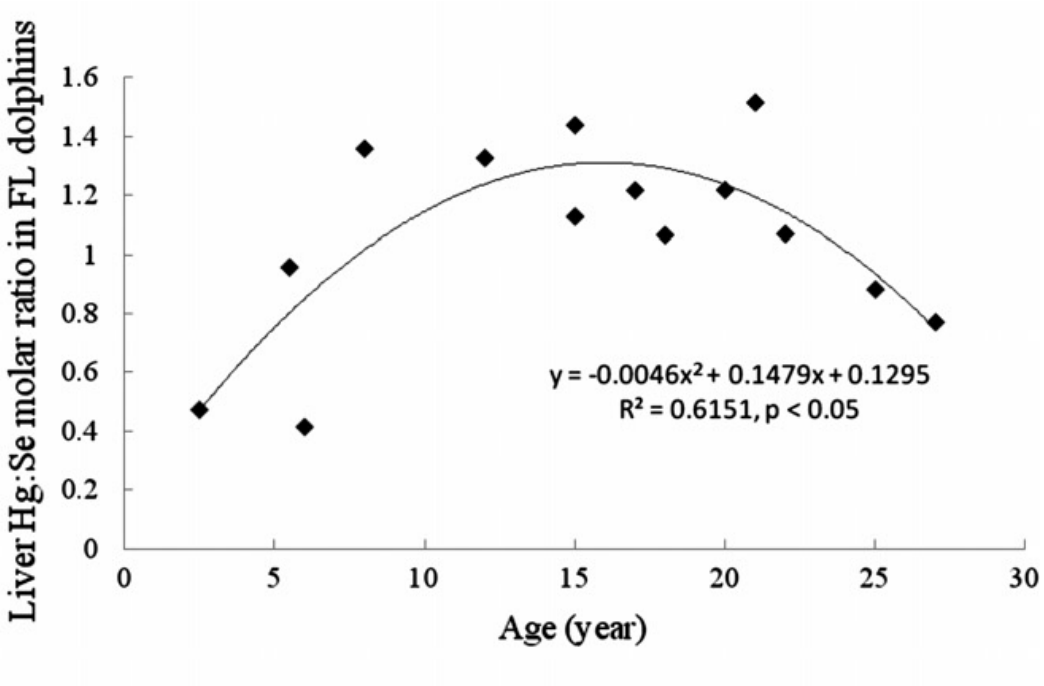
Test Statistic: _____ Rejection Region: _____

p.5.b. Use the 3rd and 4th models to test whether the weft-mass interaction is significant, controlling for all main effects.

Test Statistic: _____ Rejection Region: _____

Q.D.6. Regression analyses were fit, relating various chemical levels to age for stranded bottlenose dolphins in South Carolina and Florida.

This plot gives the quadratic fit, relating mercury/selenium molar ratio (Y) to age (X) for the Florida dolphins. Complete the following parts. Note: The data were NOT centered. The model fit was: $E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$



n = _____ Predicted value when age = 15 _____

Test H_0: $\beta_1 = \beta_2 = 0$

Test Statistic: _____ Rejection Region: _____

Q.D.7. A study was conducted to relate construction plant maintenance cost (pounds) to 3 categorical predictors (industry: coal/slate, machine type: front shovel/backacter, and attitude to used oil analysis: regular use/not regular use) and one numeric predictor (machine weight, in tons). Due to the distributions of costs and machine weight (skewed), both are modeled with (natural) logs. Thus, the variables are (based on a sample of n=33 construction plants):

- Y = ln(Costs)
- X_1 = 1 if coal industry, 0 if slate
- X_2 = 1 if machine type = front shovel, 0 if backacter
- X_3 = 1 if attitude to used oil analysis = regular use, 0 if not
- X_4 = ln(Machine Weight)

They fit 2 Models:

Model 1: $E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$

Model 2: $E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_1X_4 + \beta_6X_2X_4 + \beta_7X_3X_4$

ANOVA	Model1				
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	4	27.788	6.947	71.210	0.000
Residual	28	2.732	0.098		
Total	32	30.520			

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	-2.450	0.291	-8.426	0.000
X1	0.989	0.163	6.081	0.000
X2	-0.467	0.126	-3.708	0.001
X3	0.415	0.119	3.505	0.002
X4	1.027	0.068	15.138	0.000

ANOVA	Model2				
	df	*SS*	*MS*	*F*	*gnificance*
Regressio	7	28.425	4.061	48.465	0.000
Residual	25	2.095	0.084		
Total	32	30.520			

	Coefficients	*andard Err*	*t Stat*	*P-value*
Intercept	-2.294	2.392	-0.959	0.347
X1	0.935	2.269	0.412	0.684
X2	-1.314	0.552	-2.382	0.025
X3	1.303	0.452	2.881	0.008
X4	0.980	0.602	1.628	0.116
X1*X4	0.008	0.561	0.014	0.989
X2*X4	0.216	0.149	1.447	0.160
X3*X4	-0.229	0.120	-1.913	0.067

p.7.a. Based on Model 1, give the fitted value for a firm that is a coal industry, has machine type=backacter, is a regular user of used oil, and ln(Machine Weight) = 4.0. Note that the units of the fitted value is ln(Costs).

p.7.b. Test whether any of the categorical predictors interact with machine weight. H_0: $\beta_5 = \beta_6 = \beta_7 = 0$.

