**Math 455** **Midterm 2** **Spring 2024**

No books, no notes, no calculators.

Name: _____

1. You fit a linear regression in R and get the following summary output:

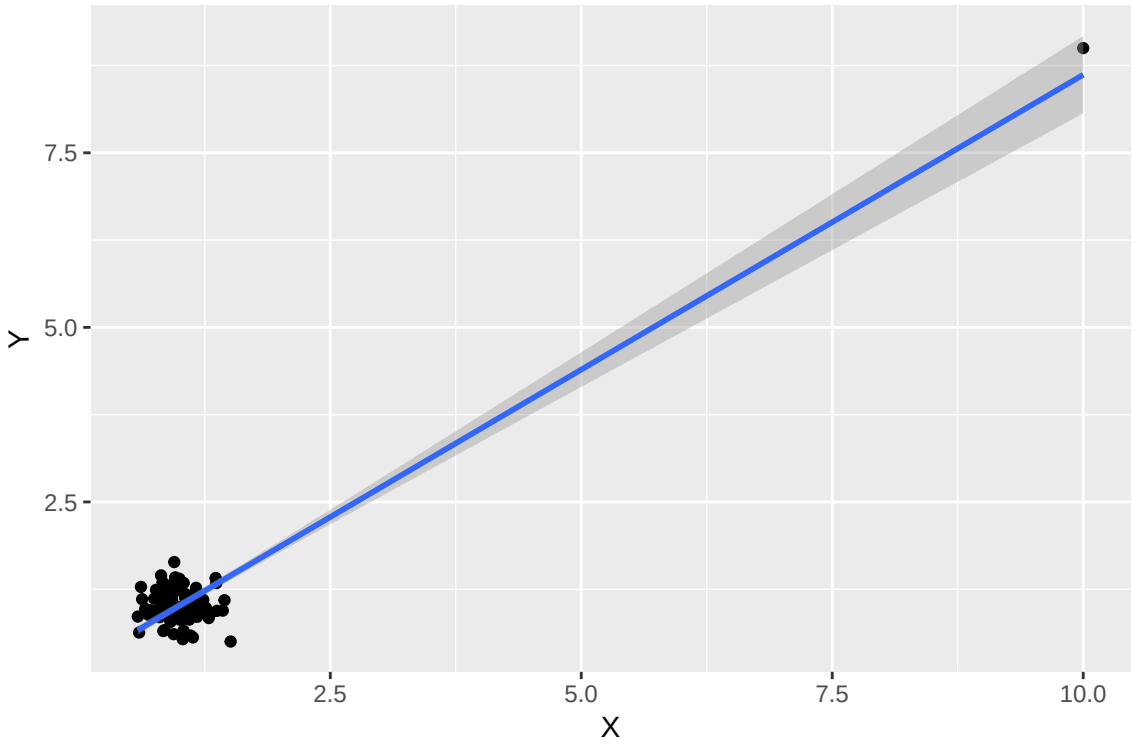
```
> summary(lmod)
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94484 -0.18748  0.00193  0.19895  0.66746

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17256    0.04469   3.861 0.000204 ***
X            0.84458    0.03118  27.086  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2853 on 97 degrees of freedom
Multiple R-squared:  0.8832,Adjusted R-squared:  0.882
F-statistic: 733.6 on 1 and 97 DF,  p-value: < 2.2e-16
```

Then you do a scatter plot of the data with the regression line and see:



