1. **SAT multiple regression**

(a) The numerical codes assigned to regions don't have any meaning. If we include this variable in the regression, then we are implicitly assuming that the difference in expected response between S and NE (2-1) is the same as the difference between MW and S (3-2) and W and MW (4-3). This assumption is almost certainly crazy. A categorical variable like this should be replaced by a series of binary indicators for the different regions.

(b) Model 1 - Here R implicitly defined northeast as the reference region and created indicators for S (a variable equal to 1 if the state is in the south region and 0 otherwise), MW, W. The coefficients of these variables can be interpreted as the difference in expected SAT score for two states identical on all predictors where one is in the specified region and the others is in the northeast region.

   i. Adding the first seven rows yields SS(Model)=228525 and df(Model)=9. The F-statistic is then \((228525/9)/437=58.10\) as given in the output.

   ii. Using a two-sided t test with 40 d.f. the cutoff for significance at the .05 level is about 2.02. Rank and South both have p-values less than .05.

   iii. The coefficient of years means that for each additional year in the average preparatory coursework of the students taking the exam in a state with all other variables held fixed, there is an expected increase of 11.6 points in the state mean SAT score.

   iv. The square root of the MSE is an estimate of the standard deviation of the regression errors or regression variation. Here we are confident that state mean SAT scores are within 42 points of the regression line (this is two standard deviations).

(c) To test the hypothesis of five coefficients being simultaneously zero we must use information from model 1 and model 2. We use SSE(2) to denote the sum of squared errors for model 2 and SSE(1) to denote the sum of squared errors for model 1 (and similarly for MSE). Our test statistic is \(F = ((SSE(2) - SSE(1))/5)/MSE(1) = (10730/5)/437.1367 = 4.91\). This should be compared to the \(F_{5,40}\) distribution which yields a p-value of .0014. This being small suggests that we reject the null hypothesis; at least one of these variables is helping the regression. Note that even though each variable individually is not significant at the .05 level, the group is making a significant contribution. This can happen when the predictors are correlated.

2. **Heart disease cost regression in SAS**

(a) The intent of this part of the question was to reinforce that it is important to think about the relationships among your variables in advance. You should comment on each variable. In this case interv is the most highly correlated with logcost (around 0.7) so this is the most important predictor (more interventions/procedures means higher cost). Other medical variables were also significantly positively correlated with logcost including ervisits, drugs, complications, comorbidities and duration (correlations between .15 and .4). These correlations are weaker so the plots do not show as obvious a trend. Also, if you look carefully at the plots you can see that the variables most closely related to cost don’t seem to have a linear relationship with cost – the impact of the covariates on cost seems to slow down as the covariates increase.

(b) Most of the variables have coefficients that are highly significantly different than zero \((p < .0001)\) with the obvious signs (i.e., positive relationships for ervisits, interventions, complications, comorbidities, and duration). Age and gender are not significant predictors. The only surprise result was that drugs had a negative sign (it had a positive correlation) but the coefficient is not significantly different than zero so this is not a big issue.

(c) The normal probability plot seems fairly linear suggesting that the log transformation (of cost) leads to roughly Gaussian errors. The residual plot shows a non-linear pattern – perhaps a quadratic pattern. We can’t tell from here which variables are involved but it is fairly clear that there are primarily negative residuals for small and large predicted values and mainly positive residuals in the middle. Several people identified non-constant variance in this plot as well. I see some signs of that but it is not too dramatic – there is less variance on the right-hand-side of the plot but there are fewer points there as well so it may be that we just haven’t seen as many large residuals out there.

(d) Several of the medical variables have significant quadratic terms and significant linear terms (interv, dur, comorb). The quadratic terms make the regression a bit harder to interpret but they definitely lead to an improved regression. The residual plot looks much better (still a touch of non-linearity but much improved), \(R^2\) has increased from .59 to .67, and the \(\sigma_e\) has decreased from 1.23 to 1.085. The latter means prediction intervals are more precise.

(e) Remember to always start with a description of the scientific problem when writing up your results and to emphasize scientific results rather than statistical steps. For example:

   The aim of this study is to understand the key factors associated with the cost of treatment for heart disease based...
on a database of 788 patients gathered from the records of a large insurance company. Previous work suggested that the linear regression model with normal errors was more appropriate for analyzing logarithm of cost, so this is the response variable used here. Note that this impacts our interpretation of coefficients. Key medical variables include the number of emergency room visits, number of interventions performed, complications recorded, drugs ordered, comorbid conditions, and the duration of hospital stays. Gender and age are also available but did not seem to have a significant impact on cost. All of the medical variables seem to be related to cost although a multiple regression suggests that drugs is not a useful predictor after controlling for the other variables. An initial analysis suggested that the relationship was still non-linear, even after transforming the response. As a result a model was fit that included quadratic effects for each variable. Number of interventions, number of comorbidities, and duration of stay all demonstrate quadratic relationships with cost; cost increases as these variables increase but the quadratic coefficients are negative suggesting that the rate of increase slows with more interventions (or comorbidities or duration). Another significant finding is that even after transformation a considerable amount of patient-to-patient variation is unexplained. For a “typical” patient having average values on the predictors (female, age 60, with 5 interventions, 0 drugs, 3 er visits, 0 complications, 4 comorbidities, and 160 days of hospital stay) the predicted (median) cost is $880 and a 95% prediction interval ranges from $100 to $7500.

