POL 213, Spring 2024
Problem Set 2 Solution

TA: Yu-Shiuan (Lily) Huang

1 Regression Mechanics

In this exercise, we will work with data on 2015 home prices. The data are from Prof. Colin Cameron (Economics, UC Davis). Load the data file `AED_HOUSE2015.RDS`. The house sale price is the response variable and other home attributes are the explanatory variables.

(a.) Compute the least squares regression of the response on the explanatory variables, interpreting the values that you obtain for the regression intercept A and slopes B_j , along with the standard error of the regression SE, and the multiple-correlation coefficient R.

For this analysis, I regress the housing price (`price`) on all the explanatory variables provided in the dataset (`size`, `bedrooms`, `bathrooms`, `days on market`, `region`). To improve interpretability, I divide the housing price by 1000 before running the regression.

As shown in Table 1, the expected housing price is \$281.39 thousand when all continuous explanatory variables are equal to 0 and the house is located in Central California. A one-square foot increase in house size is significantly associated with a \$0.31 thousand increase in housing price. Conversely, each additional day a house spends on the market before sale is associated with a \$0.79 thousand decrease in housing price. Additionally, while houses located in North California are associated with a lower housing price by \$134.26 thousand compared to houses in Central California, there is no significant difference in housing prices between East, South, and West regions and Central. Furthermore, the number of bedrooms and bathrooms does not have a significant effect on housing price.

The residual standard error is 73.52, and R^2 is around 93%, indicating that the five explanatory variables account for 93% of the variance in housing price.

Table 1: Factors Explaining California Housing Price in 2015

	Dependent variable:
	Housing Price
Size	0.31***
	(0.04)
Bedrooms	−22.24
	(33.30)
Bathrooms	−42.69
	(40.15)
Days on Market	−0.79*
	(0.34)
Region: East	−64.58
	(44.53)
Region: North	−134.26*
	(54.58)
Region: South	32.61
	(47.95)
Region: West	−34.32
	(59.14)
Constant	281.39**
	(85.11)
Observations	32
R^2	0.93
Adjusted R^2	0.90
Residual Std. Error	73.52 (df = 23)
F Statistic	35.83*** (df = 8; 23)
Note:	*p<0.05; **p<0.01; ***p<0.001

```r
# load data
house <- readRDS("/Users/yu-shiuanhuang/Desktop/method-sequence/data/AED_HOUSE2015.RDS") %>%
  mutate(region.f = factor(region,
                           level = c("central", "east",
                                     "north", "south", "west")),
         price_1000 = price/1000)

# 1-a
mod <- lm(price_1000 ~ size + bedrooms + bathrooms + daysonmarket + region.f,
          data = house)
stargazer(mod,
          title = "Table 1: Factors Explaining California
                   Housing Price in 2015",
          dep.var.labels = "Housing Price",
          covariate.labels = c("Size", "Bedrooms", "Bathrooms",
                               "Days on Market", "Region: East",
                               "Region: North", "Region: South",
                               "Region: West"),
          star.cutoffs = c(0.05, 0.01, 0.001), digits = 2)
```

(b.) Using a computer program to conveniently perform matrix computations, and working with the regression model in part (a.), compute the least squares regression coefficients as $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Please refer to the R code below, where I have performed matrix computations using the given formula. I obtained the same estimates of the intercept and coefficients as provided in Table 1.

```
     intercept      size        bed       bath        days      east
[1,]  281.3899 0.3071032 -22.24001 -42.68922 -0.7933985 -64.58102
         north   south      west
[1,] -134.2606 32.6052 -34.31912
```

```r
# 1-b
# create dummies for region
house <- fastDummies::dummy_cols(house, select_columns = "region")

# convert each variable to a vector
price_1000 <- house$price_1000
intercept <- rep(1, nrow(house)) # create this vector with 1s!
size <- house$size
bed <- house$bedrooms
bath <- house$bathrooms
days <- house$daysonmarket
east <- house$region_east
north <- house$region_north
south <- house$region_south
west <- house$region_west

X <- cbind(intercept, size, bed, bath, days, east, north, south, west)

beta.hat <- solve(t(X) %*% X) %*% t(X) %*% price_1000
t(beta.hat)

```

(c.) Verify that the least squares slope coefficients $\mathbf{b}_{new} = [B_1, B_2, ..., B_k]'$ can be computed as $\mathbf{b}_{new} = (\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}\mathbf{X}^{*'}\mathbf{y}^{*}$ where $\mathbf{X}^{*}$ and $\mathbf{y}^{*}$ contain mean deviations for the X's and Y's, respectively. What does this say about centering (transforming to mean deviations) your data before conducting a linear regression analysis?

Please refer to the updated R code below, where I have performed matrix computations using the given formula after transforming all continuous variables to mean deviation, except for the region dummy variables, which remain as binary variables specifying the location of a property. While I obtained the same coefficient estimates as provided in Table 1, the intercept differs.

The new intercept, also representing the baseline expected housing price, is now $40.26 thousand when all continuous variables are at their mean (i.e., 0 after mean deviation transformation) and the house is located in Central California. This discrepancy in intercept arises because centering changes the values of the predictors without altering their scale. Consequently, while a predictor centered at the mean has new values, the entire baseline shifts so that the mean now has a value of 0, but one unit remains equivalent to one unit. Thus, the intercept changes because we shift the baseline of the data.

Importantly, this change in intercept does not affect the interpretation of the regression coefficients, which continue to represent the effect on the mean of the response variable for each one unit difference in a predictor.

```
     intercept    size_m     bed_m    bath_m     days_m      east
[1,]  40.26223 0.3071032 -22.24001 -42.68922 -0.7933985 -64.58102
         north   south      west
[1,] -134.2606 32.6052 -34.31912
```

```r
# 1-c
# transform continuous variables to mean deviation
price_1000_m <- house$price_1000 - mean(house$price_1000 )
intercept <- rep(1, nrow(house)) # create this vector with 1s!
size_m <- house$size - mean(house$size)
bed_m <- house$bedrooms - mean(house$bedrooms)
bath_m <- house$bathrooms - mean(house$bathrooms)
days_m <- house$daysonmarket - mean(house$daysonmarket)
X_m <- cbind(intercept, size_m, bed_m, bath_m, days_m, east, north, south,
    west)

beta.hat.m <- solve(t(X_m) %*% X_m) %*% t(X_m) %*% price_1000_m
t(beta.hat.m)

```

2 Variance of Regression Parameters

(a.) Using the assumptions of linearity, constant variance, and independence, along with the fact that A and B can each be expressed as a linear function of the Y_i s, derive the sampling variances of A and B in a simple regression. [*Hint:* $V(B) = \sum m_i^2 V(Y_i)$] Show every step of your calculation.

