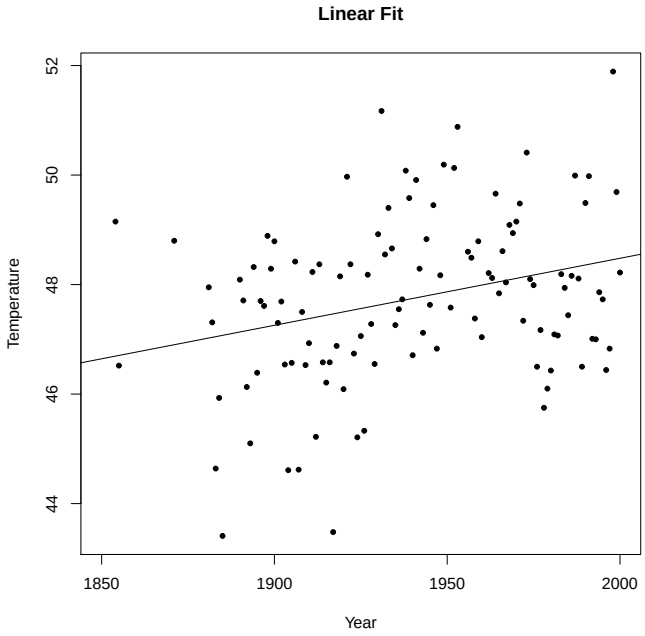
1. a) When we run a linear regression, we see get a statistically significant positive coefficient on year. This at least is a pretty good indicator for some kind of increasing trend. However, looking at the plot, it looks like there is some kind of concavity at least during much of the 20th century, which doesn't support their being a simple linear trend. So, I would conclude there is an upwards trend that probably isn't quite linear. The estimate of coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.006	7.311	3.284	0.001
year	0.012	0.004	3.247	0.002

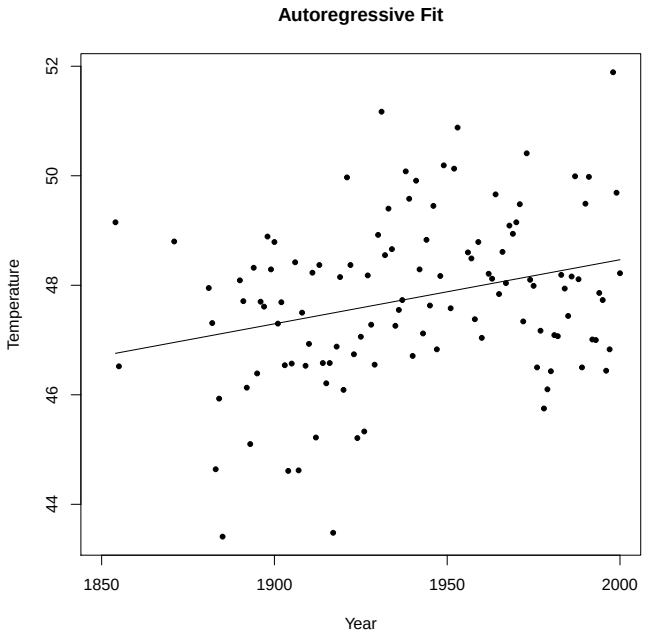


b) When we run a linear model that includes an AR(1) term, we find a statistically significant autoregressive term, with an estimate autocorrelation of .208. However, even with this, the estimates for the normal coefficients are pretty much the same as with the simple linear regression, with the coefficient on year still at .012 and only a small change in the intercept. The plot also shows how the plot barely changed. So, the autocorrelation doesn't change my opinion about the trend. The estimate of autocorrelation:

	lower	est.	upper
Phi1	0.012	0.208	0.389

The estimate of coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	25.049	8.759	2.860	0.005
year	0.012	0.005	2.592	0.011



