Draft Solutions

The original version of this exam contained typos, which have been fixed here. Students were compensated for the typos that made the exam harder to solve.

Ulf Högnäs
Course Coordinator

Answer form for multiple choice. You can make your own form, put please be clear and answer on one 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 one 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 one 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 one page. Do not 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.	□	■	□	□	□

Solutions to Multiple Choice Problems 1, 5, 9, 13, 17, 21, 25

1a.

We see that the min, max, and median are the same for all five box plots. Hence, we need to find the first and third quartiles for the data. Remember: those values are represented by the bottom and top of the box. We start by sorting the data, lowest to highest:

Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Value	30	31	33	33	40	40	42	43	44	45	45	50	58	58	60	60	64	65	65	70

Following the formula sheet:

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 $= x_{(a)} + b \cdot (x_{(a+1)} - x_{(a)})$

25th percentile:

$$(20+1)\frac{25}{100} = 5.25$$

So $a = 5$ and $b = 0.25$

So the 25th percentile must be between the 5th and the 6th smallest observation, but they are both 40, so the 25th percentile is 40. Or, we can use the formula:

$$x_{(5)} + 0.25\left(x_{(6)} - x_{(5)}\right) = 40 + 0.25(40 - 40) = 40$$

Similarly, the 75th percentile is 60.

Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Value	30	31	33	33	40	40	42	43	44	45	45	50	58	58	60	60	64	65	65	70

Reading the diagram, we see that the correct box plot is number 2.

Answer (B)

1b.

We start by finding the cumulative frequencies:

Gender	**≤15000**	**≤20000**	**≤25000**	**≤30000**	**≤35000**	**>35000**
Women	199	282	374	433	462	486
Men	144	195	278	375	442	514

We want the cumulative frequencies, relative to the total count of men and women and men respectively. We divide each of the cells on the women's row by 486 and each cell of the men's row by 514:

Gender	≤15000	≤20000	≤25000	≤30000	≤35000	>35000
Women	0.41	0.58	0.77	0.89	0.95	1.00
Men	0.28	0.38	0.54	0.73	0.86	1.00

We see that the answer is 77%, so 77% of women have an income of 25000 or less, in the sample.

Answer (E)

5a.

The problem text is a hint: we will use the Law of total probability:

Law of total probability:	$P(A) = P(A \cap E_1) + \cdots + P(A \cap E_K)$ $= P(A\|E_1)P(E_1) + \cdots + P(A\|E_K)P(E_K) = \sum_{i=1}^{K} P(A\|E_i)P(E_i)$

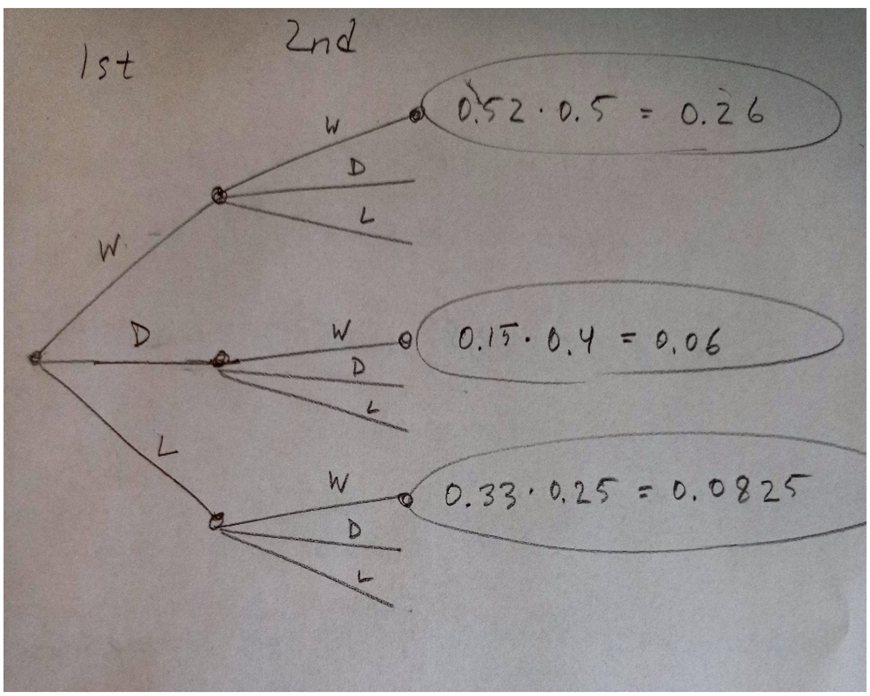
In words, the probability that she wins the second game is equal to

…the probability she wins the first game and then the win second game, plus

…the probability she draws the first game and then the wins second game, plus

…the probability she loses the first game and then the wins second game.

