1. Q
   1. MPG on Horsepower
      1. Since the p-values of both coefficients are sufficiently small we may reject the null hypothesis for this single regression. This is further supported by the fact that the F-statistic given is about 600, extremely far from the value of 1 which would support the null hypothesis.
      2. This regression has a multiple value of 0.606, indicating that about 61% of the variability in MPG is explained when Horsepower is included. Therefore, it would be safe to conclude that the relationship between these two variables is moderately strong.
      3. The relationship appears to be slightly negative, as the estimated linear term for horsepower is around -0.158, which means that as horsepower increases by 1 unit, it is expected that mpg decreases by -0.158.
      4. The predicted mpg associated with a horsepower of 98 is 24.46708, with the associated 95% confidence interval being 23.97308 to 24.98108.
   2. Plot
   3. The residual plot for the linear model displays a significant u-shape, indicative of non-linearity between the predictor and response, meaning that a linear model would likely be an inaccurate fit for the data. Points 116 and 84 of this data set have both extremely high leverage and relatively high studentized residuals, indicating that the model may be out of shape due to points such as 116 and 84.
2. Q
   1. Scatterplot Matrix
   2. Correlation Matrix
   3. Multiple Linear Regression
      1. Some of the predictors appear to be statistically significant given their p-values and corresponding estimates. Furthermore, the multiple value of this regression is 0.81, meaning that it explains 20% more of the variability in the model than the model with just horsepower as a predictor does.
      2. The predictors which could be considered statistically significant are the intercept, weight, year, and origin.
      3. The coefficient for the year variable is 0.750773, indicating that as year increases, mpg increases. Therefore, we may conclude that more modern cars have increased mileage efficiency.
   4. Once again, the residual plot of this multiple linear regression showcases a still distinct u-pattern, indicating a high chance of non-linearity. Points 327, 326, and 394 all appear to be outliers, considering that they each have studentized residuals which are greater than 3. As well, point 14 looks to be a high leverage point, as its leverage about twice the distance from the y-axis as that of all other data points.
   5. The interactions appearing to be statistically significant are: cylinders:horsepower, displacement:horsepower, displacement:year, horsepower:weight, horsepower:acceleration, horsepower:year, and weight:year.
   6. Both and result in a more accurate model that is better fitted due to its non-linearity, as evidenced by the lack of patterns in the residual plots of them both. Furthermore, the resulting residual sums of squares of both transformed models are significantly lower than that of the original, first-degree polynomial regression. Performing a transformation on all of the variables seems to have increased the value by a decent amount, from 0.81 to 0.845, and its residual plot also appears to have less discernible patterns than the original. In all 3 of these transformations, it should be noted that the severity of leverage points was reduced.
3. Q
   1. Fit model to predict Sales with Price, Urban, US
   2. Coefficients
      1. Price: How much sales increases based on the price
      2. Urban: How sales increases depending on whether or not the item is sold in a city
      3. US: How sales increases depending on whether or not the item is sold in the US
      4. Intercept: Sales if the price is 0 and the item is sold neither in a city nor in the US
   3. Formula
      1. Not Urban and Not US:
      2. Urban and Not US: :
      3. Not Urban and US:
      4. Urban and US:
   4. We may reject the null hypothesis for the intercept, price, and (subjectively) US. The p-value of urban is extremely high, 0.936, indicating a high chance that the null hypothesis is true.
   5. Fit using the significant predictors
   6. Neither of the models in a) nor in e) fit the data well, as both have a multiple R-squared value of around 0.2393, which is only about 24% of the variability in the relationship explained. This is considered very low for a field such as marketing and sales.
   7. Confidence intervals
      1. Intercept: 11.79032020 to 14.27126531
      2. Price: -0.06475984 to -0.04419543
      3. US: 0.69151957 to 1.70776632
   8. Most of the data points seem to be fairly close together in terms of both residuals and studentized residuals. A few points reach the higher of both ends, around -3 and 3 in terms of studentized residuals, but overall neither outliers nor particularly high leverage points exist.
4. Q
   1. Regression of y on x: the coefficient estimate is 1.9939 for x, almost exactly what the true relationship between them is. The standard error is only 0.1065, the t-statistic is 18.73, and the p-value is enough to reject the null-hypothesis.
   2. Regression of x on y: the coefficient estimate for y is 0.39111, which is slightly farther from the actual relationship than in a). The standard error is even less, only 0.02089, the t-value is the same, 18.73, and the p-value is just as affirmative for the negation of the null hypothesis.