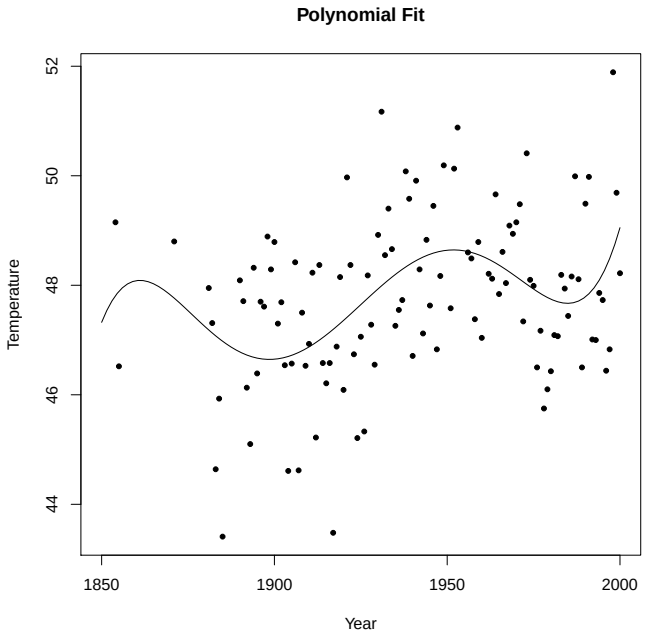
c) Starting with orthogonal polynomials of degree one through ten and backwards eliminating them with a critical value of .2, I we end up left with polynomials of degree one through five. The coefficients and plot are below. Looking at the plot, the model seems to fit quite well within the range of the data, but when we predict out to 2020, we get a prediction of 60.078, which is way outside of the range seen historically and thus not believable. This reflects the issues with polynomials only being a good representation of arbitrary functions where there is data to fit them.

Predicted value and range:

	fit	lwr	upr
1	60.078	49.841	70.315

Estimated coefficients:

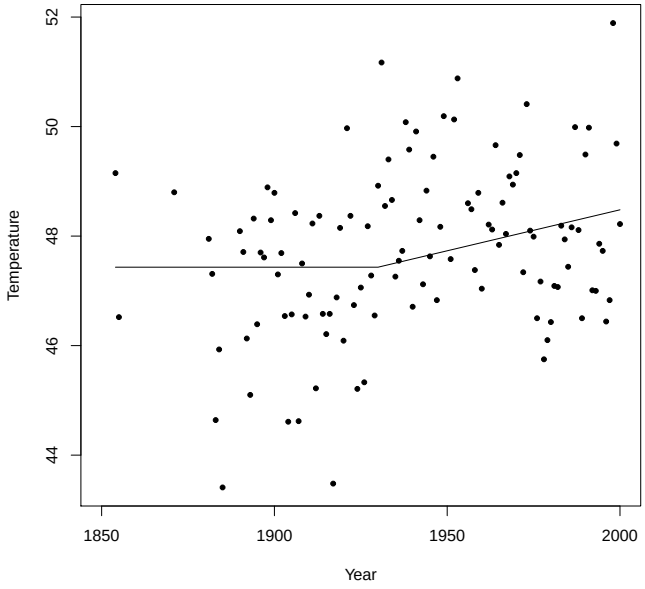
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.743	0.131	365.604	0.000
temp.poly.1	4.762	1.400	3.400	0.001
temp.poly.2	-0.907	1.400	-0.648	0.519
temp.poly.3	-3.313	1.400	-2.366	0.020
temp.poly.4	2.438	1.400	1.741	0.084
temp.poly.5	3.382	1.400	2.415	0.017



d) To test this trend, we fit a simple linear spline on the data, with an intercept and a continuous term for year which is 0 before 1930. Looking at the results, we get a statistically significant coefficient on year after 1930. However, this coefficient is pretty close to the coefficient on year we had in the simple linear model, and the fit, when plotted, looks like a less convincing fit than the original simple linear model, with the actual values not symmetrically distributed around the prediction in the flat section. We don't have a formal statistical test to reject this model, but it certainly doesn't seem like an improvement over the simple linear model.

Estimated coefficients:

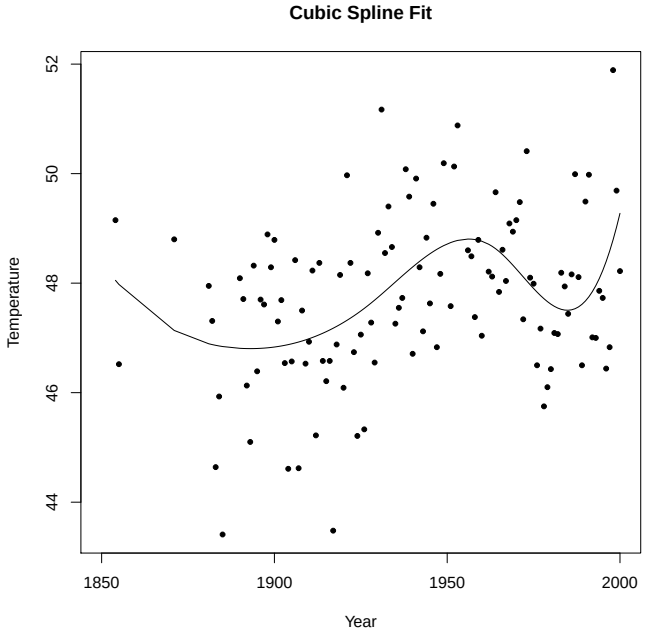
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.432	0.185	256.939	0.000
ifelse(year < 1930, 0, year - 1930)	0.015	0.006	2.556	0.012