Test Statistic: _____ Rejection Region: _____

Q.D.8. A study related Personal Best Shot Put distance (Y, in meters) to best preseason power clean lift (X, in kilograms). The following models were fit, based on a sample of n = 24 male collegiate shot putters:

Model 1: $E\{Y\} = \beta_0 + \beta_1 X$ $SSE_1 = 43.41$ $R_1^2 = .686$ $\hat{Y}(X) = 4.4353 + 0.0898X$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$ $SSE_2 = 37.41$ $R_2^2 = .729$ $\hat{Y}(X, X^2) = 12.08 + 0.3285X - 0.00084X^2$

p.8.a. Use Model 2 to test H_0: $\beta_1 = \beta_2 = 0$ (Y is not related to X)

Test Statistic: _____ Rejection Region: _____ Reject H_0? **Yes** or **No**

p.8.b. Use Models 1 and 2 to test H_0: $\beta_2 = 0$ (Y is linearly related to X)

Test Statistic: _____ Rejection Region: _____ Reject H_0? **Yes** or **No**

p.8.c. Obtain the predicted value from each model for a man with best power clean of 175 kg.

Model 1: _____ Model 2: _____

Q.D.9. A regression model was fit, relating Weight (Y, in kg) to Height (X_1, in m) and Position for English Premier League Football Players. Position has 4 levels (Forward, Midfielder, Defender, and Goalkeeper). Thus, 3 dummy variables were generated: X_2 = 1 if Forward, 0 otherwise; X_3 = 1 if Midfielder, 0 otherwise; X_4 = 1 if Defender, 0 otherwise. Goalkeepers were the "reference position." The following regression models were fit, based on data for n = 441 league players.

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X_1$

$TSS = 24847 \quad SSR_1 = 12020 \quad SSR_2 = 11842 \quad SSR_3 = 11242$

p.9.a. We wish to test whether the slopes relating weight to height is the same among the positions, allowing the intercepts to differ among positions. Conduct this test for an interaction between height and position.

H_0: _____ H_A: _____

Test Statistic: _____ Rejection Region: _____

p.9.b. Assuming the interaction is not significant, test whether there is a position effect, after controlling for height.

H_0: _____ H_A: _____

Test Statistic: _____ Rejection Region: _____

Q.D.10. A model was fit, relating optimal solar panel tilt angle (*Y*, in degrees) to a city's latitude (*X*, in degrees) for a sample of $n = 35$ cities in the Northern Hemisphere. Consider the following 3 models (the *X* values have NOT been centered):

Model 1: $E\{Y\} = \beta_0 + \beta_1 X$ $SSE_1 = 54.826$ $\hat{Y} = 9.77 + 0.7103X$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$ $SSE_2 = 54.797$ $\hat{Y} = 10.92 + 0.6465X + 0.00087X^2$

Model 3: $E\{Y\} = \beta_1 X + \beta_2 X^2$ $SSE_3 = 57.400$ $\hat{Y} = 1.2445X - 0.00715X^2$

Note that Model 3 does not have an intercept (this is the model the authors fit).

p.10.a. Give the predicted optimal solar tilt for a location at a latitude of 30 degrees for each model.

Model 1: _____ Model 2: _____ Model 3: _____

P.10.b. Use Models 2 (Full Model) and Model 3 (Reduced Model) to test H_0: $\beta_0 = 0$ (Given Lat and Lat^2 are in model).

Test Statistic: _____ Rejection Region: _____

P.10.c. Use Models 2 (Full Model) and Model 1 (Reduced Model) to test H_0: $\beta_2 = 0$ (Given an intercept is in model).

Test Statistic: _____ Rejection Region: _____

Q.D.11. A linear regression model is fit, relating Y to 3 independent variables. The researcher is interested in determining whether any interactions are significant among the 3 independent variables. She fits the following 2 models (n = 30):

Model 1: $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad R_1^2 = 0.48$

Model 2: $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 \qquad R_2^2 = 0.54$

p.11.a. Compute the test statistic for testing H_0: No interactions among X_1, X_2, and X_3 ($\beta_4 = \beta_5 = \beta_6 = 0$)

p.11.b. Reject H_0 if the test statistic falls in the range _____

Q.D.12. A regression model was fit, relating estimated cost of de-commissioning oil platforms (Y, in millions of $) to 2 predictors: Total number of piles/legs (X_1) and water depth (X_2, in 100s of feet). The model was fit, based on n = 17 oil platforms. Consider the models:

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2$ Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

ANOVA	Model 1			ANOVA	Model2	
	df	*SS*			*df*	*SS*
Regression		7397		Regression		7163
Residual		551		Residual		785
Total	16	7948		Total	16	7948
Coefficient	*tandard Error*			*Coefficient*	*tandard Err*	
Intercept	-20.399	21.040		Intercept	-26.064	5.335
totpiles	-0.357	0.821		totpiles	0.957	0.274
wtrdph	6.155	7.048		wtrdph	4.401	1.095
tp2	0.047	0.046				
wtrdph2	-0.077	0.598				
tp*wd	-0.070	0.233				

p.12.a. Test whether the linear (main effects) model is appropriate. H_0: $\beta_{11} = \beta_{22} = \beta_{12} = 0$

p.12.b. For each model, obtain the predicted value and residual for rig 17 (Y = 78.5, X_1 = 44, X_2 = 13)

Model 1: Predicted = _____ Residual = _____

Model 2: Predicted = _____ Residual = _____

Accidentally Had two Q.D.12. problems, will not re-number

Q.D.12. A study related subsidence rate (Y) to water table depth (X_1) for 3 crops: pasture ($X_2 = 0$, $X_3 = 0$), truck crop ($X_2 = 1$, $X_3 = 0$), and sugarcane ($X_2 = 0$, $X_3 = 1$). Note the total sum of squares is TSS = 35.686, and n = 24.

p.12.a. Give a model that allows **separate intercepts** for each crop type, with a **common slope for water table depth** among crop types. Sketch the graph for this model. For this model, SSE = 1.853. Give the error degrees of freedom.

p.12.b. Give a model that allows **separate intercepts** for among crop types, with **separate slopes for water table depth** among crop types. Sketch the graph for this model. For this model, SSE = 1.261. Give the error degrees of freedom.

p.12.c. Test the null hypothesis that the (simpler) model in p.2.a. is appropriate. That is, the extra parameters in the second model are not significantly different from 0.

p.12.d. For the model in p.2.a., compute R^2

Q.D.13. An experiment is conducted with 3 numeric predictors and 2 categorical predictors, one with 3 levels, the other with 2 levels. There are no interaction or polynomial terms in the model, and the sample size is n = 30. Give the degrees of freedom for Regression and Error.

