UPPSALA UNIVERSITY
Department of Mathematics
Shaobo Jin

Exam in Mathematical Statistics
Regression Analysis, 1MS555
2024-04-05

Time: 8.00-13.00. Limits for the credits 3, 4, 5 are 25, 30 and 35 points, respectively. The solutions should be well motivated.

Permitted aids: One page (both sides) hand-written cheat sheet for the course. Pocket calculator.

1. (8p) Suppose that we have a data set of three observations given by

y	x
0	-1
0	0
3	1

We want to fit a simple linear regression $y_i = \beta_0 + \beta_1 x_i + e_i$.

   (a) (2p) Find the ordinary least squares estimates of β_0 and β_1.

   (b) (2p) What are the fitted values for this data set?

   (c) (2p) Find the value of the residual sum of squares.

   (d) (2p) The variance of e_i given x_i is σ^2. Estimate σ^2.

2. (8p) Consider the model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$, where $\mathrm{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0$ and$\mathrm{Var}\left(e_i \mid \boldsymbol{x}_i\right) = \sigma^2\left(\boldsymbol{x}_i^T\boldsymbol{x}_i + 1\right)$ and $\boldsymbol{\beta}$ is a $p \times 1$ vector. We also assume that observations are independent of each other.

   (a) (2p) Derive the ordinary least squares estimator.

   (b) (2p) Is your ordinary least squares estimator unbiased for $\boldsymbol{\beta}$ if the model is correctly specified?

   (c) (2p) Find the covariance matrix of your estimator.

   (d) (2p) Do the residuals and the predicted value given $\boldsymbol{X}$ have a zero sample covariance, if the sample mean of the residuals is zero?

3. (12p) Suppose that we have measured the weight of water, the weight of fat, and the weight of protein of one type of meat. We have regressed the weight of protein on the weight of water and the weight of fat in R as follows.

```
LR <- lm(Protein ~ Water + Fat, data = Data)
summary(LR)
```

```
##
## Call:
## lm(formula = Protein ~ Water + Fat, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0331 -0.2616  0.2483  0.7249  2.1343
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.28170    2.98936   20.83   <2e-16 ***
## Water       -0.52985    0.03869  -13.69   <2e-16 ***
## Fat         -0.61068    0.03023  -20.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.348 on 497 degrees of freedom
## Multiple R-squared:  0.7827,Adjusted R-squared:  0.7818
## F-statistic:   895 on 2 and 497 DF,  p-value: < 2.2e-16
```

(a) (1p) Does Water significantly affect Protein when controlling for Fat?

(b) (1p) What is the fitted regression model?

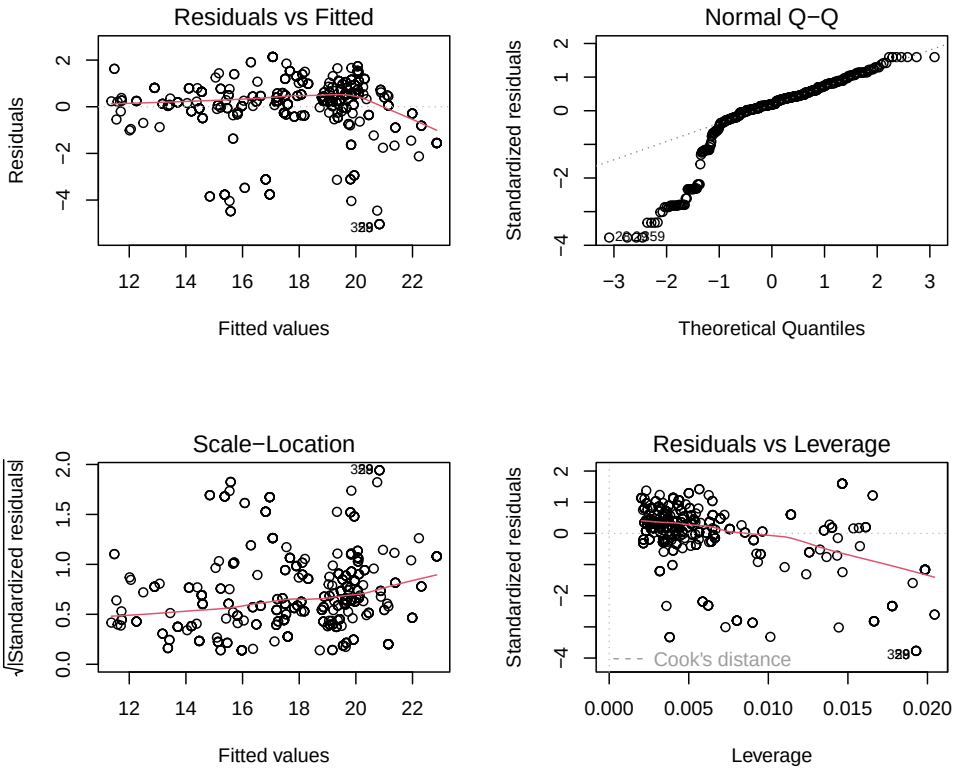
(c) (1p) How would you interpret the effect of Fat?

