This is a draft of the exam solutions. You will find answers for every problem here, but since the four versions of the exams are similar, I will only provide full solutions for one version.

Please let me know if you suspect any errors in my answers or solutions. Email me at:

ulf.hognas@stat.su.se

Have a great summer!

/Ulf

Answer form for multiple choice. You can make your own form, put please be clear and answer on <u>one</u> page. Do <u>not</u> submit solutions to the multiple-choice problems.

Number	Part	A	B	C	D	E
1	a.	☐	☐	☐	☐	■
1	b.	☐	☐	■	☐	☐
5	a.	☐	☐	■	☐	☐
5	b.	■	☐	☐	☐	☐
9	a.	☐	☐	☐	■	☐
9	b.	☐	☐	■	☐	☐
13	a.	☐	☐	☐	☐	■
13	b.	☐	☐	■	☐	☐
17	a.	☐	☐	■	☐	☐
17	b.	☐	☐	☐	☐	■

Answer form for multiple choice. You can make your own form, put please be clear and answer on <u>one</u> page. Do <u>not</u> submit solutions to the multiple-choice problems.

Number	Part	A	B	C	D	E
2	a.	☐	☐	■	☐	☐
2	b.	☐	☐	☐	■	☐
6	a.	☐	☐	■	☐	☐
6	b.	☐	■	☐	☐	☐
10	a.	☐	☐	☐	☐	■
10	b.	☐	☐	■	☐	☐
14	a.	☐	☐	☐	■	☐
14	b.	☐	☐	☐	☐	■
18	a.	☐	☐	■	☐	☐
18	b.	☐	☐	☐	■	☐

Answer form for multiple choice. You can make your own form, put please be clear and answer on <u>one</u> page. Do <u>not</u> submit solutions to the multiple-choice problems.

Number	Part	A	B	C	D	E
3	a.	☐	☐	☐	☐	■
3	b.	☐	☐	■	☐	☐
7	a.	☐	☐	■	☐	☐
7	b.	■	☐	☐	☐	☐
11	a.	■	☐	☐	☐	☐
11	b.	☐	☐	■	☐	☐
15	a.	☐	☐	☐	■	☐
15	b.	☐	☐	■	☐	☐
19	a.	☐	■	☐	☐	☐
19	b.	☐	☐	■	☐	☐

Answer form for multiple choice. You can make your own form, put please be clear and answer on <u>one</u> page. Do <u>not</u> submit solutions to the multiple-choice problems.

Number	Part	A	B	C	D	E
4	a.	☐	☐	☐	☐	■
4	b.	☐	■	☐	☐	☐
8	a.	☐	☐	■	☐	☐
8	b.	☐	☐	☐	■	☐
12	a.	■	☐	☐	☐	☐
12	b.	☐	☐	■	☐	☐
16	a.	☐	■	☐	☐	☐
16	b.	☐	☐	☐	■	☐
20	a.	■	☐	☐	☐	☐
20	b.	☐	☐	☐	■	☐

Problem 1

[…]

a) Find the 80th percentile of the students' scores, according to the method taught in the course. (5p) *Choose the alternative closest to your answer.*

Percentiles: Let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ denote the *ordered* sample, ordered by size from the smallest value $x_{(1)}$ to the largest $x_{(n)}$.

Let $a =$ integer part of $(n+1)\frac{p}{100}$

Let $b =$ decimal part of $(n+1)\frac{p}{100}$

p:te percentile $= x_{(a)} + b \cdot (x_{(a+1)} - x_{(a)})$

Ex.: 25:th percentile and $n = 9 \Rightarrow a = 2$ and $b = 0{,}5$

$\Rightarrow Q_1 = x_{(2)} + 0{,}5 \cdot (x_{(3)} - x_{(2)})$

First, sort the data, lowest to highest:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	6	8	9	11	12	12	13	15	23	26	27	27	41	54	57	69	70	84	100

Then, calculate $(n+1)\frac{p}{100} = (21+1)\frac{80}{100} = 16.8$

So, $a = 16$ and $b = 0.80$. We see that the 80th percentile is between the 16th and the 17th values.

