STAT 654 Problem Set 1

Nikko Cleri cleri@tamu.edu

March 6, 2022

All relevant R code is appended after the answers to all questions.

A

1. To prove: $E(\hat{\beta}_1) = \beta_1$, where

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \epsilon_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

With the quantity c_i defined as in the hints

$$\begin{aligned} c_i &= \frac{x_i - \bar{x}}{S_x^2} \\ &= \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

We see that $\hat{\beta}_1$ can be represented in terms of the c_i as in the hint, where

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^{n} \epsilon_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ &= \sum_{i=1}^{n} c_i (y_i - \bar{y}) \\ &= \sum_{i=1}^{n} c_i y_i - \bar{y} \sum_{i=1}^{n} c_i \end{aligned}$$

where the second term is zero:

$$\begin{aligned} \sum_{i=1}^{n} c_i &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_x^2} \\ &= \frac{1}{S_x^2} \sum_{i=1}^{n} (x_i - \bar{x}) \\ &= \frac{1}{S_x^2} (n\bar{x} - n\bar{x}) \\ &= 0 \end{aligned}$$

Returning to the original lemma, we have

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \epsilon_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i - \bar{y} \sum_{i=1}^{n} c_i \right)$$

Using the linearity of expectation, we have

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} c_i E(y_i) - \bar{y} \sum_{i=1}^{n} c_i$$

and using the definition of y_i,

$$E(\hat{\beta}_1) = \beta_1 \sum_{i=1}^{n} c_i x_i + \beta_0 \sum_{i=1}^{n} c_i - \bar{y} \sum_{i=1}^{n} c_i$$

where the second and third term are zero as shown before. The sum in the first term is

$$\begin{aligned} \sum_{i=1}^{n} c_i x_i &= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{S_x^2} \\ &= \frac{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)}{S_x^2} \end{aligned}$$

which in the same manner as the sum $\sum_{i=1}^{n} c_i$ expression before, we have

$$\sum_{i=1}^{n} c_i x_i = 1$$

So we have the desired

$$E(\hat{\beta}_1) = \beta_1$$

2. To prove: $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / (\sum_{i=1}^{n} (x_i - \bar{x})^2)$. We begin with

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\sum_{i=1}^{n} c_i y_i - \bar{y} \sum_{i=1}^{n} c_i \right) \\ &= \mathrm{Var}\left(\sum_{i=1}^{n} c_i y_i \right) \end{aligned}$$

using the results from part i. This becomes

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i^2 \mathrm{Var}(y_i) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 \end{aligned}$$

We now show

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} \right)^2$$
$$= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

so we have the desired

$$\text{Var}(\hat{\beta}_1) = \sigma^2 / (\sum_{i=1}^{n} (x_i - \bar{x})^2)$$

3. To prove: $\hat{\beta}_1$ is normally distributed. In previous parts we showed that

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$$

since we know that each Y_i is normally distributed as $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, it follows immediately that $\hat{\beta}_1$ is normally distributed, as it is the linear combination of independent normal random variables.

B

To prove: $\hat{\rho} = R^2$. We begin by showing

$$\begin{aligned} SSR &= \hat{\beta}_1^2 S_x^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 \\ &= \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \hat{\beta}_1^2 S_x^2 \end{aligned}$$

and now show

$$\begin{aligned}
\hat{\beta}_1 &= \hat{\rho}\frac{S_y}{S_x} \\
\hat{\rho} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_x S_y} \\
\hat{\rho}\frac{S_y}{S_x} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_x S_y}\frac{S_y}{S_x} \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_x^2} \\
&= \hat{\beta}_1
\end{aligned}$$

We conclude with

$$\begin{aligned}
r^2 &= \frac{SSR}{SST} \\
&= \frac{\hat{\beta}_1^2 S_x^2}{S_y^2} \\
&= \hat{\rho}^2
\end{aligned}$$

C

i. We take the derivative

$$\begin{aligned}
\mu_{Y|X}(x) &= -8.5 - 3.2x + 0.7x^2 \\
\frac{\mathrm{d}\mu_{Y|X}(x)}{\mathrm{d}x} &= -3.2 + 1.4x
\end{aligned}$$

we have $\left.\frac{\mathrm{d}\mu_{Y|X}(x)}{\mathrm{d}x}\right|_{x=0} = -3.2$, $\left.\frac{\mathrm{d}\mu_{Y|X}(x)}{\mathrm{d}x}\right|_{x=2} = -0.4$, $\left.\frac{\mathrm{d}\mu_{Y|X}(x)}{\mathrm{d}x}\right|_{x=0} = 1$

ii. In the parameterization we have

$$\begin{aligned}
\mu_{Y|X}(x) &= -8.5 - 3.2x + 0.7x^2 \\
&= \beta_0 + \beta_1(x-\mu_x) + \beta_2(x-\mu_x)^2
\end{aligned}$$

Set $\mu_x = 2$ we have $\beta_0 = -12.1$, $\beta_1 = -0.4$, $\beta_1 = 0.7$.

