MATH HW2 Solutions

Problem 1:

1. In order to prove that αT ^β is an unbiased estimator of θ, it is necessary to show that the expected value of T ^β is equivalent to θ.

The Least Squares estimate, denoted by ^β, is obtained by minimizing the sum of squared residuals, and it can be expressed as the value of ^β as

$^\beta = (X'X)^{-1}(X'Y)$

By taking the expected value of both sides, we obtain the expression for the expected value of the Least Squares estimate

$E*^\beta = E[(X'X)^{-1}(X'Y)] = (X'X)-1(X'E[Y])$

Since the expected value of Y is β*X

$E[^\beta] = (X'X)^{-1}(X'\ \beta X) = \beta$

$E[\alpha T\ ^\beta] = (\alpha T)E[^\beta] = \alpha T\beta = \theta$

2. In order to demonstrate that the variance of αT ^β is not greater than the variance of cTy. we need to compare the variances of the two variables and show that the variance of αT ^β is less than or equal to the variance of cTy.
i.e Var[αT ^β] <= Var[cTy]

The variance of αT ^β =
$Var[\alpha T\ ^\beta] = \alpha T(X'X)^{-1}X'\ Var[E]\ (X(X'X)^{-1}(X')^{-1}\alpha$

Where Var[E] is an identity matrix.

The variance of cTy =

Var[cTy] = cTVar[y]c = cTX* Var[β]X'c

The Least Squares estimate ^β is referred to as the Best Linear Unbiased Estimator of the true regression coefficient, as it is both unbiased and has the smallest variance among all unbiased linear estimators.

$Var[\alpha T\ ^\beta] = \alpha T(X'X)^{-1}X'\ Var[E]\ (X(X'X)^{-1}(X')^{-1}\alpha. <= cTX$

Var[αT ^β] = Var[cTy]

we can conclude that the variance of αT ^β is less than or equal to the variance of cTy.

Problem 2:

1. The main difference between the two methods is that the first method forms a set of points such that there is 95% confidence that the predicted value ˆf(x0) is within that set, while the second method provides a 95% confidence interval for an arbitrary point. This reflects the distinction between a pointwise approach and a global confidence estimate. The pointwise approach involves estimating the variance of individual predictions.

