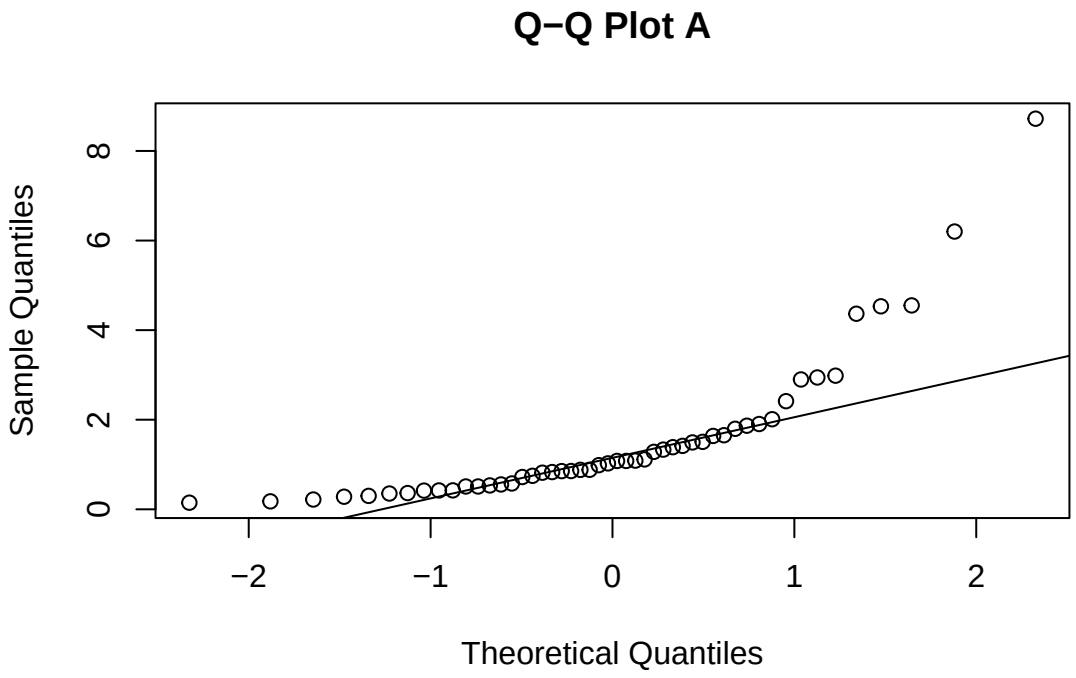
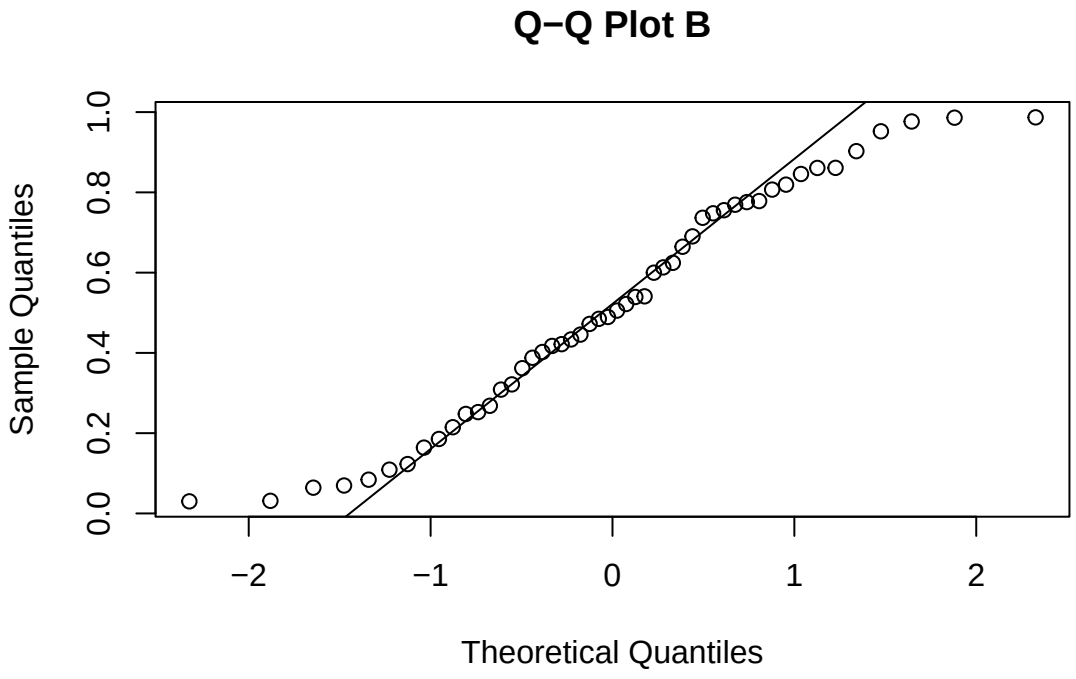
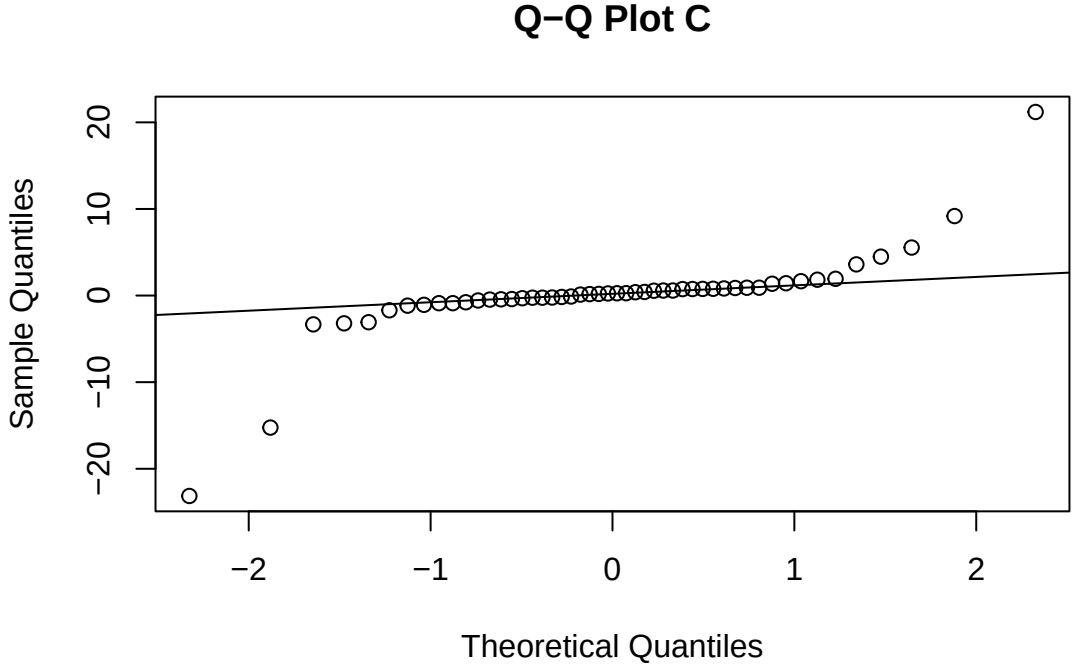
1. Each of the 3 Q-Q plots below is constructed by taking independent random samples from a probability distribution and applying the usual procedure to construct a Q-Q plot.