D

i. The adjusted R^2 for the third order polynomial fit is 0.9648. The F-statistic is 165.4 on 3 and 15 degrees of freedom, with a p-value of 1.025×10^{-11}. each regression coefficient has a p-value well below the significance threshold, all with $p < 10^{-5}$.

ii. See Figure 1. The fit is quite good, as suggested by the adjusted R^2, but the residuals suggest that it can be improved. There is a clear increase in residuals scatter at lower x.

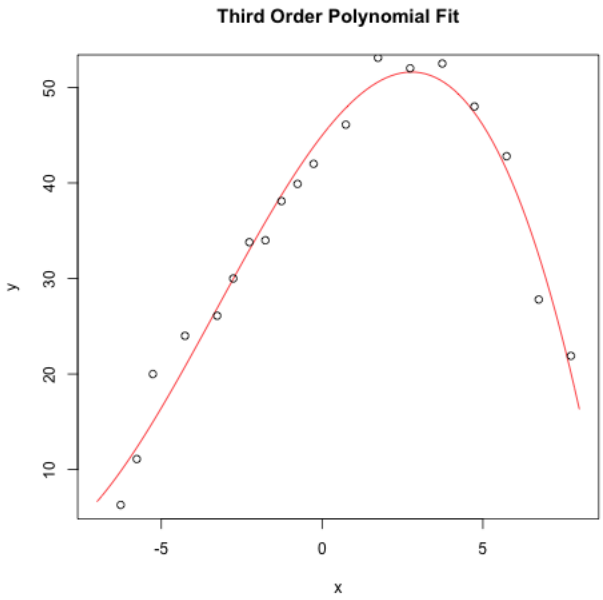

Figure 1: Question D part ii: Data with third order polynomial fit.

iii. The fifth order polynomial fit has a better adjusted R^2 than the third order fit (0.9847). The regression coefficients are all below the significance threshold of 0.01 except for the second order term, which has a p-value of 0.2157.

iv. The p-values are now all below the significance threshold of 0.01, with an adjusted R^2 of 0.984 (nominally higher than the adjusted R^2 in part i). the plot in Figure 2 shows a better fit than the third order fit in part ii.

E

i. The best AIC and BIC using the AIC/BIC function used in class is from the order 2 polynomial.

ii. The second order fit gives a p value for the coefficient of the second order term of 0.00154, significant enough to reject the null hypothesis that the coefficient is zero at a confidence of $\alpha = 0.05$.

iii. See Figure 3. The residuals visually get larger in scatter with larger predicted values.

iv. The 95% prediction interval is (18.175, 84.823).

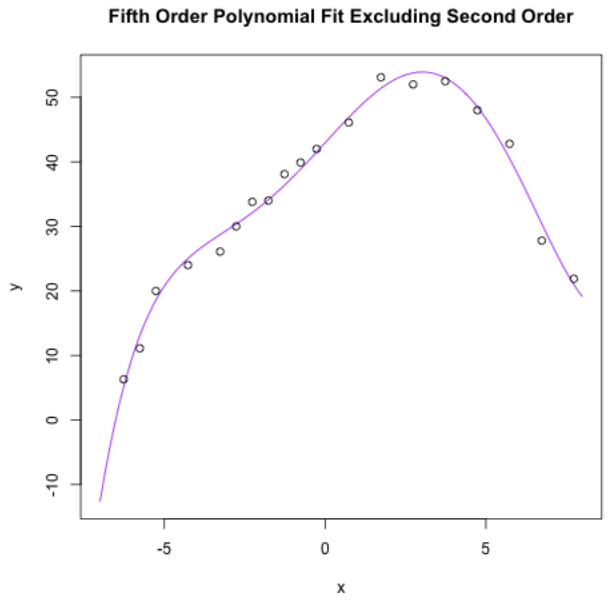

Figure 2: Question D part iv: Data with fifth order polynomial fit with second order term removed.

