Jared Yu
Module 4 Assignment
Data Mining

1. Show that, if x, y are jointly Gaussian, the regression of y on x is given by
$$E(y|x) = \frac{\alpha\sigma_y x}{\sigma_x} + \mu_y - \frac{\alpha\sigma_y\mu_x}{\sigma_x}, \text{where } \Sigma = \begin{pmatrix} \sigma_x^2 & \alpha\sigma_x\sigma_y \\ \alpha\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Ans: (Reference: [1], [2])
From the question, it can be seen that x and y follow a bivariate normal distribution,
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_y\sigma_x\sqrt{1-\alpha^2}}\exp\left\{-\frac{1}{2(1-\alpha^2)}\left[\left(\frac{y-\mu_y}{\sigma_y}\right)^2 + \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\alpha\left(\frac{y-\mu_y}{\sigma_y}\right)\left(\frac{x-\mu_x}{\sigma_x}\right)\right]\right\},$$
where
$$\alpha = \frac{E(y-\mu_y)(x-\mu_x)}{\sigma_y\sigma_x} = \frac{\sigma_{xy}}{\sigma_y\sigma_x}.$$
This is the joint p.d.f. of x and y. The conditional distribution is as follows:
$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$
The term in the denominator is the marginal density for X, which can be shown as follows:
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left[-\frac{1}{2\sigma^2}(x-\mu_x)^2\right].$$
Then the conditional distribution can be shown as follows:
$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$
$$= \frac{\frac{1}{2\pi\sigma_y\sigma_x\sqrt{1-\alpha^2}}\exp\left\{-\frac{1}{2(1-\alpha^2)}\left[\left(\frac{y-\mu_y}{\sigma_y}\right)^2 + \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\alpha\left(\frac{y-\mu_y}{\sigma_y}\right)\left(\frac{x-\mu_x}{\sigma_x}\right)\right]\right\}}{\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left[-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right]}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y}\exp\left\{-\frac{1}{2(1-\alpha^2)}[A^2 + B^2 - 2\alpha AB] + \frac{1}{2}B^2\right\}$$
Note: Let $A = \frac{y-\mu_y}{\sigma_y}$ *and* $B = \frac{x-\mu_x}{\sigma_x}$.
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y}\exp\left\{-\frac{1}{2(1-\alpha^2)}[A^2 + B^2 - B^2(1-\alpha^2) - 2\alpha AB]\right\}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y}\exp\left\{-\frac{1}{2(1-\alpha^2)}\left[\left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\alpha\left(\frac{y-\mu_y}{\sigma_y}\right)B + B^2\alpha^2\right]\right\}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y}\exp\left\{-\frac{1}{2\sigma_y^2(1-\alpha^2)}\left[(y-\mu_y)^2 - 2\alpha\sigma_y(y-\mu_y)B + B^2\alpha^2\sigma_y^2\right]\right\}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y}\exp\left\{-\frac{1}{2\sigma_y^2(1-\alpha^2)}\left[y^2 - 2y\mu_y + \mu_y^2 - 2\alpha\sigma_y yB + 2\alpha\sigma_y\mu_y B + B^2\alpha^2\sigma_y^2\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y} \exp\left\{-\frac{1}{2\sigma_y^2(1-\alpha^2)}\left[y^2 - 2y(\mu_y + \alpha\sigma_y B) + \mu_y^2 + 2\alpha\sigma_y\mu_y B + B^2\alpha^2\sigma_y^2\right]\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y} \exp\left\{-\frac{1}{2\sigma_y^2(1-\alpha^2)}\left\{\left[y - (\mu_y + \alpha\sigma_y B)\right]^2 - (\mu_y + \alpha\sigma_y B)^2 + \mu_y^2 + 2\alpha\sigma_y\mu_y B + B^2\alpha^2\sigma_y^2\right\}\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\alpha^2}\sigma_y} \exp\left\{-\frac{1}{2\sigma_y^2(1-\alpha^2)}\left\{\left[y - (\mu_y + \alpha\sigma_y B)\right]^2\right\}\right\}$$