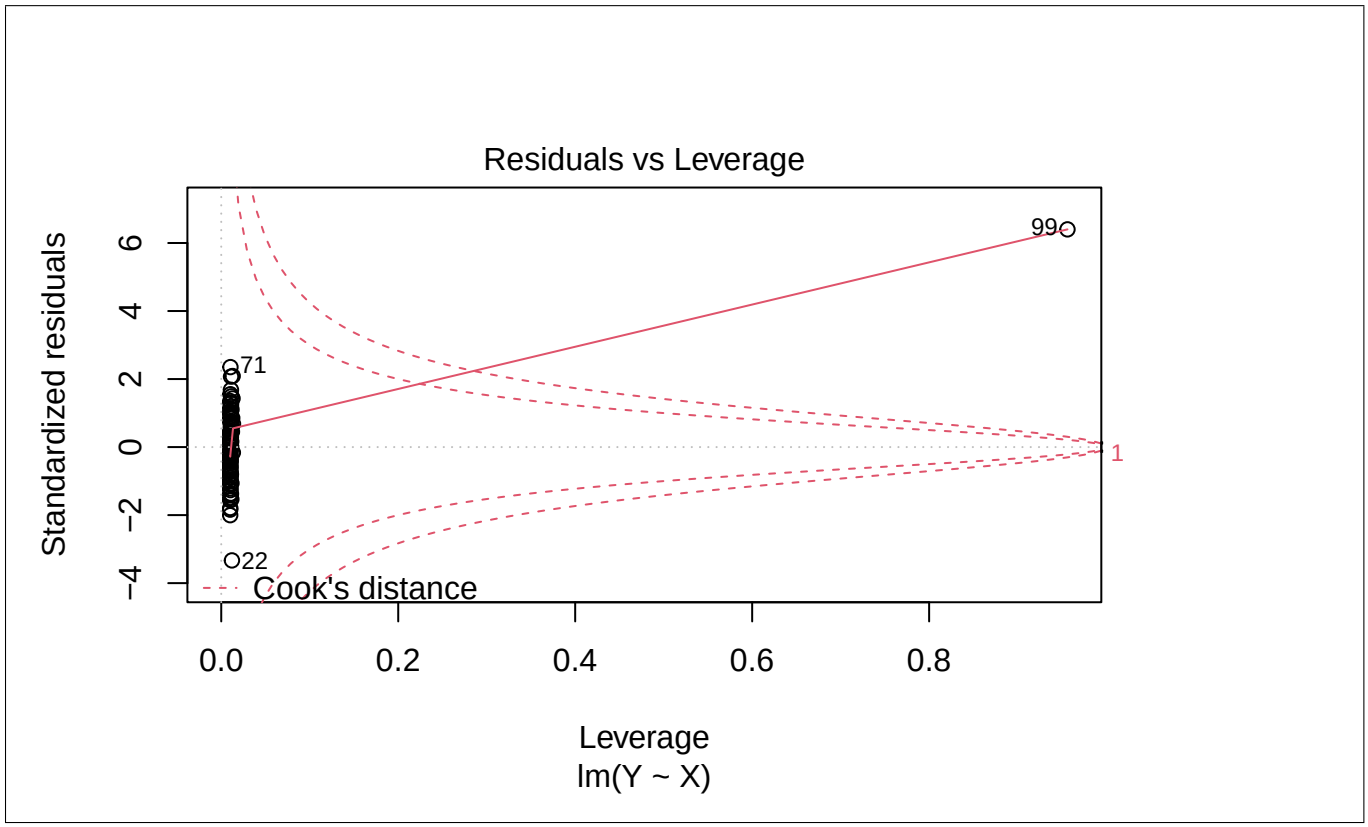

Answer the following questions about this regression:

(a) (4 points) The fourth standard diagnostic plot that R produces for a regression (using `plot(lmod)`) is sometimes called the "Cook's Distance" plot, because it has this label. What are the two variables defining the "Cook's Distance" plot? (That is, what are the labels on the axes of the "Cook's Distance" plot?)

**Solution:** Leverage and Standardized Residuals.

(b) (4 points) There is a point on the top right of the scatter plot on the previous page. Where, in the "Cook's Distance" plot, would you expect the point corresponding to this row of data to show up?

**Solution:** In the top right corner, since the residual is positive and the leverage is very high.




(c) (6 points) Suppose the row of data corresponding to the point in (b) were omitted and we refit the model. How might the intercept, the $R^2$, and the coefficient of `X` change? (At a minimum, your answer should clearly indicate whether each of these numbers increases, decreases, or remains the same.)

**Solution:** The intercept will increase, to approximately 1. The $R^2$ will decrease, to approximately 0. The coefficient of X will decrease, to approximately 0. See below, with `Xp, Yp` denoting the variables after removal of the point.

