Statistics 203
Introduction to Regression and Analysis of Variance
Assignment #1 Solutions
January 20, 2005

Q. 1) **(MP 2.7)**

(a) Let x denote the hydrocarbon percentage, and let y denote the oxygen purity. The simple linear regression model is $\hat{y} = 77.863 + 11.801x$.

```
> #MP 2.7, Oxygen
> oxygen.table <- read.table("http://www-stat/~jtaylo/courses/
  stats203/data/oxygen.table",  header=T, sep=",")
> attach(oxygen.table)
> purity.lm <- lm(purity ~ hydro)
> summary(purity.lm)

Call:
lm(formula = purity ~ hydro)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6724 -3.2113 -0.0626  2.5783  7.3037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.863      4.199  18.544 3.54e-13 ***
hydro         11.801      3.485   3.386  0.00329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.597 on 18 degrees of freedom
Multiple R-Squared: 0.3891,     Adjusted R-squared: 0.3552
F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

```
> #Use filled in circles in the plot by typing pch=21
> plot(hydro, purity, pch=21, bg='blue', main="Purity vs
  Hydrocarbon Percentage")
> abline(purity.lm$coef, lwd=2)
```

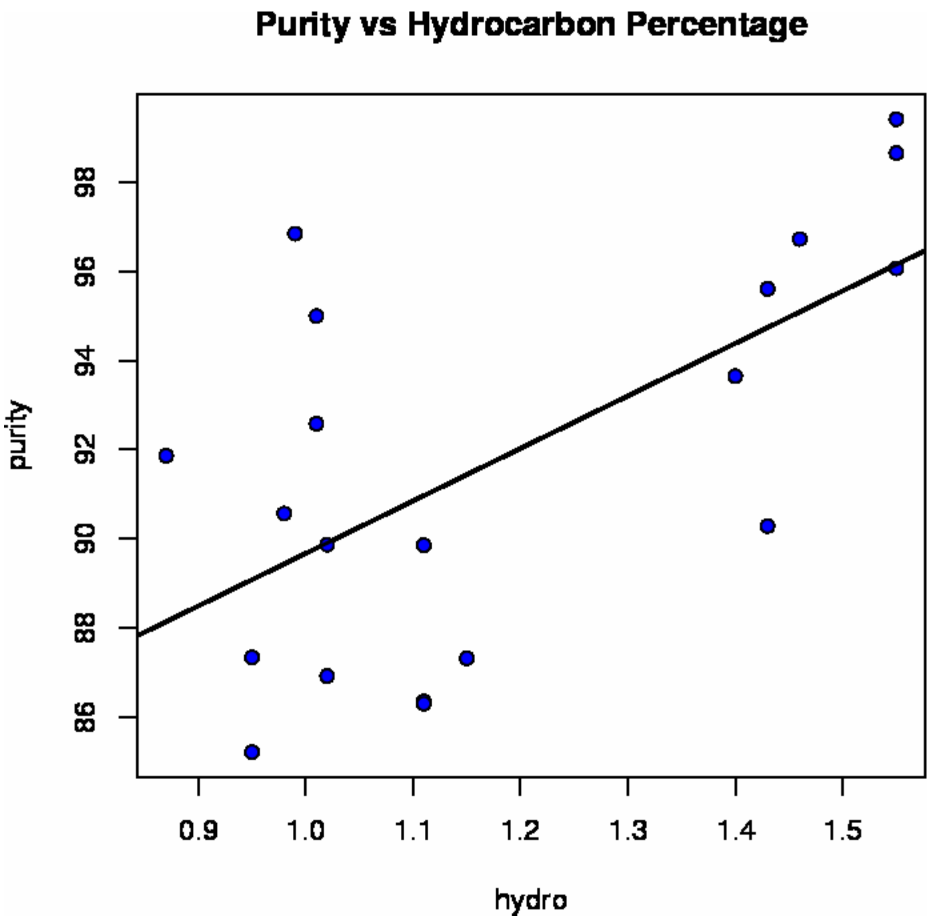

Figure 1: Plot of the purity versus hydrocarbon percentage, with the least squares line superimposed.

Figure 1 suggests a positive relationship between oxygen purity and the hydrocarbon percentage.

(b) Consider $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

We have $t_{\hat{\beta}_1} = 3.485$ on $20 - 2 = 18$ d.f., corresponding to a p-value of .00329. We therefore reject H_0 in favor of H_1, and conclude that the true slope β_1 is not zero.

(c) From `summary(purity.lm)` in part (a) above, we have $R^2 = .3891$.

(d) A 95% confidence interval for β_1 in this SLR model is given in `R` by:

