Linear Regression Problems

Q.1. A simple linear regression model is fit, relating plant growth over 1 year (y) to amount of fertilizer provided (x). Twenty five plants are selected, 5 each assigned to each of the fertilizer levels (12, 15, 18, 21, 24). The results of the model fit are given below:

Coefficients[a]

Model		Unstandardized Coefficients			t	Sig.
		B	Std. Error			
1	(Constant)	8.624	1.810		4.764	.000
	x	.527	.098		5.386	.000

a. Dependent Variable: y

Can we conclude that there is an association between fertilizer and plant growth at the 0.05 significance level? Why (be very specific).

Yes, for testing H_0: $\beta_1 = 0$ vs H_A: $\beta_1 \neq 0$, t = 5.386, p = .000

Give the estimated mean growth among plants receiving 20 units of fertilizer.

8.624 + 0.527(20) = 19.164

The estimated standard error of the estimated mean at 20 units is $2.1\sqrt{\frac{1}{25}+\frac{(20-18)^2}{450}} = 0.46$

Give a 95% CI for the mean at 20 units of fertilizer.

$t_{.025,23}$ = 2.069   19.164 ± 2.069(0.46) ≡ 19.164 ± 0.952 ≡ (18.212 , 20.116)

Q.2. A multiple regression model is fit, relating salary (Y) to the following predictor variables: experience (X_1, in years), accounts in charge of (X_2) and gender (X_3=1 if female, 0 if male). The following ANOVA table and output gives the results for fitting the model. Conduct all tests at the 0.05 significance level:

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	**3**	**2470.4**	**823.5**	**76.9**	**.0000**
Residual	**21**	**224.7**	**10.7**		
Total	**24**	**2695.1**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**39.58**	**1.89**	**21.00**	**0.0000**
experience	**3.61**	**0.36**	**10.04**	**0.0000**
accounts	**-0.28**	**0.36**	**-0.79**	**0.4389**
gender	**-3.92**	**1.48**	**-2.65**	**0.0149**

p.2.a. Test whether salary is associated with any of the predictor variables:

H_0: $\beta_1=\beta_2=\beta_3=0$ H_A: Not all β_i = 0 (i=1,2,3)

Test Statistic **F_{obs} = 76.9**

Reject H_0 if the test statistic falls in the range(s) **> $F_{.05,3,21}$ = 3.072** P-value **.0000**

p.2.b. Set-up the predicted value (all numbers, no symbols) for a male employee with 4 years of experience and 2 accounts.

39.58 + 3.61(4) + (-0.28)(2) + (-3.92)(0)

p.2.c. The following tables give the results for the full model, as well as a reduced model, containing only experience.

Test H_0: $\beta_2 = \beta_3 = 0$ vs H_A: β_2 and/or $\beta_3 \neq 0$

Complete Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	3	2470.4	823.5	76.9	.0000
Residual	21	224.7	10.7		
Total	24	2695.1			

Reduced Model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	2394.9	2394.9	183.5	0.0000
Residual	23	300.2	13.1		
Total	24	2695.1			

Test Statistic: $$F_{obs} = \frac{\left[\frac{300.2-224.7}{23-21}\right]}{\left[\frac{224.7}{21}\right]} = 3.528$$

Rejection Region: **$F_{obs} \geq F_{.05,2,21}$ = 3.467**

Conclude (Circle one): **Reject H_0** Fail to Reject H_0

Q.3. A study is conducted to determine whether students' first year GPA (Y) can be predicted by their ACT score (X). A random sample of n=120 freshmen from a small college were selected. The following EXCEL output gives the results of a simple linear regression on the data.

Regression Statistics	
Multiple R	**0.2695**
R Square	**0.0726**
Adjusted R Square	**0.0648**
Standard Error	**0.6231**
Observations	**120**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**3.5878**	**3.5878**	**9.2402**	**0.0029**
Residual	**118**	**45.8176**	**0.3883**		
Total	**119**	**49.4055**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**2.11**	**0.3209**	**6.5880**	**0.0000**	**1.4786**	**2.7495**
ACT(X)	**0.04**	**0.0128**	**3.0398**	**0.0029**	**0.0135**	**0.0641**

p.3.a. Give the fitted equation for predicting GPA as function of ACT score, and prediction for student scoring 20 on the ACT. $\hat{Y} = 2.11 + 0.04X$ $\hat{Y}_{20} = 2.11 + 0.04(20) = 2.91$

p.3.b. Test whether there is an association (positive or negative) between GPA and ACT

- Null Hypothesis: $\mathbf{H_0: \beta_1 = 0}$ Alternative Hypothesis: $\mathbf{H_A: \beta_1 \neq 0}$
- Test Statistic: $\mathbf{t_{obs} = 3.0398}$ or $F_{obs} = \mathbf{9.2402}$
- P-value **.0029**

p.3.c. What proportion of the variation in GPA is "explained" by ACT scores?