e) Fitting a cubic spline model, we certainly seem to see an improvement over the simple linear model. It picks up the apparent concavity over most of the 20th century, and also seems to capture the behavior towards the boundary reasonably, besides possibly turning up too sharply at the high end of the domain. Its fit is fairly similar to the polynomial fit though, with the exception of not having an inflection point on the lower end of the domain. There isn't much data in that section though, so it looks more like the polynomial is over fitting than the cubic splines.

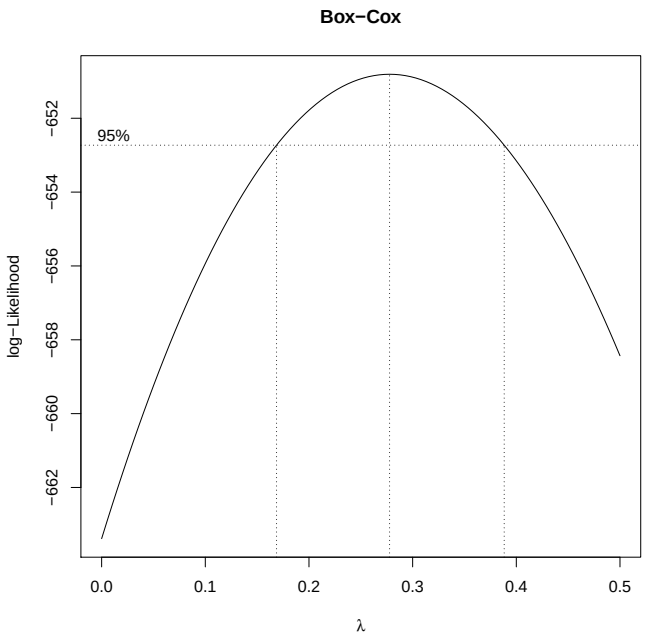
Estimated coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.275	0.704	69.954	0.000
spl1	-1.224	1.246	-0.982	0.328
spl2	-2.151	1.893	-1.136	0.259
spl3	-2.713	1.130	-2.402	0.018
spl4	-2.081	1.114	-1.868	0.064
spl5	0.910	0.907	1.003	0.318
spl6	-3.228	1.342	-2.405	0.018



2. a) After running the model on the untransformed data, we seem to see a substantial improvement from transformation. The log-likelihood for box-cox is maximized right around $\lambda = .25$, with a confidence interval that excluded 1, so I transformed the data using .25. The end result is an improvement in R^2 from .681 to .713. The untransformed linear model:

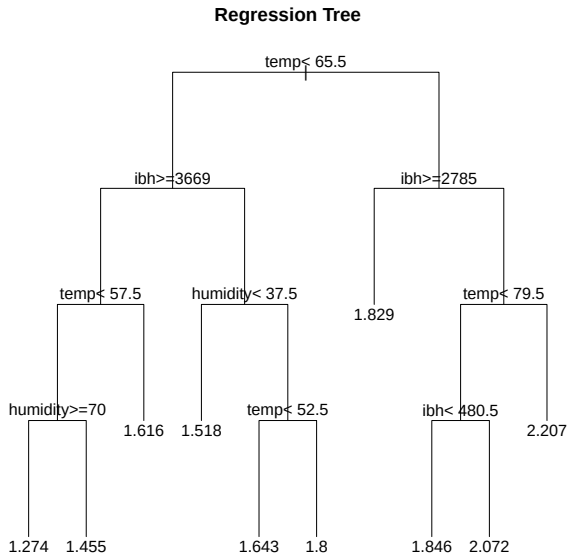
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.494	1.616	-6.492	0.000
temp	0.330	0.021	15.626	0.000
humidity	0.077	0.013	5.777	0.000
ibh	-0.001	0.000	-6.130	0.000