Solution: The examples for this problem are discussed on p. 80 of the textbook. In particular, it is noted there that graph A comes from a lognormal distribution, graph B from a uniform distribution, and graph C from a cauchy distribution.

(a) (4 points) Which, if any, of these Q-Q plots shows a distribution with a long (or heavy) right tail?

Solution: A and C.

(b) (4 points) Which, if any, of these Q-Q plots shows a distribution with a long (or heavy) left tail?

Solution: C

(c) (4 points) Which, if any, of these Q-Q plots is most likely to be constructed by taking samples from a normal distribution?

Solution: none

(d) (4 points) Which, if any, of these Q-Q plots is most likely to be constructed by taking samples from a uniform distribution?

Solution: B

(e) (4 points) Which, if any, of these Q-Q plots shows a distribution with a short (or thin) left tail?

Solution: A and B

(f) (4 points) Which, if any, of these Q-Q plots shows a distribution with a short (or thin) right tail?

Solution: B

2. Assume the linear regression model

$$Y = 1 + 3X_1 + 2X_2 + \epsilon$$

with the standard assumptions is exactly true, and that $\text{Cor}(X_1, X_2) = \rho$. Additionally, assume that ϵ is independent of X_1 and X_2, $\text{Var}(\epsilon) = \sigma^2$, and $\text{Var}(X_i) = \sigma_i^2$.

Assume now that we have no knowledge of the variable X_2 and decide to fit the simple regression model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon'$$

You can now calculate the estimated coefficient $\hat{\beta}_1$ that will be obtained in the big-data limit from the expression

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X_1)}{\text{Var}(X_1)}$$

(a) (6 points) Find $\text{Cov}(Y, X_1)$ in terms of the quantities given above.

Solution: Compute using the expression $Y = 1 + 3X_1 + 2X_2 + \epsilon$. Note: it isn't possible to compute directly with the expression $Y = \beta_0 + \beta_1 X_1 + \epsilon'$, because you aren't given anything about ϵ'.

$\text{Cov}(Y, X_1) = \text{Cov}(1+3X_1+2X_2+\epsilon, X_1) = 3\,\text{Cov}(X_1, X_1)+2\,\text{Cov}(X_2, X_1)+\text{Cov}(\epsilon, X_1) = 3\sigma_1^2+2\rho\sigma_1\sigma_2$

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

Solution:

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X_1)}{\text{Var}(X_1)} = \frac{3\sigma_1^2 + 2\rho\sigma_1\sigma_2}{\sigma_1^2)} = 3 + 2\rho\frac{\sigma_2}{\sigma_1}$$

(c) (4 points) Suppose $\rho = 1/2$ and $\sigma_1^2 = \sigma_2^2 = 1$. What will the estimated coefficient $\hat{\beta}_1$ be?

Solution: Plug in to the formula above:

$$\hat{\beta}_1 = 3 + 2\frac{1}{2} = 4$$

(d) (6 points) Is this coefficient the same as the one in the true model above? (That is, the model assumed to be "exactly true" in the problem statement.) Explain why or why not, in terms comprehensible to the non-statistician.

Solution: No, it is not. The coefficient of X_1 in the true model is 3, as given in the problem statement. Because we removed X_2 (part of the true model) from our regression, and X_1 has a nonzero correlation with X_2, the coefficient of X_1 shows not only the true effect of X_1, but also an additional term, since X_1 is acting as a proxy for the missing X_2.

(e) (4 points) What is this phenomenon called?

Solution: Omitted Variable Bias. We omitted the variable X_2 (which is in the true model) from our regression.

