1. a)

$$\begin{aligned}
\|Q\beta - y\|^2 &= (Q\beta - y)^T(Q\beta - y) \\
&= (\beta^T Q^T - y^T)(Q\beta - y) \\
&= \beta^T Q^T Q\beta - y^T Q\beta - \beta^T Q^T y + y^T y \\
&= (\beta^T\beta - y^T Q\beta - \beta^T Q^T y + y^T QQ^T y) + (y^T y - y^T QQ^T y) \\
&= (\beta^T - y^T Q)(\beta + Q^T y) + (y^T y - y^T QQ^T y - y^T QQ^T y + y^T QQ^T y) \\
&= \|\beta - Q^T y\|^2 + (y^T y - y^T QQ^T y - y^T QQ^T y + y^T QQ^T QQ^T y) \\
&= \|\beta - Q^T y\|^2 + (y^T - y^T QQ^T)(y - QQ^T y) \\
&= \|\beta - Q^T y\|^2 + \|y - QQ^T y\|^2
\end{aligned}$$

b) Since Q^T is orthogonal,

$$RSS(\beta) = \|Y - X\beta\|^2 = \|Q^T Y - Q^T X\beta\|^2 = \|Q^T Y - RP^T\beta\|^2$$

Using the result from a, we have that

$$RSS(\beta) = \|QRP^T\beta - Y\|^2 = \|RP^T\beta - Q^T Y\|^2 + \|Y - QQ^T Y\|^2$$

Since $\|y - QQ^T y\|^2$ is fixed, the RSS will be at a minimum when $\|RP^T\beta - Q^T Y\|^2$ is minimized. If $R = \mathbf{0}$, then trivially all β achieve the same RSS. Otherwise, we want to minimize the non-trivial part of the expression. Where R_{11} has m rows, let Q_1 be the first m rows of Q with Q_2 the remainder. Thus,

$$\begin{aligned}
\|RP^T\beta - Q^T Y\|^2 &= \left\| \begin{bmatrix} R_{11} & R_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} P^T\beta - [Q_1^T Q_2^T]Y \right\|^2 \\
&= \| [R_{11}R_{12}] P^T\beta - Q_1^T Y\|^2 + \|Q_2^T Y\|^2
\end{aligned}$$

Since $\|Q_2^T Y_2\|^2$ is also fixed, we can reduce our minimization problem to minimizing

$$\|[R_{11}R_{12}]P^T\beta - Q_1^T Y\|^2$$

Thus, we will get a minimum when,

$$[R_{11}R_{12}] P^T\beta = Q_1^T Y$$

c) We want to solve the expression above. Let $W = [R_{11}R_{12}] P^T$. If we have some solution β, and a vector a in the null space of W, then

$$W(\beta + a) = W\beta + Wa = Q_1^T Y + 0$$

So $\beta + a$ is also a solution. Let β^* be the vector that achieves the smallest norm and solves the equation. Assuming R isn't all 0, then there is some β perpendicular to all a. If β_1 is and another β_2 isn't, by Cauchy-Schwartz

$$\begin{aligned}
\|\beta_2 + a\|^2 &\geq \|\beta_1 + a\|^2 \\
\|\beta_2\|^2 + \|a\|^2 &> \|\beta_1\|^2 + \|a\|^2 \\
\|\beta_2\|^2 &> \|\beta_1\|^2
\end{aligned}$$

Thus, β^* must be perpendicular to all a. If b is a vector,

$$0 = b^T W a = (W^T b)^T a$$

So the range of W^T is perpendicular to all a as well. Thus, since β^* is perpendicular to the null space of W, it must be in the range of W^T (This range fills the complement of the null space because taking the transpose of a matrix doesn't reduce its rank). Accordingly, for some v,

$$\begin{aligned} \beta^* &= W^T v \\ W\beta^* &= WW^T v \\ Q_1^T Y &= WW^T v \\ W^T(WW^T)^{-1}Q_1^T Y &= W^T v \\ W^T(WW^T)^{-1}Q_1^T Y &= \beta^* \end{aligned}$$