R program

```r
library(tidyverse)
datainput <- read_csv("H://Hal//Courses/Stat210//ischemic.csv")
hospcost <- data.frame(datainput)
hospcost$logcost <- log(hospcost$cost + 1)
# plots and correlations
plot(hospcost$age,hospcost$logcost,type="p")
plot(hospcost$gender,hospcost$logcost,type="p")
plot(hospcost$interv,hospcost$logcost,type="p")
plot(hospcost$drugs,hospcost$logcost,type="p")
plot(hospcost$ervisit,hospcost$logcost,type="p")
plot(hospcost$complic,hospcost$logcost,type="p")
plot(hospcost$comorb,hospcost$logcost,type="p")
plot(hospcost$dur,hospcost$logcost,type="p")
round(cor(hospcost[,2:11]),2)
# fit regression
hospcost$logcost_reg <- lm(logcost ~ age+gender+interv+drugs+ervisit+complic+comorb+dur, data=hospcost)
summary(hospcost$logcost_reg)
# obtain and save standardized residuals
hospcost$logcost_reg$stdres <- rstandard(hospcost$logcost_reg)
# check assumptions - residuals vs fitted values
ggplot() +
  geom_point(data=hospcost, mapping=aes(x=hospcost$logcost_reg$fitted.values, y=hospcost$logcost_reg$stdres))
# check assumptions - normality
qqnorm(hospcost$logcost_reg$stdres)
# create quadratics, then repeat regression and residual analysis
hospcost$age2 <- hospcost$age*hospcost$age
hospcost$interv2 <- hospcost$interv*hospcost$interv
hospcost$drugs2 <- hospcost$drugs*hospcost$drugs
hospcost$ervisit2 <- hospcost$ervisit*hospcost$ervisit
hospcost$complic2 <- hospcost$complic*hospcost$complic
hospcost$comorb2 <- hospcost$comorb*hospcost$comorb
hospcost$dur2 <- hospcost$dur*hospcost$dur
hospcost$logcost_quadreg <- lm(logcost ~ age + age2 + gender + interv + interv2 +
  drugs + drugs2 + ervisit + ervisit2 +
  complic + complic2 + comorb + comorb2 + dur + dur2, data=hospcost)
summary(hospcost$logcost_quadreg)
hospcost$logcost_quadreg$stdres <- rstandard(hospcost$logcost_quadreg)

3. Basic multivariate regression theory

(a) $\beta_2 = \text{change in the expected value of } Y \text{ for a one-unit change in } X_2 \text{ with all other variables being held fixed.}$

(b) The simple correlation coefficient does not tell you what sign the coefficient of the variable will have in a multiple regression. The multiple regression coefficient is also effected by the correlation of $X_3$ with other variables, and the correlation of those variables with $Y$. 
(c) \( H \) matrix

i. \( H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}X^T = H \)

ii. \( HX = X(X^TX)^{-1}X^TX = X. \) Recall that \( H \) finds the regression coefficients \( \beta \) such that \( \beta^\top X \) is as close as possible to \( \beta. \) If we apply

\( H \) to any column in \( X \) then the closest point in the space spanned by \( X \) provides a perfect match (just choose

regression coefficient 1 for the column in question and 0 for the other columns).

4. Regression theory

(a) First note that \( W = (X^TX)^{-1}X^T \) and that the coefficient of \( Z \) will be obtained by seeing what happens when the bottom row of \( W^{-1} \) is multiplied by \( (X^TY) \). To apply the hint, we first identify \( D = Z^T \) s.c., \( C = Z^T X, B = X^T Z, \) and \( A = X^T X \) and then calculate \( E = Z^T Z - Z^T X(X^T X)^{-1}X^T Z = Z^T(I - H_x)Z \) where \( H_x = X(X^TX)^{-1}X^T. \)

Finally \( \hat{\beta}_z = -E^{-1}CA^{-1}X^TY + E^{-1}Z^TY = \) \(-[Z^T(I - H_x)Z]^{-1}Z^TX(X^T X)^{-1}X^T Y + [Z^T(I - H_x)Z]^{-1}Z^TY = [Z^T(I - H_x)Z]^{-1}[Z^T(I - H_x)Y]. \)

(b) Applying the result for regression without an intercept provided in the question, \( \hat{\beta} = X^TY/(X^TX), \) with \( e_y \) in place of \( Y \) and \( e_x \) in place of \( X \) yields \( \hat{\beta}_x = e_x^\top e_x/e_x^\top e_x = [Z^T(I - H_x) Y]/[Z^T(I - H_x) Y(I - H_x) Z]. \)

It turns out that \( (I - H_x) Y(I - H_x) = (I - H_x) \) and therefore the slope identified here is the same as the slope identified in (a). This provides a useful interpretation of what happens when you add a variable to the regression.

The coefficient of the new variable is the same as would be obtained by regressing residuals of \( Y \) on \( X \) (what is not yet explained) on the residuals of \( Z \) on \( X \) (the new information that \( Z \) contains).

5. Interpreting coefficients

(a) Standardized coefficients