$\mathbf{r^2 = 3.5878 / 49.4055 = 0.0726}$

Q.4. A commercial real estate company is interested in the relationship between properties' rental prices (Y), and the following predictors: building age, expenses/taxes, vacancy rates, and square footage. The results for a regression are given below.

Regression Statistics	
Multiple R	**0.7647**
R Square	**0.5847**
Standard Error	**1.1369**
Observations	**81**

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	**4**	**138.3269**	**34.5817**	**26.7555**	**0.0000**
Residual	**76**	**98.2306**	**1.2925**		
Total	**80**	**236.5575**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**12.2006**	**0.5780**	**21.1099**	**0.0000**	**11.0495**	**13.3517**
age	**-0.1420**	**0.0213**	**-6.6549**	**0.0000**	**-0.1845**	**-0.0995**
exp/tax	**0.2820**	**0.0632**	**4.4642**	**0.0000**	**0.1562**	**0.4078**
vacancy	**0.6193**	**1.0868**	**0.5699**	**0.5704**	**-1.5452**	**2.7839**
sqfoot	**0.0000**	**0.0000**	**5.7224**	**0.0000**	**0.0000**	**0.0000**

p.4.a. Can the company conclude that rental rate is associated with any of these predictors? Give the test statistic and P-value for testing:

H_0: Average rental rate is not associated with any of the 4 predictors

H_A: Average rental rate is associated with at least one of the 4 predictors

TS: F_{obs} = 26.7555 P=.0000

p.4.b. What proportion of variation in prices is "explained" by the 4 predictors? **0.5847**
p.4.c. Controlling for all other factors, we conclude age is Positively **/ Negatively /** Not associated with rental price. (Circle One)

Q.5. A study was conducted to relate weight gain in chickens (Y) to the amount of the amino acid lysine ingested by the chicken (X). A simple linear regression is fit to the data.

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	27.07	27.07	23.79	0.0012
Residual	8	9.10	1.14		
Total	9	36.18			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	12.4802	1.2637	9.8762	0.0000
lysine(X)	36.8929	7.5640	4.8774	0.0012

p.5.a. Give the fitted equation, and the predicted value for X=0.20

$$\hat{Y} = 12.4802 + 36.8929X \qquad \hat{Y}_{.20} = 12.4802 + 36.8929(.20) = 19.8588$$

p.5.b. Give a 95% Confidence Interval for the MEAN weight gain of all chickens with X=0.20 (Note: the mean of X is 0.16 and S_{XX}=0.020)

$$\hat{SE}\left\{\hat{Y}_{.20}\right\} = \sqrt{1.14\left(\frac{1}{10}+\frac{(0.20-0.16)^2}{0.020}\right)} = \sqrt{1.14(0.18)} = 0.453 \qquad t_{.025,8} = 2.306$$

$$\text{95\% CI for the Mean:} \quad 19.8588 \pm 2.306(0.453) \equiv 19.8588 \pm 1.0446 \equiv (18.8142, 20.9034)$$

p.5.c. What proportion of the variation in weight gain is "explained" by lysine intake? **0.7482**

Q.6. A researcher reports that the correlation between length (inches) and weight (pounds) of a sample of 16 male adults of a species is r=0.40.

p.6.a. Test whether she can conclude there is a POSITIVE correlation in the population of all adult males of this species:
$H_0: \rho = 0$ $\qquad$ $H_A: \rho > 0$

- Test Statistic: $t_{obs} = \dfrac{0.40}{\sqrt{\dfrac{1-0.40^2}{16-2}}} = 1.633$
- Rejection Region (α=0.05): $t_{obs} \geq t_{.05,14} = 1.761$