Thus we have our answer. We can invert WW^T since W is full row rank (since R_{11} is invertible) and similarly W^T is full column rank.

d) Here is the implemented R script:

```
################################

### Part d)
# Load the data and create the matrices
data('savings')
attach(savings)
n<-nrow(savings)
x<-as.matrix(data.frame(rep(1,n),savings[c(2,3,4,3,5)]))
y<-as.matrix(savings[1])
qr<-qr(x)

R<-qr.R(qr)[1:qr$rank,]
```

e) The two sets of coefficients are identical, except for the coefficients for pop75. In the normal regression, the coefficient was -1.6915, while in the solution to the singular version, the coefficients are both -.8457. $-.8457 + -.8457 = -1.6915$ because the total additive effect needs to be the same. For any a, $-1.6915a$ and $-1.6915(1-a)$ would have accomplished this, but since we insisted on the smallest norm, we got $a = .5$, which is the minimum of $(-1.6915a)^2 + (-1.6915(1-a))^2$.

f) It seems like a robust heuristic to compute the rank of R would be to count the number of entries on the diagonal which have a magnitude above a very small threshold (maybe $1e^{-10}$. Here, the diagonal for the second occurrence of pop75 is $-1.7e^{-15}$, despite the fact that variable is linearly dependent on the first occurrence of pop75 and should be 0. Thus, it seems like the QR has small rounding errors which we could exclude by this heuristic.
If R is not singular but extremely close, this suggests that a robust approach would be to exclude the variables which have an extremely small diagonal value in R, or, alternatively, treat them as singular and use the method developed in this problem to find β.

2. a) We know that when $A = BC$, $a_{ij} = b_i c_j$, where b_i and c_j are respectively the ith row and jth column of B and C. Further, if C is itself the product of matrices D and E, then $c_j = De_j$, where e_j is the jth column of D. Combining this, $a_{ij} = b_i De_j$. If we treat the ith and jth rows of X as column vectors on their own, this leads to the conclusion that

$$h_{ij} = x_i^T (X^T X)^{-1} x_j$$

Since H is symmetric and idempotent,

$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = \sum_{j=1}^{n} h_{ij}^2$$

Also, if $J = \{j : x_i = x_j\}$, then for all $j \in J$,

$$h_{ij} = x_j^T (X^T X)^{-1} x_i = x_i^T (X^T X)^{-1} x_i = h_{ii}$$

So, combining these,

$$\begin{aligned} h_{ii} &= \sum_{j=1}^{n} h_{ij}^2 \\ h_{ii} &= r h_{ii}^2 + \sum_{j \notin J} h_{ij}^2 \\ h_{ii} &\geq r h_{ii}^2 \\ \frac{h_{ii}}{r} &\geq h_{ii}^2 \\ \frac{1}{r} &\geq h_{ii} \end{aligned}$$

If X includes a column of 1s at the front (for an intercept), then $\mathbf{1}_n$ is in the span of H and $H\mathbf{1}_n = \mathbf{1}_n$. We may conclude that

$$\sum_{j=1}^{n} h_{ij} = 1$$

This would imply that, since $\sum_{j=1}^{n} h_{ij}^2$ is minimized when $h_{ij} = \frac{1}{n}$ for all h_{ij},

$$\sum_{j=1}^{n} h_{ij}^2 \geq \sum_{j=1}^{n} \frac{1}{n^2} = \frac{1}{n}$$

Combining this with what we showed earlier,

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 \geq \frac{1}{n}$$

Giving us the desired results that

$$\frac{1}{r} \geq h_{ii} \geq \frac{1}{n}$$

b) When X is an invertible matrix,

$$\sum_{i=1}^{n} h_{ii} = \mathrm{tr}(H) = \mathrm{rk}(H) = n$$

So $h_{ii} = 1 = \frac{1}{r}$, since $r = 1$ for all i. If X is just one row repeated n times, then $\mathrm{rk}(X) = 1$, and