Finally, we calculate $x_{(16)} + 0.8\left(x_{(17)} - x_{(16)}\right) = 57 + 0.8\,(69 - 57) = 66.6$

b) Find the probability that at least 9 on the Swedes in the sample drank alcohol during 2020. (5p) *Choose the alternative closest to your answer.*

0	1	2	3	4	5	6	7	8	9	10
10	9	8	7	6	5	4	3	2	1	0

The number of people in the sample who drank alcohol is binomially distributed:

$$X \sim Bin(10, 0.85)$$

We want to use the binomial table, but since $P > 0.50$, we need to use the transformation

$$Y = 10 - X \text{ where } Y \sim Bin(10, 0.15)$$

From the colorful table above, we see that when X at least 9, Y is less than or equal to 1. Hence, we can look up

$P(Y \leq 1) = 0.54430$

directly in the binomial table.

Problem 5

[…]

a) How many different "home teams" are possible? Order does not matter. (5p)

This is $\binom{n}{x}$ with $n = 10$ and $x = 5$.

$$\binom{10}{5} = \frac{10!}{5!\,(10-5)!} = 252$$

b) Find the probability that the home team will be all female (five women). (5p) *Choose the alternative closest to your answer.*

Method 1:
We need two draw five women in a row. At the first draw, the probability that we draw a woman is $\frac{6}{10}$. If we draw a woman the first time, the probability that we draw a woman in the next draw is $\frac{5}{9}$, since one woman is removed from the 10 people, and so on

$$\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6} \approx 0.0238$$

Method 2:
Each unique team has the same probability and there are $\binom{10}{5}$ unique teams. There are $\binom{6}{5}$ different all female teams, so

$$\frac{\binom{6}{5}}{\binom{10}{5}} = \frac{6}{252} \approx 0.0238$$

Problem 9

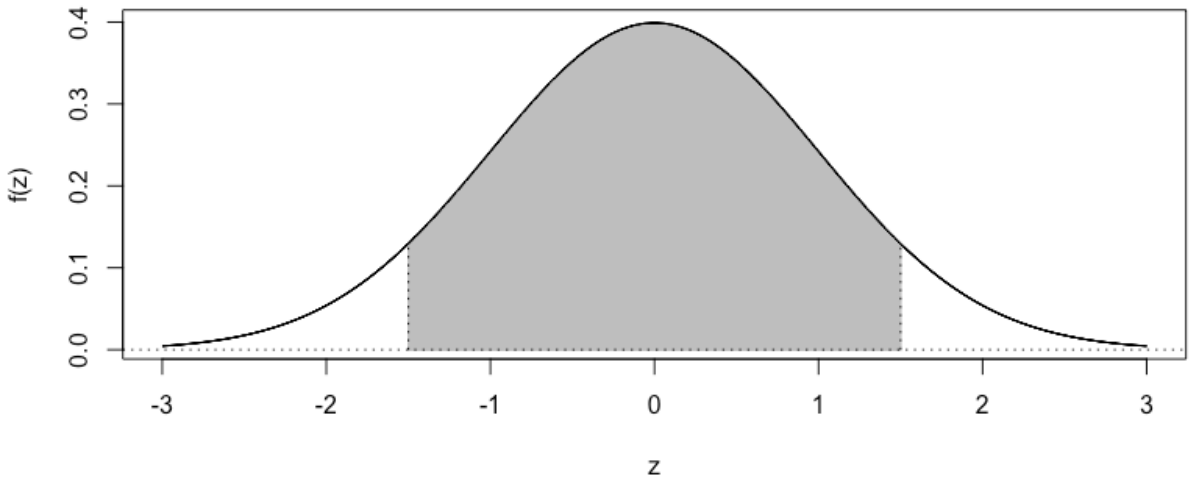
[…]

a) ***Find the probability that the June revenue from the store Vegan Vitamins will be between 170 thousand and 230 thousand SEK. (5p)*** *Choose the alternative closest to your answer.*

We seek $P(170 < Y < 230)$. We standardize

$$P(170 < Y < 230) = P\left(\frac{170-200}{20} < Z < \frac{230-200}{20}\right) = P\left(-\frac{3}{2} < Z < \frac{3}{2}\right)$$

We draw a picture:



