Recommended Exercise 5 in Statistical Linear Models, Spring 2021

alexaoh

20 februar, 2021

Problem 1 Simple linear regression

a)

```
library(MASS)
names(forbes)
```

```
#> [1] "bp"   "pres"
```

```
str(forbes)
```

```
#> 'data.frame':    17 obs. of  2 variables:
#>  $ bp  : num  194 194 198 198 199 ...
#>  $ pres: num  20.8 20.8 22.4 22.7 23.1 ...
```

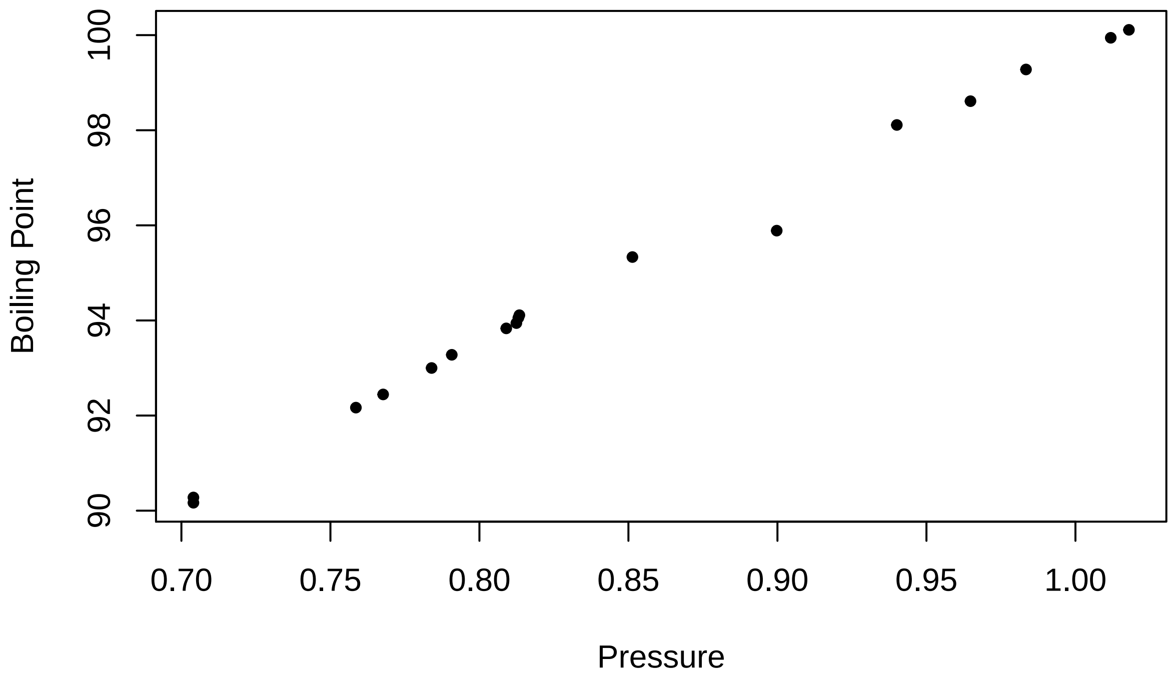
```
n <- length(forbes$bp)
Y <- matrix((forbes$bp - 32)*5/9, ncol = 1)
X <- cbind(rep(1,n), forbes$pres*0.033863882)
```

b)

The rank of X is 2, since it consists of two linearly independent columns.

c)

```
plot(X[, 2], Y, pch = 20, xlab = "Pressure", ylab = "Boiling Point")
```



It looks like there is a linear relationship between pressure and boiling point.

d)