3. We decide to fit fit a quadratic regression in R with `qmod <- lm(Y ~X + X^2)`. Answer the questions following about the summary output below:

```
Call:
lm(formula = Y ~ X + X^2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.022135   0.046148    0.48    0.633
X           0.996263   0.007228  137.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2898 on 98 degrees of freedom
Multiple R-squared:  0.9949,Adjusted R-squared:  0.9948
F-statistic: 1.9e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

(a) (4 points) Why doesn't the summary output show a coefficient estimate for the quadratic term?

Solution: The formula `Y ~X + X^2` is interpreted by R as Wilkinson-Rogers notation, and not as squaring. Thus R never sees the square of `X`.

(b) (4 points) What should you do about this?

Solution: Two possible solutions: protect the squaring from Wilkinson-Rogers notation interpretation with `Y ~X + I(X^2)`. Or define a new variable `Z <- X^2` in a separate line of code and then use `Y ~X + Z` as the regression formula.

4. A university registrar's office wants to understand the relationship between student housing and meal plan choices. They collect data on a random sample of students (n = 5,241) and record the following variables:

- Y: Total annual housing and food cost (in thousands of dollars)
- H: Housing type (1 = On-campus housing, 0 = Off-campus housing)
- M: Meal plan (1 = meal plan, 0 = No meal plan)

They fit three linear regression models to the data:

- (Model 1) $Y = \beta_0 + \beta_1 H$
- (Model 2) $Y = \beta_0 + \beta_1 H + \beta_2 M$
- (Model 3) $Y = \beta_0 + \beta_1 H + \beta_2 M + \beta_3 HM$

(a) (4 points) For model 1, suppose you obtain the coefficient estimate $\hat{\beta}_1 = 3$. Write a sentence interpreting this result in plain language, understandable to someone unfamiliar with statistics.

Solution: Students who had on-campus housing had an average total food and housing cost \$3000 higher than off-campus students.

(b) (4 points) For model 2, suppose we obtain the coefficient estimates $\hat{\beta}_1 = 3$ and $\hat{\beta}_2 = 2$. Using these estimates, write a sentence describing the difference in total costs between on-campus students having a meal plan and on-campus students not having a meal plan.

Solution: On-campus students who had meal plans had an average total food and housing cost \$2000 higher than on-campus students who did not have meal plans.

(c) (4 points) For model 3, suppose we obtain the coefficient estimates $\hat{\beta}_1 = 3$, $\hat{\beta}_2 = 2$, and $\hat{\beta}_3 = -1$. According to this model, is the change in total costs associated with having a meal plan the same for on- and off-campus students? If not, which group has a higher change?

Solution: First, read carefully. This is asking about whether there is a difference between two differences, so we must compute 4 quantities and two differences, then check if they are the same.

The answer is **no**, the change in total costs associated with having a meal plan is not the same for on- and off-campus students. The *off-campus students have a higher change in total costs.*

The change in total costs associated with having a meal plan for off-campus students is $\hat{\beta}_0 + \hat{\beta}_2 - \hat{\beta}_0 = \hat{\beta}_2 = 2$. The change in total costs associated with having a meal plan for on-campus students is $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 - (\hat{\beta}_0 + \hat{\beta}_1) = \hat{\beta}_2 + \hat{\beta}_3 = 2 - 1 = 1$.