In the above equation, $\mu_y + \alpha\sigma_y B$, is the mean of the conditional density, since it can be seen that the above equation is also a normal density with mean $\mu_y + \alpha\sigma_y B$ and variance $\sigma_y^2(1-\alpha^2)$. Therefore, it follows that the expectation of the conditional density $E(y|x) = \mu_y + \alpha\sigma_y B$. To simplify,

$$E(y|x) = \mu_y + \alpha\sigma_y B$$
$$= \mu_y + \alpha\sigma_y\left(\frac{x-\mu_x}{\sigma_x}\right)$$
$$\boxed{= \frac{\alpha\sigma_y x}{\sigma_x} + \mu_y - \frac{\alpha\sigma_y\mu_x}{\sigma_x}}$$

where $\boxed{\alpha = \frac{\sigma_{xy}}{\sigma_y\sigma_x}}$. ■

2. Given data $(X_1, Y_1), \cdots, (X_n, Y_n)$, consider the regression through the origin model $Y_i = \beta X_i + v_i$, where $E(v_i|X_i) = 0$ and $Var(v_i|X_i) = \sigma^2$.
   a. Find $\hat{\beta}$, the least squares estimate for β.

Ans:

Let $\hat{Y}_i = \hat{\beta}X_i$ be the prediction of Y_i based on x_i, the i^{th} value of X, then $v_i = y_i - \hat{y}_i$ is the i^{th} residual:

$$v_i = y_i - \hat{\beta}x_i.$$

Then, define the residual sum of squares to be

$$R = \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2.$$

The goal then is to find β to minimize R. Taking the partial derivative:

$$\frac{\partial R}{\partial \beta} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}x_i)\,x_i \overset{\text{set to}}{=} 0$$
$$\rightarrow \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)\,x_i = 0$$
$$\rightarrow \sum_{i=1}^{n} y_i x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0$$

$$\rightarrow \boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{j=1}^{n} x_j^2} = \sum_{i=1}^{n} y_i c_i}$$

where $\boxed{c_i = \frac{x_i}{\sum_{j=1}^{n} x_j^2}}$.

b. Find the standard error of the estimate, $\sqrt{Var(\hat{\beta})}$.

Ans:

Next, the standard error will be shown.

$$Var(\hat{\beta}) = Var\left(\sum_{i=1}^{n} y_i c_i\right)$$

$$= \sum_{i=1}^{n} c_i^2 Var(y_i) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{j=1}^{n} x_j^2}\right)^2 = \sigma^2 \sum_{i=1}^{n} \frac{x_i^2}{\left(\sum_{j=1}^{n} x_j^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{j=1}^{n} x_j^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

Then, the standard error can be shown as,

$$\sqrt{Var(\hat{\beta})} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}} = \boxed{\frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2}}}.$$

Two major assumptions are made. The first is that X is a fixed variable as explained in part a), which allows c_i to be treated as a constant within the variance. The other assumption is that $Var(v_i|X_i) = \sigma^2$. This second assumption will be explained as follows:

$$Var(Y_i) = Var(\beta X_i + v_i|X_i) = Var(v_i|X_i) = \sigma^2.$$

In the above steps, βX_i is considered constant and so it zeros out within the variance (e.g., $Var(aX + b) = a^2 Var(X)$ for constants a, b and random variable X). In other words, the sum of the variance is equal to the variance of the sums.

c. Find conditions that guarantee that the estimate is consistent:

$$\forall \varepsilon > 0, \qquad P\left(\left|\hat{\beta} - \beta\right| > \varepsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Ans: (Reference: [3])

In the above problem, the constant β is shown as the probability limit of the sequence, which will be abbreviated as plim. So, the statement can be rewritten as,

$$\text{plim } \hat{\beta} = \beta.$$

The statement is saying that the probability of $\hat{\beta}$ being able to differ from β by some arbitrary finite value ε tends towards zero as n grows towards ∞. This problem is looking to understand the asymptotic properties of the estimator $\hat{\beta}$. First, $\hat{\beta}$ will be restated as follows,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{j=1}^{n} x_j^2}.$$

The expectation of $\hat{\beta}$, $E\left[\hat{\beta}\right]$ is as follows,

$$E[\hat{\beta}] = E\left[\frac{\sum_{i=1}^{n} y_i x_i}{\sum_{j=1}^{n} x_j^2}\right] = \frac{\sum_{i=1}^{n} x_i E[y_i]}{\sum_{j=1}^{n} x_j^2} = \frac{\sum_{i=1}^{n} x_i E[\beta x_i + v_i]}{\sum_{j=1}^{n} x_j^2}$$
$$= \frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{j=1}^{n} x_j^2} = \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{j=1}^{n} x_j^2} = \beta.$$

In the above steps, an assumption is that $E[\beta x_i + v_i] = \beta x_i$ since $E(v_i|X_i) = 0$. It can be seen that as $n \to \infty$, the expectation of $\hat{\beta}$ remains unchanged at β.