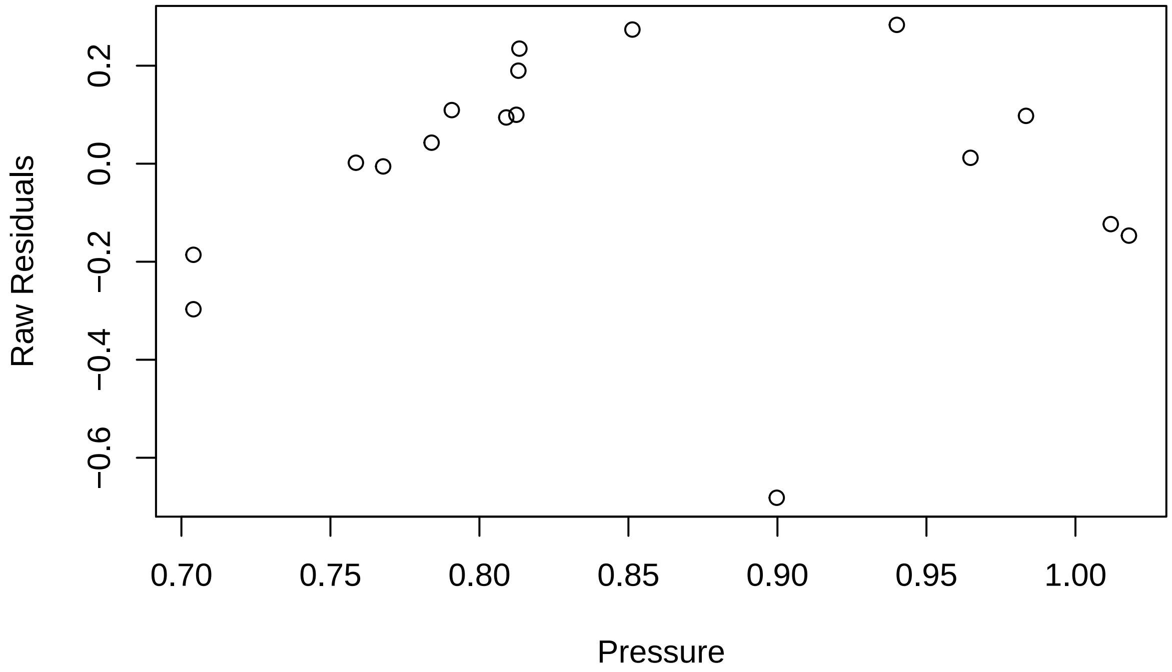
```
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%Y
beta.hat
```

```
#>          [,1]
#> [1,] 68.49805
#> [2,] 31.19980
```

To a layperson, I would say that these numbers mean the following. When the pressure is zero, the boiling point is estimated to being at ~ 68.4980464 degrees and for each unit the pressure is increased, the boiling point is estimated to increase by ~ 31.1998017 degrees.

e)

```
raw.res <- Y - X%*%beta.hat
y.hat <- X%*%beta.hat
plot(X[, 2], raw.res, xlab="Pressure", ylab="Raw Residuals")
```



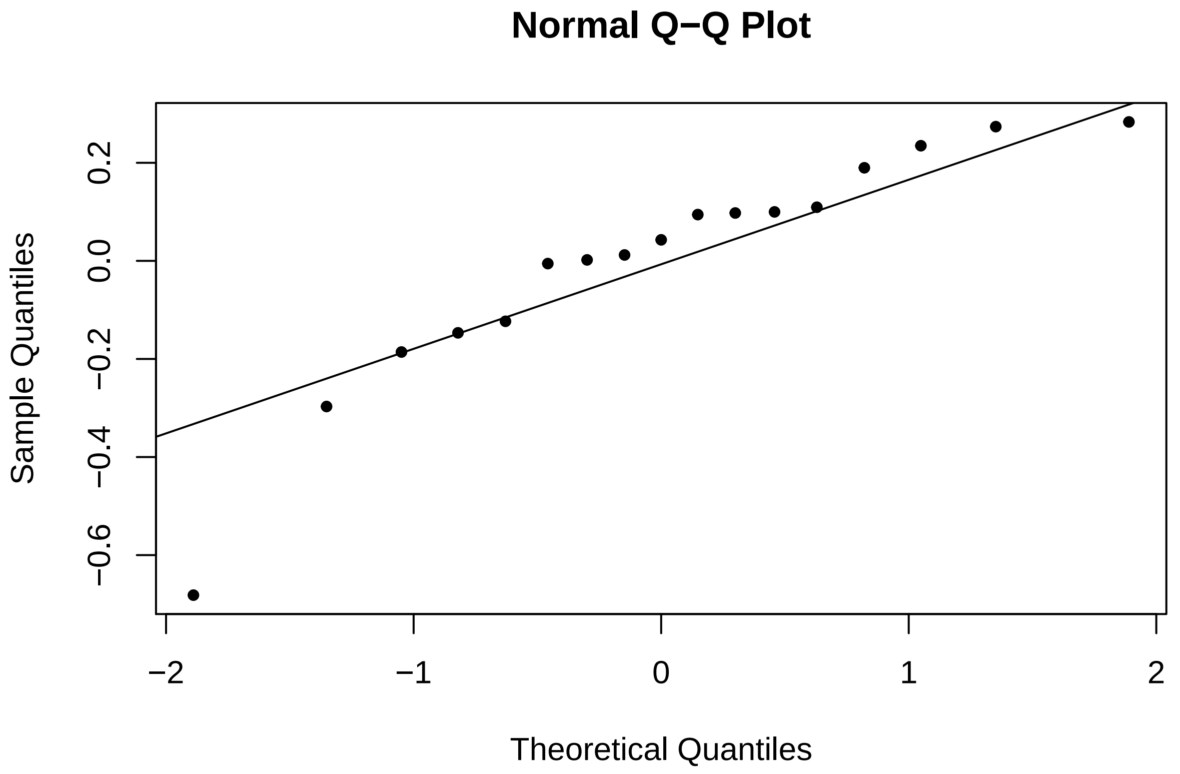
It looks like the residuals form a pattern of some sort, but it is hard to say exactly what sort of shape it is making. It looks like an inverted U-shape.