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.105607   0.116492  9.4908 1.847e-15
Xp          -0.084463   0.113429 -0.7446    0.4583

n = 98, p = 2, Residual SE = 0.21798, R-Squared = 0.01
```

(d) (4 points) Looking at the regression summary output on the previous page, another student asserts that the regression result is very reliable, because the $t$- and $F$-statistics are so significant and the $R^2$ is so high. Comment on this assertion.

**Solution:** The assertion is false; the high-leverage point in (b) shows up in all the diagnostic plots. This regression is entirely dependent on this one point.

To get full credit for this problem, you must state that the assertion "the regression result is very

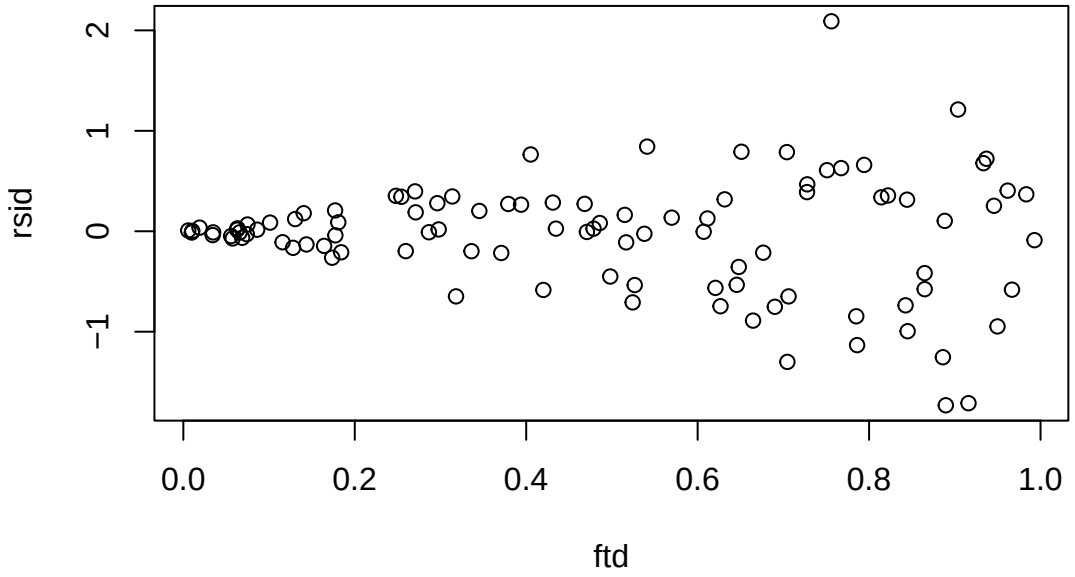
reliable" is false.

2. Suppose we fit a simple linear regression in R with `lmod <- lm(Y ~X)` and then look at the fitted vs residuals plot with `plot(lmod)`.

   (a) (4 points) What assumption of linear regression is being checked by looking at the fitted vs residuals plot?

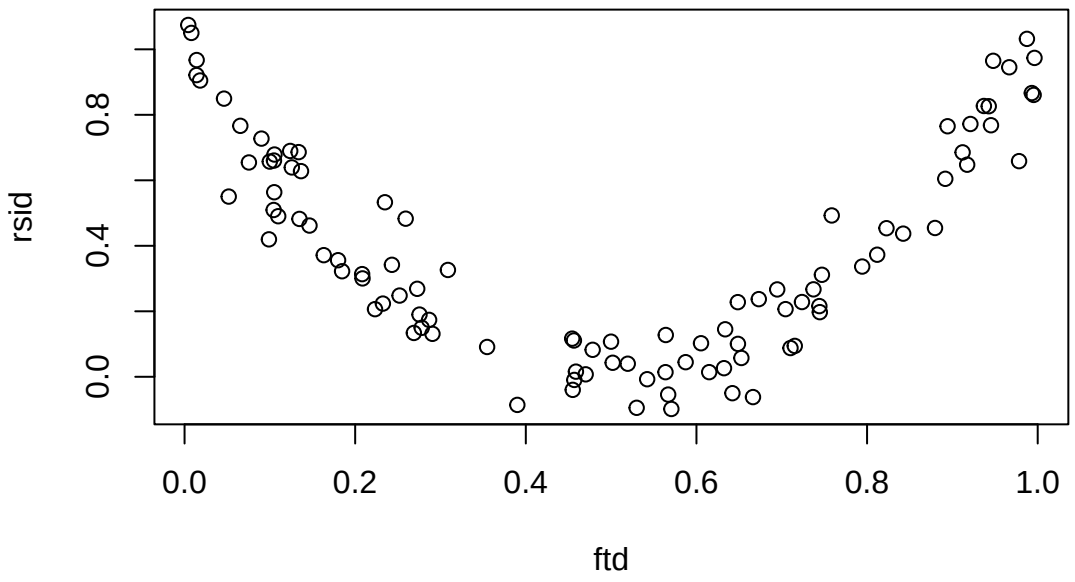
   **Solution:** Constant variance of the errors, or "homoskedasticity".

   (b) (4 points) Suppose the fitted vs residuals plot is as in the diagram below. Is the assumption satisfied? If not, what changes would you suggest to the model?




   **Solution:** The assumption is not satisfied; the standard deviation shows a linear trend, so the variance is proportional to the square of the fitted value. One standard idea here is to try replacing $Y$ with $\log(Y)$. See p. 77.

   (c) (4 points) Suppose the fitted vs residuals plot is as in the diagram below. Is the assumption satisfied? If not, what changes would you suggest to the model?




**Solution:** The assumption is not satisfied. The residuals form a parabola, so one standard idea here is to add a quadratic term with `lmod <- lm(Y ~X + I(X^2))`.

3. Assume that you have a linear regression model for the price of Honda Accords (a model of car) in terms of age in years, and miles on the odometer (measured in units of $10,000$). (So, a brand new car is 0 years old and has 0 miles, while a 3-year old car with $40,000$ miles has values of 3 and 4.) This model is given by the equation:

$$\text{price} = \alpha_0 + \alpha_1 \cdot \text{miles} + \alpha_2 \cdot \text{age} + \epsilon,$$

where $\epsilon$ is a normally distributed error. You fit this model and obtain coefficients $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2$.

Cars that are older and/or have more miles tend to have lower prices.

Cars that are older tend to have more miles on the odometer, and vice versa.

(a) (6 points) What sign(s) do you expect $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2$ to have?

**Solution:** We expect that $\widehat{\alpha}_0$ is positive: new cars cost something. We expect that $\widehat{\alpha}_1, \widehat{\alpha}_2$ are negative: cars that are older and/or have more miles tend to have lower prices.

(b) (8 points) You now fit a new model with age omitted, i.e.:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{miles} + \epsilon,$$

where $\epsilon$ is a normally distributed error, and obtain coefficients $\widehat{\beta}_0, \widehat{\beta}_1$.
Do you expect that $\widehat{\beta}_1 \approx \widehat{\alpha}_1$, $\widehat{\beta}_1 < \widehat{\alpha}_1$, or $\widehat{\beta}_1 > \widehat{\alpha}_1$?

**Solution:** We expect that $\widehat{\beta}_1 < \widehat{\alpha}_1$.

(c) (8 points) Give a reason for your answer in the previous part.

