1. Use the patient satisfaction data described in KNNL Problem 6.15.

a. Compute the pairwise correlations between the X's and between each X and Y. Which X variable appears to be the best individual predictor?

Pearson Correlation Coefficients, N = 46
Prob > |r| under H0: Rho=0

	age	severity	anxiety	satisfaction
age	1.00000	0.56795 <.0001	0.56968 <.0001	-0.78676 <.0001
severity	0.56795 <.0001	1.00000	0.67053 <.0001	-0.60294 <.0001
anxiety	0.56968 <.0001	0.67053 <.0001	1.00000	-0.64459 <.0001
satisfaction	-0.78676 <.0001	-0.60294 <.0001	-0.64459 <.0001	1.00000

Age appears to be the best individual predictor since satisfaction varies the strongest relative to age (-0.78676) than severity and anxiety (-0.60294 and -0.64459, respectively).

b. Run the linear regression with age, severity of illness and anxiety level as the explanatory variables and satisfaction as the response variable. Summarize the regression results.

The REG Procedure
Model: MODEL1
Dependent Variable: satisfaction

Number of Observations Read	46
Number of Observations Used	46

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9120.46367	3040.15456	30.05	<.0001
Error	42	4248.84068	101.16287		
Corrected Total	45	13369			

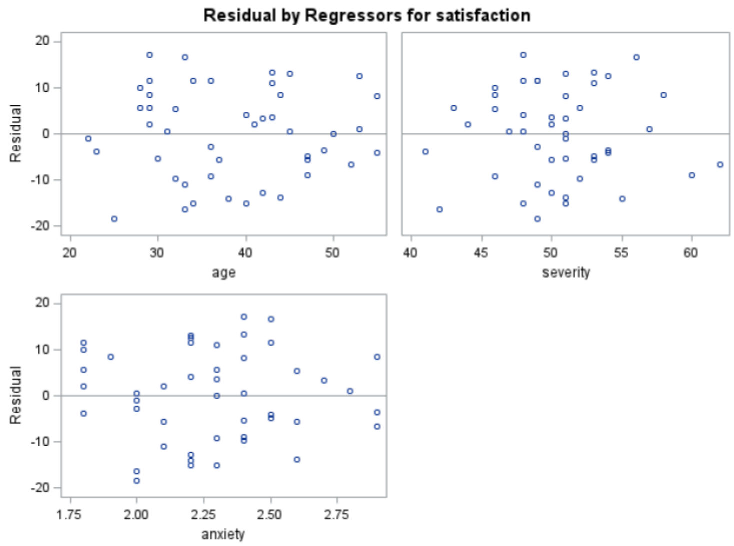
Root MSE	10.05798	R-Square	0.6822
Dependent Mean	61.56522	Adj R-Sq	0.6595
Coeff Var	16.33711		

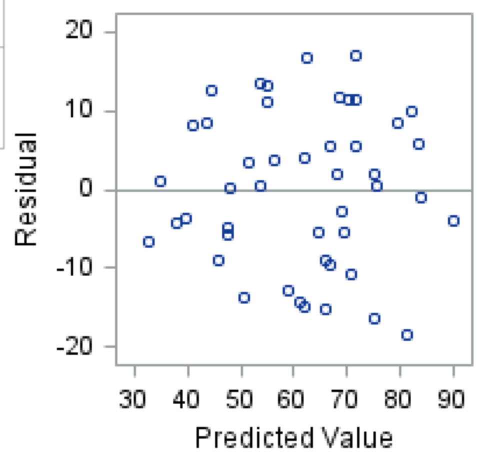
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	158.49125	18.12589	8.74	<.0001
age	1	-1.14161	0.21480	-5.31	<.0001
severity	1	-0.44200	0.49197	-0.90	0.3741
anxiety	1	-13.47016	7.09966	-1.90	0.0647

The regression model is $\hat{Y} = 158.491 - 1.142X_1 - 0.442X_2 - 13.470X_3$, but the only significant predictor seems to be age ($P < 0.001$).

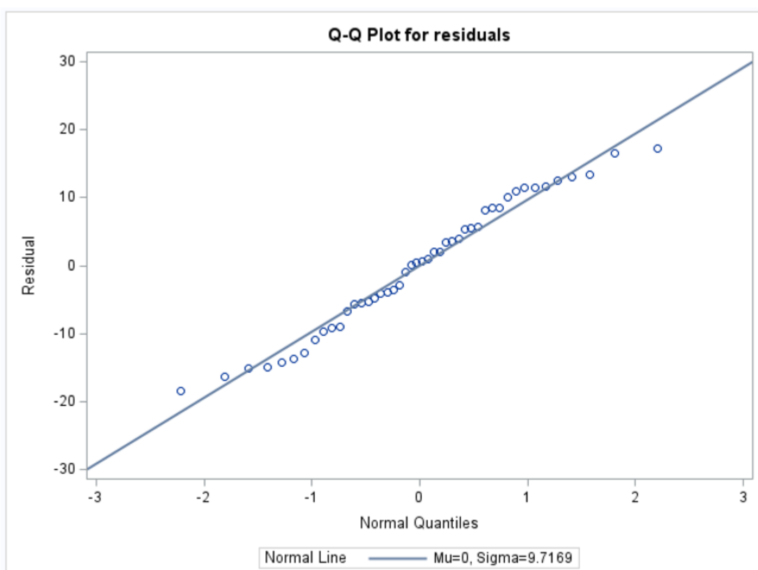
c. Plot the residuals versus the predicted satisfaction and each of the explanatory variables. Are there any unusual patterns?