You can easily see this by drawing a tree diagram.



Or formally

P(wins 2^{nd}) =
P(wins 2^{nd} | wins 1^{st})*P(wins 1^{st}) + P(wins 2^{nd} | draws 1^{st})*P(draws 1^{st}) + P(wins 2^{nd} | loses 1^{st})*P(loses 1^{st})

$$= 0.52 \cdot 0.5 + 0.15 \cdot 0.4 + 0.33 \cdot 0.25 = 0.4025$$

Answer (C)

5b.

$$\mu_X = E(X) = \sum_{x \in S_x} xP(x)$$

$$\sigma_X^2 = Var(X) = E(X^2) - \mu_X^2 = \sum_{x \in S_x} x^2 P(x) - \mu_X^2 = \sum_{x \in S_x} (x - \mu_X)^2 P(x)$$

We let X be the price money. Make a table

	Win	Draw	Lose	Sum
x	1000	500	250	
P(x)	0.2	0.4	0.4	1
P(x)*x	200	200	100	500
(x-μ)	500	0	-250	
(x-μ)^2	250000	0	62500	
P(x)*(x-μ)^2	50000	0	25000	75000

The sum of the row for P(x)*x is the expectation, so E[X] = μ = 500.

The sum of the row P(x)*(x-μ)^2 is the variance, so Var(X) = 75000.

We then take the square root of the variance to get the standard deviation: 274

Answer (A)

5b.

The revenue is a linear combination of the number of t-shirts produced. If R is the revenue and X the number of t-shirts produced then

$$R = 3 \cdot X$$

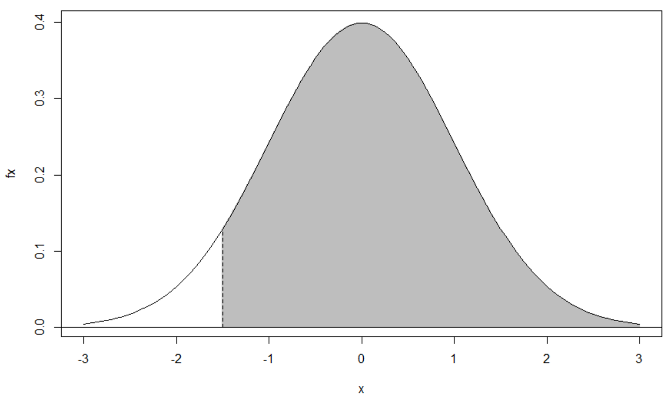
Using the formulas for linear combinations, we get

$$\mathrm{E}[R] = \mathrm{E}[3 \cdot X] = 3 \cdot \mathrm{E}[X] = 3 \cdot 2300 = 6900$$
$$\mathrm{Var}(R) = \mathrm{Var}(3 \cdot X) = 3^2\, \mathrm{Var}(X) = 3^2 \cdot 200^2$$
$$\sigma_R = \sqrt{3^2 \cdot 200^2} = 600$$

So we seek the probability that $R \geq 6000$. We standardize:

$$P(R \geq 6000) = P\left(\frac{R-6900}{600} \geq \frac{6000-6900}{600}\right) = P\left(Z \geq \frac{-900}{600}\right) = P(Z \geq -1.5)$$

We draw a picture:



We seek the area of the shaded part. Following the procedure taught in class, we use symmetry to get

$P(Z \geq -1.5) = F(1.5) = 0.93319$

Where the last value comes from table 1.

Answer (E)

Note: the true answers to 10a, 11a, and 12a round to 99%, 99%, and 100%, respectively. These exact alternatives were not listed, because of a typo. I regret the errors.

9b.