Here are some reviews before starting the below problems:

- The assumed model: $Y_i = \alpha + \beta X_i + \varepsilon_i$
- Linearity assumption: $E(\varepsilon_i) = 0$
- Normality assumption: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
- Independence assumption: $Cov(\varepsilon_i, \varepsilon_j) = 0, fori \neq j$
- The expectation of linear combination: $E(a + bY) = a + bE(Y)$
- The variance of linear combination:
  - $Var(a + bY) = b^2 Var(Y)$
  - $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- For the below demonstration, the X_i are assumed to be fixed, not random.
- $B = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma(X_i - \bar{X})Y_i - \Sigma(X_i - \bar{X})\bar{Y}}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma(X_i - \bar{X})Y_i - \bar{Y}\Sigma(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2}$
- $\Sigma(X_i - \bar{X})^2 = \Sigma(X_i - \bar{X})(X_i - \bar{X}) = \Sigma X_i(X_i - \bar{X}) - \bar{X}\Sigma(X_i - \bar{X}) = \Sigma X_i(X_i - \bar{X})$

$$\begin{aligned}
Var(B) &= Var(\frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2}) \\
&= \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} Var(\Sigma(X_i - \bar{X})Y_i) \\
&= \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} Var(\Sigma(X_i - \bar{X})(\alpha + \beta X_i + \varepsilon_i)) \\
&= \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} Var(\Sigma(X_i - \bar{X}(\alpha + \beta X_i) + \Sigma(X_i - \bar{X})\varepsilon_i) \\
&= \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} Var(\Sigma(X_i - \bar{X})\varepsilon_i) = \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} \Sigma Var((X_i - \bar{X})\varepsilon_i) \\
&= \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} (\Sigma(X_i - \bar{X})^2 Var(\varepsilon_i)) \\
&= \frac{1}{(\Sigma(X_i - \bar{X})^2)^2} (\Sigma(X_i - \bar{X})^2 \sigma_\varepsilon^2) \\
&= \frac{\sigma_\varepsilon^2}{\Sigma(X_i - \bar{X})^2} \\
Var(A) &= Var(\bar{Y} - B\bar{X}) = Var(\bar{Y}) + Var(B\bar{X}) - 2Cov(\bar{Y}, B\bar{X}) \\
&= Var(\bar{Y}) + Var(B\bar{X}) = Var(\frac{\Sigma Y_i}{n}) + \bar{X}^2 Var(B) \\
&= \frac{1}{n^2} Var(\Sigma Y_i) + \bar{X}^2 Var(B) \\
&= \frac{1}{n^2} \Sigma Var(Y_i) + \bar{X}^2 Var(B) \\
&= \frac{1}{n^2} \Sigma Var(\alpha + \beta X_i + \varepsilon_i) + \bar{X}^2 Var(B) \\
&= \frac{1}{n^2} \Sigma Var(\varepsilon_i) + \bar{X}^2 Var(B) \\
&= \frac{1}{n^2} n\sigma_\varepsilon^2 + \bar{X}(\frac{\sigma_\varepsilon^2}{\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\frac{\Sigma(X_i - \bar{X})^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\frac{\Sigma X_i^2 - 2\bar{X}\Sigma X_i + \Sigma \bar{X}^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\frac{\Sigma X_i^2 - 2\bar{X}\Sigma X_i + 2n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\frac{\Sigma X_i^2 - 2\frac{\Sigma X_i}{n}\Sigma X_i + 2n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}) \\
&= \sigma_\varepsilon^2(\frac{\Sigma X_i^2 - \frac{2\Sigma X_i^2}{n} + 2n\frac{\Sigma X_i^2}{n^2}}{n\Sigma(X_i - \bar{X})^2}) \\
&= \frac{\sigma_\varepsilon^2 \Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2}
\end{aligned}$$

(b.) The formula for the sampling variance of B in simple regression

$$V(B) = \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}$$

shows that to estimate β precisely, it helps to have spread out xs. Explain why this result is intuitively sensible, illustrating your explanation with a graph. What happens to $V(B)$ when there is *no* variation in X?

In simple linear regression, the sampling variance of the slope coefficient B reveals that having spread out values of the independent variable x contributes to estimating B more precisely. This intuitively makes sense because when the values of x are spread out, they cover a wider range of the relationship between the independent and dependent variables, providing more information to estimate the slope accurately.

Graphically, the left panel in Figure 1 presents a scatter plot where the data points are spread out across the x-axis. In this scenario, the regression line has more data points to consider when determining the slope, resulting in a more stable and reliable estimate of B. Conversely, if the data points are clustered closely together along the x-axis, as seen in the right panel of Figure 1, the regression line may not capture the true underlying relationship effectively. This occurs because the closely clustered data does not provide sufficient information about the trend pattern between x and y, leading to a less precise estimate of B.

When there is no variation in x, meaning all the data points lie on a single vertical line, the sampling variance of B becomes infinitely large. This is because without variability in the independent variable, the slope of the regression line cannot be estimated reliably. Essentially, every possible slope becomes equally plausible, leading to a highly uncertain estimate of B. Therefore, the sampling variance of B increases dramatically when there is no variation in x.

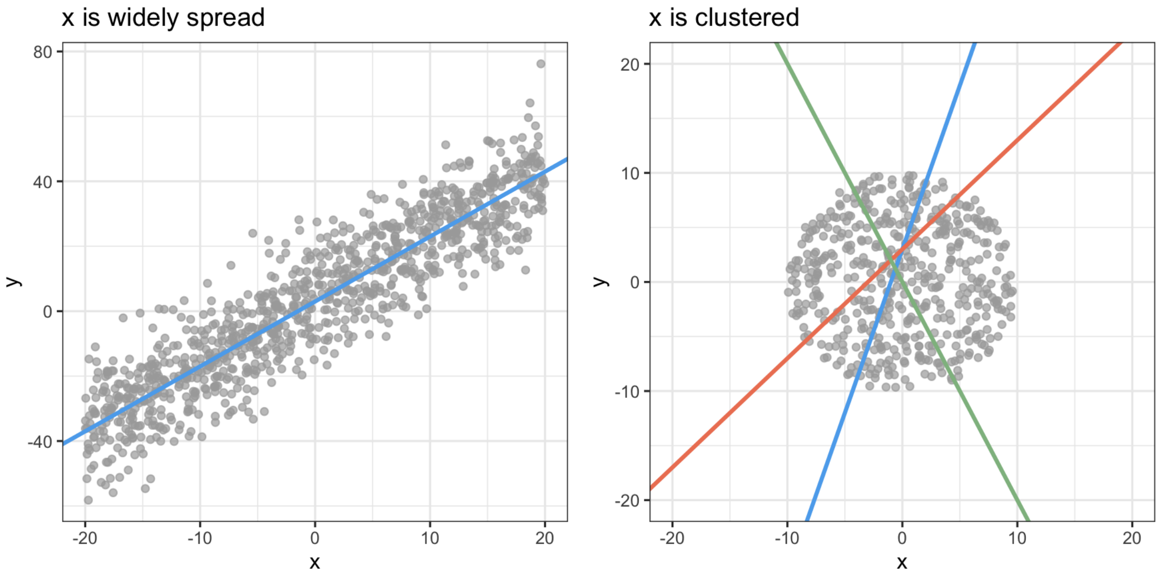

Figure 1: Two Scenarios: x is widely spread vs x is clustered

3 Nonconstant Variance and Specification Error

(a.) Perform a simulation. Generate 100 observations according to the following model:

$$Y = 10 + X + D + 2 \times X \times D + \epsilon$$

where $\epsilon \sim N(0, 10^2)$; the values of X are $[1, 2, ...50, 1, 2, ...50]$; the first 50 values of D are 0 and the last 50 values of D are 1. Here is some code to get you started:

```
eps <- rnorm(100, 0, 10)
X <- rep(1:50, 2)
D <- c(rep(0, 50), rep(1, 50))
Y <- 10 + X + D + (2*X*D) + eps
```

Regress Y on X alone (i.e. omitting D and XD) such that you estimate $Y = A + BX + E$. Then plot the residuals E from this regression against the fitted values $\hat{Y}$. Is the variance of the residuals constant? How do you account for the pattern in the plot? Explain what this implies about linear model suitability.

Table 2 presents the regression results, and Figure 2 demonstrates the residuals versus fitted values. However, since the true Y is affected by both X and D, running a regression only with X leads to an estimate of the coefficient that is far away from the true β_X (which should be 1 according to the true model). Furthermore, Figure 2 shows that the variance of the residuals is not constant, which is due to the wrong model specification in Table 2. In the estimated model, without including D and XD, residual of each observation depends on whether D is 1 or 0. When D is 1/0, residuals will be larger/smaller, as shown in the increasing/decreasing pattern in Figure 2. This incorrect model specification violates the homoskedasticity assumption, leading to inefficient estimation of coefficients (i.e., high standard error of $\hat{\beta}$).

Table 2: Ordinary least-squares estimates of X on Y

	Dependent variable:
	Y
X	2.04***
	(0.22)
Constant	9.12
	(6.46)
Observations	100
R^2	0.47
Adjusted R^2	0.46
Residual Std. Error	31.82 (df = 98)
F Statistic	85.64*** (df = 1; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01

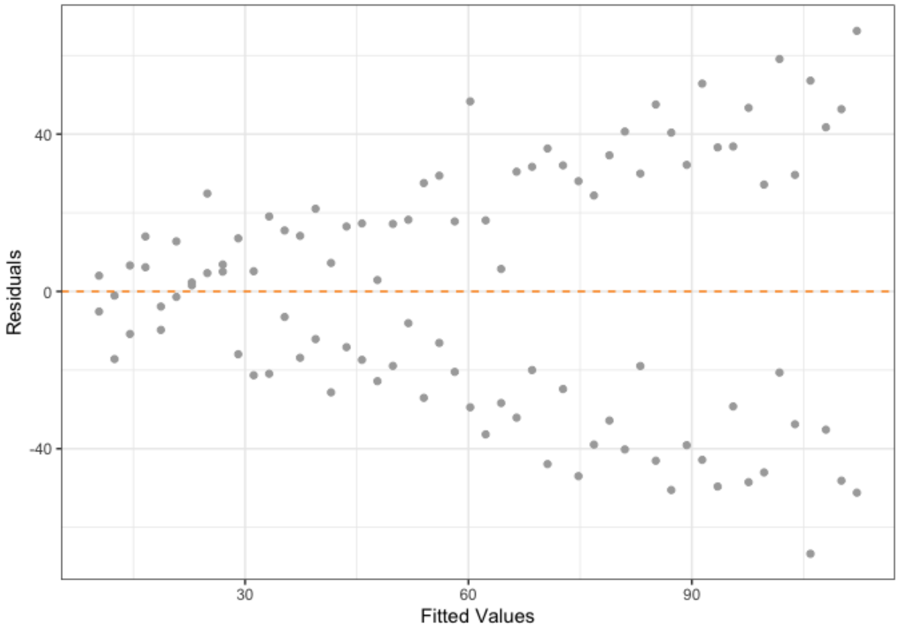

Figure 2: Residuals versus Fitted Values

```r
# generate 100 observations
eps <- rnorm(100, 0, 10)
X <- rep(1:50, 2)
D <- c(rep(0, 50), rep(1, 50))
Y <- 10 + X + D + (2*X*D) + eps

mod <- lm(Y ~ X)
stargazer(mod, digits = 2)

ggplot() +
    geom_point(aes(x = mod$fitted.values, y = mod$residuals),
               color = "darkgray") +
    geom_hline(yintercept = 0, linetype = "dashed", color = "darkorange") +
    xlab("Fitted Values") + ylab("Residuals") +
    theme_bw()
```

(b.) **Challenge Problem** (Exercise 12.4 in Fox textbook) Show that when the covariance matrix of the errors is

$$\Sigma = \sigma_\epsilon^2 \times \text{diag}\{1/\omega_1^2 \dots 1/\omega_n^2\} \equiv \sigma_\epsilon^2 \times \mathbf{W}^{-1}$$

the weighted least squares estimator

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} = \mathbf{M}\mathbf{y}$$

is the minimum-variance linear unbiased estimator of β. (*Hint:* Adapt the proof of the Gauss-Markov theorem for OLS estimation given in section 9.3.2.)

For this question, I first derive the expected value of $\hat{\beta}_{\text{WLS}}$ to prove that the weighted least squares estimator is an unbiased estimator of β. Next, I derive the variance of $\hat{\beta}_{\text{WLS}}$ and demonstrate that any other linear estimator, such as $\tilde{\beta}$, even if unbiased, possesses a larger variance than $\hat{\beta}_{\text{WLS}}$. This suggests that $\hat{\beta}_{\text{WLS}}$ is the minimum-variance linear unbiased estimator of β.