Using the symmetries of the normal distribution, we see that this is

$$2 \cdot F(1.5) - 1 = 2 \cdot 0.93319 - 1 = 0.86638$$

b) ***Find the probability that the total June revenue (the two stores combined) will be at least 330 thousand SEK. (5p)*** *Choose the alternative closest to your answer.*

The sum of X and Y is a linear combination of X and Y

$E(aX + bY + c) = aE(X) + bE(Y) + c$ $= a\mu_X + b\mu_Y + c$	$Var(aX + bY + c)$ $= a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$ $= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$

Here, $a = b = 1$ and $c = 0$. So,

$$E[X+Y] = 1 \cdot \mu_X + 1 \cdot \mu_Y = 100 + 200 = 300.$$
$$Var(X+Y) = 1^2 \cdot 10^2 + 1^2 \cdot 20^2 + 2 \cdot 1 \cdot 1 \cdot 120 = 740$$

We seek

$$P(X+Y > 330) = [standardize] = P\left(Z > \frac{330-300}{\sqrt{740}}\right) \approx P(Z > 1.10)$$

$$= 1 - P(Z < 1.10) = 1 - 0.86433 = 0.13567$$

Problem 13

[…]

a) ***Calculate a 95% confidence interval for the mean yearly restaurant and fast-food expenses of single Swedes. (5p)*** *Choose the alternative closest to your answer.*

σ_X^2 unknown: $\bar{x} \pm z_{\alpha/2} \frac{s_x}{\sqrt{n}}$

$\bar{x} = 9800$
$\alpha = 1 - 0.95 = 0.05$
$\frac{\alpha}{2} = 0.025$
$z_{0.025} = 1.96$
$s_x = 36000$
$n = 900$
$9800 \pm 1.96 \frac{36000}{\sqrt{900}} \Longrightarrow (7448, 12152)$

b) ***Based on this sample, create a 90% confidence interval for the proportion of Swedish adults who own stocks. (5p)***

– for the proportion *P*

$nP(1-P) > 5$	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\hat{p} = \frac{42}{200} = 0.21$
$\alpha = 1 - 0.90 = 0.10$
$\frac{\alpha}{2} = 0.05$
$z_{0.05} = 1.6449$
$n = 200$
$0.21 \pm 1.6449 \sqrt{\frac{0.21 \cdot (1 - 0.21)}{200}} \Longrightarrow (0.16, 0.26)$

Problem 17

[…]

a) ***Calculate the value of the test variable (5p)****. Choose the value closest to your answer.*

This is

- small sample sizes
- normal distribution
- unknown variance

σ_X^2, σ_Y^2 unknown and assumed equal:

$$t_{n_x+n_y-2} = \frac{\bar{X}-\bar{Y}-D_0}{S_p\sqrt{1/n_x+1/n_y}}$$

$$\text{where } s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$$

Note that in Problems 18, 19, and 20, the sample sizes are larger than 30. We have

- large sample sizes
- normal distribution
- unknown variance

and hence,

σ_X^2, σ_Y^2 unknown:

$$Z = \frac{\bar{X}-\bar{Y}-D_0}{\sqrt{s_x^2/n_x+s_y^2/n_y}}$$

However, whether we use a t-distribution or a normal distribution, we get the same answer for both part a) and b). Do you see why the test statistic will be the same? As for part b), remember that the t-distribution is very close to the normal distribution for large n.

$$H_0: \mu_x - \mu_y = 0$$
$$H_1: \mu_x - \mu_y > 0$$
$$D_0 = 0$$
$$\bar{x} = 8$$
$$\bar{y} = -4$$
$$s_x^2 = 67^2$$
$$s_y^2 = 70^2$$
$$n_x = 21$$
$$n_y = 21$$
$$s_p^2 = \frac{(21-1)67^2 + (21-1)70^2}{21+21-4}$$

$$= \frac{67^2 + 70^2}{2}$$

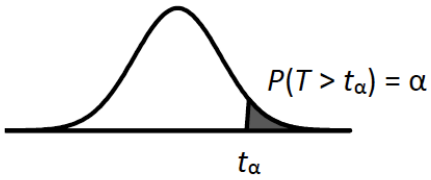
$$t_{obs} = \frac{8 - (-4) - 0}{\sqrt{\frac{67^2 + 70^2}{2}} \cdot \sqrt{\frac{1}{21} + \frac{1}{21}}} \approx 0.57$$

b) ***Use your formula sheet to find the interval that contains the p-value of their test statistic. (5p)***

We want to find the probability of the right tail beyond the German scientists' test statistic, which was 1.2. Look at the t-distribution in table 3:

$T \in t(\nu)$ where ν = degrees of freedom.