```
> confint(purity.lm, level=.95)
                2.5 %   97.5 %
(Intercept) 69.041747 86.68482
hydro        4.479066 19.12299
```

Alternatively, recall that a $100(1-\alpha)\%$ CI for β_1 is:

$$\hat{\beta}_1 \pm SE_{\hat{\beta}_1} \cdot t_{1-\alpha/2,n-2}.$$

From above, we have $SE_{\hat{\beta}_1} = 3.485$. We can now compute this in `R` by typing:

```
> t.quantiles <- qt(c(.025, .975), 18)
> 11.801 + 3.485*t.quantiles
[1]  4.479287 19.122713
```

Here, `t.quantiles` are the .025 and .975 quantiles of the t_{18} distribution.

(e) A 95% confidence interval for $E(Y|X = 1.0)$ is given by $(87.51, 91.82)$. This is computed in `R` as follows.

```
> predict(purity.lm, newdata=list(hydro = 1.0), interval="confidence",
  level=.95)
          fit      lwr      upr
[1,] 89.66431 87.51017 91.81845
```

Q. 2) **(MP 2.19)**

(a) As usual, let $SSE = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$. Then:

$$\frac{\partial SSE}{\partial \beta_1} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^1(-x_i)$$

Setting $\left.\frac{\partial SSE}{\partial \beta_1}\right|_{\hat{\beta}_1} = 0$ and dividing the above by -2, we obtain:

$$\begin{aligned} 0 &= \sum_{i=1}^{n}(x_i y_i - \beta_0 x_i - \hat{\beta}_1 x_i^2) \\ \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i \end{aligned}$$

So the least squares estimate $\hat{\beta}_1$ is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n (y_i - \beta_0) x_i}{\sum_{i=1}^n x_i^2}.$$

(b) We derive $\mathrm{Var}(\hat{\beta}_1)$ as follows:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2}\right) \\
&= (\sum_{i=1}^n x_i^2)^{-2}\,\mathrm{Var}(\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i) \\
&= (\sum_{i=1}^n x_i^2)^{-2}\,\mathrm{Var}(\sum_{i=1}^n x_i y_i) && \text{since } \beta_0 \textstyle\sum_{i=1}^n x_i \text{ is constant} \\
&= (\sum_{i=1}^n x_i^2)^{-2} \sum_{i=1}^n x_i^2\,\mathrm{Var}(y_i) && \text{since } \{y_i\} \text{ are independent} \\
&= (\sum_{i=1}^n x_i^2)^{-2} \sum_{i=1}^n x_i^2 \sigma^2 \\
&= \sigma^2 \Big/ \sum_{i=1}^n x_i^2.
\end{aligned}$$

(c) We summarize our results in the following table.

Model	$\mathrm{Var}(\hat{\beta}_1)$	SSE	$SE(\hat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}$
β_0 unknown	$\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$	$\sqrt{\frac{SSE/(n-2)}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
β_0 known	$\frac{\sigma^2}{\sum_{i=1}^n x_i^2}$	$\sum_{i=1}^n (y_i - \beta_0 - \hat{\beta}_1 x_i)^2$	$\sqrt{\frac{SSE/(n-1)}{\sum_{i=1}^n x_i^2}}$

First, notice that $\mathrm{Var}(\hat{\beta}_1^{\beta_0 \text{ known}}) \leq \mathrm{Var}(\hat{\beta}_1^{\beta_0 \text{ unknown}})$. (Equality holds if and only if $\bar{x} = 0$.) To see this, observe:

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\
&\leq \sum_{i=1}^n x_i^2 \\
\Longrightarrow \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} &\geq \frac{\sigma^2}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

Hence, confidence intervals for β_1 will be narrower when β_0 is known, regardless of sample size (unless $\bar{x} = 0$).

Furthermore, in deriving a confidence interval for β_1 when β_0 is known, it is not hard to show that:

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-1}$$

The estimator $\hat{\beta}_1$ is a linear combination of the $\{y_i\}$, and hence is normally distributed. $\hat{\beta}_1$ is also unbiased:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\left(\frac{\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2}\right) \\
&= (\sum_{i=1}^n x_i^2)^{-1}\left(\sum_{i=1}^n \mathbb{E}(x_i y_i) - \beta_0 \sum_{i=1}^n x_i\right) \\
&= (\sum_{i=1}^n x_i^2)^{-1}\left(\sum_{i=1}^n \mathbb{E}[x_i(\beta_0 + \beta_1 x_i + \varepsilon)] - \beta_0 \sum_{i=1}^n x_i\right) \\
&= (\sum_{i=1}^n x_i^2)^{-1}\left(\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 - \beta_0 \sum x_i\right) \\
&= \beta_1.
\end{aligned}$$

Again, the vector of residuals e is independent of $\hat{\beta}_1$, and so $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ is independent of $\hat{\beta}_1$. Hence,

$$\begin{aligned}
T &= \underbrace{\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum_{i=1}^n x_i^2}}\right)}_{N(0,1)} \Big/ \underbrace{\sqrt{\frac{\sum_{i=1}^n (y_i - \beta_0 - \hat{\beta}_1 x_i)^2 / \sigma^2}{n-1}}}_{\sqrt{\chi^2_{n-1}/(n-1)}} \\
&= \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \\
&\sim t_{n-1}.
\end{aligned}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for β_1 has the form:

$$\hat{\beta}_1 \pm t_{1-\alpha/2,\nu} SE(\hat{\beta}_1),$$

where the degrees of freedom $\nu = n - 1$ when β_0 is known (one less parameter to estimate) and where $\nu = n - 2$ when β_0 is unknown. This difference of 1 degree of freedom results in a *slightly* narrower

CI for β_1, but for relatively large samples, this difference is almost negligible. The major difference is attributed to the variance of the estimator $\hat{\beta}_1$.

Q. 3) **(MP 3.10)**

(a) The following R commands allow us to compute and plot the residuals and standardized residuals.