(d) (2p) Interpret Multiple R-Squared.

(e) (2p) Construct a 95% confidence interval for the regression coefficient for Fat. Use λ to denote the quantile you need. Explain also how you would find λ if applicable, e.g., from which distribution you would find λ, and what the degrees of freedom is, etc.

(f) (2p) Do the residuals of this model have a zero sample mean?

(g) (2p) The residual plots of the fitted model is



What conclusions can you draw from the residual plots?

(h) (1p) What has been calculated by the following R code

```
predict(LR, newdata = data.frame(Water = 50, Fat = 20), interval = "p")
```

4. (8p) Suppose that we have measured the weight of protein of one type of meat produced by five brands. We want to test whether different brands have the same weights. For each brand, 100 samples are measured.

(a) (2p) A statistician has done the following analysis in R.

```
LR <- lm(Protein ~ Brand, data = Data)
anova(LR)

## Analysis of Variance Table
##
## Response: Protein
##            Df Sum Sq Mean Sq F value Pr(>F)
## Brand       4   57.0 14.2480  1.7207 0.1441
## Residuals 495 4098.8  8.2805
```

Do different brands have the same weight?

(b) (2p) Another statistician has done the following analysis in R.

```
LR <- lm(Protein ~ Brand, data = Data)
summary(LR)

##
## Call:
## lm(formula = Protein ~ Brand, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3050 -2.0130  0.9445  2.3068  4.4020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3050     0.2878  63.612   <2e-16 ***
## Brand2       -1.0070     0.4070  -2.474   0.0137 *
## Brand3       -0.2920     0.4070  -0.718   0.4734
## Brand4       -0.2260     0.4070  -0.555   0.5789
## Brand5       -0.3470     0.4070  -0.853   0.3942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.878 on 495 degrees of freedom
## Multiple R-squared:  0.01371,Adjusted R-squared:  0.005744
## F-statistic: 1.721 on 4 and 495 DF,  p-value: 0.1441
```

Based on the analysis of the second statistician, do different brands have the same weight?

(c) (2p) What is the estimated difference in the weight of Protein between Brand 3 and Brand 5?

(d) (2p) Suppose that the statistician wants to test the statement: the weight of Protein of Brand 2 is twice the weight of Protein of Brand 3. Explain how you can perform such test. Note: You don't need to compute the value of your test statistics.

5. (4p) We have a data set from a study of income dynamics of married women in the US. The variables in the data set are lfp (labor-force participation; a factor with levels: no; yes), k5 (number of children 5 years old or younger), age (age of the woman in years), and inc (family income exclusive of wife's income). A model has been fitted as follows.

```
Logit <- glm(lfp ~ k5 + age + inc, family = binomial(link = "logit"),
             data = Mroz)
```

```
summary(Logit)
```

```
##
## Call:
## glm(formula = lfp ~ k5 + age + inc, family = binomial(link = "logit"),
##     data = Mroz)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.867  -1.184   0.731   1.003   1.970
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.394398   0.515576   6.584 4.59e-11 ***
## k5          -1.313316   0.187535  -7.003 2.50e-12 ***
## age         -0.056855   0.010991  -5.173 2.31e-07 ***
## inc         -0.018751   0.006889  -2.722  0.00649 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  956.75  on 749  degrees of freedom
## AIC: 964.75
##
## Number of Fisher Scoring iterations: 4
```

(a) (2p) What is the estimated probability of labor-force participation for a 30 years old women with 0 child 5 years old or younger, and the family income exclusive of wife's income is 10? It suffices to present the formula without presenting the final number.

(b) (2p) A second model has been fitted as follows.

```
LR <- lm(lfp ~ k5 + age + inc, data = Mroz)
```

Which model is more plausible? Motivate your answer.