The value of t_α if $P(T > t_\alpha) = \alpha$ where α is a given probability. Also, use that $P(T < -t_\alpha) = P(T > t_\alpha)$.



ν	α = 0.1	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
35	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
45	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520

We have $\upsilon = 21 + 21 - 2 = 40$ degrees of freedom. We see that the lowest value in the table for 40 degrees of freedom is 1.303. When the value on the number line is 1.303, the probability of the tail is 0.1. But our value is to the left of 1.303; what does that mean for the area of the tail? It means that the tail must be bigger than 0.1, so the correct answer is the interval:

$(0.1, 1)$

---END OF MULTIPLE-CHOICE PART---

Problem 21

[…]

This is a goodness-of-fit test. John assumes in the null hypothesis that the probabilities of each of the names are the same. His alternative hypothesis is that the probabilities are not all the same, because the lottery is rigged or not fair.

a) State hypotheses and the test variable. (5p)

$$H_0: P_{John} = P_{Mike} = P_{Jane} = P_{Marcy} = 0.25$$
$$H_1: At\ least\ two\ of\ the\ probabilities\ are\ not\ 0.25$$

Test variable:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \quad \text{where } E_i = nP_i$$

b) State the critical value and decision rule. (5p)

There are four categories, so $K = 4$ and $4 - 1 = 3$, so there are 3 degrees of freedom.

v	α = 0.999	0.995	0.99	0.975	0.95	0.05	0.025
1	0.000	0.000	0.000	0.001	0.004	3.841	5.024
2	0.002	0.010	0.020	0.051	0.103	5.991	7.378
3	0.024	0.072	0.115	0.216	0.352	7.815	9.348

Critical value: 7.815
Decision Rule: We reject the null of $X^2_{obs} > 7.815$.

c) Calculate the test statistic and draw conclusion. (5p)

To calculate the sum in the test variable, follow the table method described in the course:

	John	Mike	Jane	Marcy	Sum
Pi	0.25	0.25	0.25	0.25	1
Expected Ei	25	25	25	25	100
Observed Oi	34	15	27	24	100
Oi-Ei	9	-10	2	-1	0
(Oi-Ei)^2	81	100	4	1	186
(Oi-Ei)^2/Ei	3.24	4	0.16	0.04	**7.44**

Alternatively, you can write out the expression for each of the four terms in the sum.

Conclusion: Since $X^2_{obs} = 7.44 \not> 7.815$, we fail to reject the null. We do not have sufficient evidence, at the 5% level, for the hypothesis that the probabilities are not all 0.25.

d) Given that the true probability that John's name is drawn is 25% each week, find the approximate probability that his name will be drawn more than 30 times in the next 100 weeks. (5p)

Each week, either John's name gets drawn or it does not. This "experiment" is repeated 100 times, the experiments are independent, and the probability is 0.25 each time.

Therefore, the number of times that John's name gets drawn is binomially distributed. If X is the number of times, then

$$X \sim Bin(n = 100, P = 0.25)$$

We seek $P(X > 30)$. Since n is larger than 20, we cannot use the table. We use the normal approximation of the binomial distribution taught in class.

$$\mathrm{E}[X] = nP = 100 \cdot 0.25 = 25$$
$$Var(X) = nP(1-P) = 25 \cdot 0.75 = 18.75$$

We can check the rule of thumb (not required for full score) and indeed,

$$nP(1-P) = 18.75 > 5$$

This means that we can approximate the number of "successes" with the normal distribution:

$$P(X > 30) = P\left(Z > \frac{30-25}{\sqrt{18.75}}\right) \approx P(Z > 1.15) = 1 - F(1.15)$$
$$= 1 - 0.87493 = 0.12507 \approx 0.13$$

e) ***Was it possible for the test in part a)-c) to result in a type-II error? Explain. (5p)***

A type-II error is the error of failing to reject the null when the null is false. But it says in the problem text: *In reality, the name drawings are not rigged – John has just been unlucky.* Since the null is true, it was impossible for the test to result in a type-II error.