```
> softdrink.table <-
  read.table("http://www-stat/~jtaylo/courses/stats203/
  data/softdrink.table", header=T, sep=" ")
> attach(softdrink.table)

> #Compute residuals
> softdrink.lm <- lm(y ~ x1 + x2)
> softdrink.resid <- softdrink.lm$residuals

> #Compute standardized residuals
> softdrink.st.resid <- rstandard(softdrink.lm)

> #Combine residuals & standardized residuals using cbind
> print(cbind(softdrink.resid, softdrink.st.resid))

   softdrink.resid softdrink.st.resid
1       -5.0280843        -1.62767993
2        1.1463854         0.36484267
3       -0.0497937        -0.01609165
4        4.9243539         1.57972040
5       -0.4443983        -0.14176094
6       -0.2895743        -0.09080847
7        0.8446235         0.27042496
8        1.1566049         0.36672118
9        7.4197062         3.21376278
10       2.3764129         0.81325432
11       2.2374930         0.71807970
12      -0.5930409        -0.19325733
13       1.0270093         0.32517935
14       1.0675359         0.34113547
15       0.6712018         0.21029137
16      -0.6629284        -0.22270023
17       0.4363603         0.13803929
18       3.4486213         1.11295196
19       1.7931935         0.57876634
20      -5.7879699        -1.87354643
```

```
21      -2.6141789        -0.87784258
22      -3.6865279        -1.44999541
23      -4.6075679        -1.44368977
24      -4.5728535        -1.49605875
25      -0.2125839        -0.06750861