Residuals appear to have constant variance and more or less evenly distributed about 0, so we conclude that so far, none of our assumptions about regression have been violated.

d. Examine the assumption of normality for the residuals using a qqplot or histogram. State your conclusions.



We see that the residuals do not significantly deviate from the 45-degree line, so we conclude our assumption of normality is appropriate.

e. Predict the satisfaction for a 55 year old patient with illness severity 50 and anxiety level 2.8. Provide a 95% prediction interval with your prediction.

$\widehat{Y}_h = 158.491 - 1.142(55) - 0.442(50) - 13.470(2.8) = 35.886$. $MSE = 101.163$, $\boldsymbol{X}_h^T = [1\ 55\ 50\ 2.8]^T, \boldsymbol{X}_h = [1\ 55\ 50\ 2.8]$, and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is singular… how do I proceed?

$$t\left(1 - \frac{.05}{2}, 46 - 4\right) = t(0.975, 42) = 2.0181$$

2. Refer to Patient satisfaction Problem 6.15.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_2; with X_1, given X_2; and with X_3, given X_2 and X_1.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	158.49125	18.12589	8.74	<.0001	174353	7734.51573
age	1	-1.14161	0.21480	-5.31	<.0001	8275.38885	2857.55338
severity	1	-0.44200	0.49197	-0.90	0.3741	480.91529	81.65905
anxiety	1	-13.47016	7.09966	-1.90	0.0647	364.15952	364.15952

b. Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained. Use the F^* test statistic and level of significance $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Hypotheses: $H_0\!: \beta_3 = 0$, $H_a\!: \beta_3 \neq 0$. We want $F^* = \frac{SSE(X_1,X_2) - SSE(X_1,X_2,X_3)}{1} \div \frac{SSE(X_1,X_2,X_3)}{42} \leq F(0.975,1,42)$ to conclude H_0, and $F^* > F(0.975,1,42)$ to conclude H_a. Then $F^* = (4613.00 - 4248.84) \div \frac{4248.84}{42} = 3.59974$. This is less than $F(0.975,1,42) = 5.404$, so we conclude H_0; $\beta_3 =$ anxiety not useful in predicting satisfaction and can be dropped.

3. Refer to Patient satisfaction Problem 6.15. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 is retained. Use $\alpha = 0.025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

The REG Procedure
Model: MODEL1

Test thimid Results for Dependent Variable satisfaction				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	422.53741	4.18	0.0222
Denominator	42	101.16287		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	158.49125	18.12589	8.74	<.0001	174353
age	1	-1.14161	0.21480	-5.31	<.0001	8275.38885
severity	1	-0.44200	0.49197	-0.90	0.3741	480.91529
anxiety	1	-13.47016	7.09966	-1.90	0.0647	364.15952

Hypotheses: $H_0: \beta_2 = \beta_3 = 0$, H_a: At least one $\beta_i \neq 0$ for $i \in \{2,3\}$. The test statistic is $F^* = \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{1} \div \frac{SSE(X_1, X_2, X_3)}{42} = \frac{SSE(X_2, X_3 | X_1)}{SSE(X_1, X_2, X_3)/42} = \frac{(480.91529)+(364.15952)}{4248.841/42} = 8.3536$. This is greater than $F(0.0975, 1, 42) = 5.404$, so we conclude H_a; among the predictor variables severity and anxiety, at least one of them is statistically useful in predicting satisfaction. We can further justify our conclusion using "thimid test severity, anxiety" in SAS. After doing so, we get a P-value of $0.0222 > 0.025$, so we also conclude H_a.

4. Derive the equation (7.56) on page 281 (HINT: Recall that b_1 can be obtained by regressing the residuals of $Y|X_2$ vs the residuals of $X_1|X_2$).