Problem 22

[…]

a) State the hypotheses, test variable, critical value, and decision rule. (5p)

$H_0: \beta_3 \geq -1000 \mid \beta_1 \neq 0, \beta_2 \neq 0$
$H_1: \beta_3 < -1000 \mid \beta_1 \neq 0, \beta_2 \neq 0$

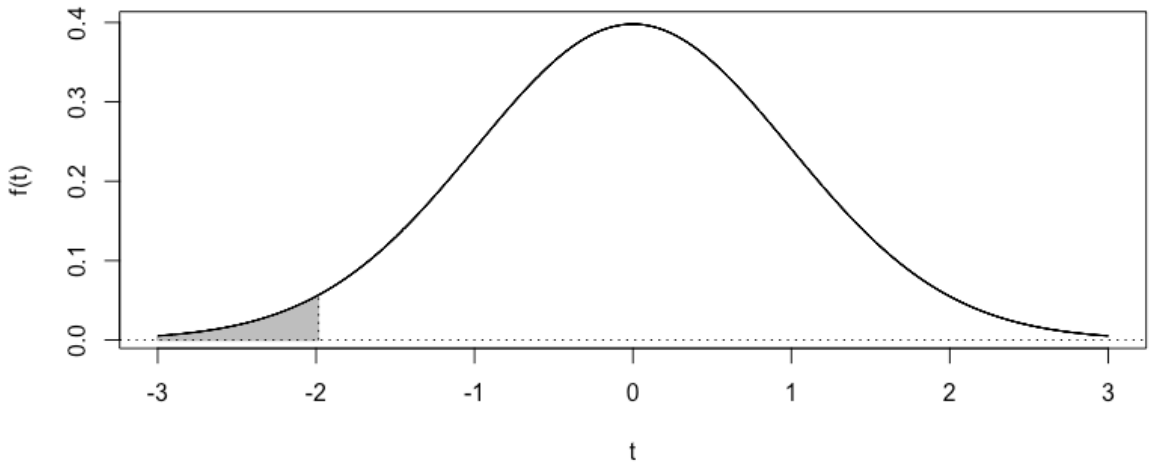
Test variable:

$$t_{n-K-1} = \frac{b_j - \beta_j^*}{s_{b_j}}$$

Critical value:

$$n - K - 1 = 100 - 3 - 1 = 96$$
$$\alpha = 0.05$$
$$\frac{\alpha}{2} = 0.025$$
$$t_{0.025;96} \approx t_{0.025;95} = 1.985$$

We can draw a picture (also not required):



Decision Rule: We reject the null if $t_{obs} < 1.985$.

b) Calculate the value of the test statistic and state your conclusions. (5p)

$$t_{n-K-1} = \frac{b_j - \beta_j^*}{s_{b_j}}$$

We get the estimated slope and the standard error from the output. The slope under the null is given in the problem.

$$b_3 = -1940.6245$$
$$s_{b_3} = 411.6950657$$
$$\beta_3^* = -1000$$
$$t_{obs} = \frac{-1940.6245 - (-1000)}{411.6950657} \approx -2.28$$

Conclusion: Since $t_{obs} = -2.28 < 1.985$, we reject the null. We have evidence at the 5% level for the hypothesis that manual transmission is associated with an expected price that is more than $1000 lower than cars with automatic transmission, all else being equal.

c) **Calculate the coefficient of determination for model 1. Name two possible variables that could be added to the model to increase the coefficient of determination, if we had the data. (5p)**

Method 1:

Coefficient of determination: $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

We can use either of these formulas. For example, we can get SSE and SST from the output:

$$SSE = 329247149.3$$
$$SST = 2735466877$$
$$R^2 = 1 - \frac{329247149.3}{2735466877} = 0.8796377$$

Method 2:

From the output:

Model 1

Regression Statistics	
Multiple R	0.93789001
R Square	[redacted]