1. $\hat{\beta}_{\text{WLS}}$ is unbiased.

   $$\begin{aligned} E(\hat{\beta}_{\text{WLS}}) = E(\mathbf{My}) = \mathbf{M}E(\mathbf{y}) \\ &= (\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W}(\mathbf{X}\beta) \\ &= \mathbf{X^{-1}W^{-1}}(\mathbf{X'})^{\mathbf{-1}}\mathbf{X'WX}\beta \\ &= \mathbf{X^{-1}W^{-1}WX}\beta \\ &= \mathbf{X^{-1}X}\beta \\ &= \beta \end{aligned}$$

   In the third line, we used identity matrix: $1 = (\mathbf{X'})^{\mathbf{-1}}\mathbf{X'}$
   In the fourth line, we used identity matrix: $1 = \mathbf{W^{-1}W'}$
   In the fifth line, we used identity matrix: $1 = \mathbf{X^{-1}X'}$

2. $Var(\hat{\beta}_{\text{WLS}})$ is the linear unbiased estimator that possesses the minimum variance.

   First,
   $$\begin{aligned} Var(\hat{\beta}_{\text{WLS}}) &= \mathbf{M}V(y)\mathbf{M'} \\ &= [(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W}]\sigma_\epsilon^2\mathbf{W}^{-1}[(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W}]' \\ &= \sigma_\epsilon^2(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'WW^{-1}}([(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W}]') \\ &= \sigma_\epsilon^2(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'}([(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W}]') \\ &= \sigma_\epsilon^2(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'}(\mathbf{W'X''}(\mathbf{X'WX})^{\mathbf{-1'}}) \\ &= \sigma_\epsilon^2(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W'X}(\mathbf{X'WX})^{\mathbf{-1'}} \\ &= \sigma_\epsilon^2(\mathbf{X'XW})^{\mathbf{-1}} \end{aligned}$$

   where we used the fact that $(\mathbf{X'WX})^{\mathbf{-1'}} = (\mathbf{X'WX})^{\mathbf{-1}}$ due to symmetry.

   Next, any linear estimator, say $\tilde{\beta}$, could be written as $\tilde{\beta} = (\mathbf{M}+\mathbf{A})y$. We will show that if $\tilde{\beta}$ is unbiased, then it has larger variance than $\hat{\beta}_{\text{WLS}}$.

   For $\tilde{\beta}$ to be an unbiased estimator, the matrix product $\mathbf{AX}\beta$ must be 0. Thus, regardless of the value of β, $\mathbf{AX}$ must be 0.
   $\beta = E(\tilde{\beta}) = E[(\mathbf{M}+\mathbf{A})y] = E(\mathbf{M}y) + E(\mathbf{A}y) = E(\hat{\beta}_{\text{WLS}}) + \mathbf{A}E(y) = \beta + \mathbf{AX}\beta$

   The covariance matrix of $\tilde{\beta}$ is given by:
   $$\begin{aligned} Var(\tilde{\beta}) &= (\mathbf{M}+\mathbf{A})V(y)(\mathbf{M}+\mathbf{A})' = (\mathbf{M}+\mathbf{A})\sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{M}+\mathbf{A})' \\ &= \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{MM'}+\mathbf{MA'}+\mathbf{AM'}+\mathbf{AA'}) \end{aligned}$$

   Since we have shown that $\mathbf{AX} = 0$, consequently, $\mathbf{AM'}$ and its transpose $\mathbf{MA'}$ are 0, for $\mathbf{AM'} = \mathbf{A}[(\mathbf{X'WX})^{\mathbf{-1}}\mathbf{X'W}]' = \mathbf{AX}(\mathbf{XW'X'})^{\mathbf{-1}}\mathbf{W'} = 0 \times (\mathbf{XW'X'})^{\mathbf{-1}}\mathbf{W'} = 0$. Thus,

$$\begin{aligned} Var(\tilde{\beta}) &= \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{MM}' + \mathbf{AA}') = \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{MM}') + \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{AA}') \\ &= \sigma_\epsilon^2\mathbf{W}^{-1}[(\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{W}][(\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{W}]' + \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{AA}') \\ &= \sigma_\epsilon^2(\mathbf{X}\mathbf{X}'\mathbf{W})^{-\mathbf{1}} + \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{AA}') \\ &= Var(\hat{\beta}_{\text{WLS}}) + \sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{AA}') \end{aligned}$$

As $\sigma_\epsilon^2\mathbf{W}^{-1}(\mathbf{AA}')$ is a positive definite term, we can conclude that $Var(\tilde{\beta}) > Var(\hat{\beta}_{\text{WLS}})$. Hence, $\hat{\beta}_{\text{WLS}}$ has the least variance among all possible linear, unbiased estimators of the regression coefficients.

4 Unusual and Influential Data

Choose your favorite data set. This can be data you worked on in POL 212 or another political science data set in which the dependent variable is continuous (not categorical). For the greatest educational value, choose a data set where the number of observations is roughly between 40 and 250. If you don't have a favorite data set, pick something from the Fox textbook website.

Perform the following steps.

(a.) Run a multivariate regression with at least 3 predictor variables. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the associated standard errors.

For this question, I use the `beauty.csv` data from Hamermesh and Parker (2005), which documented on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

Table 3 presents the results of a multivariate linear regression analysis that examines the relationship between course evaluations (`eval`) and the main explanatory variable, `beauty`, while controlling for several other variables (instructors' sex, age, minority status, whether he/she received an undergraduate education in an English-speaking country, and whether he/she is assigned to lower-division courses).

As depicted in Table 3, the coefficient of the composite standardized beauty measure is statistically significant at a level of 0.01, indicating that an instructor's average course rating increases by 0.14 units for every one-unit increase in their beauty rating. This represents a close to 25% increase in the standard deviation of the average class rating. Figure 3 shows a scatterplot between course evaluations and beauty ratings, as well as a predicted fitted class rating based on different beauty ratings while holding all other variables at their mean or median values. The predicted line has a positive slope, suggesting that an instructor's appearance is positively associated with their average class rating.

Additionally, I find that female and non-native English speaking instructors receive lower course ratings than their male and native English speaking counterparts, with a difference of 0.2 and 0.27 units, respectively. Furthermore, instructors who teach lower-division courses tend to receive higher course evaluations than those who teach higher-division courses, although this difference is only statistically significant at a level of 0.1. Finally, my analysis reveals no evidence of any significant effect of instructors' age or minority status on their average class rating. The residual standard error of this model, which measures the average vertical distance between the regression line and each observation, is 0.53.

Table 3: Ordinary least-squares estimates of the determinants of course evaluations.