$$\sum_{i=1}^{n} h_{ii} = \mathrm{tr}(H) = \mathrm{rk}(H) = 1$$

So it must be the case that $h_{ii} = 1 = \frac{1}{n}$

3. a) Under the normal assumption,

$$\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$$

So we can also assume

$$R\hat{\beta} \sim N_q(R\beta, \sigma^2 R (X^T X)^{-1} R^T)$$

In turn since, z is independent of Y, and thus independent of $\hat{\beta}$,

$$\begin{aligned} z &\sim N_q(R\beta, \sigma^2 \mathbb{I}_q) \\ \Rightarrow R\hat{\beta} - z &\sim N_q(R\beta - R\beta, \sigma^2 R(X^T X)^{-1} R^T + \sigma^2 \mathbb{I}_q) \\ \Rightarrow \frac{R\hat{\beta} - z}{\sigma} &\sim N_q(\mathbf{0}, (R(X^T X)^{-1} R^T + \mathbb{I}_q)) \end{aligned}$$

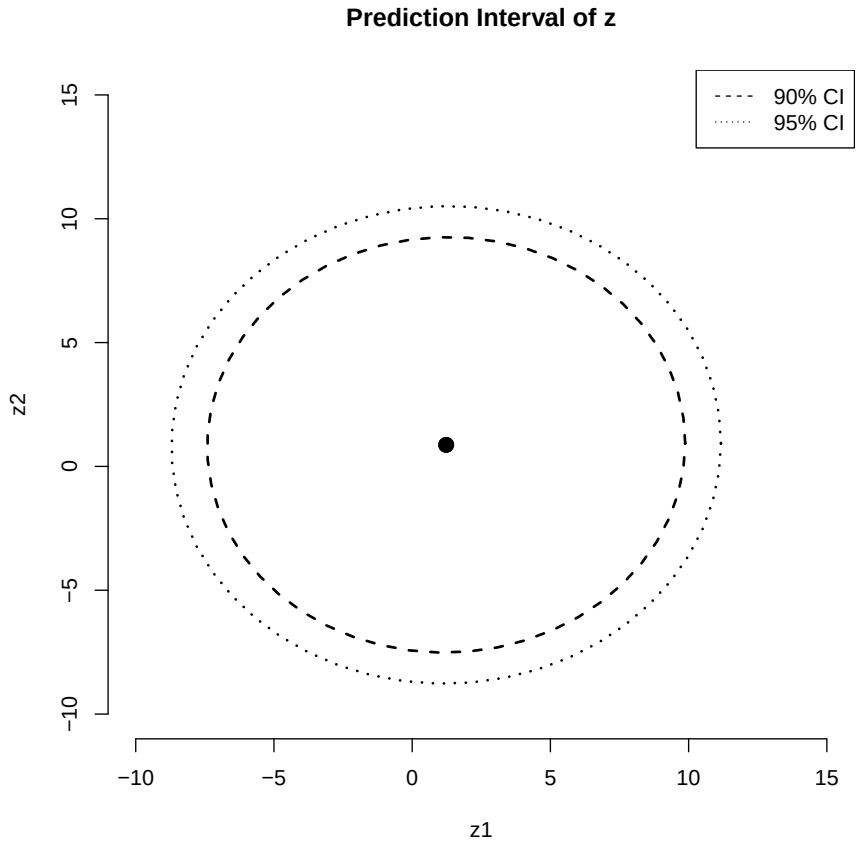
Accordingly, we would expect that

$$\frac{\frac{1}{q}(R\hat{\beta} - z)^T (R(X^T X)^{-1} R^T + \mathbb{I}_q)^{-1} (R\hat{\beta} - z)}{\hat{\sigma}^2} \sim F_{q,n-p}$$