f)

- Linearity of covariate effects: The covariate-response plot shows that the relationship looks to be quite linear.
- Homoscedasticity of errors: The residuals do not look homoscedastic, since they follow an inverted U-shape. This means that this assumption looks to be violated by the data.
- Uncorrelated errors: Based on the residual-covariate plot, the errors look to be correlated, which is another violation of the model assessments.
- Additivity of errors: Not sure how to check this from the two plots, but several of the other assumptions seem to be violated, so it does not matter too much in this case.

How can we investigate if the errors are normally distributed? Can plot a Q-Q-plot of residuals, i.e. theoretical quantiles of the normal distribution and the residuals from the model. If they seem to be linear, the errors seem to be normally distributed. Have a look in the plots in the next task. This can also be done by

```
qqnorm(raw.res, pch = 20)
qqline(raw.res)
```

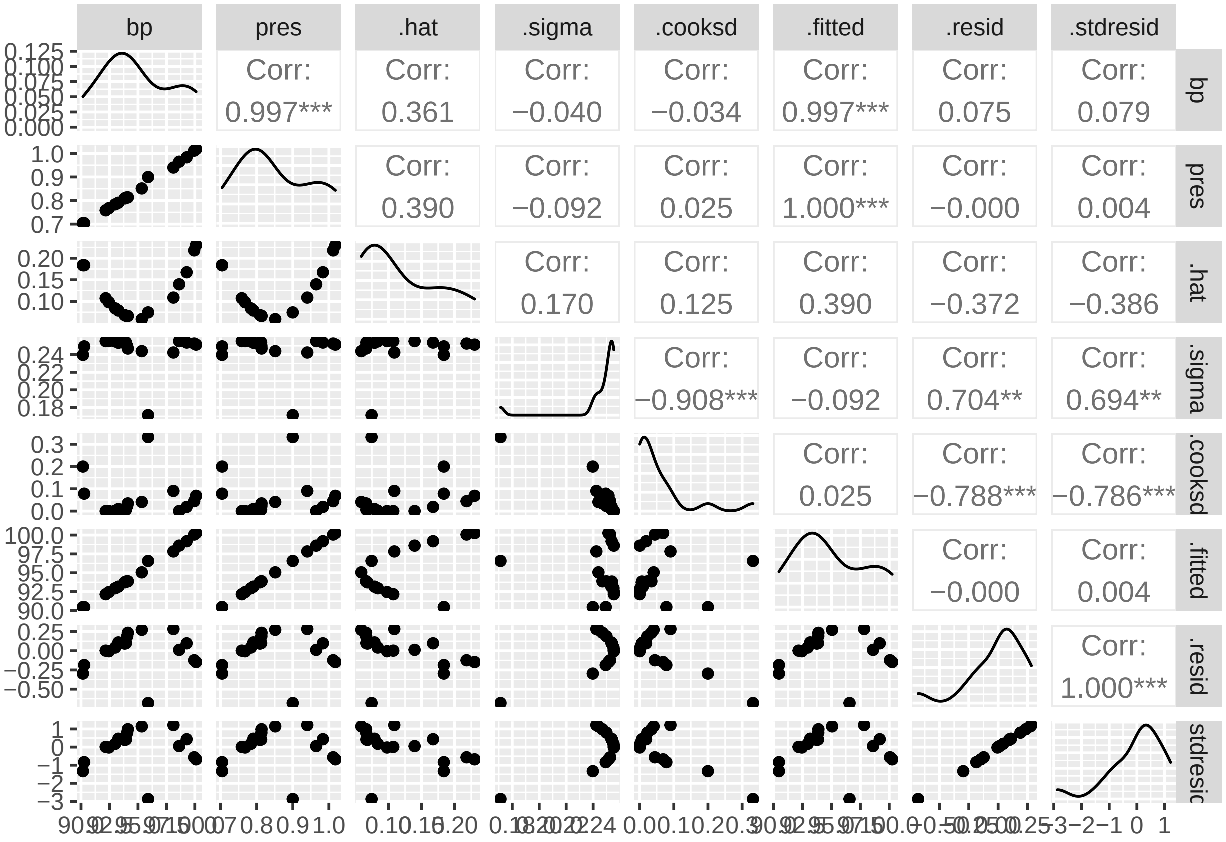


g)

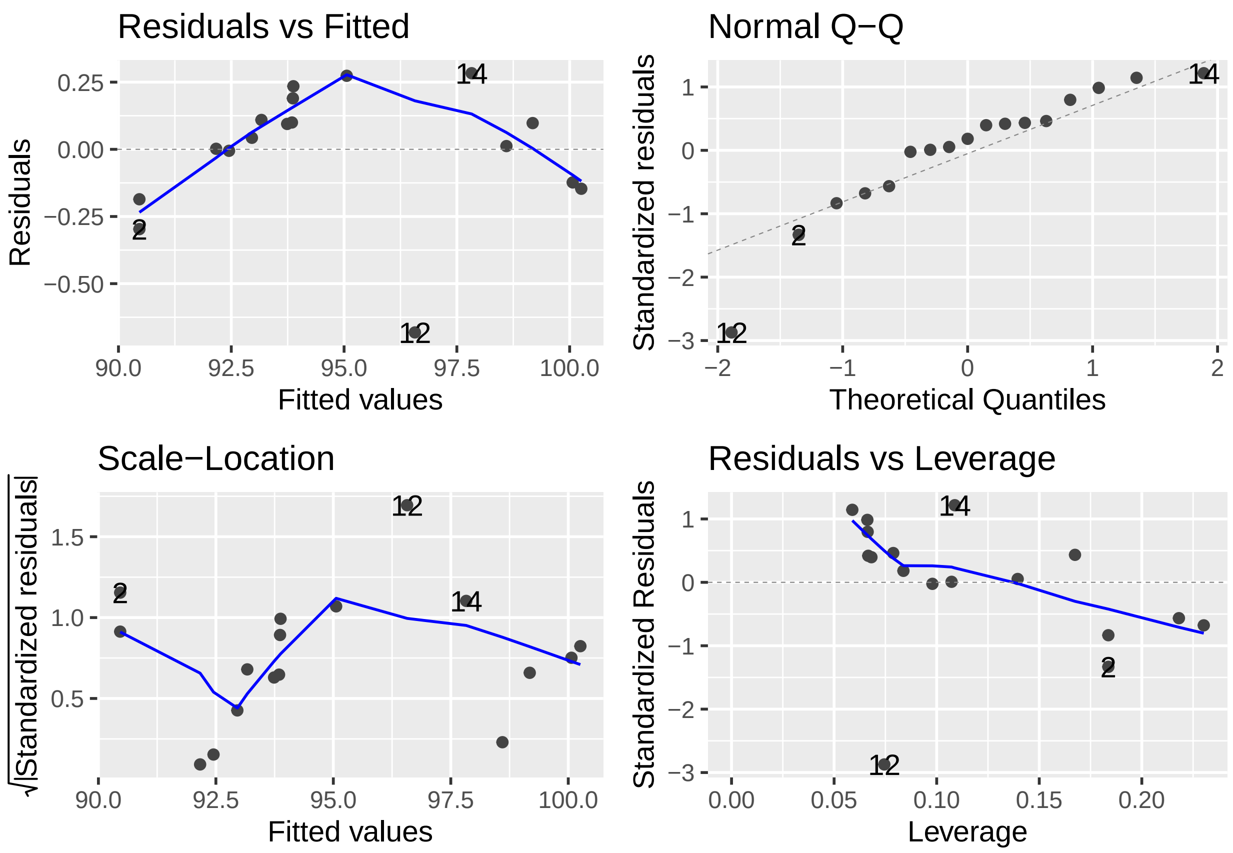
```
newds <- data.frame(bp = (forbes$bp - 32) * 5 / 9,pres = forbes$pres * 0.033863882)
lm1 <- lm(bp ~ pres, data = newds)
summary(lm1)
```

```
#>
#> Call:
#> lm(formula = bp ~ pres, data = newds)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.6816 -0.1232  0.0429  0.1094  0.2833
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  68.4980     0.5152  132.96   <2e-16 ***
#> pres         31.1998     0.6030   51.74   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.2467 on 15 degrees of freedom
#> Multiple R-squared:  0.9944, Adjusted R-squared:  0.9941
#> F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16
```