	Dependent variable:
	Course Evaluations
Composite Standardized Beauty	0.14^{***}
	(0.03)
Female	-0.20^{***}
	(0.05)
Age	-0.002
	(0.003)
Minority	-0.07
	(0.08)
Non-native English	-0.27^{**}
	(0.11)
Lower division	0.10^{*}
	(0.05)
Constant	4.19^{***}
	(0.15)
Observations	463
R^2	0.10
Adjusted R^2	0.08
Residual Std. Error	0.53 (df = 456)
F Statistic	8.05^{***} (df = 6; 456)
Note:	*p<0.1; **p<0.05; ***p<0.01

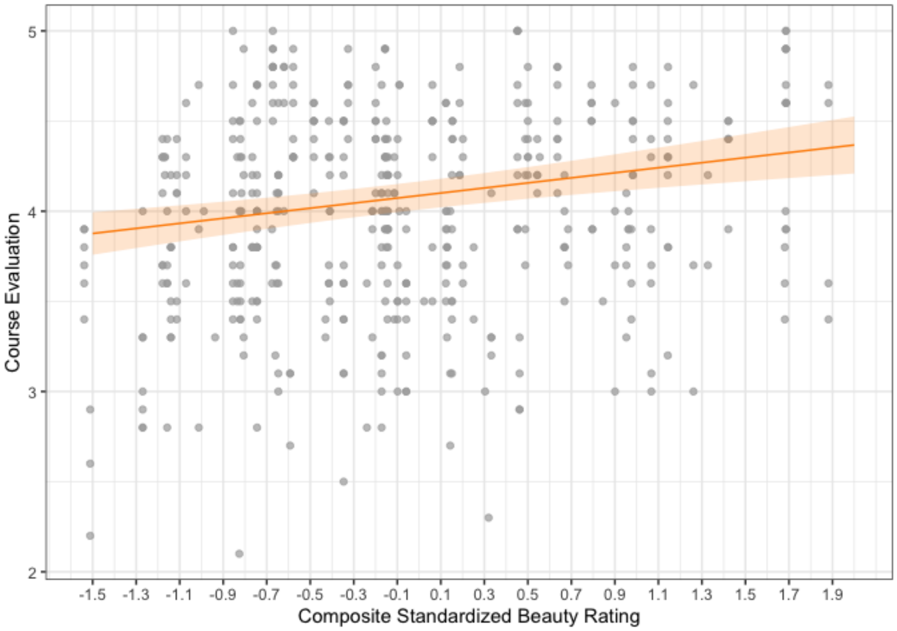

Figure 3: Predicted Course Evaluation

```r
## load data
df <- read.csv("data/beauty.csv")

## regression
mod1 <- lm(eval ~ beauty + as.factor(female) + age +
           as.factor(minority) + as.factor(nonenglish) +
           as.factor(lower), data = df)

library(stargazer)
stargazer(mod1, digits = 2)

## display the fitted model graphically
summary(df)

beauty <- seq(-1.5, 2, 0.05) # beauty: -1.5~2
pred.df <- data.frame(beauty = beauty,
                      female = rep(0, length(beauty)),
                      age = rep(48.37, length(beauty)),
                      minority = rep(0, length(beauty)),
                      nonenglish = rep(0, length(beauty)),
                      lower = rep(0, length(beauty)))
plot.df <- data.frame(beauty = beauty,
                      predict(mod1, newdata = pred.df,
                              interval = "confidence", level = 0.95))
library(tidyverse)
library(ggrepel)
ggplot() +
    geom_point(data = df, aes(x = beauty, y = eval),
               alpha = 0.7, color = "darkgray") +
    geom_ribbon(data = plot.df, aes(x = beauty, ymin = lwr, ymax = upr),
                fill = "darkorange", alpha = 0.2) +
    geom_line(data = plot.df, aes(x = beauty, y = fit),
              color = "darkorange") +
    scale_x_continuous(name = "Composite Standardized Beauty Rating",
                       breaks = seq(-1.5, 2, 0.2)) +
    ylab("Course Evaluation") +
    theme_bw()

```

(b.) Plot the residuals versus fitted values and interpret your results.

I examined the Residuals vs Fitted plot (Figure 4) to assess the linearity and homoskedasticity assumptions. The plot displays the predicted values of the outcome variable on the horizontal axis and the residuals (i.e., the differences between the observed and predicted values) on the vertical axis. Ideally, the residual plot should show no discernible pattern, with the red line being approximately horizontal at zero, indicating that the expected (mean) value of the disturbance term is zero. If a pattern emerges, it could signify issues with the linear model. According to Figure 4, the red line exhibits no clear pattern, suggesting that the linearity assumption is not violated. Furthermore, the residuals form a roughly horizontal band around the zero line, indicating constant variances of the error terms and supporting the homoskedasticity assumption.

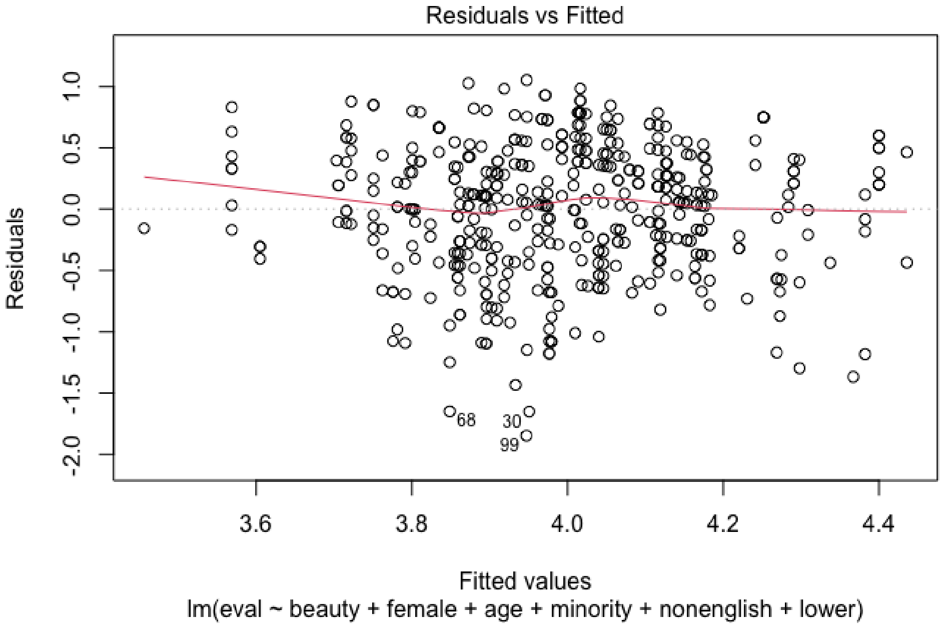

Figure 4: Residuals versus Fitted Values

```r
## asessing Gauss-Markov assumption
plot(mod1)

```

(c.) Evaluate the leverage of your data points and interpret your results. Be sure to explain the measure of leverage you use.