**Solution:** Let's explain this with an example using simplified numbers. Assume that the average car with 30,000 miles on the odometer is 3 years old, and that $\widehat{\alpha}_1 = -1000$, $\widehat{\alpha}_2 = -2000$. Then the (fitted) reduction in price for a 3-year old car with 30,000 miles on the odometer will be \$9,000. If we eliminate age from the model, then, in order to get this same reduction in price, we must have $\widehat{\beta}_1 = -3000$. That is, $\widehat{\beta}_1 < \widehat{\alpha}_1$.

This happens because $\widehat{\alpha}_1$ is negative, $\widehat{\alpha}_2$ is negative, and the correlation of age and miles is positive, so when miles acts as a proxy for age, it increases the magnitude of the coefficient estimate for miles. Note also that the order of magnitude of age and miles is the same in the given units, so the effect is significant.

4. Suppose we fit a simple (one explanatory variable) linear regression to a data set with 50 data points. We obtain residuals $e_i$, then define ordered residuals $e_{(i)}$ so that $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(50)}$. Then we define points in the plane $(a_i, e_{(i)})$ where $a_i = \Phi^{-1}(\frac{i}{51})$ and $i = 1, 2, \ldots, 50$, and $\Phi$ is the cumulative distribution function of the standard normal distribution. Finally we make a scatter plot of these 50 points in the plane.

(a) (4 points) This scatter plot is part of one of the standard diagnostic plots for the regression. What is this plot called?

**Solution:** The Q-Q plot or the quantile-quantile plot.

(b) (4 points) What is this plot intended to diagnose, or test?

**Solution:** It is intended to check whether the residuals are normally distributed and roughly uncorrelated, or since the residuals bear a close relationship to the errors, whether the errors are normally distributed and independent (which is a standard assumption of linear regression).

(c) (4 points) If the standard assumptions of linear regression are satisfied, these points should (roughly) form a certain pattern. Describe this pattern.

**Solution:** The points should look like an approximate version of the line $y = x$. This line appears in the full Q-Q plot.

(d) (6 points) If the standard assumptions of linear regression are satisfied, why would we expect this pattern to appear?

**Solution:** The residuals satisfy $\hat{\epsilon} = (I - H)\epsilon$, and this means that the residuals will be normal and typically will be roughly uncorrelated (see p. 84). Thus, the ordered residuals will be a

lot like order statistics for the normal distribution. The pdf of the i-th order statistic $e_{(i)}$ is very sharply peaked around $\Phi^{-1}(\frac{i}{51})$, so we expect the coordinates $a_i, e_{(i)}$ of the i-th point to be roughly equal, so the points should be close to the line $y = x$.

To get full credit for this problem, it is not enough to say "because the residuals are normal". We can see why this is by simulating $n = 50$ (say) normal random variables in $R$ using the `mvrnorm` function in the `MASS` package. Take these variables to all have mean 0, variance 1, and covariance 0.9 (say). These variables will stand in for "highly correlated residuals". Generate these variables, sort them, and plot them using `qqnorm`, and `abline(0,1)`. You will see that they are not on the line $y = x$.

5. Suppose we fit a linear regression in R and get the following summary output.