Df_{Reg} = _____ Df_{Error} = _____

Q.D.14. A linear regression model is fit, relating salary (Y) to experience (X_1), gender (X_2=1 if female, 0 if male) and an experience/gender interaction term to employees in a large law firm. The fitted equation is $\hat{Y} = 50000 + 2000X_1 + 1000X_2 - 100X_1X_2$. Give the predicted salaries for the following groups of individuals.

Males with 0 experience _____ Females with 0 experience _____

Males with 10 experience _____ Females with 10 experience _____

Q.D.15. A regression model was fit, relating blood alcohol elimination rate measurements (Y, in grams/litre/hour) to Gender (X_1=1 if female, 0 if male), breath alcohol elimination measurements (X_2 in mg/l/h) and a gender/breath interaction term. The sample was 59 adult Austrians.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

p.15.a. Complete the following Analysis of Variance Table and test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

ANOVA					
	df	*SS*	*MS*	*F*	*F(.05)*
Regression		0.0478			
Residual				#N/A	#N/A
Total		0.0624	#N/A	#N/A	#N/A

p.15.b. Is the P-value for the test **Larger** or **Smaller** than 0.05?

p.15.c. What proportion of the variation in Blood alcohol elimination rate measurements is "explained" by the model?

p.15.d. The regression coefficient estimates are given below. Test $H_0 : \beta_i = 0 \quad H_A : \beta_i \neq 0$ for each coefficient.

	Coefficients	*Standard Error*	t_obs	*t(.025)*	*Reject H0?*
Intercept	0.0427	0.0154	#N/A	#N/A	#N/A
female	-0.0335	0.0229			
breath	1.5349	0.1951			
f*breath	0.4213	0.2744			

Q.D.16. A response surface model is fit, relating potato chip moistness (Y) to 3 factors: drying time (X_1), frying temperature (X_2), and frying time (X_3). There were n = 20 experimental runs (observations). The following 3 models were fit:

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad SSR_1 = 475.2 \quad SSE_1 = 145.2$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 \quad SSR_2 = 558.3 \quad SSE_2 = 62.1$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 \quad SSR_3 = 599.0 \quad SSE_3 = 21.4$

p.16.a. Test whether any of the 2-way interaction effects are significantly different from 0 , controlling for main effects.

H_0:

Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

p.16.b. Test whether any of the quadratic effects are significantly different from 0, controlling for main effects and 2-factor interactions.

H_0:

Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

Q.D.17. A study of "legitimacy theory" was conducted after the Exxon Valdez oil spill in the late 1980s among a sample of n = 21 firms in Alaska. The response variable was the change in the number of environmental disclosures (Y, 1989-1988) and the predictor variables were the firms' log revenues for 1989 (X_1) and an indicator variable of whether the firm was part owner of the Aleyska pipeline (X_2 = 1 if Yes, 0 if No).

$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

p.17.a. Complete the following Analysis of Variance Table and test H_0: $\beta_1 = \beta_2 = 0$.

ANOVA					
Source	df	SS	MS	F	F(.05)
Regression		13.72			
Error				#N/A	#N/A
Total		27.95	#N/A	#N/A	#N/A

p.17.b. Is the P-value for the test **Larger** or **Smaller** than 0.05?

p.17.c. Give the residual standard error for the model (estimate of σ).

p.17.d. The theory states that the regression coefficients for both firm size and part ownership of the pipeline should be positive. The regression coefficient estimates are given below. Test H_0: $\beta_i = 0$ vs H_A: $\beta_i > 0$ for each coefficient.

	Coeffs	Std. Err.	t_obs	t(.05)	Reject H0?
Intercept	-8.142	3.642	#N/A	#N/A	#N/A
X1	0.932	0.373			
X2	1.107	0.485			

Q.D.18. A response surface model is fit, relating lipstick color characteristic b* (yellow/blue) (Y) to 3 factors: liposoluble dye amount (X_1), titanium dioxide (X_2), and pigment (X_3). There were n = 17 experimental runs (observations). The following 3 models were fit:

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad SSR_1 = 187.71 \quad SSE_1 = 21.15$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 \quad SSR_2 = 197.02 \quad SSE_2 = 11.84$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 \quad SSR_3 = 199.77 \quad SSE_3 = 9.09$

p.18.a. Test whether any of the 2-way interaction effects or quadratic terms are significantly different from 0 , controlling for main effects.

H_0:

Test Statistic: _____ Rejection Region: _____ P-value: > or < 0.05

p.18.b. What proportion of variation in Y is "explained" by each model?

Model 1 _____ Model 2 _____ Model 3 _____

Q.D.19. A multiple regression model is fit, relating antioxidant activity (Y, DPPH scavenging activity, %) to total phenolic content (X_1, mg GAE/g dw) and Flavonoids content (X_2, mg RE/g dw) for a sample of $n = 68$ Chinese Herbs. The following 4 models are fit, with the following resulting sums of squares.