X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

We are looking for the probability for the orange area, i.e. zero to eighteen people out of twenty. The number of people in the sample who have at least one computer is binomially distributed with $P = 0.95$ and $n = 20$. Since the table only goes to $P = 0.50$ we create a new variable Y which is the number of people with no computer.

$$Y \sim Binom(n = 20, P = 0.05)$$

$$P(X \leq 18) = P(Y \geq 2) = 1 - P(Y \leq 1)$$

We get $P(Y \leq 1)$ from tabell 7 page 19:

$$1 - P(Y \leq 1) = 1 - 0.73584 \approx 26\%$$

Answer (C)

13a

Paired data, small sample size, unknown variance, normal distribution:

– for the difference $\mu_D = \mu_X - \mu_Y$ (paired dependent samples, $D_i = X_i - Y_i$)

$n < 30$, D_i normally distributed:	σ_D^2 unknown	$\bar{d} \pm t_{n-1;\alpha/2} \frac{s_d}{\sqrt{n}}$

We need $\bar{d}, \frac{\alpha}{2}, t_{n-1;\,\alpha/2}$, och s_d

We start by finding the differences for each year. We calculate the mean difference

$\bar{d} = \frac{6+6+20}{3} = 10.6667$

	2018	2019	2020	Sum
A	89	140	150	
B	83	134	130	
d	6	6	20	32
(d-d_bar)	-4.66667	-4.66667	9.333333	
(d-d_bar)^2	21.77778	21.77778	87.11111	130.6667

We find the sample variance.

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s_d^2 = \frac{130.6667}{3-1}$$

$$s_d = \sqrt{\frac{130.6667}{3-1}} = 8.0829$$

$$\frac{\alpha}{2} = 0.025$$

$t_{2;\,0.025} = 4.303$

Putting it all together we get

$(-9.41, 30.7)$

Answer (E)

13b.

Here we have known variance and normal distribution. In those cases, it does not matter if we have large or small sample. We get the following formula:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$$

This part if the margin of error:

$$z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}$$

$z_{0.025} = 1.96$

We get the following equation

$10 = 1.96\ \frac{100}{\sqrt{n}}$
$\sqrt{n} = 1.96\ \frac{100}{10}$
$\sqrt{n} = 19.6$
$n =\ 19.6^2 =\ 384.16$
So we need a sample size of at least 385

Answer (C)

17a.

Since the probability that the self-proclaimed expert is assumed to be some fixed P and we have 1000 "experiments" that can succeed or fail, the number of correct guesses is binomially distributed. We can check the rule of thumb:

$$n\,P_0(1-P_0) = 1000\frac{1}{37}\left(1-\frac{1}{37}\right) = 26.29657$$

Which is way more than five. Accoring to the central limit theorem, then, both the number of correct guesses and the proportion of correct guesses are approximnately normal distributed.

So the critical value will be

$$z_{0.1} = 1.28$$

Answer: (A)

17b.

We find the correct test variable

- for the proportion *P*

$nP_0(1-P_0) > 5$	$Z = \frac{\hat{p}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$

$$P_0 = \frac{1}{37}$$
$$n = 1000$$
$$\hat{p} = \frac{33}{1000}$$

So,

$$z_{obs} = \frac{\frac{33}{1000} - \frac{1}{37}}{\sqrt{\frac{\frac{1}{37}\left(1-\frac{1}{37}\right)}{1000}}} = 1.16$$

(We fail to reject the null)
Answer (A)

21a.

This is an independence test.

Assumptions: independent, identically distributed sample

Hypotheses:

$$H_0\text{: being part of organized sports and home city are } \textbf{independent}$$
$$H_1\text{: being part of organized sports and home city are } \textbf{dependent}$$

Test variable:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = \frac{R_i C_j}{n}$$

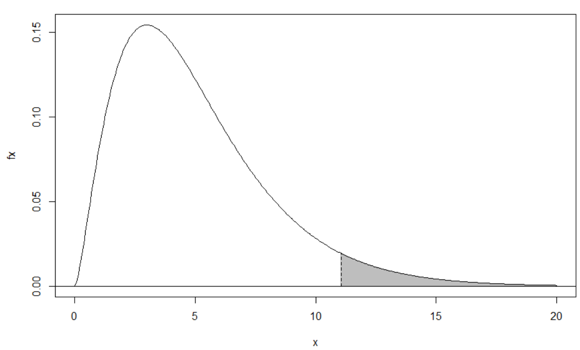
21b.

The degrees of freedom are $(\#rows - 1)(\#columns - 1) = 5$.

So the testvariable is χ^2distributed with 5 degrees of freedom. We look up the critical value in the table:

$$\chi^2_{0.05;5} = 11.070$$

We can draw a picture if we want (but you certainly did not have to):



Rule: Reject the null hypothesis if $\chi^2_{obs} > 11.070$.

21c.