To determine whether there are any influential outliers in this dataset, I assess the leverage, discrepancy, and influence of each observation. Figure 5 displays the hat values of each observation, which measure leverage using the diagonal entries of the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The dashed orange and blue lines indicate the cutoff values at twice and three times the average hat value $\left(\bar{h} = \frac{6+1}{463} \approx 0.015\right)$, respectively. Leverage quantifies the extent to which a predictor value differs from the centroid of all predictors, while simultaneously controlling the correlational and variational structure of all predictors. As shown in Figure 5, several observations exceed these cutoff hat values, indicating that they may have a disproportionate influence on the regression results and warrant closer examination.

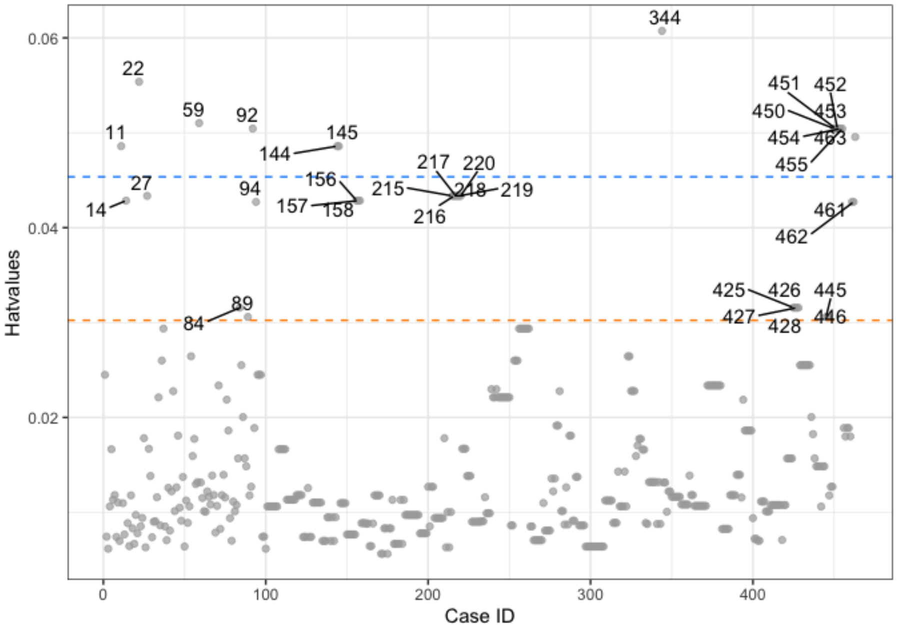

Figure 5: Hatvalues

```r
## assessing leverage
hat.df <- data.frame(case_id = as.numeric(names(hatvalues(mod1))),
                     hatvalues = hatvalues(mod1))
mean_hat <- (6+1)/nrow(df)

ggplot(hat.df, aes(x = case_id, y = hatvalues, label = case_id)) +
    geom_point(alpha = 0.7, color = "darkgray") +
    geom_hline(yintercept = 2*mean_hat, linetype = "dashed",
               color = "darkorange") +
    geom_hline(yintercept = 3*mean_hat, linetype = "dashed",
               color = "dodgerblue") +
    geom_text_repel(aes(label = ifelse(hatvalues >  2*mean_hat,
                                       as.character(case_id), "")),
                    hjust = -0.8, vjust = 0) +
    xlab("Case ID") + ylab("Hatvalues") +
    theme_bw()

```

(d.) Examine the data for any outliers. Which observations have the largest studentized residuals? What does this mean? Use words and equations to interpret.

I then proceed to evaluate the discrepancy of each observation. To this end, I plot the residuals and studentized residuals versus fitted course evaluations, respectively, in Figure 6. This approach allows me to assess the extent to which each observation deviates from the regression line. Since high-leverage observations have a tendency to produce small residuals, which can distort the regression surface, I complement the analysis with studentized residuals. Studentized residuals are calculated by fitting a model without the case for which the residual is calculated, and then scaling the resulting residual (E_i) by an estimate of the standard deviation of the residuals ($S_{E(-i)}$) and the point's hat value (h_i):

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1 - h_i}}$$

As shown in the right panel of Figure 6, case 99 has the highest studentized residuals with a value of -3.54. I also performed a Bonferroni adjustment to test the statistical significance of the studentized residuals for case 99 to determine if it can confidently be classified as an outlier. According to the test, the Bonferroni p-value for case 99 is no less than 0.05, leading to the conclusion that case 99 is not an outlier.

```
    No Studentized residuals with Bonferroni p < 0.05
    Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
99  -3.534709         0.00045001      0.20835
```