$TSS = \sum\left(Y_i - \overline{Y}\right)^2 = 52011.4$

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 \quad \hat{Y}^1 = 19.295 + 0.437 X_1 \quad SSE_1 = 29483.4 \quad SSR_1 = 22528.0$

Model 2: $E\{Y\} = \beta_0 + \beta_2 X_2 \quad \hat{Y}^2 = 12.617 + 0.806 X_2 \quad SSE_2 = 12091.8 \quad SSR_2 = 39919.6$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \hat{Y}^3 = 12.468 + 0.054 X_1 + 0.754 X_2 \quad SSE_3 = 11920.9 \quad SSR_3 = 40090.5$

Model 4: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad \hat{Y}^4 = 5.569 + 0.502 X_1 + 0.911 X_2 - 0.005 X_1 X_2 \quad SSE_4 = 6738.3 \quad SSR_4 = 45273.1$

p.19.a. Test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Give the Test Statistic, Rejection Region, and give the P-value relative to .05.

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

p.19.b. What proportion of the total variation in Y is "explained" by the predictors in Model 3.

p.19.c. Give the following sequential sums of squares.

$SSR(X_1)$ = _____ $SSR(X_2 \mid X_1)$ = _____ $SSR(X_1X_2 \mid X_1,X_2)$ = _____

$SSR(X_2)$ = _____ $SSR(X_1 \mid X_2)$ = _____ $SSR(X_1X_2 \mid X_1,X_2)$ = _____

p.19.d. Compute $R^2_{YX_2 \bullet X_1}$ (the coefficient of partial determination between Y and X_2, given X_1).

Q.D.20. A regression model was fit, relating adjusted costs of $n = 37$ Hong Kong office towers (Y) to the following predictors: Average Floor Area (X_1), Total Floor Area (X_2), Average Storey Height (X_3), and a Steel Indicator (X_4=1 if Steel, 0 if Reinforced Concrete). Consider the following two models (the first includes whether the building is Steel, as well as interactions with Steel and the three size variables, the second ignores steel completely).

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_4 + \beta_6 X_2 X_4 + \beta_7 X_3 X_4 \quad SSE_1 = 30808$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad SSE_2 = 35714$

Test whether the steel term, and all interaction terms involving steel are all 0. That is, above and beyond the 3 size variables (X_1, X_2, and X_3), none of the remaining predictors are associated with costs.

p.20.a Give the null hypothesis. _____

p.20.b. Give the Test Statistic, Rejection Region, and P-value relative to .05

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

p.20.c. R^2 for Model 1 is .9965. Give the Total Sum of Squares.

Q.D.21. An antioxidant study of n = 24 varieties of Chinese Herbs related FRAP Value (Y) to the following predictors: Total Phenolic Content (X_1) and Flavonoids Content (X_2). Consider the following 3 models. $TSS = 238.058$

Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$ independent $SSE_1 = 2.788$

Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$ independent $SSE_2 = 2.497$

Model 3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_{11}^2 + \beta_{22} X_{22}^2 + \beta_{12} X_1 X_2 + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$ independent $SSE_3 = 0.641$

p.21.a. What proportion of the variation in FRAP is "explained" by Model 1?

p.21.b. Give the coefficient of partial determination between FRAP and Flavonoids given Total Phenolic Content: $R^2_{Y2\bullet 1}$

p.21.c. Test $H_0 : \beta_2 = \beta_{11} = \beta_{22} = \beta_{12} = 0$

Test Statistic _____ Rejection Region _____ P-value < or > 0.05

Q.D.22. A regression model was fit relating time for nuclear power plant workers to complete emergency tasks (Y, in minutes) with predictors: TACOM complexity (X_1) and an indicator of whether the employee was from the US (X_2 = 1 if Yes, 0 if No). The sample size was n = 70, with 35 US operator and 35 non-US operator measurements. The models fit are given here and results for Model 3 are given below.

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 \quad SSR_1 = 631.9 \quad SSE_1 = 517.5$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad SSR_2 = 835.6 \quad SSE_2 = 313.9$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad SSR_3 = 898.4 \quad SSE_3 = 251.0$

	Coefficients	*Standard Err*	*t Stat*	*P-value*
Intercept	-6.77	1.74	-3.88	0.0002
complex	1.95	0.31	6.24	0.0000
US	-6.43	2.47	-2.61	0.0112
US*comp	1.80	0.44	4.06	0.0001

p.22.a. Compute R^2 for each model.

R_1^2 = _____ R_2^2 = _____ R_3^2 = _____

p.22.b. Fill in the following table of predicted values and their difference by group (non-US, US)

TACOM\US	non-US	US
5		
6		
6-5		

p.22.c. Test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Test Statistic _____ Rejection Region _____ P-value < or > 0.05

Q.D.23. A researcher wants to fit a response surface with 3 factors and wishes to include: intercept, all linear effects, 2-factor interactions, and quadratic terms. She wants to have 25 degrees of freedom for Error. How many experimental runs will she need?

Q.D.24. A regression model was fit, relating total evaporative resistance (Y) to air gap size (X) for n = 39 Chinese male clothing ensembles. The data were centered to reduce multicollinearity, as linear, quadratic, and cubic models were fit. The following summary results were obtained from the three regression models. Note that *TSS* = 340.09

Model 1: $\hat{Y}_i = 21.836 + 2.523\left(X_i - \overline{X}\right)$ $SSE_1 = 225.31$

Model 2: $\hat{Y}_i = 21.277 + 2.523\left(X_i - \overline{X}\right) - 0.476\left(X_i - \overline{X}\right)^2$ $SSE_2 = 216.17$

Model 3: $\hat{Y}_i = 21.289 + 2.513\left(X_i - \overline{X}\right) - 0.511\left(X_i - \overline{X}\right)^2 + 0.014\left(X_i - \overline{X}\right)^3$ $SSE_3 = 216.16$

p.24.a. Test whether evaporative resistance is related to air gap size in any of the polynomials.
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Test Statistic _____ Rejection Region _____ P > or < .05

p.24.b. Test whether the quadratic and/or cubic coefficients are significant, controlling for the linear term.