$\sigma_0^2 = \mathrm{Var}(f\hat{}(x_0)|x_0)$

$= \mathrm{Var}(X_0^T \hat{}\beta|x_0)$

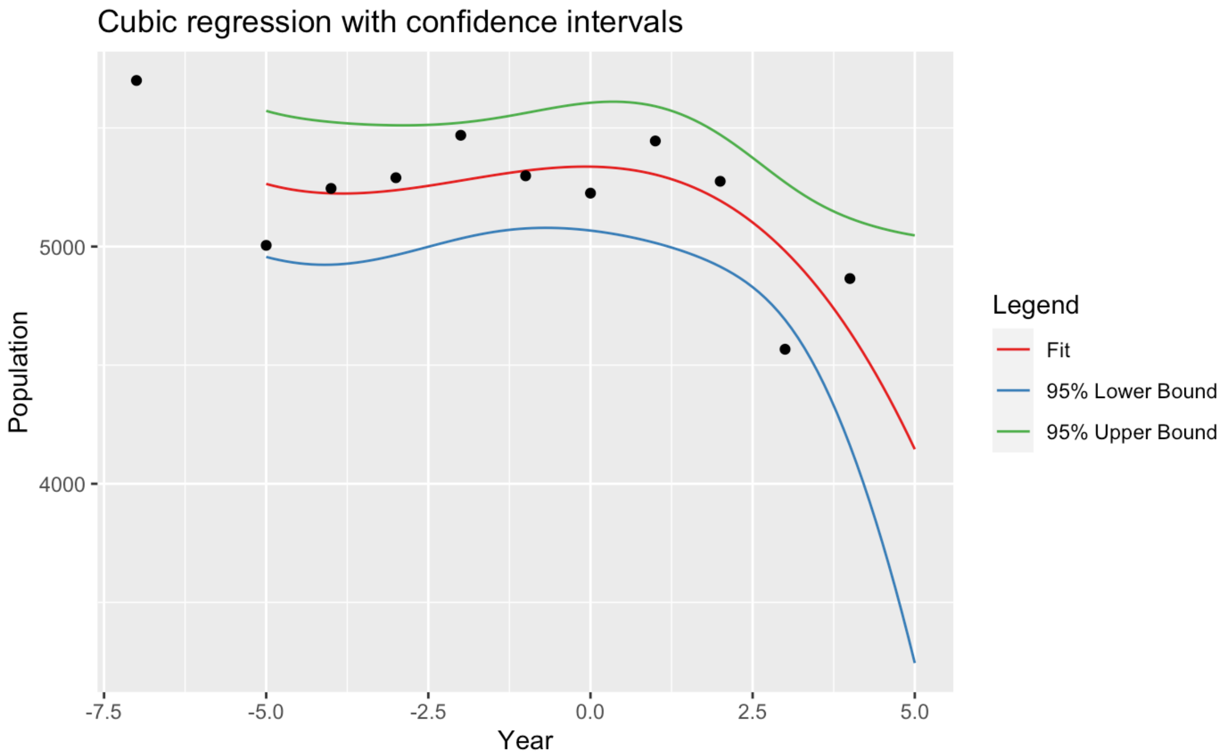
$= X_0^T \mathrm{Var}(\hat{}\beta)x_0$

$= \hat{}\ \sigma_0^2 X_0^T(X^TX)^{-1}x_0$

R code:

```
library(reshape2)
simulation.xs <- c(1949, 1950, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1939)
simulation.ys <- c(4567, 4865, 5005, 5245, 5290, 5469, 5298, 5225, 5445, 5275, 5700)
simulation.df <- data.frame(pop = simulation.ys, year = simulation.xs)

# Rescale years
simulation.df$year <- simulation.df$year - 1946

# Generate regression, construct confidence intervals
fit <- lm(pop ~ year + I(year^2) + I(year^3), data = simulation.df)
xs <- seq(-5, 5, 0.1)
fit.confidence <- predict(fit, data.frame(year = xs), interval = "confidence", level = 0.95)

# Create data frame containing variables of interest
df <- as.data.frame(fit.confidence)
df$year <- xs
df <- melt(df, id.vars = "year")

# Create the plot
p <- ggplot() +
  geom_line(aes(x = year, y = value, colour = variable), df) +
  geom_point(aes(x = simulation.df$year, y = simulation.df$pop)) +
  scale_x_continuous("Year") +
  scale_y_continuous("Population") +
  ggtitle("Cubic regression with confidence intervals") +
  scale_color_brewer(name = "Legend", labels = c("Fit", "95% Lower Bound", "95% Upper Bound"), palette = 
"Set1")

print(p)
```

result:



Cubic regression with confidence intervals
Population
5000
4000
-7.5
-5.0
-2.5
0.0
2.5
5.0
Year
Legend
Fit
95% Lower Bound
95% Upper Bound

Problem 3:

a. Best Subset (M)
In the orthogonal case, the matrix X^TX is an identity matrix (I).
For the best subset (size M), we can write X as X = X.I.

The estimator $\hat{}\beta$ is given by $\hat{}\beta = (XY)(X^TX)\text{-}1 = XY$.

Using the concepts of QR decomposition, for each step q, we choose K such that:

$K = \text{argmax}(X^T{}_kY)$ where $q < k <= p$

which is equivalent to $K = \text{argmax}(\hat{}\beta)$ where $q < k <= p$

The best subset with k(M) predictors gives the smallest residual sum of squares, which is equivalent to finding the largest M(k) coefficient.

$$r_j = (Y\text{-}\ xj\ \hat{}\beta_j)^T\ (Y\text{-}\ xj\ \hat{}\beta_j)$$

$$= Y^TY\ \text{-}2\hat{}\beta_j\ xj^Ty + \hat{}\beta_j{}^2$$

$$= Y^TY\ \text{-}2(xj^Ty)^2 + xj^Ty^2$$

$$= Y^TY\ \text{-}\ |\hat{}\beta_j|^2$$

Which can be minimized by having $|\hat{}\beta_j|$ as large as possible.

b. Ridge Regression

$\hat{}\beta = (X^TY)(XTX + \lambda I)^{-1}$
$\quad = (X^TY)(1/1+\lambda)$
$\quad = (\hat{}\beta)(1/1+\lambda)$

c. Lasso

$\min \frac{1}{2}\ |y\text{-}\ x\beta|_2{}^2 + \lambda|\beta|_1$

Taking derivative above equations and putting $\hat{}\beta\ != 0$
Which gives $-x_j^T\ (Y\text{-}x_j\hat{}\beta) + \text{sign}(\hat{}\beta)\ \lambda = 0$