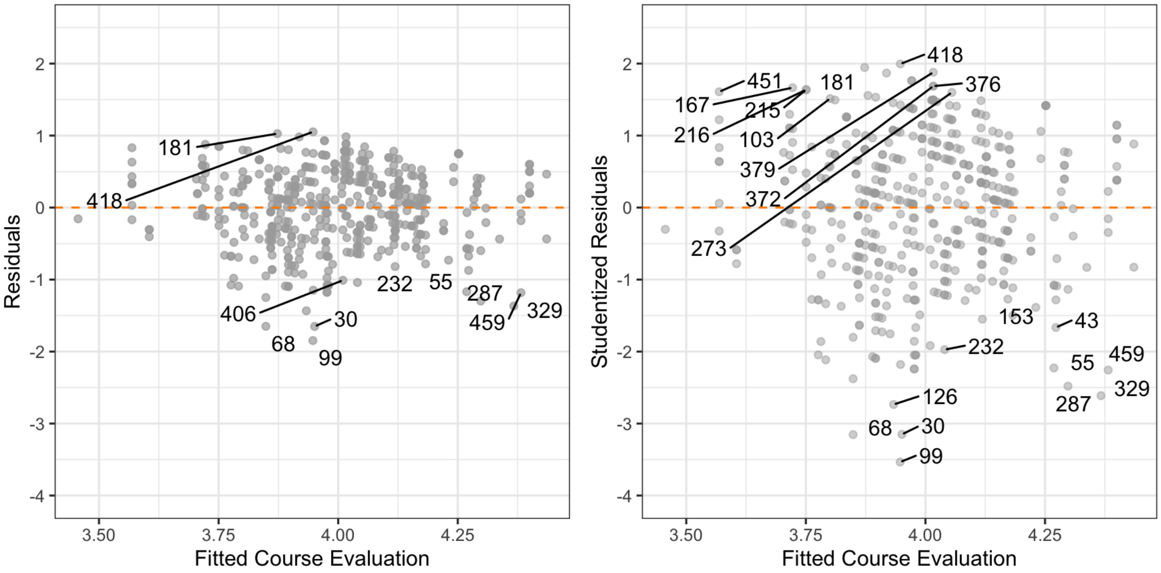

Figure 6: Residuals versus Fitted Values

```r
## assessing discrepancy
df2 <- data.frame(fit = mod1$fitted.values, residual = mod1$residuals,
                  rstudent = rstudent(mod1), case_id = c(1:nrow(df)))

outlierTest(mod1)

p1 <- ggplot(df2, aes(x = fit, y = residual, label = case_id)) +
    geom_point(alpha = 0.7, color = "darkgray") +
    geom_hline(yintercept = 0, linetype = "dashed",
               color = "darkorange") +
    geom_text_repel(aes(label = ifelse(residual < -1 | residual > 1,
                                       as.character(case_id), "")),
              hjust = -0.8, vjust = 0) +
    scale_y_continuous(breaks = seq(-3, 2.5, 1), limits = c(-3, 2.5)) +
    xlab("Fitted Course Evaluation") +
    ylab("Residuals") +
    theme_bw()


p2 <- ggplot(df2, aes(x = fit, y = rstudent, label = case_id)) +
    geom_jitter(alpha = 0.5, color = "darkgray") +
    geom_hline(yintercept = 0, linetype = "dashed",
               color = "darkorange") +
    geom_text_repel(aes(label = ifelse(rstudent < -1.5 | rstudent > 1.5,
                                       as.character(case_id), "")),
              hjust = -0.8, vjust = 0) +
    scale_y_continuous(breaks = seq(-3, 2.5, 1), limits = c(-3, 2.5)) +
    xlab("Fitted Course Evaluation") +
    ylab("Studentized Residuals") +
    theme_bw()

ggpubr::ggarrange(p1, p2)

```

(e.) Evaluate influence of these outlier observations with the Cook's D statistic. Which observations have the greatest influence? What does this mean? Use words and equations to interpret.

To assess influence, I calculate Cook's distance for each observation, which considers both hat values and studentized residuals (see below formula). Observations with high leverage and large studentized residuals can significantly impact the regression coefficients, suggesting that their exclusion could lead to major changes in the fitted regression function.

$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

Table 4 and Figure 7 display the six observations with the largest Cook's distance statistics in the data set. I use a cutoff of 0.5 to determine if any observations require further investigation. According to Table 4, observation 144 has the highest Cook's distance value but is still below the threshold of 0.5.

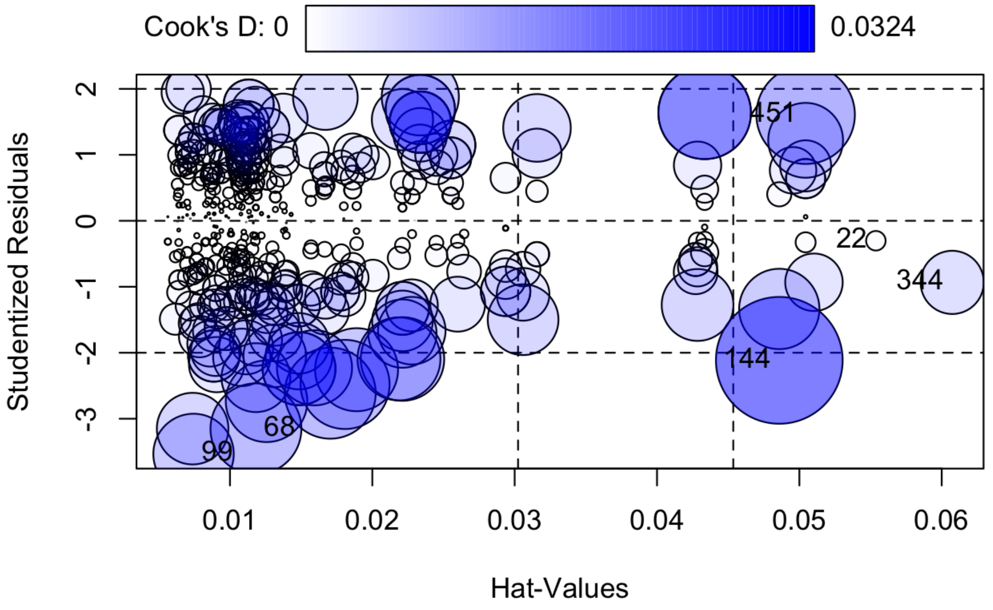

Figure 7: Influential Observations

Table 4: Cook’s Distance

Obs	Studentized Residuals	Hat Values	Cook’s Distance
22	-0.3025	0.0554	0.0008
68	-3.1536	0.0118	0.0167
99	-3.5347	0.0074	0.0130
144	-2.1152	0.0486	0.0324
344	-0.9335	0.0607	0.0081
451	1.6087	0.0504	0.0196

```r
## assessing influence
library(car)
influencePlot(mod1)

```

(f.) Test for nonlinearity with component plus residual plots. Transform the most problematic of your explanatory variables or alter the model specification to improve model fit. Explain what you did and why.

In Figure 8, the magenta line represents a smooth curve through the Component and Residual (C+R) plots, while the blue dashed line represents the regression coefficients for each explanatory variable from the multivariate regression model. A notable disparity between the component line and the residual line would suggest that the predictor does not exhibit a linear relationship with the response variable. The advantage of using C+R plots is that they provide insights into the partial relationship between the response variable and each explanatory variable while controlling for all other variables. As shown in Figure 8, among all variables included in the model, there is no significant difference between the magenta and blue dashed lines, suggesting that our model specification systematically captures the relationships. Thus, I did not perform any variable transformations or alter the model specification.