Conclude: Positive Association or **No Positive Association**

p.6.b. A colleague from Europe transforms the data from length in inches to centimeters (1 inch=2.54 cm) and weight from pounds to kilograms (1 pound=2.2 kg). What is the colleague's estimate of the correlation? **0.40**

Q.7. Late at night you find the following SPSS output in your department's computer lab. The data represent numbers of emigrants from Japanese regions, as well as a set of predictor variables from each region.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.525(a)	.275	.222	181.89029

a Predictors: (Constant), PIONEERS, LANDCULT, AREAFARM

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	514814.087	3	171604.696	5.187	.004(a)
	Residual	1356447.158	41	33084.077		
	Total	1871261.244	44			

a Predictors: (Constant), PIONEERS, LANDCULT, AREAFARM

b Dependent Variable: EMGRANTS

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	407.070	226.341		1.798	.079
	LANDCULT	-1.685	3.567	-.069	-.472	.639
	AREAFARM	-2.132	1.056	-.299	-2.019	.050
	PIONEERS	175.968	61.222	.391	2.874	.006

a Dependent Variable: EMGRANTS

a) How many regions are there in the analysis? **45**
b) Give the test statistic and P-value for testing (H0) that none of the predictors are associated with EMGRANTS **F_{obs} = 5.187 p = .004**

c) Give the test statistic and P-value for testing whether LANDCULT is associated with EMGRANTS, after controlling for AREAFARM and PIONEERS **t_{obs} = -0.472 p = .639**

d) What proportion of the variation in EMGRANTS is "explained" by the model? **.275**

e) Give the estimated regression equation $\hat{Y} = 407.070 - 1.685L - 2.132A + 175.968P$

Q.8. A realtor is interested in the determinants of home selling prices in his territory. He takes a random sample of 24 homes that have sold in this area during the past 18 months, observing: selling PRICE (Y), AREA (X_1), BEDrooms (X_2), BATHrooms (X_3), POOL dummy (X_4=1 if Yes, 0 if No), and AGE (X_5). He fits the following models (predictor variables to be included in model are given for each model):

Model 1: AREA, BED, BATH, POOL, AGE $SSE_1 = 250$, $SSR_1 = 450$

Model 2: AREA, BATH, POOL $SSE_2 = 325$, $SSR_2 = 375$

a) Test whether neither BED or AGE are associated with PRICE, after adjusting for AREA, BATH, and POOL at the α=0.05 significance level. That is, test:

$$H_0 : \beta_2 = \beta_5 = 0 \qquad vs \qquad H_A : \beta_2 \neq 0 \; and / or \; \beta_5 \neq 0$$

$$TS : F_{obs} = 2.700 \quad RR : F_{obs} \geq 3.555$$

b) What statement best describes β_4 in Model 1?
   a) **Added value (on average) for a POOL, controlling for AREA, BED, BATH, AGE**
   b) Effect of increasing AREA by 1 unit, controlling for other factors
   c) Effect of increasing BED by 1 unit, controlling for other factors
   d) Effect of increasing BATH by 1 unit, controlling for other factors
   e) Average price for a house with a POOL

Q.9. In linear regression, it is possible for an independent variable to be significant at the 0.05 significance level when it is the only independent variable, and not be significant when it is included in a regression with other independent variables. **T**/F

Q.10. A simple linear regression is fit, and we get a fitted equation of $Y = 50 + 10X$. Our estimate of the increase in the mean of *Y* for unit increase in *X* is 60. **False**

Q.11. In simple regression, if X is temperature (in Fahrenheit) and Y is distance (in Yards) and a colleague wishes to transform X to Celsius and Y to Meters, the regression coefficient for X will remain the same for the two regressions, but the correlation coefficient will change. T/**F**

Q.12. In multiple regression, it is possible for the error sum of squares to increase when we add an independent variable to an existing model. **FALSE**

Q.13. A multiple regression model is fit, relating Gainesville House Prices (Y, in $1000s) to 4 predictors: BEDrooms, BATHrooms, an indicator (dummy) variable for NEW, and SIZE (ft^2). A subset of the results are given in the following tables.