So we would get a confidence set of

$$E_\alpha = \left\{ \beta \in \mathbb{R}^q : (R\hat{\beta} - z)^T (R(X^T X)^{-1} R^T + \mathbb{I}_q)^{-1} (R\hat{\beta} - z) \leq q\hat{\sigma}^2 f_{q,n-p,\alpha} \right\}$$

b) I have plotted the confidence intervals for where z will land:



You will notice that the confidence interval is very close to a circle. If we create just a confidence interval for $R\hat{\beta}$, it is a very elongated ellipse, with the main axis not parallel to the $z1$ or $z2$ axis. However, the noise of z around$R\beta$ has no covariance, and there is equal variance in the direction of $z1$ and $z2$, so if it we created a confidence interval for it conditioned on a $R\beta$, we would get a circle. Since the magnitude variance of z around $X\beta$ is much higher than the variance of the estimate of $R\hat{\beta}$, this makes the confidence interval close to a circle. Here is the R code I used:

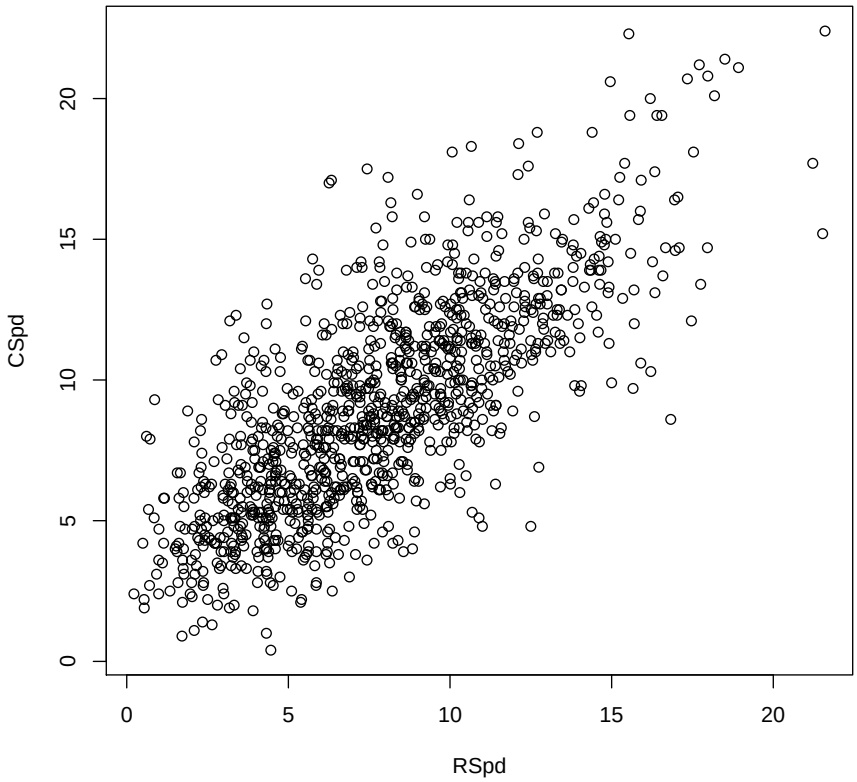
```
# Linear model and output
model<-lm(sr ~ pop15 + pop75 + dpi + ddpi)
X <- model.matrix(model)
beta<-model$coef
```

```
sigma2 <- summary(model)$sigma^2
df <- summary(model)$df[2]

# Set up for ellipse
R<-matrix(c(0,0,1,-1,-1,0,0,0,0,1),nrow=2,ncol=5)
variance.pred <- R %*% solve(t(X) %*% X) %*% t(R) + diag(2)
variance.mean <- R %*% solve(t(X) %*% X) %*% t(R)

mycenter <- R%*%beta
radius9<- sqrt(2* sigma2 * qf(.1,2,df, lower.tail=FALSE))
radius95<- sqrt(2* sigma2 * qf(.05,2,df, lower.tail=FALSE))

# Draw the ellipses
# Ellipse for 90% CI
pdf("hw4_3_b_ci.pdf")
plot.new()
plot.window(xlim=c(-10,15),ylim=c(-10,15))
title(main='Prediction Interval of z', xlab='z1', ylab='z2')
car::ellipse(center=c(mycenter), shape=variance.pred, radius=radius9, lty=2, mai
car::ellipse(center=c(mycenter), shape=variance.pred, radius=radius95,lty=3, add=
legend('topright', legend=c('90% CI','95% CI'), lty=c(2,3))
axis(1)
axis(2)
dev.off()
```