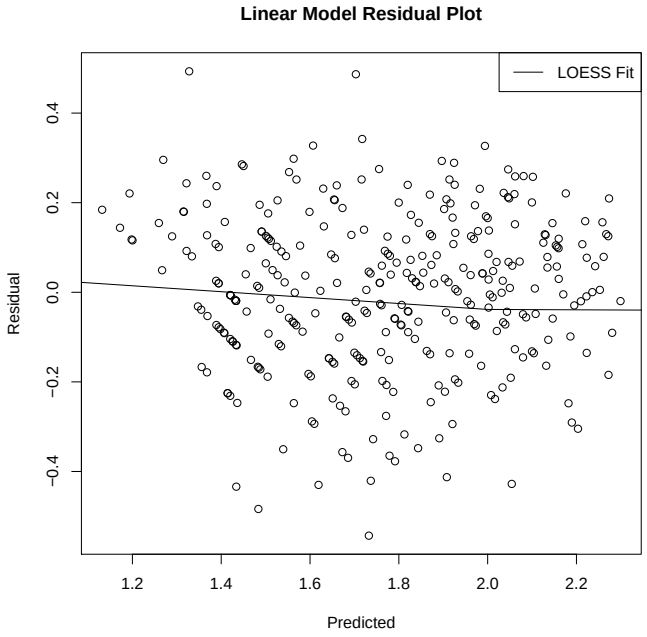
The transformed linear model:

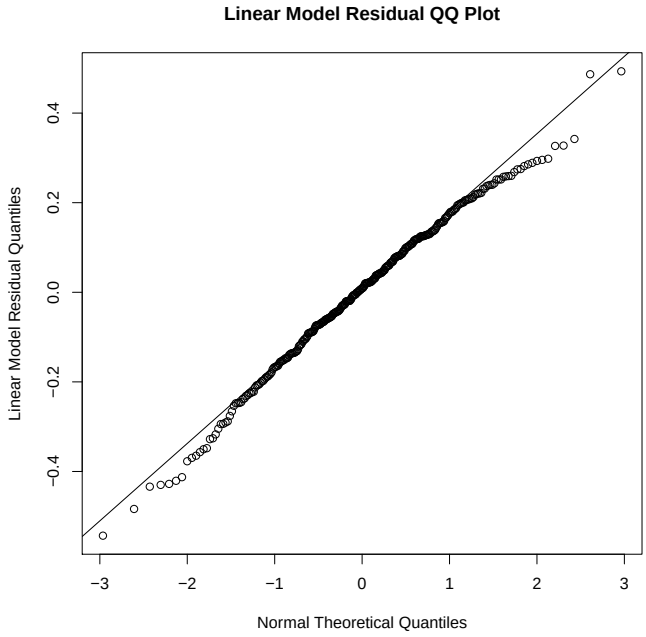
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.914	0.062	14.774	0.000
temp	0.013	0.001	16.192	0.000
humidity	0.003	0.001	5.810	0.000
ibh	-0.000	0.000	-7.734	0.000

b) Comparing the linear fit to the regression tree and the random forest, it appears that the latter two fit slightly better than the former. As we look at the residuals, both the regression tree and the random forest are nearly perfectly distributed around 0 through out the predicted range (for the random tree this is of course by construction), while the residuals for the linear model do not quite seem to be at the lower end of the predicted range. As well, while the linear model residuals deviate from a normal distribution towards the upper end of their range, the regression tree and random forest have residuals that are normally distributed pretty well throughout.
This is the regression tree I fit:

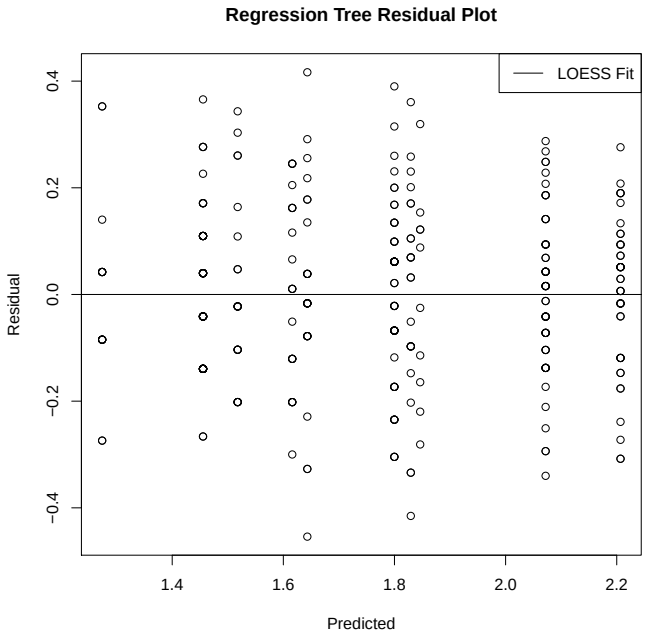


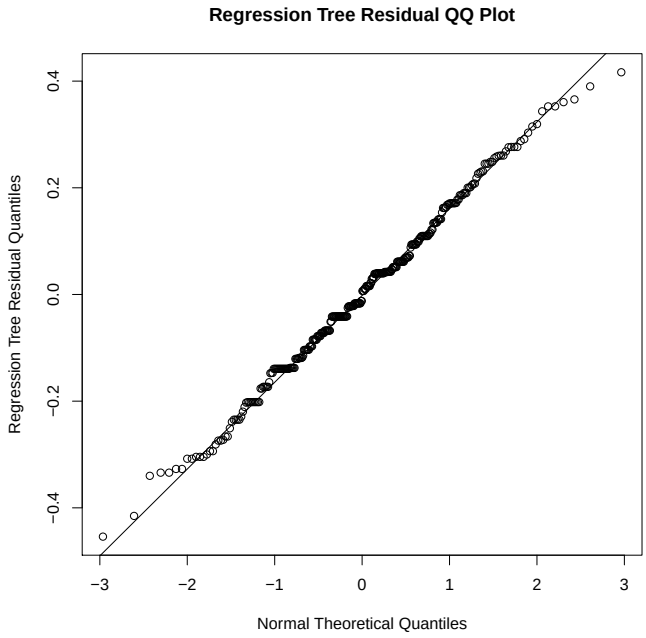
The transformed linear regression:



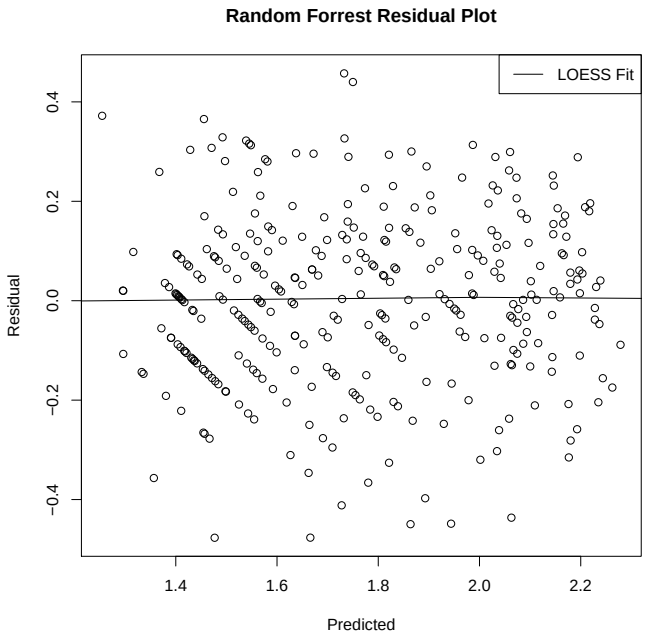


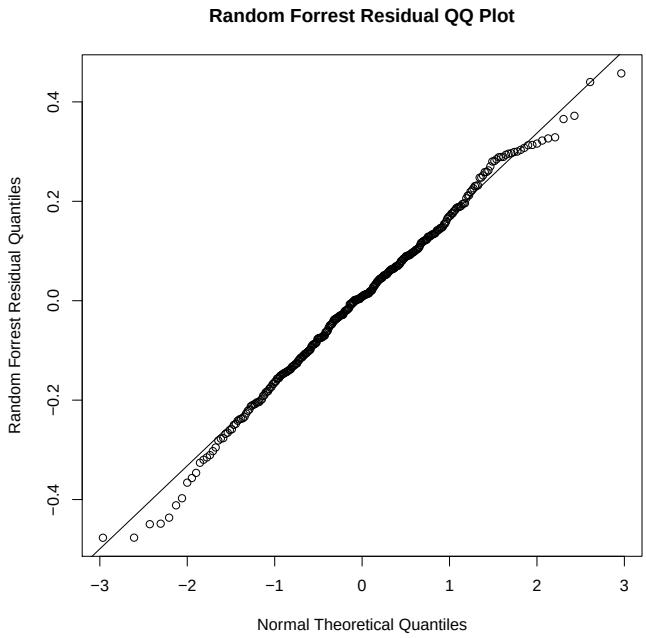
The regression tree:





The random forest:





3.

a) I applied backwards elimination, working with a critical value of .2. Doing this, I eliminated three variables and kept the rest. This was the order I eliminated variables in, and their p-value when I eliminated them:

	Order	p-value
gleason	1.000	0.775
lcp	2.000	0.251
pgg45	3.000	0.253

The end result was this regression:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.951	0.832	1.143	0.256
lcavol	0.566	0.075	7.583	0.000
lweight	0.424	0.167	2.539	0.013
age	-0.015	0.011	-1.385	0.170
lbph	0.112	0.058	1.927	0.057
svi	0.721	0.209	3.449	0.001

b) Going by forward selection, again with a critical value of .2, I ended up with the same model gotten previously by backwards selection. This was the order I included variables in, and their p-value when I included them:

	Order	p-value
lcavol	1.000	0.000
lweight	2.000	0.002
svi	3.000	0.002
lbph	4.000	0.112
age	5.000	0.170

The end result was this regression:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.951	0.832	1.143	0.256
lcavol	0.566	0.075	7.583	0.000
lweight	0.424	0.167	2.539	0.013
svi	0.721	0.209	3.449	0.001
lbph	0.112	0.058	1.927	0.057
age	-0.015	0.011	-1.385	0.170

4. As I think we had to show on the midterm,

$$
\begin{aligned}
H &= X(X^TX)^{-1}X^T \\
H &= QR(R^TQ^TQR)^{-1}R^TQ^T \\
H &= QR(R^TR)^{-1}R^TQ^T \\
H &= QRR^{-1}(R^T)^{-1}R^TQ^T \\
H &= QQ^T
\end{aligned}
$$

As was discussed with added variable plots, for an additional variable X_i added to a regression, if $\hat{e}_{(-i)}$ are the residuals from the regression excluding X_i, $\hat{e}$ are the residuals from the regression including it, and β_i is what the coefficient would be on X_i if it was included in the regression, then

$$
\begin{aligned}
\hat{e}_{(-i)} &= \hat{\beta}_i(I-H)X_i + \hat{e} \\
\hat{e}_{(-i)} &= \hat{\beta}_i(I-QQ^T)X_i + \hat{e}
\end{aligned}
$$

From simple linear regression, since $\hat{e}_{(-i)}$ has mean 0, it must be the case that $\hat{\beta}_i = \frac{\text{Cov}((I-QQ^T)X_i, \hat{e}_{(-i)})}{\text{Var}((I-QQ^T)X_i)}$.

Plugging this in, we get

$$\hat{e}_{(-i)} = \frac{\text{Cov}((I - QQ^T)X_i, e_{(-i)})}{\text{Var}((I - QQ^T)X_i)}(I - QQ^T)X_i + \hat{e}$$

$$\hat{e} = \hat{e}_{(-i)} - \frac{\text{Cov}\left((I - QQ^T)X_i, \hat{e}_{(-i)}\right)}{\text{Var}\left((I - QQ^T)X_i\right)}(I - QQ^T)X_i$$

$$\|\hat{e}\|^2 = \left\|\hat{e}_{(-i)} - \frac{\text{Cov}\left((I - QQ^T)X_i, \hat{e}_{(-i)}\right)}{\text{Var}\left((I - QQ^T)X_i\right)}(I - QQ^T)X_i\right\|^2$$

$$RSS = \left\|\hat{e}_{(-i)} - \frac{\text{Cov}\left((I - QQ^T)X_i, \hat{e}_{(-i)}\right)}{\text{Var}\left((I - QQ^T)X_i\right)}(I - QQ^T)X_i\right\|^2$$

Thus, we can minimize the RSS by choosing the variable X_i such that the quantity on the right is minimized.

5. RSS will be increased the least by removing the variable which has the highest p-value.
Explanation: If $RSS_{(-i)}$ is the residual sum of squares for a model excluding variable i, we want to pick i such that $RSS_{(-i)} - RSS$ is smallest. Obviously, the same i will minimize

$$F = \frac{\frac{RSS_{(-i)} - RSS}{p-(p-1)}}{\frac{RSS}{n-p}}$$

This will in turn maximize the p-value for the F test comparing the model to the model excluding variable i. However, as we know, this p-value is equivalent to the p-value from the t test in the regression summary for variable i. Thus, this variable, when removed, will result in the smallest increase in RSS.