ANOVA					
	df	*SS*	*MS*	*F*	*F(.05)*
Regression	**4**	735525.457	**183881.4**	**62.47**	**2.467**
Residual	**95**	**279624.1**	**2943.4**		
Total	99	1015149.53			

	Coefficients	*andard Erro*	*t Stat*	*P-value*
Intercept	-28.8492	27.2612	-1.0583	0.2926
Beds	-8.2024	10.4498	-0.7849	0.4344
Baths	5.2738	13.0802	0.4032	0.6877
New	54.5624	19.2149	2.8396	0.0055
Size	0.1181	0.0123	9.6016	<.0001

p.7.a. Complete the ANOVA table.

p.7.b. Complete the Coefficients Table.

p.13.c. Compute R^2 **0.7245**

p.13.d. Compute the predicted price (in $1000s) for a 3 Bedroom, 3 Bathroom, not new home that is 3000 ft^2. **316.665**

p.13.e. We fit a reduced model with only NEW and SIZE, and obtain R^2=0.7226, and SSR=733543.3.

Test H_0: $\beta_{Bed} = \beta_{Bath} = 0$ at the $\alpha = 0.05$ significance level:

p.13.e.i. Test Statistic: $\mathbf{F_{obs} = 0.33}$

p.13.e.ii. Reject H_0 if the test statistic falls in the following range $\geq$ **3.092**

p.13.e.iii True/**False**: After controlling for NEW and SIZE, neither BEDS nor BATHS is associated with house prices.

Q.14. A simple linear regression is to be fit, relating fuel efficiency (*Y* in gallons/100 miles) to cars weight (*X*, in pounds), based on a sample of *n*=45 cars. You are given the following information:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = 13069326 \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right)(Y_i - \overline{Y}) = 13385 \quad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 16.5$$

$$\overline{X} = 2739 \quad \overline{Y} = 3.4 \quad \sum\left(Y - Y\right)^2 = 2.835$$

Compute the following quantities:

p.14.a. β_1 **= 13385 / 13069326 = 0.001024**

p.14.b. β_0 **= 3.4 – 0.001024(2739) = 0.5953**

p.14.c. Residual Std. Deviation s_e **= sqrt(2.835 / (45-2)) = 0.2568**

p.14.d. Estimate of mean efficiency for all cars of x*=2000 pounds **2.6433**

p.14.e. 95% Confidence Interval for all cars of *x**=2000 pounds

$$2.6433 \pm 2.017(0.2568)\sqrt{\frac{1}{45} + \frac{(2000-2739)^2}{13069326}} \equiv 2.6433 \pm 2.017(0.2568)(0.2530) \equiv$$
$$2.6433 \pm 0.1310 \equiv (2.5123, 2.7743)$$

p.14.f. Regression Sum of Squares ***SSR* = 16.5 – 2.835 = 13.665**

p.14.g. Proportion of Variation in Efficiency "Explained" by Weight **13.665 / 16.5 = 0.8282**

Q.15. A multiple regression equation was fit for n = 36 observations using 5 independent variables X1, X2,…, X5 gave SS(Residual) = 900. What is the residual standard deviation (standard error of estimate)?

sqrt(900 / (36-6)) = 5.477

Q.16. A multiple regression equation was fit for n = 21 observations using 5 independent variables X1, X2,…, X5 gave SS(Total) = 1500 and SS(Residual) = 375.

p.16.a. Calculate the value of the coefficient of determination. **0.75**

p.16.b. Test the hypothesis that all the slopes are zero. $H_0: \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0$

Test Statistic $\mathbf{F_{obs} = 9.00}$ Rejection Region: $\mathbf{F_{obs} \geq 2.901}$

Q.17. Write the multiple regression equations needed to be fit for determining if the linear relationship of Y = response time as a function of X_1 = strength of signal has the same slope for three groups (clearly define all independent variables).

p.17.a. Complete (Full) Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon \qquad X_2 = \begin{cases} 1 \text{ if group 2} \\ 0 \text{ otherwise} \end{cases} \quad X_3 = \begin{cases} 1 \text{ if group 3} \\ 0 \text{ otherwise} \end{cases}$$

p.17.b. Reduced Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad X_2 = \begin{cases} 1 \text{ if group 2} \\ 0 \text{ otherwise} \end{cases} \quad X_3 = \begin{cases} 1 \text{ if group 3} \\ 0 \text{ otherwise} \end{cases}$$

Q.18. The ANOVA tables for fitting Y as a linear function of X are shown below. In the first table the "independent variables" include X1, the continuous variable, X2, X3, and X4 as dummy variables to denote the four groups, and X12, X13, and X14 representing the crossproducts of X1 and the dummy variables. The second table is the ANOVA table for fitting Y as a linear function of X1, X2, X3, and X4.