We want
$$b_1 = \frac{\dfrac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} - \left[\dfrac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right]^{1/2} r_{Y2} r_{12}}{1 - r_{12}^2}.$$

Consider $\varepsilon_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$ and $\varepsilon_i(X_1|X_2) = X_{i1} - (\hat{X}_{i1}|X_2)$.
Now regress $\varepsilon_i(Y|X_2) \sim \varepsilon_i(X_1|X_2)$:

$$Y_i - \hat{Y}_i(X_2) = b_0 + b_1[X_{i1} - (\hat{X}_{i1}|X_2)]$$

Then $b_1 = \dfrac{\sum(\varepsilon_i(X_1) - \varepsilon_i(\hat{X}_1|X_2))(\varepsilon_i(Y|X_1) - \varepsilon_i(Y|X_2))}{\sum(\bar{\varepsilon}(X_1|X_1) - \bar{\varepsilon}(X_2|X_1))^2}$ ... help

5. Steroid level. An endocrinologist was interested in exploring the relationship between the level of a steroid (Y) and age (X) in healthy female subjects whose ages ranged from 8 to 25 years. She collected a sample of 27 healthy females in this age range.

a. Fit regression model (8.2). Plot the fitted regression function and the data. Does the quadratic regression function appear to be a good fit here? Find R^2.

The REG Procedure
Model: MODEL1
Dependent Variable: steroid_level

Number of Observations Read	27
Number of Observations Used	27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			

Root MSE	3.15265	R-Square	0.8143
Dependent Mean	17.64444	Adj R-Sq	0.7989
Coeff Var	17.86766		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-26.32541	5.88154	-4.48	0.0002
age	1	4.87357	0.77515	6.29	<.0001
age2	1	-0.11840	0.02347	-5.05	<.0001

$R^2 = 0.8143$ and the predictors "age" and "age2" are significant, so the quadratic regression function appears to be a good fit.

b. Test whether or not there is a regression relation; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Hypotheses: $H_0: \beta_1 = \beta_2 = 0$, H_a: At least one $\beta_i \neq 0$ for $i \in \{1,2\}$. The decision rule is as follows: $F^* > F(0.99, 1, 27-3)$, conclude H_a, and $F^* \leq F(0.99, 1, 27-3)$, conclude H_0. Then $F^* = 52.63$ and $F(0.99, 1, 24) = 2.927$, so we conclude that there exists a regression relationship between age and steroid level.

c. Obtain joint interval estimates for the mean steroid level of females aged 10, 15, and 20 respectively. Use the most efficient simultaneous estimation procedure and a 99 percent family confidence coefficient. Interpret your intervals.

We know $\hat{Y}_h = -26.32541 + 4.87357X_h - 0.11840X_h^2$, so $\hat{Y}_{10} = 10.570$, $\hat{Y}_{15} = 20.138$, $\hat{Y}_{20} = 23.786$. Consider the Bonferroni simultaneous prediction limits for $g = 3$: $B = t\left(1 - \frac{0.01}{2(3)}, 27 - 3\right) = t(0.9833, 24) = 2.2568$. Then $s^2\{\hat{Y}_h\} = \boldsymbol{X}_h^T \boldsymbol{s}^2\{\boldsymbol{b}\}\boldsymbol{X}_h = MSE(\boldsymbol{X}_h^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_h) = 9.93920([1\ 10\ 15\ 20]^T \ldots$ once again, $\boldsymbol{X}^T\boldsymbol{X}$ is singular and I can't proceed. If I know the value of $s^2\{\hat{Y}_h\}$, then I know the simultaneous mean response CIs are $\hat{Y}_i \pm t(0.9833, 24)s^2\{\hat{Y}_h\} = \hat{Y}_i \pm (2.2568)s^2\{\hat{Y}_i\}$ for $i = 10, 15$, and 20. The way we would interpret these intervals is that we are 99% confident the true mean steroid level is within those intervals for ages $10, 15$, and 20.

d. Predict the steroid levels of females aged 15 using a 99 percent prediction interval. Interpret your interval.

We know $s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$, $\hat{Y}_{15} = 20.138$, and the 99% PI is given by $\hat{Y}_{15} \pm t\left(1 - \frac{0.01}{2}, 27 - 3\right)s^2\{pred\} = 20.138 \pm (2.794)s^2\{pred\}$. We are 99% confident that a new observation of a steroid level for a 15-year-old woman will lie within this interval.

e. Test whether the quadratic term can be dropped from the model; use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-26.32541	5.88154	-4.48	0.0002	8405.81333
age	1	4.87357	0.77515	6.29	<.0001	793.28051
age2	1	-0.11840	0.02347	-5.05	<.0001	252.98535

Hypotheses: $H_0: \beta_2 = 0$, $H_a: \beta_2 \neq 0$. The decision rule is as follows: $F^* > F(0.99, 1, 27 - 3)$, conclude H_a, and $F^* \leq F(0.99, 1, 27 - 3)$, conclude H_0. $F^* = \frac{SSE(X_2|X_1)}{SSE(X_2,X_1)/24} = \frac{252.985}{238.541/24} = 25.453$, which is greater than $F(0.99, 1, 24) = 2.927$, so we conclude that the quadratic term has statistically significant predictive power in the model.

f. Express the fitted regression function obtained in part (a) in terms of the original variable X.

