# Section 1: Questions 1-5

Download the Auto data set, from the course Blackboard page to answer questions 1 - 5.

1. Which of the predictors are quantitative, and which are qualitative? [Hint: Use the codes: Auto <- read.csv("Auto.csv", na.strings="?"), summary(Auto)]

Quantitative predictors:

- mpg
- cylinders
- displacement
- horsepower
- weight
- acceleration
- year
- origin

Qualitative predictors:

- name

Note:

- "origin" is essentially a qualitative predictor that has been coded as a numeric value.

2. What is the range and median of each quantitative predictor? [Hint: Use the codes: require(psych), describe(Auto)]

| | Variable | Min | Max | Median |
|---|---|---|---|---|
| 1 | mpg | 9 | 46.6 | 23 |
| 2 | cylinders | 3 | 8 | 4 |
| 3 | displacement | 68 | 455 | 146 |
| 4 | horsepower | 46 | 230 | 93.5 |
| 5 | weight | 1613 | 5140 | 2800 |
| 6 | acceleration | 8 | 24.8 | 15.5 |
| 7 | year | 70 | 82 | 76 |
| 8 | origin | 1 | 3 | 1 |

3. What is the mean and standard deviation of each quantitative predictor?

| | Variable | Mean | Std |
|---|---|---|---|
| 1 | mpg | 23.5158690176322 | 7.82580392894656 |
| 2 | cylinders | 5.45843828715365 | 1.70157698079185 |
| 3 | displacement | 193.53274559194 | 104.37958329993 |

| | | | |
|---|---|---|---|
| 4 | horsepower | 104.469387755102 | 38.4911599328285 |
| 5 | weight | 2970.26196473552 | 847.904119489725 |
| 6 | acceleration | 15.5556675062972 | 2.74999529297615 |
| 7 | year | 75.9949622166247 | 3.69000490146168 |
| 8 | origin | 1.57430730478589 | 0.802549495797039 |

4. Remove the 25th through 115th observations. What is the range, median, mean, and standard deviation of each predictor in the subset of the data that remains?

| | Variable | Min | Max | Median | Mean | Std |
|---|---|---|---|---|---|---|
| 1 | mpg | 11 | 46.6 | 24.4 | 25.0483660130719 | 7.67368169550678 |
| 2 | cylinders | 3 | 8 | 4 | 5.27124183006536 | 1.60989453692315 |
| 3 | displacement | 68 | 455 | 140 | 180.37908496732 | 95.9156396099349 |
| 4 | horsepower | 46 | 230 | 90 | 98.7880794701987 | 34.402466355131 |
| 5 | weight | 1649 | 4699 | 2722.5 | 2867.70261437908 | 755.640837439492 |
| 6 | acceleration | 8 | 24.8 | 15.65 | 15.7388888888889 | 2.74549241438579 |
| 7 | year | 70 | 82 | 77.5 | 77.2091503267974 | 3.31247734496922 |
| 8 | origin | 1 | 3 | 1 | 1.63071895424837 | 0.820378245089793 |

5. Suppose that we wish to predict gas mileage (mpg) on the basis of other variables. Using the full data set which variables do you believe will be useful in predicting mpg? Explain your answer using plots and correlation coefficients of the data.

The order of usefulness of predicting "mpg" is:

1. weight
2. year
3. origin
4. displacement
5. horsepower
6. cylinders
7. acceleration

This is based on the forward selection stepwise process. By definition, the most predictive variable is added into the model with each iteration, thus this order is indicative of the predictive value.

However, generating the full linear model and checking variable significance, we see that only the following variables have a significant relationship with "mpg":

- displacement
- weight
- year
- origin

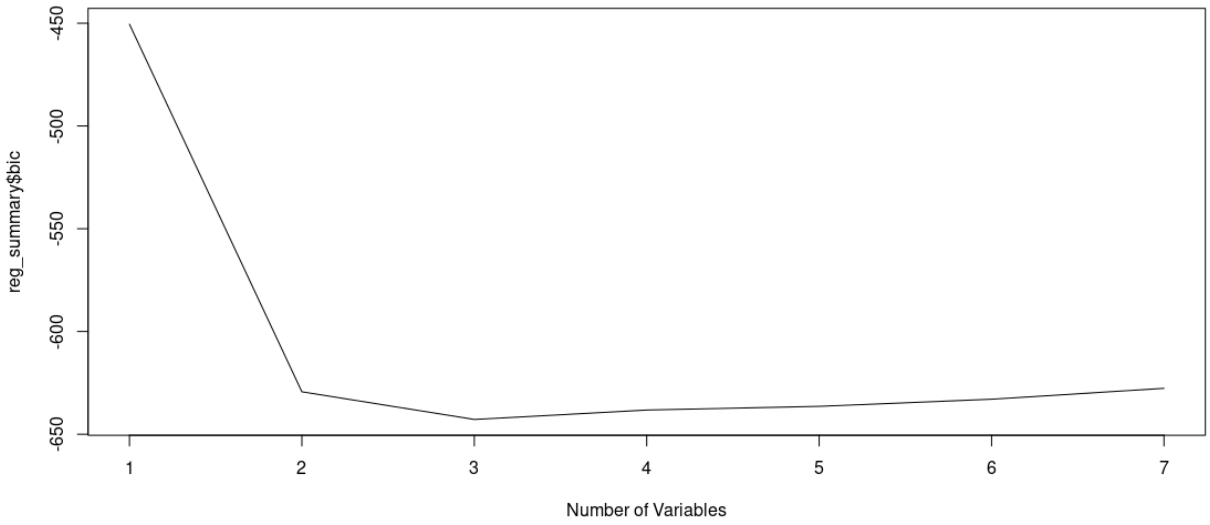
Call:
lm(formula = mpg ~ ., data = Auto3)