Model:(X1,X2,X3,X4,X12,X13,X14)		
Source	**df**	**SS**
Regression	**7**	**28000**
Error	**12**	**7000**
Total	**19**	**35000**
Model:(X1,X2,X3,X4)		
Source	**df**	**SS**
Regression	**4**	**21000**
Error	**15**	**14000**
Total	**19**	**35000**

p.18.a. Complete the tables.
p.18.b. For the second model, test H_0: $\beta_1=\beta_2=\beta_3=\beta_4 = 0$

Test Statistic **F_{obs} = 5.625** Rejection Region **$F_{obs} \geq 3.056$**

p.18.b. Is there significant evidence the slopes are not equal among the 4 groups?

H_0: $\beta_{12} = \beta_{13} = \beta_{14} = 0$

Test Statistic **F_{obs} = 4.000** Rejection Region **$F_{obs} \geq 3.490$**

Q.19. You obtain the following spreadsheet from a regression model. The fitted equation is $\hat{Y} = -4.67 + 4.00X$
Conduct the F-test for Lack-of-Fit. n = **6** c = **3**

X	Y	Ybar_grp	Y-hat_grp	Pure Error	Lack of Fit
2	3	4	3.33	-1	0.67
2	5	4	3.33	1	0.67
4	8	10	11.33	-2	-1.33
4	12	10	11.33	2	-1.33
6	18	20	19.33	-2	0.67
6	22	20	19.33	2	0.67

Source	df	SS	MS	F	F(0.05)
Lack-of-Fit	3-2=1	5.33	5.33	0.89	10.13
Pure Error	6-3=3	18	6		

Q.20. Bob fits a regression model relating weight (Y) to weight (X_1) for professional basketball players, with a dummy variable for males (X_2 = 1 if Male, 0 if Female). Cathy fits a model **on the same dataset**, but she defines X_2 = 1 if Female, 0 if Male.

$$\text{Bob: } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \qquad X_2 = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

$$\text{Cathy: } \hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 X_1 + \hat{\gamma}_2 X_2 \qquad X_2 = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if Male} \end{cases}$$

What are the relationships among $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$

$$\hat{\beta}_0 = \hat{\gamma}_0 + \hat{\gamma}_2 \qquad \hat{\beta}_1 = +\hat{\gamma}_1 \qquad \hat{\beta}_2 = \hat{\gamma}_0 - \hat{\beta}_0 = \hat{\gamma}_0 - \left(\hat{\gamma}_0 + \hat{\gamma}_2\right) = -\hat{\gamma}_2$$

Q.21. The ANOVA tables for fitting Y as a linear function of X are shown below. In the first table the "independent variables" include X1, the continuous variable, X2 and X3 as dummy variables to denote the three groups, and X12 and X13 representing the cross-products of X1 and the two dummy variables. The second table is the ANOVA table for fitting Y as a linear function of X1, X2, X3.

Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_{12} + \beta_{13} X_{13}$ Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Model 1:(X1,X2,X3,X12,X13)		
Source	df	SS
Regression	5	32000
Error	18	8000
Total	23	40000
Model 2:(X1,X2,X3)		
Source	df	SS
Regression	3	28000
Error	20	12000
Total	23	40000

p.21.a. Complete the tables.
p.21.b. For the second model, test H_0: $\beta_1=\beta_2=\beta_3 = 0$

Test Statistic **F_{obs} = 15.556** Rejection Region **$F_{obs} \geq 3.098$**

p.21.c. Is there significant evidence the slopes are not equal among the 3 groups?