> #Plot the residuals & standard residuals in one window
> par(mfrow = c(1,2))
> plot(softdrink.lm$residuals, pch=23, bg='blue', cex=2, lwd=2,
  main="Residuals")
> plot(rstandard(softdrink.lm), pch=23, bg='red', cex=2, lwd=2,
  main="Standardized Residuals")
```

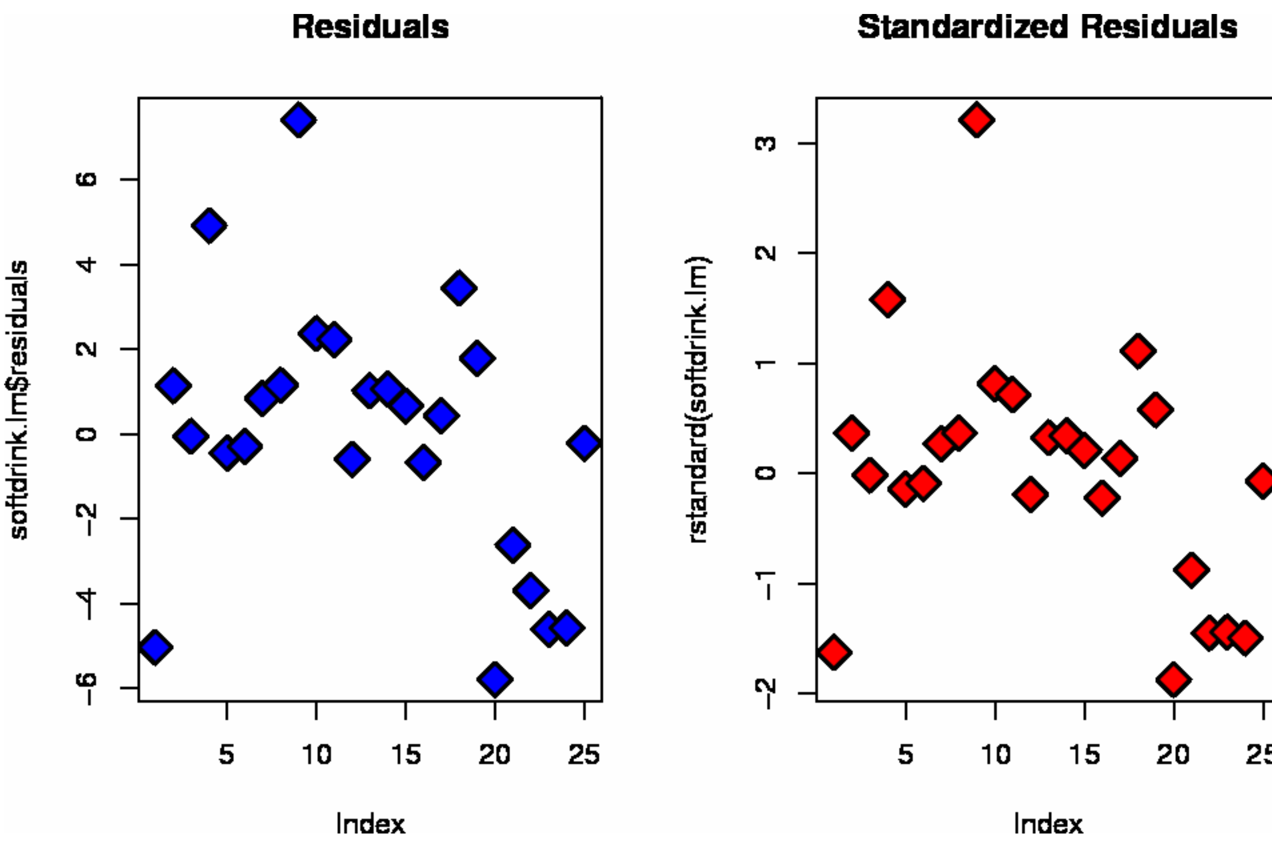

Figure 2: Plots of the residuals (blue) and standardized residuals (red) for the soft drink data.

(b) From Table 4.2 of Montgomery and Peck, we notice that the x_1 and