Residuals:
| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -9.5903 | -2.1565 | -0.1169 | 1.8690 | 13.0604 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -17.218435 | 4.644294 | -3.707 | 0.00024 *** |
| cylinders | -0.493376 | 0.323282 | -1.526 | 0.12780 |
| displacement | 0.019896 | 0.007515 | 2.647 | 0.00844 ** |
| horsepower | -0.016951 | 0.013787 | -1.230 | 0.21963 |
| weight | -0.006474 | 0.000652 | -9.929 | < 2e-16 *** |
| acceleration | 0.080576 | 0.098845 | 0.815 | 0.41548 |
| year | 0.750773 | 0.050973 | 14.729 | < 2e-16 *** |
| origin | 1.426141 | 0.278136 | 5.127 | 4.67e-07 *** |

Finally, examining the marginal value gained from each additional predictor, we generate the plot:




This indicates that the best (most generalizable) model, likely only uses two or three predictors, specifically:

- weight
- year
- origin

# Section 2: Questions 6-8

GPA's (Grade Point Averages) for 16 graduating MBA students, and their GMAT scores taken before entering the MBA program are given below. Use this data to respond to questions 6-8.

6. Create a linear regression model that uses GMAT scores as a predictor of GPA. Obtain and interpret the coefficient of determination $R^2$.

The R^2 value for the model is 0.0818, indicating that the model explains approximately 8% of the variance observed within gpa. It is worth being skeptical of any conclusions based on this observation, since the model does not show a significant relationship between GPA and GMAT.

Call:
lm(formula = gpa ~ gmat, data = data6)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -1.07025 | -0.21217 | 0.04427 | 0.35855 | 0.54359 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.571451 | 1.588509 | 0.989 | 0.339 |
| gmat | 0.003048 | 0.002729 | 1.117 | 0.283 |

Residual standard error: 0.4707 on 14 degrees of freedom
Multiple R-squared: 0.0818, Adjusted R-squared: 0.01622
F-statistic: 1.247 on 1 and 14 DF, p-value: 0.2829

7. Calculate the fitted value for the fourth student on the list (GMAT = 580).

The predicted GPA for a student with a GMAT score of 580 is 3.339291.

8. Test whether GMAT is an important variable using a significance level of 0.05.

GMAT is not significant at the level of 0.05, as seen from the results of the model summary (shown in Q6).

# Section 3: Questions 9-10

Use the rnorm() function to create a vector of 150 observations drawn from a N(0,1) distribution (call this vector x), and another vector of 150 observations drawn from a N(0, 0.2) distribution (call this vector Error). Use these to create a vector *y* according to the model *Y = -1.5 + 0.8 X + Error.* Consider this data to answer questions 9 and 10.

9. Fit a least squares linear model to predict y using x. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to the actual values of $\beta_0$ and $\beta_1$ ?

The estimated value for B0_hat is -1.48. The estimated value for B1_hat is 0.81. These values are very close approximations to the true values of -1.5 and 0.8.

Call:
lm(formula = y ~ x)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -0.60818 | -0.13057 | -0.02169 | 0.13646 | 0.53198 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.48218 | 0.01691 | -87.67 | <2e-16 *** |
| x | 0.81574 | 0.01891 | 43.13 | <2e-16 *** |

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2067 on 148 degrees of freedom
Multiple R-squared: 0.9263, Adjusted R-squared: 0.9258
F-statistic: 1860 on 1 and 148 DF, p-value: < 2.2e-16

10. What are the 95% confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$ and what is the prediction interval for a case with x = 1?

The confidence intervals for B0_hat and B1_hat are:
> confint(lm9)

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -1.5155923 | -1.4487730 |
| x | 0.7783667 | 0.8531157 |

The prediction interval for an observation of x=1 is:

```
> predict(lm9, data.frame(x=1), interval="prediction")
        fit       lwr       upr
1 -0.6664415 -1.078096 -0.2547867
```