5. Suppose we are studying the relationship between years of work experience (denoted X) and annual salary (denoted Y, and measured in thousands of dollars) for statisticians. We have collected data from a sample of employees, and our final regression model is as follows:

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

We obtain coefficient estimates $\hat{\beta}_0 = 4$, $\hat{\beta}_1 = 0.05$.

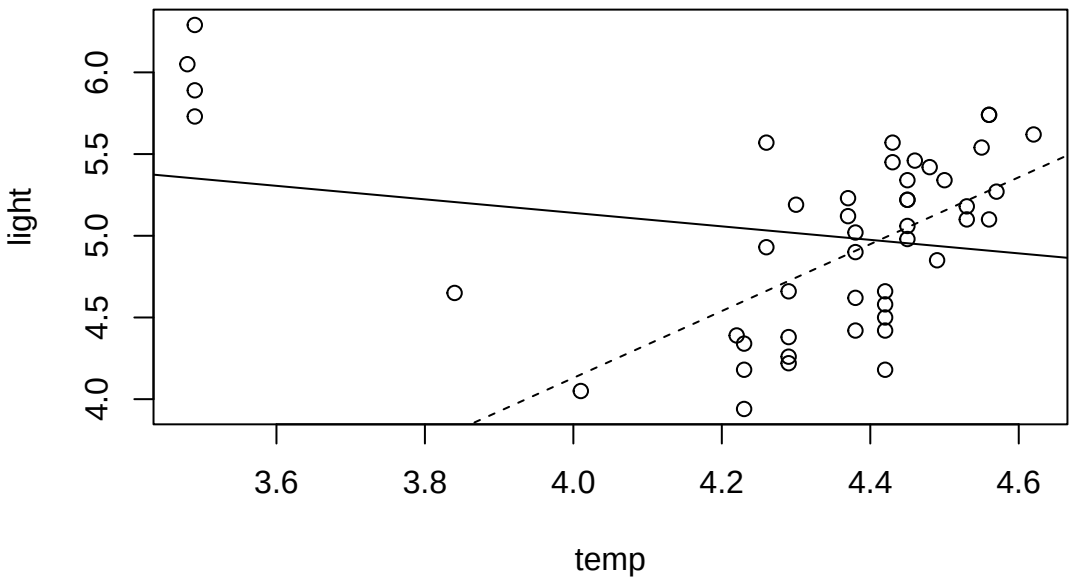
(a) (5 points) Write a sentence interpreting the coefficient estimate $\hat{\beta}_0 = 4$, comprehensible to the non-statistician.

> **Solution:** The estimated annual salary for a statistician with no work experience is about \$54,000.
>
> Algebra and the problem statement gives you "e^4 thousand dollars"; to get to an actual number, note that $e \approx 2.7$, so $e^4 \approx (3 - 0.3)^4$. The first three terms of the binomial expansion give you $e^4 \approx (3-0.3)^4$. The first three terms of the binomial expansion give $3^4 - 4 \cdot 3^3 \cdot (0.3) + 6 \cdot 3^2 \cdot (0.3)^2 = 81 - 4 \cdot 8.1 + 6 \cdot 0.81 = 53.46$; we should add a little because e is a little more than 2.7.
>
> Alternatively, if you know the occasionally-useful approximation $e^3 \approx 20$, then you can just say $e^4 \approx 2.7 \cdot 20 \approx 54$.

(b) (5 points) Write a sentence interpreting the coefficient estimate $\hat{\beta}_1 = 0.05$, comprehensible to the non-statistician.

> **Solution:** Statisticians with one extra year of work experience tended to have a salaries 5% higher.
>
> Recommendation: you should read the advice on interpretation of such models in the textbook on p. 134, which was discussed in class.

6. The following graph is based on the data set `star`, which records log of the surface temperature and the log of the light intensity of 47 stars. The solid line is the regression line for `lm(light ~ temp, star)` and the dotted line is for the same regression with the four points in the upper-left-hand part of the graph removed from the data set.



(a) (1 point) Is the slope of the dotted line positive or negative?

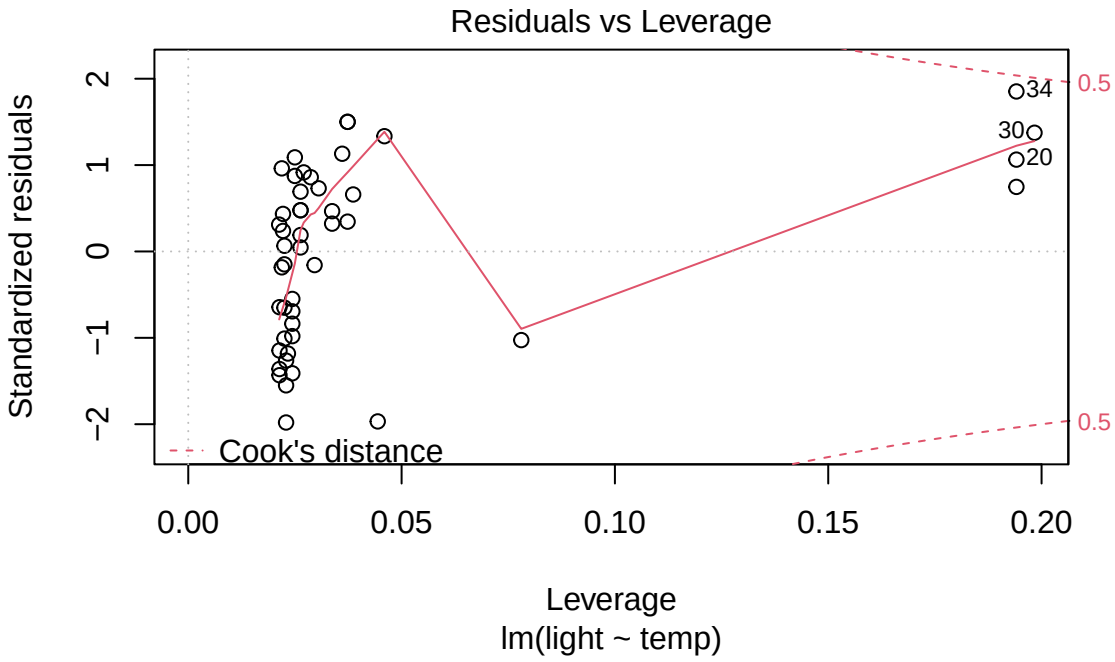
Solution: Positive.

(b) (1 point) Is the slope of the solid line positive or negative?

Solution: Negative.

(c) (4 points) How would you explain the difference, if any, between the two models?

Solution: The solid line is fitted via least-squares to a data set including the four points in the upper-left-hand corner. These points "pull" the left part of the line up, making the slope negative.



(d) (4 points) In the "Cook's Distance" plot provided above, which is a standard diagnostic plot provided by R for a linear regression, we usually say that a point is *influential* if the Cook's Distance associated to that point is greater than 0.5. Are any of the four points in the upper-left-hand part of the graph before part (a) influential? Why or why not?

Solution: No, none of these points are influential by the Cook's distance measurement.

(e) (8 points) Explain the apparent conflict between your answers to parts (c) and (d).