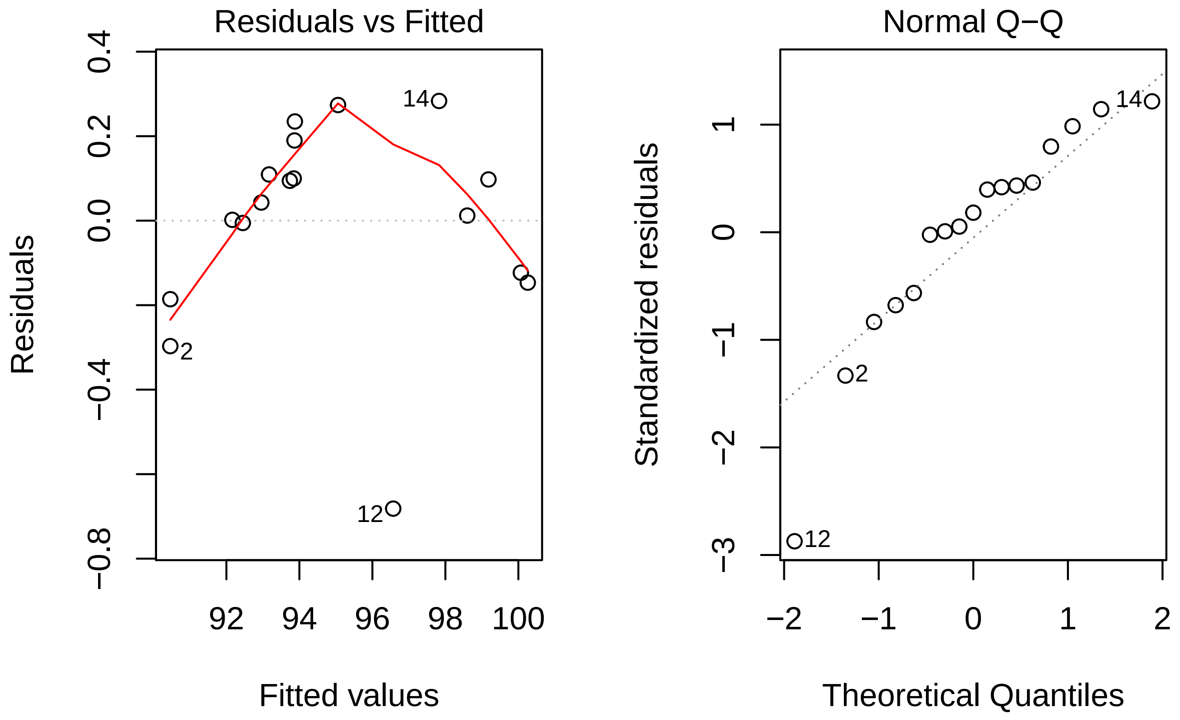
```
library(GGally)
ggpairs(lm1)
```



```
library(ggfortify)
autoplot(lm1)
```



```r
par(mfrow = c(1, 2)) # change number of subplots in a window
plot(lm1, which = c(1, 2))
```



Problem 2 Results on $\hat{\beta}$ and SSE in multiple linear regression

a)

H is idempotent and symmetric, since

$$H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}X^T = H$$

and

$$H^T = (X(X^TX)^{-1}X^T)^T = (X((X^TX)^{-1})^TX^T) = (X((X^TX)^T)^{-1}X^T) = (X(X^TX)^{-1}X^T) = H.$$

Since H is idempotent we know that $\text{rank}H = \text{tr}H = \text{tr}(X(X^TX)^{-1}X^T) = \text{tr}((X^TX)^{-1}X^TX) = \text{tr}I = p$, where the third equality follows from the fact that the trace is invariant under cyclic permutations. I would interpret the vector $H\boldsymbol{Y}$ as being the orthogonal projection of $\boldsymbol{Y}$ onto $\text{Col}X$.