$H_0 : \beta_{12} = \beta_{13} = 0$

Test Statistic **F_{obs} = 4.50** Rejection Region **$F_{obs} \geq 3.555$**

Q.22. In the production of a certain chemical it is believed the yield, Y, can be increased by increasing the amount of a particular catalytic agent. Twenty trials were made with different amounts of the catalyst. Analysis of the yields, measured in grams, and amounts of the catalyst, X, in milligrams gave based on the following model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma_e^2\right)$$

$\bar{x} = 10$ $\bar{y} = 139.0$ $\Sigma(x - \bar{x})^2 = 500$ $\Sigma(y - \bar{y})^2 = 895$ $\Sigma(x - \bar{x})(y - \bar{y}) = 350$

p.22.a. Compute the estimated slope. **0.700**

p.22.b. Compute the estimated y-intercept. **132**

p.22.c. $SSE = \sum\left(Y - \hat{Y}\right)^2 = 650$ Compute the estimate of the residual standard deviation: S_e **6.009**

p.22.d. Compute a 95% Confidence Interval for β_1: $0.700 \pm 2.101\left(\frac{6.009}{\sqrt{500}}\right) \equiv 0.700 \pm 0.565 \equiv (0.135, 1.265)$

Q.23. A regression model was fit, relating revenues (Y) to total cost of production and distribution (X) for a random sample of n=30 RKO films from the 1930s (the total cost ranged from 79 to 1530):

$$n=30 \quad \overline{X}=685.2 \quad S_{xx}=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=6126371 \quad \hat{Y}=55.23+0.92X \quad S_e^2=\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{n-2}=40067$$

p.23.a. Obtain a 95% Confidence Interval for the **mean revenues for all movies** with total costs of $x^*= 1000$

$$\text{Note: }\left[\frac{1}{30}+\frac{(1000-685.2)^2}{6126371}\right]=0.0495$$

$$\hat{\mu}_y=975.23 \quad SE_{\hat{\mu}}=44.53 \quad 95\%\text{CI}: 975.23\pm 91.21\equiv(884.02,1066.44)$$

p.23.b. Obtain a 95% Prediction Interval for **tomorrow's new film** that had total costs of $x^*= 1000$

$$\hat{\mu}_y=975.23 \quad SE_{\hat{y}}=205.06 \quad 95\%\text{PI}: 975.23\pm 419.97\equiv(555.26,1395.20)$$

Q.24. A researcher is interested in the correlation between height (X) and weight (Y) among 12 year old male children. He selects a random sample of n = 18 male 12-year olds from a school district, and intends on testing H_0: $\rho = 0$ versus H_A: $\rho \neq 0$, where ρ is the population correlation coefficient.

His sample correlation is $r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}=0.60$

Test H_0: $\rho = 0$ versus H_A: $\rho \neq 0$:

Test Statistic $\mathbf{t_{obs} = 3.00}$ Rejection Region: $\mathbf{|t_{obs}| \geq 2.120}$

Q.25. In order to help estimate the peak power demand of a generating plant, data was collected to see if there was a linear relationship between the forecast high temperature for a day and the peak load demand for that day. Computer analysis of the data gave the following (abbreviated) results:

Variable	N	Mean	Std Dev
temp	10	91.400	6.687
load	10	195.000	45.225

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16735	16735	80.01	<.0001
Error	8	1673	209		
Corrected Total	9	18408			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error
Intercept	1	-394.42097	66.05542
temp	1	6.44881	0.72097

p.25.a. Give a 95% confidence interval for the increase in expected peak load for a 1 degree increase in predicted high temperature. **6.449 ± 2.306(0.721) ≡ 6.449 ± 1.663 ≡ (4.786 , 8.112)**

p.25.b. Give a point estimate and 95% confidence interval for the mean peak load when the forecast high is 84.

$$s_e\sqrt{\frac{1}{n}+\frac{\left(x^*-\bar{x}\right)^2}{S_{xx}}}=14.46\sqrt{\frac{1}{10}+\frac{(84-91.4)^2}{402.44}}=7.03$$

Point Estimate: **147.295** 95% CI: **(131.084 , 163.506)**

p.25.c. Compute r^2, the coefficient of determination **0.9091**

Q.26. A regression model is fit, relating height (Y, in cm) to hand length (X_1, in cm) and foot length (X_2, in cm) for a sample of n=75 adult females. The following results are obtained from a regression analysis of:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad \varepsilon \sim NID(0,\sigma^2)$$

Regression Statistics	
R Square	0.616

ANOVA	*df*	*SS*	*MS*	*F**	*F(0.05)*
Regression	2	1105.52	552.76	57.82	3.1240
Residual	72	688.33	9.56		
Total	74	1793.85			

	Coeff	*StdErr*	t*	t(.025)
Intercept	74.41	7.97		
X1	2.38	0.49	4.857	1.993
X2	1.73	0.37	4.676	1.993

p.26.a. Complete the tables.

p.26.b. The first woman in the sample had a hand length of 19.56cm, a foot length of 25.70cm, and a height of 160.60cm. Obtain her fitted value and residual.