Solution: In (c) we said that the four points were influential in the sense that they changed the slope of the line from positive to negative. In (d) we said that they were not influential. This is the apparent conflict.

The conflict is resolved by noting that the two senses of the word influential are not the same. The sense of influential in (c) is based on the change in sign of the slope, and is a collective measure of influence. The sense of influence in (d) is an individual measure of influence, based on Cook's distance. There is no reason these two senses of influence need to agree with one another.

One could also state that the points are "collectively influential" but not individually influential.

7. Answer the following questions about Box-Cox procedure for selecting a transformation in the regression $Y = \beta_0 + \beta_1 X + \epsilon$.

(a) (1 point) Does the Box-Cox procedure try to select a transformation of the response Y, or the explanatory variable X?

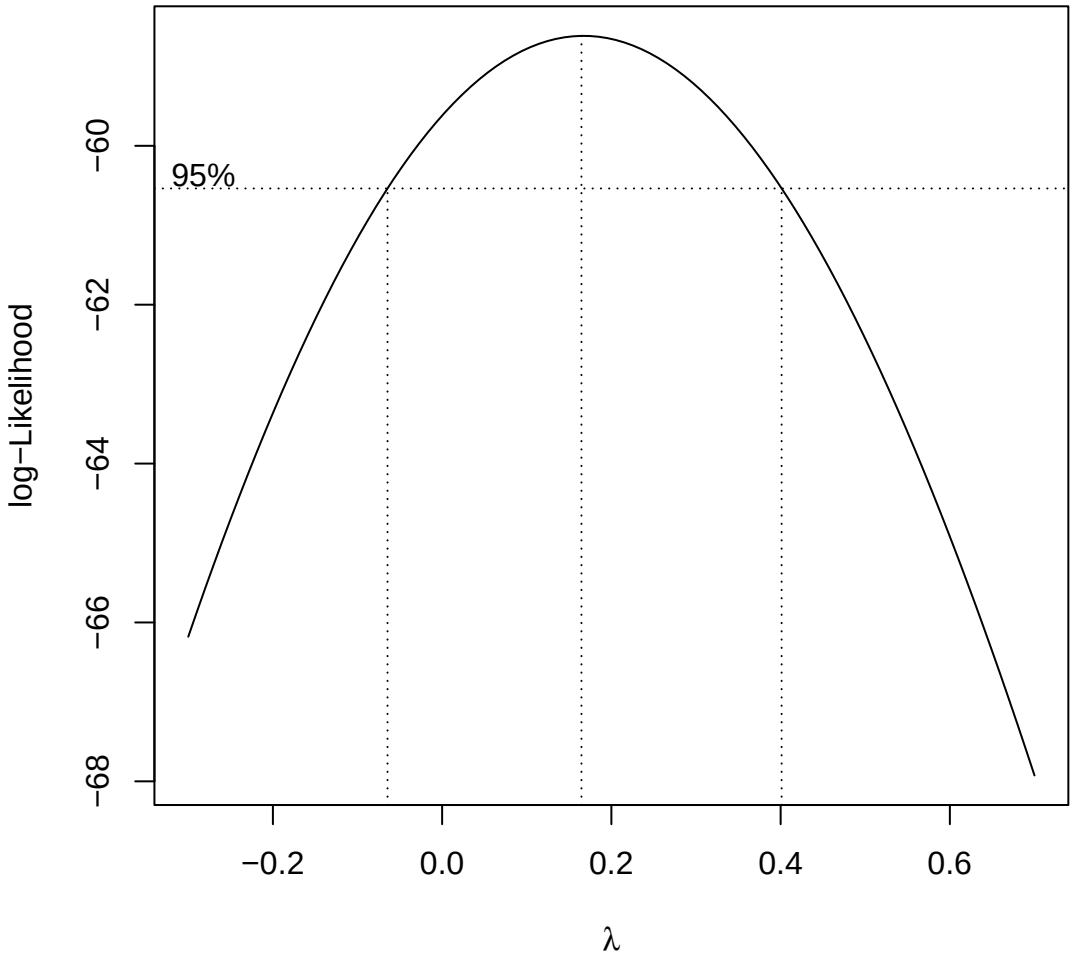
Solution: For the response Y.

(b) (4 points) What types of transformations are considered?

Solution: Transformations $Y \mapsto Y^\lambda$, where λ is a real number, plus $Y \mapsto \log(Y)$. Technically we consider $Y \mapsto (Y^\lambda - 1)/\lambda$, for consistency with $\log(Y)$ in the case $\lambda = 0$, but this detail is not required to get full credit.

(c) (3 points) Are there any restrictions on the variables that might cause problems for the procedure?

Solution: The response variable Y cannot have negative values, as we have no interpretation of, for example, $(-1)^{\frac{1}{2}}$ as a real number.



(d) (6 points) Suppose the graph produced by the Box-Cox procedure is as shown above. What transformation would you apply to your data after seeing this graph, and why?

Solution: We need to choose the closest *intepretable* value of λ to the maximum of the log-likelihood. The natural choice here is $Y \mapsto \log(Y)$. Of course $\lambda = 0.2$ is closer to the peak, but $\sqrt[5]{Y}$ is hard to interpret in most contexts.

Multiple Choice: Circle the item corresponding to the best answer. *There is no penalty for guessing.*

8. (4 points) Which model selection procedure starts with an empty model and iteratively adds the most significant feature at each step until a stopping criterion is met?

   A. Backward Elimination
   B. Forward Selection
   C. Best Subsets Regression
   D. Adjusted R-squared

Solution: Correct answer: (b) Forward Selection

Forward Selection is a stepwise procedure that starts with no predictors and adds variables incrementally based on statistical significance or a model selection criterion like AIC or BIC.

Why the others are wrong:

- (a) Backward Elimination removes predictors from a full model.
- (c) Best Subsets Regression evaluates all possible combinations of predictors.
- (d) Adjusted R-squared is a model evaluation metric, not a selection procedure.

9. (4 points) When using Best Subsets Regression, you consider all possible combinations of features for your model. This can become computationally expensive for datasets with:

   A. A small number of features and a large sample size
   B. A large number of features and a small sample size
   C. A small number of features and a small sample size
   D. All datasets can be efficiently analyzed with Best Subsets Regression.

Solution: Correct answer: (b) A large number of features and a small sample size