c) Having repeated the simulation 1000 times, z fell within the 90% confidence interval 91.3% of the time, which indicates that my expression for the confidence interval was pretty accurate. I calculated whether or not the simulated data point fell in the confidence interval using exactly the formulation from part a. I've attached the code:

```
n<-1000
predictions <- X %*% beta
inv.variance.pred <- solve(variance.pred)
in.range <- function(u) {
    m<-lm(predictions+rnorm(50,0,sqrt(sigma2)) ~ 0 + X)
    p<-mvrnorm(1,R%*%m$coef,summary(m)$sigma^2 * diag(2))
    e<-mycenter - p
    return(t(e) %*% inv.variance.pred %*% e <= radius9^2)
}
boots<-sapply(1:n,in.range)
boot.prob<-sum(boots)/n
boot.prob
```

4.2.13. 1. Looking over this plot, simple linear regression looks extremely plausible for this model. We see what looks like a linear relationship between the two variables with a multivariate normal noise distribution, which is exactly what we would want.



2. For the coefficients, we get

	Estimate	Std. Error	t value	Prob
Intercept	3.141	.120	18.52	<2e-16
RSpd	.756	.020	38.50	<2e-16

These coefficients are both quite meaningful, the t values indicating that there is a negligible probability of the estimates being positive by chance if each coefficient was actually 0. The most interesting is the coefficient on RSpd, which indicates that their is a strong correlation between the wind speed at the two sites. Additionally, we have a residual standard error on 2.466 degrees of freedom, an R-squared of .5709, and a F-statistic of 1482 on 1 and 1114 degrees of freedom, resulting in a p-value of $< 2e-16$. This speaks very strongly that this regression is meaningful; it explains 57% of the original variance, and the probability that we would see these results if we just had a mean model is negligible.

3. From the model, we get a prediction that when the wind at the reference site was 7.4285, we would see wind at the test site of $\hat{y}_0 = [1\ 7.4285]\beta = 8.7552$. We predict with 95% certainty, based on the normality assumption, that the actual test site wind speed would be in the range $\hat{y}_0 \pm t_{1114}^{.025}\hat{\sigma}\sqrt{1 + [1\ 7.4285]^T(X^TX)[1\ 7.4285]} = (3.914, 13.570)$.

4. Starting with the mean of the prediction,

$$\frac{1}{m}\sum_{i=1}^{m}\hat{y}_{*i}x = \frac{1}{m}\sum_{i=1}^{m}(x_{*i}\beta_1 + \beta_0)$$
$$\frac{1}{m}\sum_{i=1}^{m}\hat{y}_{*i}x = \hat{\beta}_0 + \frac{1}{m}\sum_{i=1}^{m}(x_{*i}\beta_1)$$
$$\frac{1}{m}\sum_{i=1}^{m}\hat{y}_{*i}x = \hat{\beta}_0 + \frac{\hat{\beta}_1}{m}\sum_{i=1}^{m}(x_{*i})$$
$$\frac{1}{m}\sum_{i=1}^{m}\hat{y}_{*i}x = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$$

So we conclude the prediction for the mean is equal to the mean of the predictions. If we call the prediction at $\bar{x}_*$ $\bar{y}_*$ and the expectation $\hat{y}_*$, then

$$\text{Var}(\bar{y}_*) = \text{Var}(\bar{y}_* - \hat{y}_* + \hat{y}_m)$$
$$\text{Var}(\bar{y}_*) = \text{Var}(\bar{y}_* - \hat{y}_*) + \text{Var}(\hat{y}_*) - 2\text{Cov}(\bar{y}_* - \hat{y}_*, \hat{y}_*)$$
$$\text{Var}(\bar{y}_*) = \frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{SXX}\right)$$
$$se\ \bar{y}_* = \sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{SXX}\right)}$$

We know the covariance term is equal to 0 since the residuals and the predictions are orthogonal, and thus have 0 covariance.