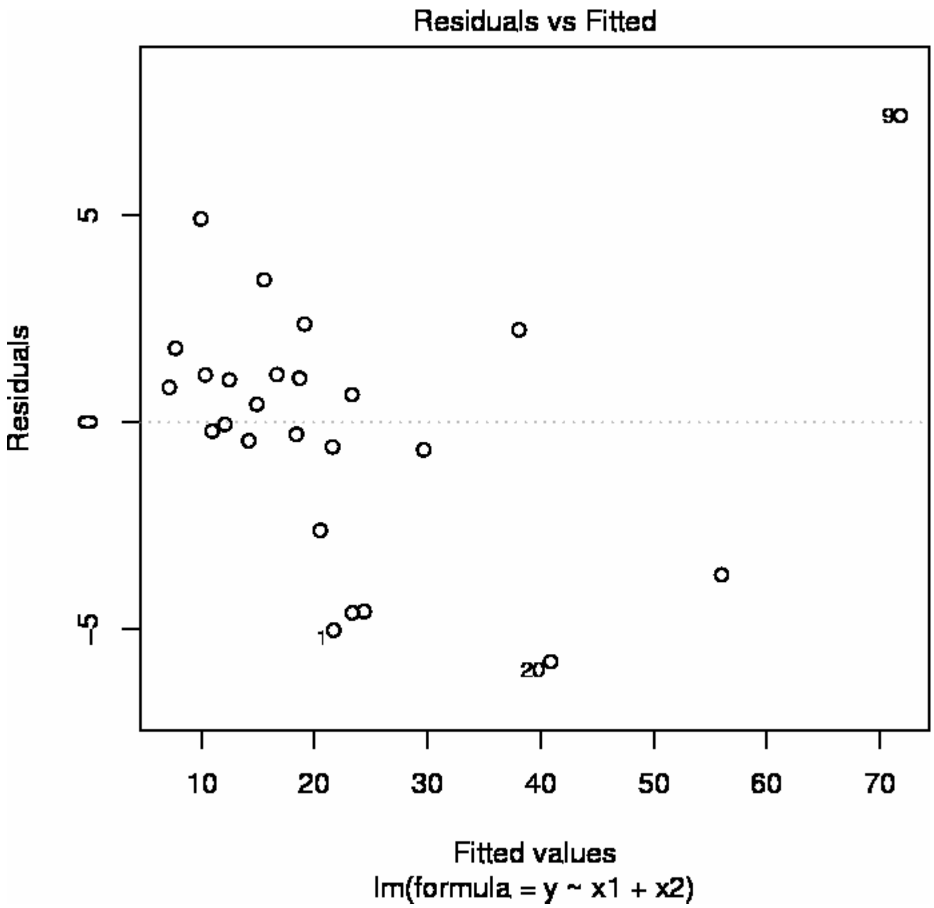

Figure 3: Plot of the residuals versus fits, suggesting that case number 9 is an outlying observation.

x_2 values for Observation 9 are much higher than what appears to be typical, suggesting that Observation 9 is an unusual observation. We can use the `plot` command in `R` to obtain the residuals versus fits and a Cook's distance plot. Both plots suggest that case number 9 is an outlying observation. Refer to Figures 3 and 4.