The number of possible models grows exponentially with the number of predictors. When the number of predictors is large, especially with a small sample size, the computational cost and risk of overfitting become significant.

Why the others are wrong:

- (a), (c) Small numbers of features don't pose a computational issue.
- (d) Best Subsets Regression is not computationally efficient for all datasets.

10. (4 points) Compared to the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) generally:
    A. Penalizes model complexity more heavily.
    B. Penalizes model complexity less heavily.
    C. Focuses solely on R-squared for comparison.
    D. Does not consider model complexity in its calculation.

Solution: Correct answer: (a) Penalizes model complexity more heavily.

BIC adds a heavier penalty for the number of parameters, especially with large sample sizes, making it more conservative than AIC.

Why the others are wrong:

- (b) AIC penalizes complexity less than BIC.
- (c), (d) Neither criterion is based solely on R-squared; both balance fit and complexity.

11. (4 points) Which of the following statements is most accurate regarding Adjusted R-squared?
    A. It is always higher than R-squared for the same model.
    B. It is calculated the same way as R-squared, but for the training data only.
    C. It penalizes adding features to the model, unlike R-squared.
    D. It is the difference between R-squared and the BIC value.

Solution: Correct answer: (c) It penalizes adding features to the model, unlike R-squared.

Adjusted R-squared adjusts the traditional R^2 value downward when unnecessary predictors are added, to prevent overfitting.

Why the others are wrong:

- (a) Adjusted R-squared can be lower than R-squared.
- (b) It adjusts for the number of predictors, not just training data.
- (d) It is not mathematically related to BIC.

12. (4 points) While building a regression model, you identify a group of highly correlated variables. Including all these features simultaneously in the model might lead to a problem called:

    A. Error Persistence

    B. Multicollinearity

    C. Underfitting

    D. Residual Heteroscedasticity

Solution: Correct answer: (b) Multicollinearity

Multicollinearity inflates variance of coefficient estimates and makes them unstable when predictor variables are highly correlated.

Why the others are wrong:

- (a) Not a standard term in regression.
- (c) Underfitting occurs when a model is too simple.
- (d) Heteroscedasticity relates to non-constant residual variance, not predictor correlation.

13. (4 points) Cook's distance is a diagnostic measure used to identify:

    A. Non-linearity in the relationship between the independent and dependent variables.

    B. A violation of the assumption of homoscedasticity.

    C. Influential observations that can significantly impact the regression line.

    D. The presence of multicollinearity among the independent variables.

Solution: Correct answer: (c) Influential observations that can significantly impact the regression line.

Cook's distance identifies points that significantly affect the regression coefficients when removed.

Why the others are wrong:

- (a) Non-linearity is assessed with residual plots, not Cook's distance.
- (b) Homoscedasticity is tested with tests like Breusch–Pagan.