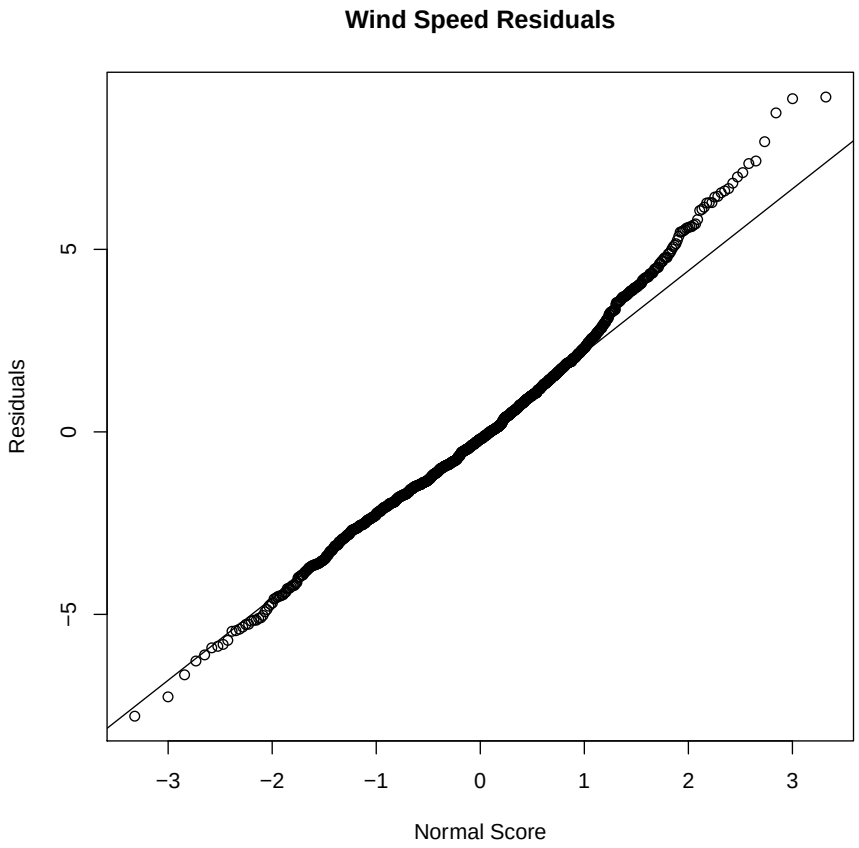
5. We can plug the values given and from the data set into the above equation to get the standard error. We get a $\sigma^2 = 6.082$ from the regression, $\bar{x} = 7.778$, and $SXX = 15785.56$ from the data. Thus, the standard error is:

$$se\ \bar{y}_* = \sqrt{\frac{6.082}{62039} + 6.082\left(\frac{1}{1116} + \frac{(7.429 - 7.778)^2}{15785.56}\right)} = .075$$

Combining this with the model's prediction of the expectation of $\bar{y}_* = 8.7552$ from earlier, we get a 95% confidence interval of

$$\bar{y}_* \pm se(\bar{y}_*)t^{.025}_{1114} = (8.609, 8.901)$$

4.6. When I bootstrapped the mean wind speed at the candidate site, I got a 95% of the means landing in $(8.800, 9.22)$. This is surprising, because its in general higher, and only partially overlaps with the analytic predicted confidence interval. So, I checked the assumption of normality:



Here, the qq plot shows that the distribution of the residuals has a thicker tail than the normal distribution. This supports the notion that there is more variability than we assumed, and in particular the upper bound of the analytic confidence interval is too low, supporting the bootstrapped confidence interval.