Fitted value = **165.42** Residual = **+1.18**

Q.27. In simple linear regression, if the estimated slope is 0, then the correlation will be **0**

Q.28. In multiple regression with 2 predictors, it is possible to reject H_0: $\beta_1 = \beta_2 = 0$ but fail to reject either H_0: $\beta_1 = 0$ or H_0: $\beta_2 = 0$ **True** / False

Q.29. A regression model is fit, based on n = 25 subjects and k=4 predictor variables. How large will R^2 need to be to reject H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$?

$$\text{Test Statistic: } F_{obs} = \frac{R^2/4}{\left(1-R^2\right)/(25-5)} = 5\left(\frac{R^2}{1-R^2}\right) \quad \text{Rejection Region: } F_{obs} \geq F_{.05,4,20} = 2.866$$

$$\Rightarrow \left(\frac{R^2}{1-R^2}\right) \geq \frac{2.866}{5} \quad \Rightarrow \quad \frac{1-R^2}{R^2} \leq \frac{5}{2.866} = 1.744 \quad \Rightarrow \quad \frac{1}{R^2} \leq 2.744 \quad \Rightarrow \quad R^2 \geq \frac{1}{2.744} = 0.364$$

Q.30. In a regression model, if R^2 = 1, then SSE = **0**

Q.31. A study was conducted to determine the effects of daily temperature (X, in °C) on Electricity Consumption (Y, in 1000s of Wh) in an experimental house over a period of n = 31 days. Consider the following model:

$$E\{Y\} = \beta_0 + \beta_1 X \quad SSR = 594.0 \quad SSE = 241.4 \quad TSS = 835.4 \quad S_{XX} = 158.5 \quad \overline{X} = 27.0 \quad \hat{Y} = -30.179 + 1.936X$$

p.31.a. What proportion of the variation in Electricity consumption is “explained” by daily temperature (X)? **0.7110**

p.31.b. Compute the residual standard deviation, s_e **2.885**

p.31.c. Obtain the estimated **mean** electricity consumption when x^* = 27.0 degrees, and the 95% Confidence Interval.

Estimated Mean: **22.093** 95% CI: **22.093 ± 1.060 ≡ (21.033 , 23.15)**

Q.32. A study involved measuring head size (x, in cm^3) and brain weight (y, in grams) among a sample of n = 77 adult males over 45 years old. The following summary statistics were obtained:

$$\sum\left(x-\overline{x}\right)\left(y-\overline{y}\right) = 180.85 \qquad \sum\left(x-\overline{x}\right)^2 = 751.47 \qquad \sum\left(y-\overline{y}\right)^2 = 91.25 \qquad \overline{x} = 37.49 \qquad \overline{y} = 13.07$$

p.32.a. Compute the sample correlation between head size and brain weight, r_{yx}: **0.691**

p.32.b. Test whether there is a positive association in the corresponding population: H_0: $\rho_{yx} \leq 0$ H_A: $\rho_{yx} > 0$

Test Statistic: **t_{obs} = 8.279** Rejection Region: **$t_{obs} \geq 1.665$**

p.32.c. If the measurements had been made in ounces (1 ounce = 28.35 grams) and inches^3 (1 inch = 2.54cm) , what would be the sample correlation between brain weight and head size? **0.691**

Q.33. 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.33.a. The following model allows **separate intercepts** for each crop type, with a **common slope for water table depth** among crop types. For this model, SSE = 1.853. Give SSR, R^2, and the error degrees of freedom.

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

SSR_1 = **33.833** R_1^2 = **0.9481** df_{ERR1} = **20**

p.33.b. The following model allows **separate intercepts** for each crop types, with **separate slopes for water table depth** among crop types. For this model, SSE = 1.261. Give SSR, R^2, and the error degrees of freedom.

Model 2: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3$

SSR_2 = **34.425** R_2^2 = **0.9647** df_{ERR2} = **18**

p.33.c. Test whether the slopes are the same for the 3 crop types. H_0: $\beta_{12} = \beta_{13} = 0$

Test Statistic: $\mathbf{F_{obs} = 4.225}$ Rejection Region: $\mathbf{F_{obs} \geq 3.555}$