Multiple R is the square root of R^2, so we can get R^2 by squaring this value:

$$0.93789001^2 = 0.8796377$$

The coefficient of determination is a measure of how much of the variation in the dependent variable that the model explains. Here it is 88%. This is high since all the cars are of the same brand and model. How can we get it even closer to 100%? Some examples, if we had the data we could:
add the number of previous owner of the used car as a variable
add a dummy variable for condition of the exterior (e.g. great/not great)
add a dummy variable for extra equipment/no extra equipment
add a dummy variable for Diesel/Gasoline fuel type

d) ***Figure 3 shows a histogram of the residuals from Model 1. A friend of the student remarks "six of the residuals are more than two standard deviations from the mean, so according to the empirical rule, the residuals are likely not normally distributed." Are both parts of the friend's remark correct? Explain. (5p) (tip: the standard deviation of the residuals can be found in the output)***

First part: *six of the residuals are more than two standard deviations from the mean*

The standard deviation of the residuals can be found as "Standard Error" in the top box of the output, so the standard deviation is:

Regression Statistics	
Multiple R	0.93789001
R Square	[redacted]
Adjusted R Square	
Standard Error	1851.93353
Observations	100

In the histogram, we see that six of the residuals are at a distance 4000 or more from the mean. Two standard deviations is $1852 \cdot 2 = 3704$, so this part if correct.

Second part: *so according to the empirical rule, the residuals are likely not normally distributed*

According to the empirical rule (from L2):

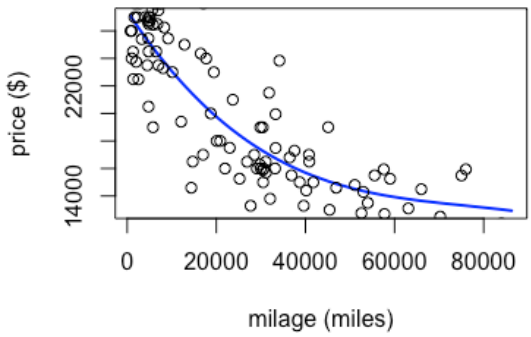
Rule	$\mu \pm \sigma$	$\mu \pm 2\sigma$	$\mu \pm 3\sigma$	
Chebyshev:	0 %	75 %	88,89 %	Guaranteed
Empirical:	ca 68 %	ca 95 %	ca 100 %	Under some conditions

We did not know it during L2, but now we know that "some conditions" means an approximately normal distribution.
Six of the observations are outside of the $\pm 2\sigma$ interval and there are 100 observations total. This means that 94% of the observations (residuals) ended up inside this interval. This is consistent with the empirical rule. For more see:
https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

e) Figure 4 shows pair-wise scatterplots of all the variables. Notice the scatterplot between PRICE and MILAGE. Is this pattern a problem for the validity of Model 1? Explain. (5p)

This pattern shows correlation between price and milage. But this is good! We want the independent variables to be correlated with the dependent variable. This means that PRICE can be explained by MILAGE, which is what we hope for with Model 1, so this is not very concerning.

However, if we look carefully, we can detect a bow shaped pattern in the scatter plot, as illustrated with a blue line here:



This is a sign that the relationship between milage and price might not be linear, which would be a violation of one of the assumptions of linear regression. (This could be resolved. For example, one could try adding quadratic and cubic terms to the model, but that is outside the scope of this course.)