i. The usual regression coefficients are measured in units of \( Y \) divided by units of \( X \) (e.g., insurance policies per additional fire per 1000 housing units). The s.d.of \( X \) is in the same units as \( X \) and the s.d.of \( Y \) is in the same units as \( Y \) thus in forming the standardized coefficient all of the units cancel and that is why it is a unitless quantity.

ii. The standardized coefficient is the least squares regression coefficient multiplied by the s.d. of the predictor and divided by the s.d. of the response. For these predictors we get: pctmin .0091*.32.59/.63 = .47; fires .57; thefts -.36; pctold .30; income .09. Thus fires has the largest standardized coefficient (as well as the most significant t-statistic).

iii. In a simple linear regression the standardized slope is equal to the simple Pearson correlation of \( X \) and \( Y \).

This can help us interpret standardized coefficients – they are something like a correlation (but not precisely so because of the impact of other variables).

(b) The positive simple correlation reflects the estimated negative direct effect as well as some positive indirect effects. If you apply the formula given in class, then you can get some insight into the indirect effects. Specifically, we find thefts is highly positively correlated with fires and pctmin which both have important positive effects. Here I’d argue that the positive correlation matches our intuition more than the negative coefficient – the latter is tricky to interpret because it involves imagining that we can keep all variables the same while increasing thefts (but of course its not really possible to do this).

6. Case diagnostics

(a) Case 1 - The residual is easiest to obtain from \( Y_1 - \hat{Y}_1 = 1088 - 1057 = 31. \) This is not the most precise answer because the predicted value is rounded off but it is sufficient. Another option is to use \( e_i = r_i \sqrt{MSE(1 - \hat{h}_{ii})} = 1.250 \times \sqrt{694} \times (1 - .1161) = 30.96. \)

Case 2 - The standardized residuals is \( r_i = e_i/\sqrt{MSE(1 - \hat{h}_{ii})} = 37.3739/\sqrt{694} \times (1 - .1693) = 1.56. \)

Case 3 - The externally studentized residual is \( t_i = e_i \sqrt{n - (k+1)} \times [SSE(1 - \hat{h}_{ii}) - e_i^2]. \)

Here \( t_i = 26.2569 \times \sqrt{29842+4.42 - 26.2569^2} = 1.057. \)

Case 4 - Leverage can be easily found from the standardized residual formula \( r_i = e_i/\sqrt{MSE(1 - \hat{h}_{ii})}. \) This yields \( (1 - \hat{h}_{ii}) = 24.0234^2/.928^2(694) = 9656 \) so the leverage is .0344.

(b) The average leverage is \( (k+1)/n = .14. \) We can use two or three times the average leverage as a criterion. Cases LA (.36) and AK (.58) are high leverage cases. These have unusual \( X \)'s and may have a big impact on the regression. A review of the data shows that LA is high on mean income (this surprises me) and low on percentage of public school students , while AK is high on income, high on percentage of public school students, and especially high on expenditures per student.
(c) The traditional terminology is that “outliers” refers to unusual values of the response for the given values of the predictors. It is most common to assess this by looking at standardized or externally studentized residuals. There are 3 studentized residuals out of 50 that are larger than 2 which is not terribly unusual. One of these (Alaska) is larger than 3 which is quite unusual. You can also use the externally studentized residuals which tell the same basic story (except one more inches above 2). Note that using the Bonferroni correction for examining 50 externally studentized residuals suggests a critical value of 3.5 (which none of the points surpass).

(d) Not surprisingly Alaska shows up as being influential. It has high leverage and a large residual. The single best measure of influence is Cook’s Distance. It would certainly be worthwhile to re-run the regression without Alaska and see how things change (I did this – the same variables remain significant but coefficients change quite a bit with rank going from 8.47 to 9.78, expend going from 2.2 to 3.7 and years going from 22.6 to 16.5). One possible resolution is to argue that Alaska is a special case and therefore that we should build a model only for the continental U.S. (likely removing Hawaii also).