From part b), it has been shown that,

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

The numerator is a constant, σ^2. The denominator is a summation of terms indexed by i for $i = 1, \cdots, n$. Then it can be seen that as $n \to \infty$, $Var(\hat{\beta}) = 0$.

Therefore, it has been shown that since as $n \to \infty$ that the plim of $\hat{\beta}$ approaches its true value β, then it is a consistent estimator. ■

3. The columns in the file `polynomial_data.txt` are the X and Y values of a polynomial function $y = \sum_{k=0}^{n} a_k x^k$ with added Gaussian noise.
    a. For each $n \in \{3,4,5\}$, fit an n^{th} degree polynomial to the data. What would you say is the most likely value of n?

Ans:

The first step was to create the relevant design matrices, $\boldsymbol{X}$, for each $n \in \{3,4,5\}$. This involves creating a column of ones, $\mathbf{1}_n$, to represent the intercept term in the model. The terms $X^2, \cdots, X^5$ are created simply by taking X to the n^{th} power for $n \in \{3,4,5\}$. The design matrices are then simply the concatenation of the $\mathbf{1}_n$ column and the corresponding X^n variables.

The next step is to calculate the $\widehat{\boldsymbol{\beta}}$ term, which is as follows,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Then, the corresponding estimates $\widehat{\boldsymbol{y}}$, can be calculated as follows,

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{\beta}} \boldsymbol{X}.$$

From here, the R_{Adj}^2 is calculated, which has the following formula,

$$R_{Adj}^2 = 1 - \frac{\frac{SS_{res}}{df_e}}{\frac{SS_{tot}}{df_t}},$$

where $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$, $SS_{tot} = \sum_i (y_i - \bar{y})^2$, $\frac{1}{n}\sum_i y_i$, $df_e = n - p - 1$, $df_t = n - 1$ (where here n refers to the number of observations in the dataset i.e. 50), and p is the number of parameters (e.g., for polynomial $n = 3$, $p = 3$, etc.).

Based on the above formulas, they were plugged into RStudio to calculate the R_{Adj}^2 for each of the n^{th} degree polynomial models.

	$n = 3$	$n = 4$	$n = 5$
R^2_{Adj}	≈ 0.9637	≈ 0.9784	≈ 0.9787

Table 1 The above table shows the corresponding R^2_{Adj} for each of the n^{th} degree polynomial models.

From the above table, it can be seen that $n = 5$ has the best R^2_{Adj}. However, comparing the three models, the jump from $n = 3$ to $n = 4$ seems relatively large. The difference between $n = 4$ and $n = 5$ though is comparatively quite small. Since adding more degrees to the model increases the complexity of the model that is based on the sample dataset, the generalization ability tends to weaken. Therefore, it seems that the most likely value of n is 4.

b. Estimate the level of the noise.

Ans:

With linear regression, often times it is assumed that the random error, ε is normally distributed with mean zero and some unknown variance σ^2, i.e., $\varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. To estimate the level of noise in the dataset, then we can try to look at this random error term based on the $n = 4$ polynomial model calculated in part a). We can do this by analyzing the residuals, $e_i = y_i - \hat{y}_i$, from the sample data. These terms are analogous to the random error, ε.

First, we can try to check the normality of these residuals using a Q-Q plot (a.k.a. quantile-quantile plot). Below in Figure 1, it shows that the residuals based on the $n = 4$ model fall quite closely on the line. The sorted residuals are plotted against a normal distribution with mean and standard deviation based on the residuals that are derived from the sample data. This seems to indicate that there's a strong probability that the residuals themselves are normally distributed, with mean equal to the mean of the residuals and variance equal to the variance of the residuals.

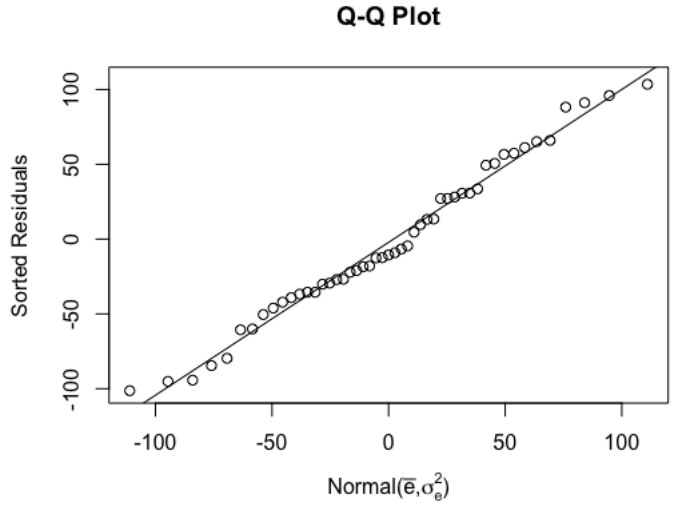

Figure 1 The above figure shows a Q-Q plot of the residuals from the $n = 4$ degree polynomial model. They are compared to a normal distribution with a mean and standard deviation based on the set of residuals. Additionally, a fitted line is plotted through the data to show how linear the data are.