H_0:

Test Statistic _____ Rejection Region _____ P > or < .05

Q.D.25. A response surface was fit, relating Acidity of mango wine (Y, in g/L) to three predictors: Temperature (X_1, in C),

pH (X_2), and Inoculum Size (X_3, in %). The model fit is given below.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2$$

The output from the **rsm** package in R is given below.

```
> mango1 <- rsm(acidity ~ SO(tempC,pH,inoc))
               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 14.00194497  2.72249268  5.1431 0.0004358 ***
tempC       -0.23492437  0.06323471 -3.7151 0.0040074 **
pH          -6.19083239  1.23480611 -5.0136 0.0005267 ***
inoc         0.12499074  0.07149478  1.7482 0.1109947
tempC:pH     0.04166667  0.01384648  3.0092 0.0131358 *
tempC:inoc  -0.00300000  0.00110772 -2.7083 0.0219989 *
pH:inoc     -0.04375000  0.01661578 -2.6330 0.0250312 *
tempC^2      0.00248887  0.00068764  3.6194 0.0046941 **
pH^2         0.75881958  0.15471924  4.9045 0.0006190 ***
inoc^2       0.00457367  0.00099020  4.6189 0.0009522 ***

Multiple R-squared:  0.9128,    Adjusted R-squared:  0.8343
F-statistic: 11.63 on 9 and 10 DF,  p-value: 0.000331

Analysis of Variance Table
                     Df  Sum Sq  Mean Sq F value    Pr(>F)
FO(tempC, pH, inoc)   3 0.28560 0.095200 10.7757 0.0017852
TWI(tempC, pH, inoc)  3 0.20605 0.068683  7.7743 0.0057033
PQ(tempC, pH, inoc)   3 0.43326 0.144419 16.3468 0.0003497
Residuals            10 0.08835 0.008835
Lack of fit           5 0.02981 0.005963  0.5093 0.7615901
Pure error            5 0.05853 0.011707

Stationary point of response surface:
    tempC        pH      inoc
22.536788  3.804183 11.921781
```

p.25.a. Give the number of experimental runs: _____

p.25.b. Give the test statistic and P-value for H_0: Second-order model is appropriate:
TS: _____ P-Value _____

p.25.c. The sequential sums of squares for the "blocks" of terms are given in the ANOVA table. That is:

$$SSR(X_1, X_2, X_3) = 0.28560 \quad SSR(X_1X_2, X_1X_3, X_2X_3 \mid X_1, X_2, X_3) = 0.20605$$

$$SSR(X_1^2, X_2^2, X_3^2 \mid X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3) = 0.43326$$

Give R^2 for models with: i) Linear Terms only, ii) Linear and 2-Way Interactions, iii) Linear, 2-Way, and Quadratics.

Q.D.26. Two multiple linear regression models were fit relating price of art works (Y = log(sale price)) to the following predictors: surface area (SA) of the object, the medium of the object (collage, drawing, painting*, photograph, print, sculptures). There were 5 dummy variables for medium ($M_1,\ldots,M_5$), with painting being the reference category. The first model had a linear trend for year (t), while the second model had 12 dummy variables ($Yr_1,\ldots,Yr_{12}$) for the 13 individual years (thus not forcing the trend to be linear). The models and results are given below, based on a sample of n = 518 artworks sold during the 13 year period 1997-2009.

Model 1: $E\{Y\} = \beta_0 + \beta_{SA}SA + \sum_{i=1}^{5}\beta_{M_i}M_i + \beta_t t \quad R_1^2 = .502$

Model 2: $E\{Y\} = \beta_0 + \beta_{SA}SA + \sum_{i=1}^{5}\beta_{M_i}M_i + \sum_{i=1}^{12}\beta_{Yr_i}Yr_i \quad R_2^2 = .555$

p.26.a. Give the number of parameters for the models. Model 1: _____ Model 2: _____

p.26.b. For Model 1, test H_0: $\beta_{SA} = \beta_{M1} = \beta_{M2} = \beta_{M3} = \beta_{M4} = \beta_{M5} = \beta_t = 0$

Test Statistic _____ Rejection Region _____ P < or > 0.05

p.26.c. Model 1 is a special case of Model 2, with the yearly trend being a straight line, while Model 2 allows any structure for the year effects. Based on comparing Complete and Reduced models, test between the following hypotheses.

H_0: Model 1 is appropriate (linear trend) versus H_A: Model 2 is appropriate (trend is not linear)

Test Statistic _____ Rejection Region _____ P < or > 0.05

Q.D.27. A regression model was fit, relating the heat capacity of solid hydrogen bromide (Y, in cal/(mol*K)) to Temperature (X, in degrees Kelvin) based on n=18 experimental runs. The temperatures were centered (for computational reasons), but this has no effect on predicted values or Sums of Squares. The following 3 models are fit where the mean temperature was 145.16.

Model 1: $E\{Y\} = \beta_0 + \beta_1(X - \overline{X})$ $\hat{Y}^1 = 11.2756 + 0.0216(X - 145.16)$ $SSE_1 = 0.1945$ $SSR_1 = 3.3889$

Model 2: $E\{Y\} = \beta_0 + \beta_1(X - \overline{X}) + \beta_2(X - \overline{X})^2$ $\hat{Y}^2 = 11.1596 + 0.0192(X - 145.16) + 0.00029(X - 145.16)^2$ $SSE_2 = 0.0370$ $SSR_2 = 3.5465$

Model 3: $E\{Y\} = \beta_0 + \beta_1(X - \overline{X}) + \beta_2(X - \overline{X})^2 + \beta_3(X - \overline{X})^3$

$\hat{Y}^3 = 11.1718 + 0.0155(X - 145.16) + 0.00021(X - 145.16)^2 + 0.0000059(X - 145.16)^3$ $SSE_3 = 0.0172$ $SSR_3 = 3.5662$

p.27.a. For Model 3, Test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

p.27.b. What proportion of the total variation in Y is "explained" by the predictors in Model 2.

p.27.c. Give the predicted heat capacities for temperatures X=125.16, 145.16, and 165.16 for each Model.