The fitted regression function is $\hat{Y}_i = b_0 + b_1X_i + b_2X_i^2$. Let $\tilde{X}_i = X_i - \bar{X}, X_i = \tilde{X}_i + \bar{X}$. Then $\hat{Y}_i = b_0 + b_1X_i + b_2X_i^2 = b_0 + b_1(\tilde{X}_i + \bar{X}) + b_2\left(\tilde{X}_i + \bar{X}\right)^2$.

6. Refer to Copier maintenance Problem 1.20. The users of the copiers are either training institutions that use a small model, or business firms that use a large, commercial model An analyst at Tri-City wishes to fit a regression model including both number of copiers serviced (X_1) and type of copier (X_2) as predictor variables and estimate the effect of copier model (S-small, L-large) on number of minutes spent on the service call. Assume that regression model (8.33) is appropriate, and let $X_2 = 1$ if small model and 0 if large, commercial model.

a. Explain the meaning of all regression coefficients in the model.

(8.33), the regression model is $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$. β_0 is the expected number of service minutes when there are no copiers to be serviced and the user is a business firm. β_1 is the expected increase in service minutes for each additional copier that requires work. β_2 is the expected increase in service minutes for a training institution. Finally, ϵ_i is the error term.

b. Fit the regression model and state the estimated regression function.

Number of Observations Read	45
Number of Observations Used	45

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	77059	38530	487.81	<.0001
Error	42	3317.38664	78.98540		
Corrected Total	44	80377			

Root MSE	8.88737	R-Square	0.9587
Dependent Mean	76.26667	Adj R-Sq	0.9568
Coeff Var	11.65302		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.11143	3.11239	-0.68	0.5012
number_serviced	1	15.10676	0.48589	31.09	<.0001
type	1	3.08592	2.75652	1.12	0.2693

The regression function is $\hat{Y} = -2.11143 + 15.10676X_1 + 3.08592X_2$.

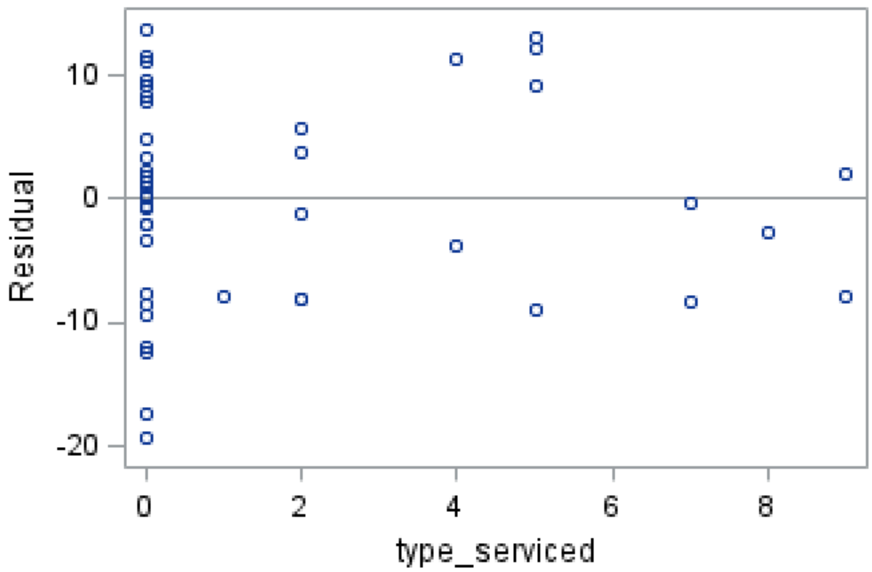
c. Estimate the effect of copier model on mean service time with a 95 percent confidence interval. Interpret your interval estimate.

We want $b_2 \pm t\left(1 - \frac{0.05}{2}, 45 - 3\right) s\{b_2\} = 3.083592 \pm t(0.975{,}42)(2.75652) = 3.083592 \pm (2.01808)(2.75652) = (-2.47928, 8.64646)$. We are 95% confident that the mean service time for a small copier lies within the interval $(-2.47928, 8.64646)$.

d. Why would the analyst wish to include X_1, number of copiers, in the regression model when interest is in estimating the effect of type of copier model on service time?

Because the number of copiers is also statistically significant in predicting service time, so accounting for this variable lets us make better estimates based on type of copier.

e. Obtain the residuals and plot them against X_1X_2. Is there any indication that an interaction term in the regression model would be helpful?



The "type_serviced" variable is the interaction term X_1X_2. Indeed, an interaction term in the regression model would appear helpful as evidenced by the distribution of residuals.