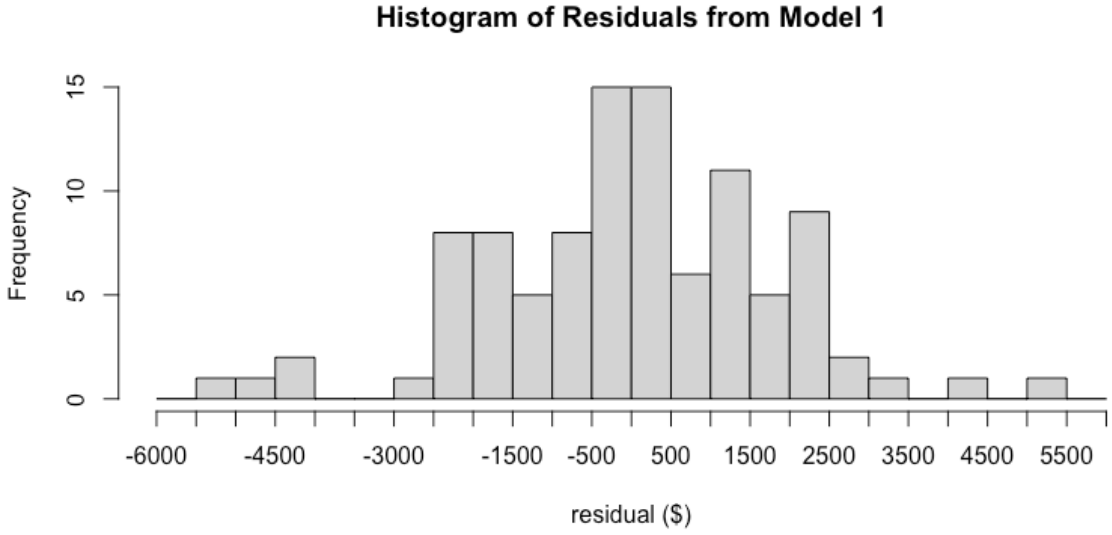


Figure 3: Histogram

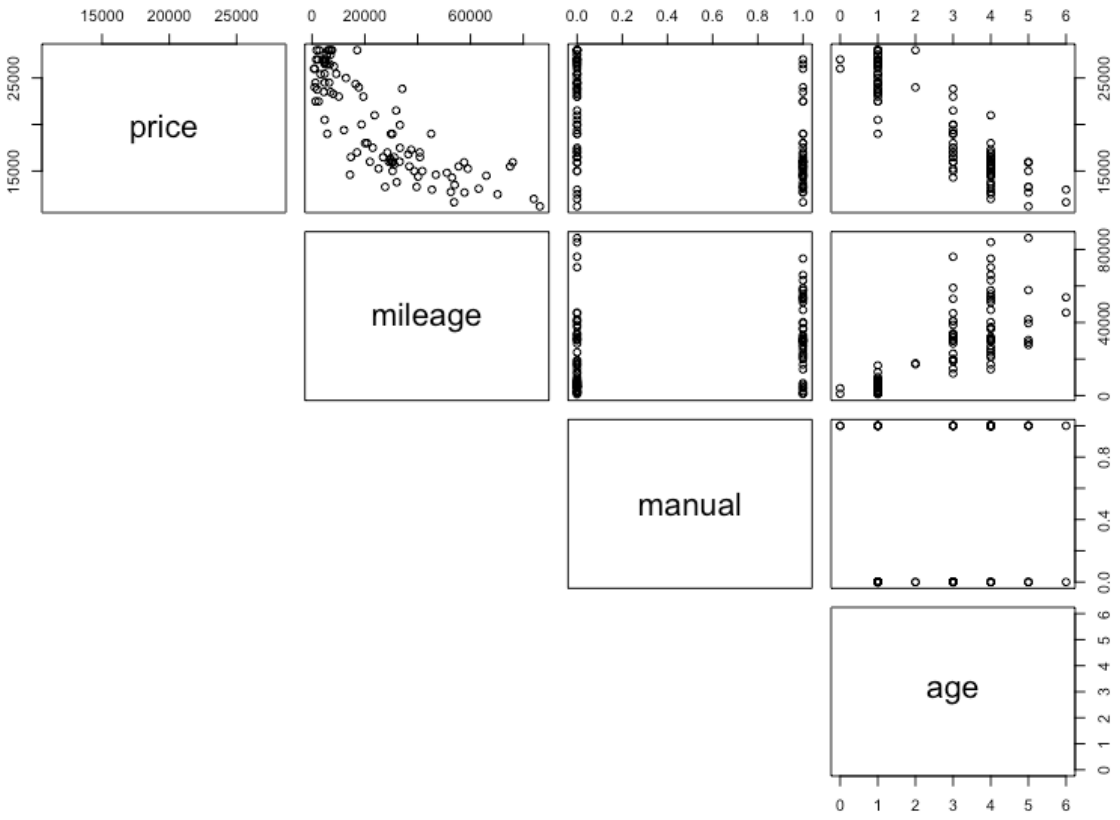


Figure 4: Scatter Plots

---END OF EXAM---