$|\hat{}\beta|=1$ of $\hat{}\beta > 0$ else -1

$\hat{}\beta = X_j^TY - \text{sign}(\hat{}\beta)\ \lambda$

$\hat{}\beta = \hat{}\beta_j - \text{sign}(\hat{}\beta)\ \lambda$

At this point we have two scenarios
1. If $\text{sign}(\hat{}\beta) < 0$, then $\hat{}\beta_j + \lambda > 0$
Here, Here, lasso estimation is given by $\hat{}\beta = \hat{}\beta j - \lambda = \text{sign}(|\hat{}\beta j| + \lambda)(\hat{}\beta)$
2. If $\text{sign}(\hat{}\beta) > 0$, then $\hat{}\beta_j - \lambda > 0$
Here, lasso estimation is given by $\hat{}\beta = \hat{}\beta j - \lambda = \text{sign}(|\hat{}\beta j|\text{-}\ \lambda)(\hat{}\beta)$

Problem 4:

To obtain the least square estimate of the coefficient $\beta 1$, we can minimize the sum of squared residuals S, which is given by the equation:

$S = \sum(Yi - \beta 1Xi)^2$

By taking the derivative of S with respect to $\beta 1$ and setting it to zero, we can solve for $\beta 1$ as follows:

$dS/d\beta 1 = 2\sum -Xi(Yi - \beta 1Xi) = 0$

The resulting value of $\beta 1$ is given by:

$\beta 1 = \sum(XiYi)/\sum(Xi^2)$

To demonstrate that the vector $(Y - \hat{Y})$ is orthogonal to the vector X for the training set (X, Y), we start with the definition of orthogonality:

$(Y - \hat{Y})X = 0$

Expanding the dot product, we get:

$(Y - \beta 1X)\cdot X = Y\cdot X - \beta 1X\cdot X = Y\cdot X$

As $\beta 1$ is a constant, the last term simplifies to: $YX = 0$
This implies that the vector $(Y - \hat{Y})$ is orthogonal to the vector X.

Extra Credit:

To show that SSE/σ^2 is distributed as a chi-squared random variable with N-p-1 degrees of freedom, we first note that SSE can be written as:

$SSE = (Y - X\beta)^T(Y - X\beta)$

where Y is the $N \times 1$ vector of responses, X is the $N \times (p + 1)$ design matrix with the first column being all ones, and β is the $(p + 1) \times 1$ vector of parameters to be estimated. The hat over the Y indicates that it is the predicted value of Y from the model.
Expanding SSE, we get:

$SSE = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$

Taking the expectation of SSE, we have:
$E[SSE] = E[Y^TY] - 2\beta^TX^TE[Y] + \beta^TX^TX\beta$

Since $E[Y] = X\beta$, we can simplify this to:
$E[SSE] = E[Y^TY] - \beta^TX^TX\beta$

Now, since $Y \sim N(X\beta, \sigma^2I)$, we have:
$E[Y^TY] = E[(X\beta + \epsilon)^T(X\beta + \epsilon)] = E[\beta^TX^TX\beta] + \sigma^2N$

Substituting this into the above equation, we get:
$E[SSE] = \sigma^2N$

Next, we compute the covariance matrix of the residuals $e = Y - X\beta$:
$Cov(e) = E[ee^T] - E[e]E[e^T]$

Since $E[e] = 0$, this simplifies to:
$Cov(e) = E[ee^T]$

Now, we can write SSE/σ^2 as:
$SSE/\sigma^2 = e^Te/\sigma^2$

Taking the transpose of both sides, we have:
$(SSE/\sigma^2)^T = e^T(e^T)^T/\sigma^2 = e^Te/\sigma^2$

Therefore, SSE/σ^2 is a symmetric matrix, and we can use the spectral decomposition to show that it is distributed as a chi-squared random variable with N-p-1 degrees of freedom.

Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the eigenvalues of X^TX. Then, X^TX can be decomposed as $X^TX = Q\Lambda Q^T$, where Q is an orthonormal matrix whose columns are the eigenvectors of X^TX, and Λ is a diagonal matrix whose diagonal entries are the eigenvalues.

Now, consider the matrix Q^TY. Since Q is orthonormal, Q^TY has the same distribution as Y. Also, since Q is orthonormal, $QQ^T = I$, so we have:

$SSE = (Y - X\beta)^T(Y - X\beta) = (Q^TY - \Lambda\beta)^T(Q^TY - \Lambda\beta)$

Expanding this out, we get:
$SSE = Y^TY - 2\beta^T\Lambda^TQ^TY + \beta^T\Lambda^T\Lambda\beta$

$SSE = (Q^TY)^T(Q^TY) - 2\beta^T\Lambda^T(Q^TY) + \beta^T\Lambda^T\Lambda\beta$

Now, note that $Q^TY \sim N(0, \sigma^2 I)$ and $Q^TQ = I$. Therefore, the vector Q^TY has independent components that are normally distributed with mean 0 and variance σ^2. Since the distribution of Q^TY is invariant