The mean and variance calculated from the sample of residuals, e_i, can be seen in the table below. The sample mean is quite close to zero, and the sample variance is roughly 2,922.326. The calculations used are,

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{50} e_i \text{ and } s^2 = \frac{1}{n-1}\sum_{i=1}^{50}(e_i - \bar{e})^2,$$

for the sample mean and sample variance respectively. These are then used as the unbiased estimates, for the population mean ($\hat{\mu}$) and population variance ($\hat{\sigma}^2$). Therefore, it is estimated that the level of the noise is $\varepsilon \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, where μ is estimated with $\hat{\mu}$ and σ^2 is estimated with $\hat{\sigma}^2$.

Sample Mean	*Sample Variance*
$\approx -2.3019 \times 10^{-12}$	2922.326

Table 2 The above table shows the sample mean and sample variance for the residuals, e_i, calculated from the sample data using the polynomial model with degree $n = 4$.

References

[1] Montgomery, D.C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

[2] Piazza post 109. Can be found at https://piazza.com/class/kc0jkwru805u1?cid=109.

[3] Dougherty, C. (2016). Introduction to Econometrics (5th ed.). Oxford University Press.

Code Appendix

```
# Load data
polynomial_data <- read.csv('polynomial_data.txt', header = FALSE, sep = "")
colnames(polynomial_data) <- c("X", "Y")

# Create initial variables
n <- nrow(polynomial_data)
X <- polynomial_data[,1]
Y <- polynomial_data[,2]

# Calculate the nth degree for X
nth_degree <- function(x, n) {
  return(x^n)
}

x2 <- nth_degree(x = X, n = 2)
x3 <- nth_degree(x = X, n = 3)
x4 <- nth_degree(x = X, n = 4)
x5 <- nth_degree(x = X, n = 5)

# Create the matrices for the nth degree polynomials
ones <- rep(1, n)
X3 <- cbind(ones, X, x2, x3)
X4 <- cbind(ones, X, x2, x3, x4)
X5 <- cbind(ones, X, x2, x3, x4, x5)

# Calculate the beta hat
beta_hat <- function(X, y) {
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta_hat)
}

beta_hat_3 <- beta_hat(X = X3, y = Y)
beta_hat_4 <- beta_hat(X = X4, y = Y)
beta_hat_5 <- beta_hat(X = X5, y = Y)

# Calculate the nth degree y-hat
y3 <- X3 %*% beta_hat_3
```

```r
y4 <- X4 %*% beta_hat_4
y5 <- X5 %*% beta_hat_5

# Calculate the adjusted R-squared
adj_r_squared <- function(X, betahat, y) {
  y_bar <- mean(y)
  y_hat <- X %*% betahat

  SS_tot <- sum((y - y_bar)^2)
  SS_res <- sum((y - y_hat)^2)

  n <- length(y)
  p <- ncol(X) - 1
  df_e <- n - p - 1
  df_t <- n - 1

  adj_r_square <- 1 - ((SS_res / df_e) / (SS_tot / df_t))
  return(adj_r_square)
}

adj_r_squared_3 <- adj_r_squared(X = X3, betahat = beta_hat_3, y = Y)
adj_r_squared_4 <- adj_r_squared(X = X4, betahat = beta_hat_4, y = Y)
adj_r_squared_5 <- adj_r_squared(X = X5, betahat = beta_hat_5, y = Y)

### Chosen model: n=4
y_hat_4 <- X4 %*% beta_hat_4
e_4 <- Y- y_hat_4
mean(e_4)
sd(e_4)

sorted_residuals <- sort(e_4)
e_quantile <- qnorm((1:n)/n, mean = mean(e_4), sd = sd(e_4))
plot(e_quantile, sorted_residuals, main = 'Q-Q Plot',
     xlab = latex2exp::TeX('$Normal(\\bar{e}, \\sigma_{e}^2)$'),
     ylab = latex2exp::TeX('Sorted Residuals'))
df <- data.frame(x = e_quantile, y = sorted_residuals)
fit <- lm(formula = y~x, data = df[1:(n-1),])
abline(fit)
```