We follow the procedure from the course and create a series of tables. The sum of the cells in the last table is precisely the double sum that we seek:

Observed Oij	Malmö	Göteborg	Stockholm	Uppsala	Västerås	Örebro	Ri
Part	6	5	38	12	6	8	75
Not Part	14	15	52	16	14	14	125
Ci	20	20	90	28	20	22	200

…

Expected Eij	C1	C2	C3	C4	C5	C6
R1	7.5	7.5	33.75	10.5	7.5	8.25
R2	12.5	12.5	56.25	17.5	12.5	13.75

Difference Oij-Eij						
	-1.5	-2.5	4.25	1.5	-1.5	-0.25
	1.5	2.5	-4.25	-1.5	1.5	0.25

(Oij-Eij)^2/Eij						
	0.3	0.833333	0.535185	0.214286	0.3	0.007576
	0.18	0.5	0.321111	0.128571	0.18	0.004545

Going through the double sum is the same as going to every cell in the last table and adding up the values. We get the sum of 3.50

Conclusion:

Since $\chi^2_{obs} = 3.50 \ngtr 11.070$ we **fail to reject the null**. We have **not** found statistically significant evidence at the 5% level for the hypothesis that propensity to be part of organized sports is dependent on home city.

21d.

To perform a χ^2test, we need to have a large enough sample. How do we check that we have a large enough sample? We check that the expected count E_{ij} is at least 5 in every cell:

Cannabis Use	Malmö	Göteborg	Stockholm	Uppsala	Västerås	Örebro	
Yes	3	2	15	6	6	3	35
No	17	18	75	22	14	19	165
	20	20	90	28	20	22	200

Expected Eij	C1	C2	C3	C4	C5	C6	

R1	3.5	3.5	15.75	4.9	3.5	3.85	35
R2	16.5	16.5	74.25	23.1	16.5	18.15	165
	10	20	42	12	8	8	100

The cannabis data fails this in five of the six cells. We could, for example, combine five cities to the category "not Stockholm," but I say we collect a larger sample or that we just forget about the whole test.

21e.

This is repeated draws without replacement. There are 90 students from Stockholm in the sample of 200. Probability of getting a student not from stockholm on the first draw is

$$\left(1-\frac{90}{200}\right)=\frac{110}{200}$$

The probability of getting a student not from stockholm on the second draw, given that we drew somone not from Stockholm on the first draw is

$$\left(1-\frac{90}{199}\right)=\frac{109}{199}$$

And similarly, if we got two students not from Stockholm on the first draw, the third probability is

$$\left(1-\frac{90}{198}\right)=\frac{108}{198}$$

We multiply these probabilities to get the answer

$$\frac{110}{200}\cdot\frac{109}{199}\cdot\frac{108}{198}\approx 0.164$$

It may help to draw a tree diagram in this case, if you find this confusing.

25a.

We start by finding the mean of x and y

$$\bar{y}=\frac{11}{11}=1$$
$$\bar{x}=\frac{16.5}{11}=1.5$$

To find the variance x and covariance of x and y, we use the following formulas:

Variance: $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Covariance: $s_{xy} = Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

We can create a table

												Sum
Returns (y)	-1	0	1	-2	0	1	2	1	8	1	0	11
Bllshnss (x)	0.6	1.8	1.2	1.6	1.1	1.3	1.7	1.7	2.6	0.7	2.2	16.5
x-μ_x	-0.9	0.3	-0.3	0.1	-0.4	-0.2	0.2	0.2	1.1	-0.8	0.7	0
(x-μ_x)^2	0.81	0.09	0.09	0.01	0.16	0.04	0.04	0.04	1.21	0.64	0.49	3.62
y-μ_y	-2	-1	0	-3	-1	0	1	0	7	0	-1	0
(y-μ_y)*(x-μ_x)	1.8	-0.3	0	-0.3	0.4	0	0.2	0	7.7	0	-0.7	8.8

The variance of x is

$$\text{Var}(X) = \frac{3.62}{11-1} = 0.362$$

The covariance of x and y is

$$\text{Cov}(X) = \frac{8.8}{11-1} = 0.88$$

25b.

To estimate the simple linear regression model, we use

$$b_1 = \frac{Cov(x, y)}{s_x^2} = r_{xy}\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

We found these values in part a:

$$b_1 = \frac{0.88}{0.362} = 2.43$$
$$b_0 = 1 - 2.43 \cdot 1.5 = -2.645$$

The estimated model is

$$\hat{y} = -2.645 + 2.43\,x$$

25c.