$I - H$ is also idempotent and symmetric, which can be checked via direct calculation also

$$(I-H)^2 = (I-H)(I-H) = I - IH - HI + H^2 = I - H - H + H = I - H$$

and

$$(I-H)^T = I^T - H^T = I - H.$$

Since $I - H$ is idempotent, we know that $\text{rank}(I - H) = \text{tr}(I - H) = \text{tr}I - \text{tr}H = n - p$. I would interpret the vector $(I - H)\boldsymbol{Y}$ as being the part of $\boldsymbol{Y}$ that is perpendicular to $\text{Col}X$ after an orthogonal projection of $\boldsymbol{Y}$ onto $\text{Col}X$. In other words $(I - H)\boldsymbol{Y}$ is the orthogonal projection onto the orthogonal complement of $\text{Col}X$.

b)

Let $\text{SSE} := \boldsymbol{Y}^T(I - H)\boldsymbol{Y}$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. We know that $\frac{1}{\sigma^2}\boldsymbol{Z}^T D\boldsymbol{Z} \sim \chi^2_r$, when D is an idempotent and symmetric matrix with rank r and $\boldsymbol{Z} \sim N(0, \sigma^2 I)$. This result can be applied here, since $\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$ and, hence, $\boldsymbol{Y} - X\boldsymbol{\beta} \sim N(0, \sigma^2 I)$. This can be used because $(I - H)X\boldsymbol{\beta} = 0$, which gives

$$\frac{\text{SSE}}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{Y}^T(I - H)\boldsymbol{Y} = \frac{1}{\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(I - H)(\boldsymbol{Y} - X\boldsymbol{\beta}) \sim \chi^2_{n-p},$$

since, as noted in a), $I - H$ is idempotent and symmetric and $\text{rank}(I - H) = n - p$. Since the expected value of a chi-squared distributed random variable is its degrees of freedom

$$\text{E}\left(\frac{\text{SSE}}{\sigma^2}\right) = \frac{1}{\sigma^2}\text{E}(\text{SSE}) = n - p,$$

which leads to the following suggestion of an unbiased estimator of σ^2

$$\widehat{\sigma^2} = \frac{\text{SSE}}{n - p}.$$

The variance of $\widehat{\sigma^2}$ is

$$\text{Var}(\widehat{\sigma^2}) = \text{Var}\left(\frac{\text{SSE}}{n - p}\right) = \text{Var}\left(\frac{\sigma^2}{n - p}\frac{\text{SSE}}{\sigma^2}\right) = \frac{\sigma^4}{(n - p)^2}\text{Var}\left(\frac{\text{SSE}}{\sigma^2}\right) = \frac{\sigma^4}{(n - p)^2} \cdot 2(n - p) = \frac{2\sigma^4}{n - p},$$

since we know that the variance of a chi-squared distributed random variable is two times its degrees of freedom.

c)

Let $A := (X^T X)^{-1}X^T$ and $B := I - H$. The dimensions of these matrices are $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{n\times n}$. $A\boldsymbol{Y}$ and $B\boldsymbol{Y}$ are independent random vectors because $\text{Cov}(B\boldsymbol{Y}, A\boldsymbol{Y}) = B\text{Cov}(\boldsymbol{Y})A^T = \sigma^2(I - H)((X^T X)^{-1}X^T)^T = \sigma^2(I - H)X(X^T X)^{-1} = 0$, since $(I - H)X = X - X = 0$. This means that $\hat{\boldsymbol{\beta}}$ and SSE are independent random variables because $A\boldsymbol{Y} = (X^T X)^{-1}X^T\boldsymbol{Y} = \hat{\boldsymbol{\beta}}$ and $(B\boldsymbol{Y})^T(B\boldsymbol{Y}) = \boldsymbol{Y}^T B\boldsymbol{Y} = \boldsymbol{Y}^T(I - H)\boldsymbol{Y} = \text{SSE}$ must be independent also. This independence is used to construct t-tests for the components β_j of $\boldsymbol{\beta}$.