   3. The t-statistics of the lone variables in both of these regressions are equivalent.
   4. Algebraic proof of the -statistic being presentable in the given form is saved on phone. (to be added at some point)
   5. The t-statistics of both regressions are equal. This is because of the fact that, as in the final form, the only determining factors of the result are the and values, and the number of data points . Since the values and values are all the same in both regressions, as is the number of data points, it would be logical to conclude that the returned -statistics for both would be equivalent, which they are. It should be noted that all and values appear in pairs together, meaning that even if the response and predictor were swapped, the outputs would be equal as it would simply be swapping the positions of where the values went.
   6. With intercepts:
      1. : -stat of
      2. : -stat of
5. Q
   1. Only if the values for y and x are all exactly equal will the coefficient estimates of Y onto X be equivalent to that of X onto Y.
   2. Estimates
   3. Estimates
6. Q
   1. Feature
   2. Feature
   3. The length of vector y is the same as that of X and , or 100. and in this model.
   4. After plotting the relationship between and, it appears as if the line of best fit would simply be the line .
   5. The estimates for and appear to be within a 0.01 error radius of the true values (-1 and 0.5 respectively).
   6. Display lines on plot
   7. There is no evidence that the quadratic term improves the model fit. The p-value of the quadratic term in the new model is 0.165, far too large to reject the null hypothesis. Furthermore, the model itself does not have an RSS that is low enough when compared to the original that can be considered and improvement.
   8. Less noise
   9. More noise
   10. Confidence Intervals
       1. Original
          1. : -1.1150804 to -0.9226122
          2. : 0.3925794 to 0.6063602
       2. Less Noise
          1. : -1.0570515 to -0.9256389
          2. : 0.4337114 to 0.5796757
       3. More Noise
          1. : -1.0999424 to -0.8185064
          2. : 0.3043238 to 0.6169242
   11. It is evident from the resulting confidence intervals that the larger the random error, or “noise”, the wider and more inaccurate the confidence intervals will be. This is to account for the increased uncertainty in the possible range for the true response value as a result of the increased random error terms.
7. Q
   1. The regression coefficients are 2, 2, and 0.3.
   2. The correlation between and is positive: as one increases, so does the other.
   3. Estimates
      1. Intercept: 2.1305
      2. The estimate of intercept is somewhat close to the true value, but the other two are significantly off. We cannot reject either null hypothesis, not B1=0 nor B2=0, as the p-values of both variables are simply too large.
   4. The p-value of when only is used as a predictor is significantly lower: it is subjectively possible to now reject the null hypothesis of .
   5. The same can be said of a regression with only : it is entirely possible to reject the null hypothesis for as well upon seeing its lone p-value.
   6. The results in c)-e) do not actually contradict each other despite what it seems at first glance. While we cannot reject the null hypothesis for either in c), but we can reject them each when regressing on them individually, this is entirely possible due to the issue of collinearity. is entirely dependent on , meaning that they are closely related, and it is difficult to separate the effects of each individually on the response due to the increase in standard error resulting from their causal relationship. Therefore, while both and may be closely related to the response , regressing onto both at once masks both of their importance by significantly inflating the -value by factors of 200 or more!
   7. After adding an additional observation, it appears to be a very high leverage point , as well as a possible outlier, in the regression with both and . With just , it appears to be an outlier with a studentized residual of about 3.3. With just , it appears to be a high leverage point as well, as evidenced by the studentized residual vs. leverage plot. In particular, the -values of the variables in every single regression were increased upon adding this new point.
8. Q
   1. We may reject the null hypothesis for the following variables after performing systematic single-variable regression from crim onto each variable:
      1. Age
      2. Black
      3. Dis
      4. Indus
      5. Lstat
      6. Medv
      7. Nox
      8. Rad
      9. Tax
   2. In a multiple regression with all variables included, it would appear as if essentially none of the variables can have their null hypothesis rejected due to the high p-value, other than rad.
   3. Overall, the coefficient estimates in the multiple regression are far lower in absolute value than those in the single regressions, as evidenced by the resulting plot of the single regression values versus the multiple regression values.
   4. The following variables show evidence of non-linearity through a polynomial regression being an improvement over a standard linear regression:
      1. Age
      2. Dis
      3. Indus
      4. Medv (strong)
      5. Nox
      6. Tax