Homework 3

P1

A

Show that $\text{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

We begin with our identities that are given:

$$\begin{aligned}
\hat{\beta} &= (X^TX)^{-1}X^Ty \\
&= (X^TX)^{-1}X^T(X\beta + \epsilon) \\
&= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon \\
&= \beta + (X^TX)^{-1}X^T\epsilon \\
Var(\hat{\beta}) &= Var(\beta + (X^TX)^{-1}X^T\epsilon) \\
&= [(X^TX)^{-1}X^T]Var(\epsilon)[(X^TX)^{-1}X^T]^T \\
&= (X^TX)^{-1}X^T(\sigma^2I)X((X^TX)^{-1})^T \\
&= \sigma^2(X^TX)^{-1}X^TX((X^TX)^{-1}) \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

B

Let rank(X) = r. Show that r $\leq$ min(n, p).

This is quite simple to show. We know that since X has n rows, with each row being a vector in $\mathbb{R}^p$, and that there cannot be more linearly independent vectors than the total number of vectors available. Thus, the rank of X must be less than or equal to n. Next, we have p columns, each column is a vector in $\mathbb{R}^n$ and again, the rank of X cannot exceed the total number of columns p. Thus,

$$\text{rank}(X) \leq \min(n, p)$$

C

When p > n, show that X^TX has (p-r) zero eigenvalues

Let rank(X) = r. Then we know then that the rank and nullity of (X^TX) is equal to that of X.

Furthermore, the rank-nullity theorem also states that for any p by p matrix M,

$$\text{rank}(M) + \text{nullity}(M) = p$$

Let $M = (X^T X)$. Then we find

$$r + \text{nullity}(X^T X) = p$$

$$\text{nullity}(X^T X) = p - r$$

Since $\text{nullity}(X^T X) = p - r$, we know that the null space has dimension $p - r$. This means that we can find $p - r$ linearly independent vectors such that each vector v can satisfy $(X^T X)v = 0$.

We are now get that there are at least $p - r$ linearly independent vectors that satisfy $(X^T X)v = 0 = 0 \cdot v$. This means we have $p - r$ eigenvectors with eigenvalue $\lambda = 0$.

D

> Show that the least squares estimate does not exist when p > n.

The least squares estimate requires use to compute $\hat{\beta} = (X^T X)^{-1} X^T y$. The main issue we are worried about here is inverting $(X^T X)$. This is because a matrix is only invertible only if it has full rank. In our case of $p > n$, we need $\text{rank}(X^T X) =$ p.

However, we know from step C that $rank(X^T X) = rank(X) = r \leq min(n, p)$. Since $n < p$, we know $rank(X^T X) \leq n < p$. Since the rank is not p, the matrix $(X^T X)$ is singular and we cannot solve for the inverse. Thus, we cannot find $\hat{\beta}$ and cannot make a least squares estimate.

E

> Take n = 70 and p = 100. Simulate X and from i.i.d. standard normals. Take to be a vector of all ones and set y = X + ϵ. Verify C. and D. above in R. The commands eigen() for calculating eigenvalues and lm() for tting linear models may be useful.

```r
# We setup parameters
set.seed(123)
n <- 70
p <- 100

# Simulate data
# X and epsilon from i.i.d. Standard Normals
X <- matrix(rnorm(n * p), nrow = n, ncol = p)
epsilon <- rnorm(n)
# Beta of only 1's
beta_true <- rep(1, p)

# Now we calculate y
y <- X %*% beta_true + epsilon


# C: Zero Eigenvalues
# Calculate X^T X (the Gram matrix)
XTX <- t(X) %*% X

# Calculate eigenvalues
```

```r
ev <- eigen(XTX)$values

# Count eigenvalues that are basically zero
# (using a small tolerance for numerical precision because its R)
zero_threshold <- 1e-10
num_zero_eigenvalues <- sum(abs(ev) < zero_threshold)

cat("Number of our zero eigenvalues:", num_zero_eigenvalues, "\n")
```

```
## Number of our zero eigenvalues: 30
```

```r
cat("Theoretical expected number of zero eigenvalues (p - n):", p - n, "\n")
```

```
## Theoretical expected number of zero eigenvalues (p - n): 30
```

```r
# D: OLS Existence/Uniqueness
model <- lm(y ~ X - 1)
# '- 1' removes the intercept because we don't have intercept in our simulation

# Check how many coefficients were successfully estimated
estimated_coefs <- sum(!is.na(coef(model))) # Count the non NA

cat("Number of coefficients estimated:", estimated_coefs, "\n")
```

```
## Number of coefficients estimated: 70
```

```r
cat("Number of coefficients that are NA:", sum(is.na(coef(model))), "\n")
```

```
## Number of coefficients that are NA: 30
```

We see our results. We expected 100-70=30 zero eigenvalues and the actual number is also 30.

We also expected that the least squares estimate for $\hat{\beta}$ to not exist, and we can see that this is again true as our $\hat{\beta}$ contains 30 NA values. R probably estimated by dropping the linearly dependent variables until it could compute the inverse and that is why we only have 70 estimated coefficients.