```
> plot(softdrink.lm)
```

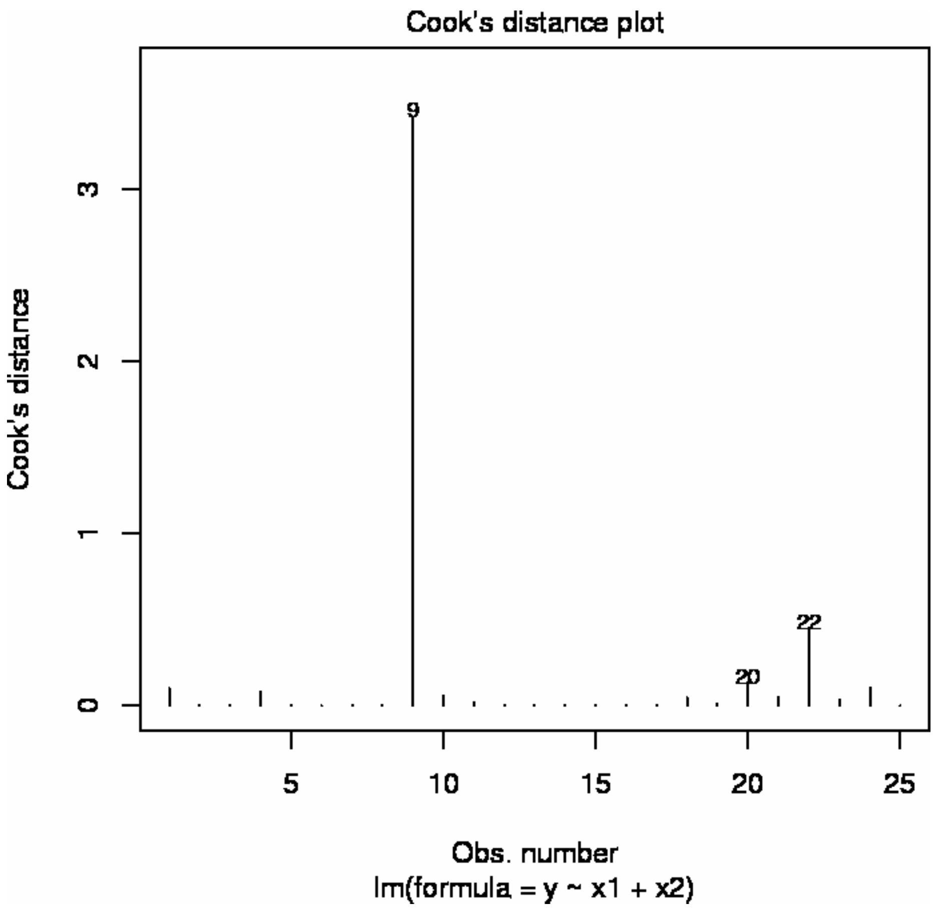

Figure 4: The Cooks distance plot suggests that Observation 9 is an outlier.

Q. 4) **(MP 4.24)**

Method 1: First note that:

- For a constant matrix A and a random vector Z, we have $\mathrm{Var}(AZ) = A\,\mathrm{Var}(Z)A^t$.
- The hat matrix H is only a function of X, which we treat as fixed. Hence, H is a constant matrix.
- Under the multiple regression model, we assume $\mathrm{Var}(Y) = \sigma^2 I$ is a constant matrix.
- For a symmetric matrix U, that is, $U = U^t$, we have $U^{-1} = (U^{-1})^t$. To see this, observe:

$$\begin{aligned} U^{-1}U = &\quad I \\ (U^{-1})^t U^t = &\quad I \end{aligned}$$

  Since $U = U^t$, we must have $U^{-1} = (U^{-1})^t$.

With $\hat{Y} = HY$, we therefore have:

$$\begin{aligned} \mathrm{Var}(\hat{Y}) &= \mathrm{Var}(HY) \\ &= H\,\mathrm{Var}(Y)H^t \\ &= [X(X^tX)^{-1}X^t](\sigma^2 I)[X(X^tX)^{-1}X^t]^t \\ &= \sigma^2[X(X^tX)^{-1}X^t][(X^t)^t((X^tX)^{-1})^tX^t] \\ &= \sigma^2 X(X^tX)^{-1}(X^tX)((X^tX)^{-1})^tX^t \\ &= \sigma^2 XI((X^tX)^{-1})^tX^t \\ &= \sigma^2 X(X^tX)^{-1}X^t \\ &= \sigma^2 H \end{aligned}$$

The fact that $((X^tX)^{-1})^t = (X^tX)^{-1}$ follows, since X^tX is a symmetric matrix.

Method 2: Alternatively, notice that $H = H^t$ (provide a proof) and use the result of the next problem (that $H = H^2$) to see:

$$\begin{aligned} \mathrm{Var}(\hat{Y}) &= \mathrm{Var}(HY) \\ &= H\,\mathrm{Var}(Y)H^t \\ &= \sigma^2 HH^t \\ &= \sigma^2 HH \\ &= \sigma^2 H \end{aligned}$$

Q. 5) **(MP 4.25)**

$$\begin{aligned} H^2 &= [X(X^tX)^{-1}X^t][X(X^tX)^{-1}X^t] \\ &= X(X^tX)^{-1}(X^tX)(X^tX)^{-1}X^t \\ &= XI(X^tX)^{-1}X^t \\ &= H. \end{aligned}$$

$$\begin{aligned} (I-H)^2 &= I^2 - 2IH + H^2 \\ &= I - 2H + H \\ &= I - H. \end{aligned}$$