M1: 125.16: 145.16: 165.16:

M2: 125.16: 145.16: 165.16:

M3: 125.16: 145.16: 165.16:

Q.D.28. A study related height (Y, in cm) to foot length (X, in cm) among n = 5195 adult South Koreans of ages 20 to 59. A dummy variable (M = 1 if male, 0 if female) is created to reflect subject's gender. Three models are fit (each assuming independent, normally distributed errors with constant variance).

Model 1: $E\{Y\} = \beta_0 + \beta_1 X$ $\hat{Y}^1 = 45.609 + 4.947X$ $SSE_1 = 116921$ $SSR_1 = 313347$

Model 2: $E\{Y\} = \beta_0 + \beta_1 X + \gamma_1 M$ $\hat{Y}^2 = 65.574 + 4.031X + 3.857M$ $SSE_2 = 108416.5$ $SSR_2 = 321851.5$

Model 3: $E\{Y\} = \beta_0 + \beta_1 X + \gamma_1 M + \delta_1 XM$ $\hat{Y}^3 = 66.910 + 3.972X + 1.577M + 0.096XM$ $SSE_3 = 108404$ $SSR_3 = 321864$

p.28.a. Give the predicted heights for females and males with foot lengths of 23 and 25 cm based on model 3.

F/23: _____ F/25: _____ M/23: _____ M/25: _____

p.28.b. Based on models 1 and 2 test whether males and females differ in mean height, controlling for foot length.

H_0: $\gamma_1 = 0$ H_A: $\gamma_1 \neq 0$

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

p.28.c. Based on models 2 and 3 test whether the slopes with respect to foot length differ for males and females.

H_0: $\delta_1 = 0$ H_A:$\delta_1 \neq 0$

Test Statistic: _____ Rejection Region: _____ P-value > or < .05

Q.D.29. A study was conducted, relating Total Medical Waste (Y, in kg/day) to hospital type (Government: X_1=1, Education and Non-Education: X_2=1, University: X_3=1, Private is the reference type), Hospital Capacity (X_4 = # of beds), and Occupancy Rate (X_5 = % of Beds in Use). Consider the following two models:

Model 1: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

Model 2: $E\{Y\} = \beta_0 + \beta_4 X_4 + \beta_5 X_5$

SUMMARY OUTPUT		Model 1			
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regression	5	3723252	744650	19.78	0.0000
Residual	44	1656130	37639		
Total	49	5379382			
	Coefficients	*andard Err*	*t Stat*	*P-value*	
Intercept	17.07	88.20	0.19	0.8475	
govtype	-94.75	128.92	-0.74	0.4662	
eduntype	-16.63	83.76	-0.20	0.8435	
univtype	-90.48	70.78	-1.28	0.2078	
beds	2.42	0.29	8.49	0.0000	
occ_rate	-0.85	1.57	-0.54	0.5927	

SUMMARY OUTPUT		Model 2			
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regression	2	3637224	1818612	49.06	0.0000
Residual	47	1742158	37067		
Total	49	5379382			
	Coefficients	*andard Err*	*t Stat*	*P-value*	
Intercept	-42.86	67.22	-0.64	0.5268	
beds	2.38	0.26	9.09	0.0000	
occ_rate	-0.53	1.48	-0.36	0.7240	

p.29.a. Test whether there is a Hospital type effect (controlling for beds and occupancy rate):
H_0:$\beta_1 = \beta_2 = \beta_3 = 0$

Test Statistic: _____ Rejection Region: _____ Reject H_0? **Yes** or **No**

p.29.b. The first hospital in the sample is University type, has 215 beds, and an occupancy rate of 47%. Based on Model 1, compute its fitted (predicted) value. Its observed value was 302. Compute its residual.

Fitted Value: _____ Residual: _____

p.29.c. What proportion (or percentage) of the Variation in Total Medical waste is explained by number of beds and occupancy rate?

Part E: Model Building

Q.E.1. A study was conducted to determine which factors were associated with percent release (Y) of hydroxypropyl methylcellulose (HPMC) tablets. The factors were:

X1 = Carr's compressibility index,
X2 = angle of repose,
X3 = solubility,
X4 =molecular weight,
X5 = compression force
X6 = apparent viscosity of 4% (w/v) HPMC.

The sample size was n=18, and the authors reported the fit of the following models.

p.1.a. Complete the table in terms of AIC and SBC (BIC).

Predictors	p'	SSE	AIC	SBC
X1,X2,X3,X4,X5,X6	7	42.62	29.5151	35.7477
X1,X2,X3,X4,X5	6	42.62	27.5151	32.8574
X1,X2,X3,X4	5	48.58		
X1,X3,X4,X6	5	48.58		
X1,X3,X4	4	52.86	27.3910	30.9524
X2,X3,X4	4	75.31	33.7623	37.3238
X3,X4,X6	4	48.85		

p.1.b. Which model is "best" based on AIC: _____ BIC: _____

p.1.c. R^2 for the complete model was 0.9278. Compute the total (corrected) sum of squares (TSS):

Q.E.2. Consider the following models, relating Stature (Y) to foot dimensions (RFL = Right Foot Length, RFB = Right Foot Breadth) and Age among the Rajbanshi of North Bengal. We observe the following statistics, based on several regressions among n = 175 adult males. Complete the following table. Note: The total sum of squares is TSS = 5633.4, and for C_p, use s^2 = MSE(RFL,RFB,Age) = 19.1. All models contain an intercept (β_0).