- (d) Multicollinearity is diagnosed with VIF or condition number.

14. (4 points) Leverage is a measure in regression analysis that represents:

    A. The distance of a data point from the fitted regression line.

    B. The influence of an observation on the estimated regression coefficients.

C. How well a specific data point is fitted by the regression model.

D. The degree to which an independent variable explains the dependent variable.

Solution: Correct answer: (b) The influence of an observation on the estimated regression coefficients.

Leverage reflects how far a predictor value is from the mean predictor values; high-leverage points have the potential to influence the model's fit substantially.

Why the others are wrong:

- (a) Describes residuals.
- (c) Describes prediction error.
- (d) Describes R^2, not leverage.

15. (4 points) A key limitation of implementing GLS in regression analysis is:

A. The need for a large sample size to ensure reliable estimates.

B. The requirement for specifying the exact form of the heteroscedasticity in the data.

C. Its inability to address the presence of influential outliers.

D. The computational complexity compared to the simpler Ordinary Least Squares (OLS) method.

Solution: Correct answer: (b) The requirement for specifying the exact form of the heteroscedasticity in the data.

GLS adjusts for heteroscedasticity or correlation in the errors, but requires the correct error covariance structure to be known or estimated accurately.

Why the others are wrong:

- (a) Sample size helps, but is not an issue specific to GLS.
- (c) GLS can reduce the effect of outliers if a relevant variable is designated as higher-variance.
- (d) GLS can be reduced to OLS with some matrix computations.

16. (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

Solution: Correct answer: (b) The stability of parameter estimates in the regression model A high condition number indicates near-linear-dependence of the predictors. From p. 107, "signs of collinearity ... none of the individual predictors are significant". A lack of significance is another way of saying "unstable estimates".

Why the others are wrong:

- (a) Heteroscedasticity concerns residuals, not the design matrix.
- (c) Condition number reflects correlation among *linear combinations of* the predictors.
- (d) The proportion of variance explained is measured by R^2.

17. (4 points) Not controlling for a relevant covariate in a regression analysis can lead to:

    A. A more precise estimate of the coefficient of an independent variable.

    B. Collinearity and higher coefficient standard errors.

    C. Biased estimates of the coefficients, potentially masking or inflating the true effect of an independent variable.

    D. An inability to calculate the standard errors of the coefficients.

Solution: Correct answer: (c) Biased estimates of the coefficients, potentially masking or inflating the true effect of an independent variable.

Omitted variable bias occurs when a relevant covariate is not included, leading to confounded and biased coefficient estimates.

Why the others are wrong:

- (a) Omission leads to bias, not precision.
- (b) Collinearity concerns correlation among included variables.
- (d) Standard errors can still be computed, but they may be biased.