```
Call:
lm(formula = nhtemp ~ wusa + jasper + urals, data = globwarm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47105 -0.11109  0.00358  0.08590  0.69373

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.22664    0.02181 -10.393  < 2e-16 ***
wusa         0.11447    0.03161   3.622 0.000408 ***
jasper      -0.09218    0.07768  -1.187 0.237324
urals        0.21731    0.07718   2.816 0.005567 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2001 on 141 degrees of freedom
  (856 observations deleted due to missingness)
Multiple R-squared:  0.2965,Adjusted R-squared:  0.2816
F-statistic: 19.81 on 3 and 141 DF,  p-value: 8.975e-11
```

Suppose that we decide to replace `nhtemp` by `1.8 * nhtemp + 32` and run the following regression instead:

```
lmod2 <- lm(I(1.8*nhtemp+32) ~ wusa + jasper + urals)
```

Of the following quantities in the regression output, which change, and how do they change? You need not give a numerical answer, just circle "changes" or "does not change" and give a clear description of the change, if any.

(a) (2 points) The $R^2$ **changes** | **does not change**

(b) (2 points) The intercept **changes** | **does not change**

(c) (2 points) The standard error of the intercept **changes** | **does not change**

(d) (2 points) The $p$-value of the coefficient of `jasper` **changes** | **does not change**

(e) (2 points) The residual standard error **changes** | **does not change**

Suppose that we decide in the original regression to replace `jasper` by `jasper*100` and run the following regression instead:

```
lmod3 <- lm(nhtemp ~ wusa + I(100*jasper) + urals, data=globwarm)
```

Of the following quantities in the regression output above, which change, and how do they change? (Give a numerical answer for the changed values.)

(f) (2 points) The $R^2$ **changes** | **does not change**

(g) (2 points) The intercept **changes** | **does not change**

(h) (2 points) The standard error of `jasper` **changes** | **does not change**

(i) (2 points) The $p$-value of the coefficient of `jasper` **changes** | **does not change**

(j) (2 points) The residual standard error **changes** | **does not change**

**Solution:** For the first regression:

```
Call:
lm(formula = I(1.8 * nhtemp + 32) ~ wusa + jasper + urals, data = globwarm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84789 -0.19996  0.00644  0.15461  1.24872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.59204    0.03925 804.826  < 2e-16 ***
wusa         0.20605    0.05689   3.622 0.000408 ***
jasper      -0.16593    0.13982  -1.187 0.237324
urals        0.39115    0.13893   2.816 0.005567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3602 on 141 degrees of freedom
  (856 observations deleted due to missingness)
Multiple R-squared:  0.2965,Adjusted R-squared:  0.2816
F-statistic: 19.81 on 3 and 141 DF,  p-value: 8.975e-11
```

We see that:

- The $R^2$ does not change.

- The intercept changes: it increases by approximately 32. Technically, the new intercept is $1.8 * (-0.22) + 32$, where the old intercept is $\approx -0.22$, but $0.8 * (-0.22)$ is small, hence the simpler statement of the previous sentence.
- The standard error of the intercept changes: it increases by a factor of 1.8.
- The $p$-value of the coefficient of `jasper` does not change.
- The residual standard error changes: it increases by a factor of 1.8.

For the second regression:

```
Call:
lm(formula = nhtemp ~ wusa + I(100 * jasper) + urals, data = globwarm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47105 -0.11109  0.00358  0.08590  0.69373

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.2266426  0.0218074 -10.393  < 2e-16 ***
wusa             0.1144714  0.0316077   3.622 0.000408 ***
I(100 * jasper) -0.0009218  0.0007768  -1.187 0.237324
urals            0.2173072  0.0771811   2.816 0.005567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2001 on 141 degrees of freedom
  (856 observations deleted due to missingness)
Multiple R-squared:  0.2965,Adjusted R-squared:  0.2816
F-statistic: 19.81 on 3 and 141 DF,  p-value: 8.975e-11
```

We see that

- The $R^2$ does not change.
- The intercept does not change.
- The standard error of the coefficient of `jasper` changes: it decreases by a factor of 100.
- The $p$-value of the coefficient of `jasper` does not change.
- The residual standard error does not change.

6. Assume the linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with $X$ a random variable such that $\text{Var}(X) = \sigma_X^2$ and errors $\epsilon$ iid $N(0, \sigma^2)$ is exactly true, but that we can only observe $X' = X + \eta$, where $\eta$ is an error such that $\text{Var}(\eta) = \sigma_\eta^2$ and $\text{Cov}(X, \eta) = \kappa$. Further, $\text{Cov}(X, \epsilon) = \text{Cov}(\epsilon, \eta) = 0$. We now fit the regression model

$$Y = \beta_0 + \beta_1 X' + \epsilon'$$

which in terms of the variables above is

$$Y = \beta_0 + \beta_1(X + \eta) + (\epsilon - \beta_1 \eta)$$

(I have done the algebra to compute $\epsilon'$ in terms of $\epsilon$ and $\eta$ for you and you may assume the expression above is correct.)

Suppose now that we compute

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X')}{\text{Var}(X')}$$

as we normally do.

(a) (6 points) Find $\text{Cov}(Y, X')$ in terms of the quantities $(\beta_0, \beta_1, \sigma_X, \kappa, \sigma_\eta)$ given above.

**Solution:** Write $\text{Cov}(Y, X') = \text{Cov}(\beta_0 + \beta_1(X + \eta) + (\epsilon - \beta_1\eta), X + \eta)$ and compute using bilinearity of covariance and the assumptions above to get $\text{Cov}(Y, X') = \beta_1(\sigma_X^2 + \kappa)$.

(b) (4 points) Find $\hat{\beta}_1$ in terms of the quantities given above.

**Solution:** Use bilinearity of covariance and the assumptions above to get

$$\text{Var}(X') = \text{Cov}(X + \eta, X + \eta) = \sigma_X^2 + \sigma_\eta^2 + 2\kappa$$

Then divide to get

$$\hat{\beta}_1 = \beta_1\left(\frac{\sigma_X^2 + \kappa}{\sigma_X^2 + \sigma_\eta^2 + 2\kappa}\right)$$

(c) (4 points) Suppose $\beta_0 = 1, \beta_1 = 3, \kappa = 1$ and $\sigma_X = \sigma_\eta = 1$. What will the estimated coefficient $\hat{\beta}_1$ be?

**Solution:** Plug in the numbers and get $\hat{\beta}_1 = 3/2$.

(d) (4 points) Explain, in terms comprehensible to the non-statistician, why the answer to part (c) differs from the true value of $\beta_1$.

**Solution:** The observed value $X'$ is an exaggeration of the value of $X$, since the correlation is positive. Thus the coefficient of $X'$ has to be reduced to compensate for this increase.

7. (6 points) Give an example of a data set on which the use of the Bootstrap would be inappropriate, and explain why the use of the Bootstrap would be inappropriate.

**Solution:** Many examples are possible. If the data set has a structure which is not preserved by random sampling, then the use of the Bootstrap is not appropriate, because the randomly generated data sets will not have this structure. This is common in time series. One example is the airline passengers data set (see p. 55 in the textbook) which shows a strong seasonality.

The letter corresponding to the correct answer of each multiple choice question is in **boldface**.

8. (4 points) Which statement is most accurate about the relationship between WLS and GLS?

   a) WLS is a special case of GLS where the weights are based on the squared residuals.

   **b)** GLS is a more general approach that can incorporate various weighting schemes, including those used in WLS.

   c) There is no practical difference between the two methods.

   d) WLS is always preferable due to its simpler implementation.

9. (4 points) If you suspect autocorrelation (correlated errors) in your data, applying WLS would be:

   a) An appropriate solution, as long as the specific correlation structure is known.

   **b)** Not helpful, as WLS only addresses heteroscedasticity.

   c) Potentially beneficial if combined with other techniques to address autocorrelation.

   d) Impossible, as WLS cannot handle correlated errors.

10. (4 points) What does the condition number of the design matrix indicate about multicollinearity?

    a) The level of heteroscedasticity in the regression model

    **b)** The stability of parameter estimates in the regression model

    c) The degree of correlation between the predictor variables

    d) The proportion of variance explained by the regression model