Predictors	p'	SSE	R-square	Cp	AIC	BIC(SBC)
RFL		3439.9		9.00	525.22	
RFB		4191.1	0.26	48.31	559.79	566.12
Age		5560.2	0.01	119.95		615.59
RFL,RFB		3282.5	0.42		519.03	528.52
RFL,RFB,Age		3267.9	0.42	4.00	520.25	532.90

p.2.a. What is the best model based on C_p?

p.2.b. What is the best model based on AIC?

p.2.c. What is the best model based on BIC (SBC)?

Q.E.3. A data set consisted of n = 32 observations on the variables Y, X1, X2, X3, and X4. Error Sum of Squares = SSE for each of all possible models. For each model, the variables that are in the model are also shown. Use this information to answer the questions following the table. The Total Sum of Squares = SSTO = 1150.

#Variables	SSE	Cp	AIC	SBC=BIC	Vars in Model
1	330				X2
1	448	#N/A	#N/A	#N/A	X3
1	505	#N/A	#N/A	#N/A	X1
1	785	#N/A	#N/A	#N/A	X4
2	255				X2,X4
2	284	#N/A	#N/A	#N/A	X2,X3
2	290	#N/A	#N/A	#N/A	X1,X3
2	295	#N/A	#N/A	#N/A	X1,X2
2	402	#N/A	#N/A	#N/A	X1,X4
2	445	#N/A	#N/A	#N/A	X3,X4
3	245				X1,X2,X4
3	253	#N/A	#N/A	#N/A	X1,X2,X3
3	255	#N/A	#N/A	#N/A	X2,X3,X4
3	290	#N/A	#N/A	#N/A	X1,X3,X4
4	243				X1,X2,X3,X4

p.3.a. Complete the table by computing C_p, AIC, and SBC=BIC for the best models with 1,2,3, and 4 independent variables.

p.3.b. Give the best model (in terms of which independent variables to be included), based on each criteria.

C_p: _____ AIC: _____ SBC=BIC: _____

Q.E.4. A study considered failure times of tools (Y, in minutes) for n=24 tools. Variables used to predict Y were: cutting Speed (X_1 feet/minute), feed rate (X_2), and Depth (X_3). The following 2 models were fit (SSTO=3618):

Model 1: $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$ $SSE_1 = 874$

Model 2: $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1*X_2 + \beta_5X_1*X_3 + \beta_6X_2*X_3 + \beta_7X_1*X_2*X_3$ $SSE_2 = 483$

Compute C_p, AIC, and SBC=BIC for each model. Based on each criteria, which model is selected?

Criteria	Model1	Model2	Best Model
Cp			
AIC			
SBC=BIC			

Q.E.5. A study investigated meteorological effects on condition of wheat yield in Ohio (Y), based on a series of n = 24 years. The predictors were: Average October/November temperature (X_1), September precipitation (X_2), October/November precipitation (X_3), and percent September sunshine (X_4). The best 1-,2-,3-, and 4-variable models (minimum SSE) are given below.

Model	p'	SSE	Cp	AIC	BIC
X3	2	1738	4.27	106.78	109.13
X2,X3	3	1531		105.73	109.27
X1,X2,X3	4	1395	3.48	105.50	
X1,X2,X3,X4	5	1361	5.00		112.80

p.5.a. Complete the table. Use MSE of the full (X_1,X_2,X_3,X_4) models as the estimate of σ^2 when computing C_p

p.5.b. Give the best model based on each criteria. C_p _____ AIC _____ BIC _____

p.5.c. The following output gives the regression coefficients for the X_1,X_2,X_3 model. Give the fitted value and residual for the first year (Y = 92, X_1 = 46, X_2 = 1.6, X_3 = 6.3).

	Coefficients
Intercept	16.59
tempon.x1	1.06
rains.x2	2.38
rainon.x3	2.96

Fitted Value _____ Residual _____

p.5.d. For the model in p.5.c., we obtain:

$$\sum_{t=2}^{24}\left(e_t - e_{t-1}\right)^2 = 3206 \qquad d_L\left(n=24, p=3\right)=1.10 \quad d_U\left(n=24, p=3\right)=1.66$$

Test H_0: Errors are not autocorrelated versus H_A: Errors are autocorrelated

Circle the Best Answer Reject H_0 Accept H_0 Test is inconclusive

Q.E.6. A regression model is fit, relating energy consumption (Y) to 3 predictors: area (X_1), age (X_2), and effective number of guest rooms (X_3 = rooms*occupancy rate) for a sample of n = 19 hotels.

p.6.a. Complete the following table of C_p, AIC, and BIC for all possible regressions involving X_1, X_2, and X_3.

Model	p*	SSE	C_p	AIC	BIC
X1		75.13		30.12	32.01
X2		327.08	57.31	58.07	59.96
X3		187.85	26.53	47.53	49.42
X1,X2		70.84	2.66		33.84
X1,X3		71.04	2.71	31.06	33.89
X2,X3		186.24	28.18	49.37	52.20
X1,X2,X3		67.85	4.00	32.18	

p.6.b. Which model is selected based on each criteria?

C_p: _____ AIC: _____ BIC: _____

p.6.c. To check for issues of multicollinearity, a regression relating each predictor on the other 2 predictors is fit. The largest R^2 of the 3 regressions is when X_1 is regressed on X_2 and X_3. That R_1^2 value is 0.468. Compute the Variance Inflation Factor VIF for X_1, where $VIF_1 = 1/\left(1-R_i^2\right)$. Does it exceed 10?