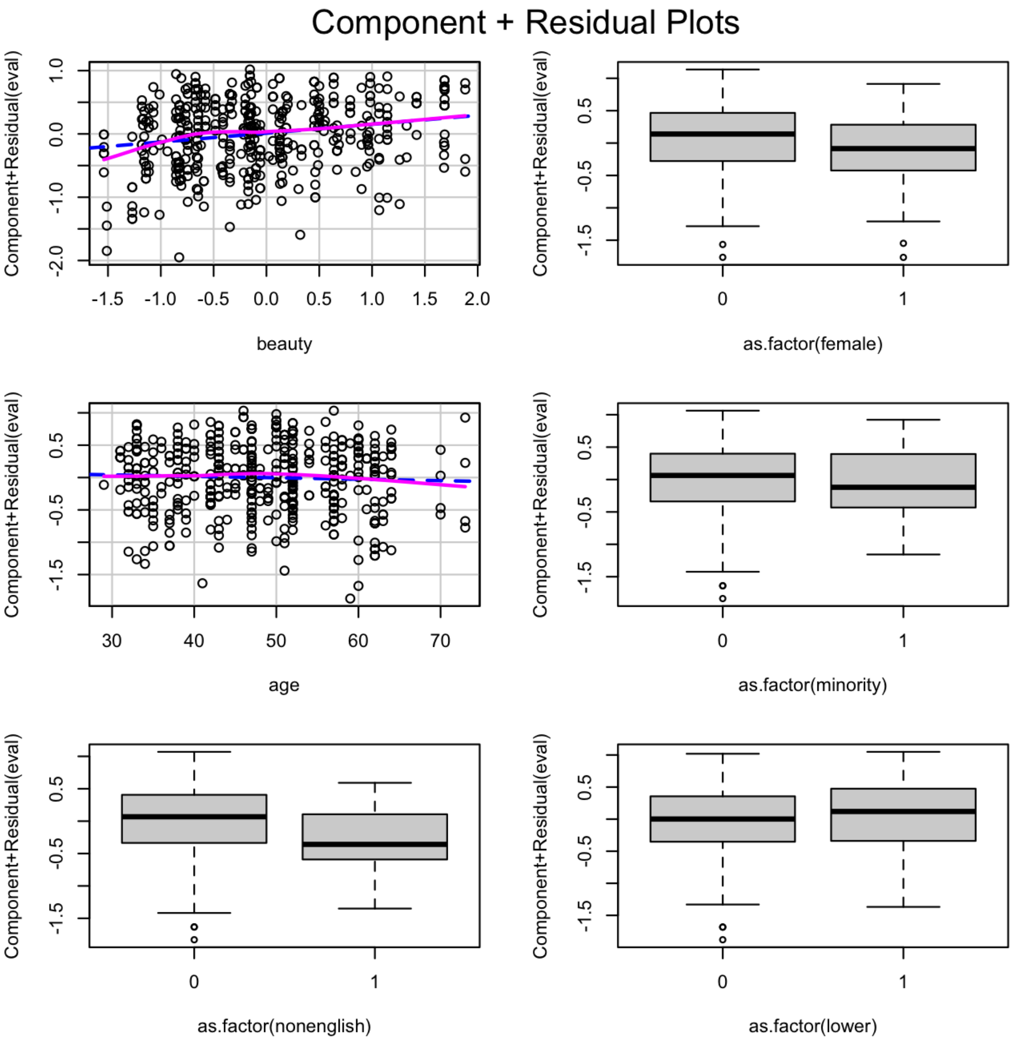

Figure 8: Component + Residual Plots

```
## assess linearity assumption using C+R plots
crPlots(mod1)

```

(g.) **Challenge Problem** The hat matrix transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. It is a projection matrix because it projects $\mathbf{y}$ orthogonally onto the subspace spanned by the columns of $\mathbf{X}$. Demonstrate that the hat-matrix $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}$ is symmetric ($\mathbf{H} = \mathbf{H'}$) and idempotent($\mathbf{H} = \mathbf{H^2}$). Explain in words what these properties for the hat matrix mean conceptually for interpreting the individual data points' contributions to the fitted model.

First, to demonstrate that $\mathbf{H}$ is symmetric, we need to show that $\mathbf{H} = \mathbf{H'}$.[1].

$$\mathbf{H'} = [\mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}]' = \mathbf{X}[(\mathbf{X'X})^{-\mathbf{1}}]'\mathbf{X'} = \mathbf{X}[(\mathbf{X'X})']^{-\mathbf{1}}\mathbf{X'} = \mathbf{X}(\mathbf{XX'})^{-\mathbf{1}}\mathbf{X'} = H$$

Second, to demonstrate that $\mathbf{H}$ is idempotent, we need to show that $\mathbf{H^2} = \mathbf{H}$.

$$\mathbf{H^2} = [\mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}][\mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}] = \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}(\mathbf{X'X})(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'} = \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'} = \mathbf{H}$$

[1] Here, we apply one of the properties of transpose: $(ABC)' = C'B'A'$

The symmetry of the hat matrix in multivariate regression means that it acts like a mirror, reflecting itself across a diagonal line. This property ensures that the order in which data points are entered does not affect the outcomes, promoting fairness and uniformity in how each data point is treated. Additionally, it maintains a balance in the influence that each observation has on itself and on others, ensuring that no single point skews the regression results disproportionately. On the other hand, the idempotence of the hat matrix means that once the observed responses are projected onto the space defined by the predictor variables to produce fitted values, applying the matrix again makes no further changes. This one-time application prevents over-adjustment and reassures us that the model predictions are stable and not overfitted to the sample data. These properties—symmetry ensuring consistent and impartial predictions regardless of data order, and idempotence confirming that predictions are definitive and resistant to overfitting—help create a regression model that is robust, reliable, and fair.

5 Challenge Problem

Under the assumptions of the linear model, the least-squares estimator **b** is also the maximum likelihood estimator of β. Review Section 9.3.3 in the Fox textbook to understand equation (9.10):

$$p(\mathbf{y}) = \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}} exp\left[-\frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma_\epsilon^2}\right]$$

Show that the maximum likelihood for the linear model can be written as:

$$L = \left[2\pi e \frac{\mathbf{e}'\mathbf{e}}{n}\right]^{-n/2}.$$

Hint: Solutions can be found in Fox's online resources. Don't just copy; make sure to explain each step to demonstrate you fully understand the derivation.

Following Section 9.3.3 in the Fox textbook, we understand that $p(\mathbf{y})$ represents the joint probability density function of the observations in the dataset. To identify parameters that maximize the likelihood of the observed data originating from the specified distribution, we first take the logarithm of the likelihood function. Subsequently, we perform partial differentiation of this log-likelihood function to derive estimators, $\hat{\beta}$ and $\hat{\sigma}_\epsilon^2$, that maximize the log-likelihood. Setting these partial derivatives to zero, we find optimal solutions where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}_\epsilon^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$, maximizing the log-likelihood. We then substitute $\widehat{\sigma_\epsilon^2} = \frac{\mathbf{e}'\mathbf{e}}{n}$ back into the likelihood function $p(\mathbf{y})$:

$$\begin{aligned}
L = p(\mathbf{y}) &= \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}} exp\left[-\frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma_\epsilon^2}\right] \\
&= \frac{1}{(2\pi\frac{\mathbf{e}'\mathbf{e}}{n})^{n/2}} exp\left[-\frac{\mathbf{e}'\mathbf{e}}{2\frac{\mathbf{e}'\mathbf{e}}{n}}\right] \\
&= (2\pi\frac{\mathbf{e}'\mathbf{e}}{n}) exp(-\frac{n}{2}) \\
&= \left[2\pi e \frac{\mathbf{e}'\mathbf{e}}{n}\right]^{-n/2}
\end{aligned}$$