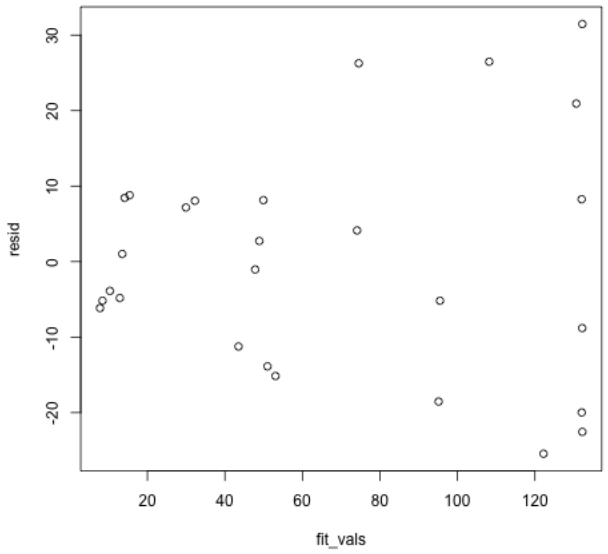

Figure 3: Question E part iii: Residuals of the second order polynomial fit

```
# 654 Homework 1
# Question D
hc = read.table('/Users/alvis/Classes/STAT_654/HW_1/
HardwoodTensileStr.txt', header = TRUE)
# summary(hc)
attach(hc)

x = Concentration
y = Strength
x = x-mean(x)

hc3 = lm(y~poly(x, 3, raw=TRUE))
hc5 = lm(y~poly(x, 5, raw=TRUE))
resid3 = hc3$residuals
predict3 = hc3$fitted.values
coeff3 = hc3$coefficients
resid5 = hc5$residuals
predict5 = hc5$fitted.values
coeff5 = hc5$coefficients
summary(hc3)
summary(hc5)
hc5_2 = lm(y ~ I(x) + I(x^3) + I(x^4) + I(x^5))
summary(hc5_2)
coeff5_2 = hc5_2$coefficients

#third order fit
png("/Users/alvis/Classes/STAT_654/HW_1/D_ii.png")
curve(coeff3[[1]] + coeff3[[2]]*x + coeff3[[3]]*x^2 + coeff3[[4]]*x^3,
from = -7, to =8,
      ylab = 'y', col = 'red', main = "Third Order Polynomial Fit")
points(x, y)
dev.off()
plot(x, resid3, xlab="x", ylab="Residuals from Third Order Polynomial
Fit")

#fifth order fit without x^2 term
png("/Users/alvis/Classes/STAT_654/HW_1/D_iv.png")
curve(coeff5_2[[1]] + coeff5_2[[2]]*x + coeff5_2[[3]]*x^3 +
coeff5_2[[4]]*x^4 + coeff5_2[[5]]*x^5, from = -7, to =8,
      ylab = 'y', col = 'purple', main = "Fifth Order Polynomial Fit
Excluding Second Order")
points(x, y)
dev.off()

#Question E

#The AIC function from class
AICpoly=function(x,y,kmax){
  # Given regression data (x,y), this function calculates AIC, BIC, and
  # adjusted R^2 for polynomial models of degree 1 to kmax.
  #
  # The AIC and BIC criteria are plotted, and the minimizers indicated
  # by vertical lines.  If there is only one vertical line, then both
  # criteria are minimized at the same k.
  #
  # The output is a list containing the AIC values, BIC values, values of
  # adjusted R^2, and the respective optimizers of the three criteria.
  #
```

```
  adjr2=1:kmax
  aic=1:kmax
  bic=1:kmax
  n=length(y)
  for(k in 1:kmax){
    fit=lm(y~poly(x,k))
    aic[k]=AIC(fit)
    bic[k]=BIC(fit)
    adjr2[k]=summary(fit)$adj
  }
  kr2=(1:kmax)[adjr2==max(adjr2)]
  kaic=(1:kmax)[aic==min(aic)]
  kbic=(1:kmax)[bic==min(bic)]
  ylim=range(c(aic,bic))
  plot(1:k,aic,xlab='k',ylab='aic/bic',ylim=ylim,col='red',pch = 19)
  points(1:k,bic,col='blue',pch=19)
  abline(v=kaic,col='red')
  abline(v=kbic,col='blue')
  title("Red points: AIC  Blue points: BIC")
  list(aic,bic,adjr2,kaic,kbic,kr2)
}

tree_data = read.table('/Users/alvis/Classes/STAT_654/HW_1/
TreeAgeDiamSugarMaple.txt', header = TRUE)
print(tree_data)
attach(tree_data)
x=Diamet
y=Age

AICpoly(x,y,8)
fit=lm(y~poly(x,2))
summary(fit)
resid = fit$residuals
fit_vals = fit$fitted.values

# Plot of residuals for E (3)
png("/Users/alvis/Classes/STAT_654/HW_1/E_iii.png")
plot(fit_vals,resid)
dev.off()

# Prediction interval for E (4)
predict(fit, newdata=data.frame(x = 110), interval="prediction",
level=0.95)
```