Q.E.7. A regression model for language diversity was fit, relating log of the number of languages spoken (Y) to 3 predictors: log of area (X_1), log of population (X_2), and mean growing season (X_3 , in months) for n = 54 countries.

p.7.a. Complete the following table of C_p, AIC, and BIC for all possible regressions involving X_1, X_2, and X_3.

Model	p*	SSE	C_p	AIC	BIC
X1		88.57		30.72	34.70
X2		84.27	32.40	28.03	32.01
X3		77.21	25.49	23.31	27.29
X1,X2		82.72	32.87		34.99
X1,X3		53.14	3.96	5.13	11.10
X2,X3		62.51	13.12	13.90	19.87
X1,X2,X3		51.14	4.00	5.06	

p.7.b. Which model is selected based on each criteria?

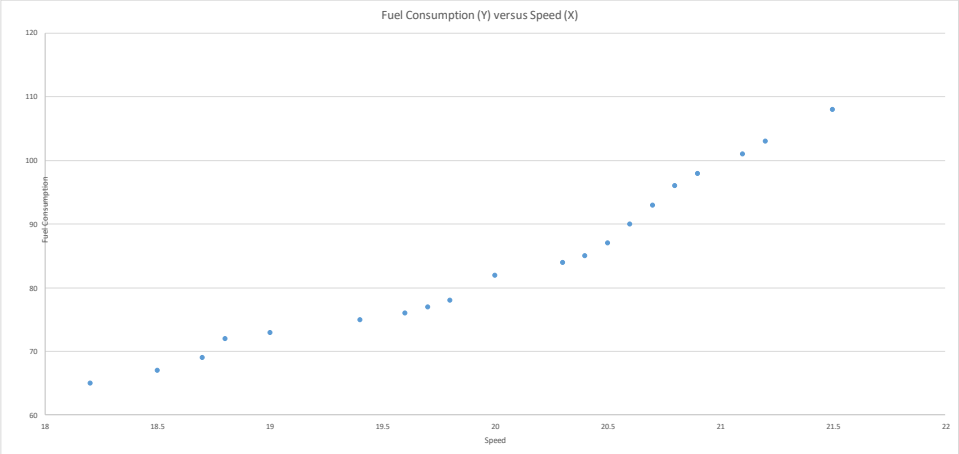
C_p: _____ AIC: _____ BIC: _____

p.7.c. To check for issues of multicollinearity, a regression relating each predictor on the other 2 predictors is fit. The largest R^2 of the 3 regressions is when X_1 is regressed on X_2 and X_3. That R_1^2 value is 0.385. Compute the Variance Inflation Factor VIF for X_1, where $VIF_1 = 1/\left(1-R_i^2\right)$. Does it exceed 10?

Q.E.8. Backward Elimination and Forward Selection based on using the model AIC will always "choose" the same model.

True / False

Q.E.9. An experiment considered the relationship between fuel consumption (Y) and speed (X) for container ships carrying 3000TEUs between Singapore and Kaohsiung. Measurements were based on various speeds over n = 20 runs. A plot of the data are given below.



p.9.a. Due to the "bends" in the plot, we consider fitting 4 models: a linear, quadratic, cubic, and quartic model. Complete the following table for various model selection measures.

Predictors	p'	SSE	Cp	AIC	BIC(SBC)
X		173.78		47.24	49.23
X,X^2		38.81	15.61		22.25
X,X^2,X^3		33.95	13.90	18.58	22.56
X,X^2,X^3,X^4		19.66	5.00	9.66	

p.9.b. Which model is selected based on each criteria?

C_p: _____ AIC: _____ BIC: _____

Q.E.10. An experiment was conducted relating air conditioner power performance (Y) to the following set of predictor variables: **T**emperature, **A**irFlux, **H**umidity, **W**heelSpeed, and **R**egenerationTemp. There were n = 18 experimental runs. Consider the following progression of models. Note that *TSS* = 66.57

p.10.a. Complete the following table.

Model	p*	SSE	C_p	AIC	BIC
T		37.21	39.86	17.07	18.85
T,A		16.88	12.43		7.51
T,A,H		10.75	5.56	-1.28	
T,A,H,W		9.12	5.20	-2.24	2.21
T,A,H,W,R		8.29			

p.10.b. Which model is selected based on each criteria?

C_p: _____ AIC: _____ BIC: _____

p.10.c. What proportion of the variation in power performance is explained by the one variable model? By the 5 variable model?

1 Variable: _____ 5 Variable: _____

Q.E.11. For any regression model, Adjusted-R^2 will be larger than R^2. _____

Q.E.12. A linear regression model is fit relating energy consumption in cars (Y, in MJ) to a set of 3 potential predictor variables: temperature (**tempC**), fiber space velocity (**fsv**, in mph), and stretching ratio (**stretch**, in %). The following models were fit (only main effects, no interactions or quadratic terms) and the error sums of squares as well as some of the model-based measures are given in the following table. Note, the experiment had n = 30 runs and TSS = 48.736.

Predictors	p'	SSE	R-square	Cp	AIC	BIC
tempC	2	48.618	0.002	856.825	18.484	21.286
fsv	2	1.630		3.600	-83.376	-80.574
stretch	2	47.713	0.021		17.920	20.722
tempC,fsv	3	1.439	0.970	2.130	-85.117	-80.914
tempC,stretch	3	47.585	0.024	840.067		24.043
fsv,stretch	3	1.624	0.967	5.493	-81.485	
tempC,fsv,stretch	4	1.432	0.971	4.000	-83.267	-77.662

p.12.a. Complete the table.

p.12.b. Based on C_p, AIC, and BIC, which model(s) are selected?