11. (4 points) Why might one examine the correlation matrix of predictors in regression analysis? Choose the most accurate answer.

    a) To identify outliers in the dataset

    b) To assess the assumption of a linear relationship between the predictors and the response

c) To determine the statistical significance of each predictor

**d)** None of the above

12. (4 points) When interpreting the correlation matrix to assess collinearity, a high absolute value (close to 1 or -1) between two predictors suggests:

a) There is no linear relationship between the variables.

b) There is a weak linear relationship between the variables.

**c)** There is a very strong linear relationship between the variables.

d) The specific value doesn't matter, any correlation indicates collinearity.

13. (4 points) Robust regression techniques are particularly beneficial for analyzing data that exhibit:

a) Strong non-linear relationships between variables.

b) Multicollinearity among the independent variables.

c) Heteroscedasticity (unequal variance of errors).

**d)** The presence of outliers or extreme data points.

14. (4 points) Compared to Ordinary Least Squares (OLS) regression, robust regression methods are primarily designed to be:

a) More efficient in estimating regression coefficients.

b) Less sensitive to the influence of outliers in the data.

c) Applicable to a wider range of data distributions.

**d)** Both b and c.

15. (4 points) Which of the following best describes the Bootstrap?

a) It involves randomly selecting subsets of data without replacement

**b)** It involves randomly selecting subsets of data with replacement

c) It involves systematically rearranging the data

d) It involves generating entirely new data points based on the original dataset

16. (4 points) In Bootstrap resampling, what does "with replacement" mean?

a) Each data point is used exactly once in each resampled dataset

**b)** Each data point can be used multiple times in each resampled dataset

c) Data points are replaced with new values in each resampled dataset

d) Data points are removed from the original dataset after being used once

17. (4 points) According to the Hill Criteria, which of the following statements is *not* indicative of a causal relationship between an explanatory variable and the response variable?

**a)** A statistically significant association is observed between the explanatory variable and the response variable.

b) The observed association persists when the analysis is repeated in different populations or under varying conditions.

c) Increasing the level of the explanatory variable is consistently linked to an increase in the risk of the response.

d) There is a plausible biological or mechanistic explanation for how the explanatory variable might influence the response.

18. (4 points) The concept of temporality in the Hill Criteria refers to:

a) The possibility of confounding variables influencing the observed association.

b) The need for a strong and statistically significant relationship between variables.

**c)** The requirement that the change in the explanatory variable precedes the change in the response variable.

d) The existence of a dose-response relationship between the explanatory and response variables.

19. (4 points) When interpreting the Hill Criteria, it's important to remember that:

a) All criteria must be definitively met to establish a causal link.

**b)** Each criterion provides evidence for causality, but a strong overall picture is needed.

c) The importance of each criterion is equal, regardless of the specific study context.

d) The Hill Criteria are only applicable to studies with randomized controlled trials.

20. (4 points) In the context of the Hill Criteria, which of the following is a limitation of using regression analysis alone to assess causality?

a) Regression analysis cannot account for the possibility of confounding variables.

b) Regression is unable to establish the temporal order between the explanatory and response variables.

c) Regression is limited to studying only linear relationships between variables.

**d)** All of the above.

21. (4 points) The primary purpose of applying the Box-Cox transformation in regression analysis is to:

a) Reduce the influence of outliers in the data.

b) Improve the normality of the residuals in the model.

**c)** Account for non-linear relationships between the independent and dependent variables.

d) Decrease the multicollinearity among the independent variables.

22. (4 points) When interpreting the value of lambda after applying the Box-Cox transformation:

a) A lambda close to 1 indicates a strong positive linear relationship in the original data.

b) A lambda of 0 implies the data already follows a normal distribution.

c) A negative lambda value suggests the data needs to be squared for normality.

**d)** The specific interpretation of lambda depends on the context of the data and analysis.