We will use the following formula:

Prediction interval for the prediction of y given $X = x$: $\quad (b_0 + b_1 x) \pm t_{n-2,\alpha/2}\sqrt{s_e^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$

$$b_1 = 2.43$$
$$b_0 = -2.645$$
$$\frac{\alpha}{2} = 0.05$$
$$t_{9;0.025} = 2.262$$
$$s_e^2 = 4.96$$
$$x = 2$$
$$\bar{x} = 1$$
$$s_x^2 = 0.362$$

All these values can be taken from the problem or from part (a) and part (b). Putting it all together, we get

(-3.675, 8.105)

Interpretaion: According to our model, there is a 95% probability that the percentage return of Bitcoin will be betweem -3.209% and 7.640%, given that the Bullishness value was 2 the previous day.

2d.

This was meant to be the hardest question on the exam. We are asked to calculate the test variable.

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

Since the David tests whether the slope is zero, $\beta_j^* = 0$. We also know that $b_1 = 2.43$. We still need

$$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2} = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

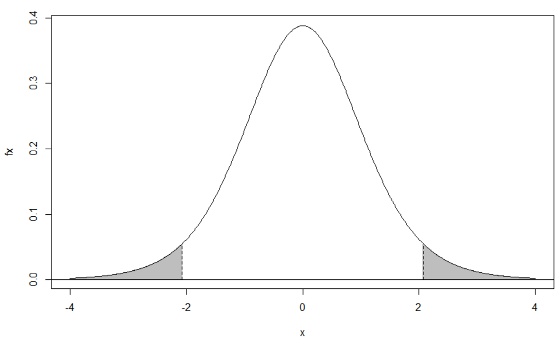
We get the denominator from our work in part (a). The numerator is given earlier in the problem:

$$s_{b_1}^2 = \frac{4.96}{3.62}$$

So we get (with some minor rounding error; I have rounded on the way to make the solution easy to read)

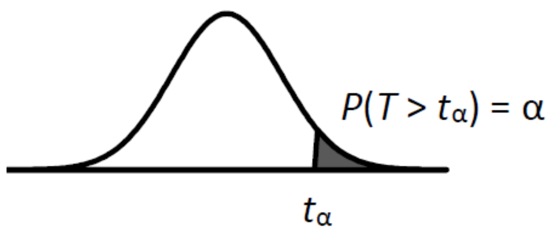
$$t_{obs} = \frac{2.43 - 0}{\sqrt{\frac{4.96}{3.62}}} = 2.076$$

Now the p-value. Let us draw a picture:



So one formulation of the p-value is this: if the null is true and we could repeat the sample, how likely is it what we would get a value like t_{obs} or something more extreme? This probability of the same as the areas of the two tails. We cannot get this are exactly from table 3, but we can get something close:

v	α = 0.1	0.05	0.025
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571
6	1.440	1.943	2.447
7	1.415	1.895	2.365
8	1.397	1.860	2.306
9	1.383	1.833	2.262



We see that the area of the right tail must be between 0.05 and 0.025. We know for sure that the right tail is larger than 0.025. Since the p-value is the area of the two tails, we know that the p-value must be between 0.10 and 0.05.

We reject the null when the p-value is smaller than alpha, but we see that the p-value is larger than alpha. Hence David would fail to reject the null. If you are interested, you can get the p-value exactly with any of these commands in Excel:

=T.DIST.2T(2.076, 9) (this gives us two tails)

=2*T.DIST.RT(2.076, 9) (we can take two times the right tail)

=2*(1-T.DIST(2.076, 9,1)) (we can take two times the complement of the left tail)

But this is not necessary. I will grade this one generously if you have shown some understanding of the problem.

25e.

When we create a scatter plot of the residuals like in the left plot, we want the residual to form a somewhat even cloud around the x-axis (see chapter 13.6 from the book or my lecture L17). This is not what we see here. We see that the residuals on form a funnel, where all the largest residuals are those for high estimated y-values. In the right plot, the plot with square residuals makes this even more clear. This problem is called ***Heteroscedasticity***.

Heteroscedasticty is a problem. One of the assumptions of the linear regression model is *homoscedasticity*, which is the opposite of heteroscedasticity. Homoscedasticity means that the variance is the same for for all values of x and does not depend on size of the independent variable or the size of the estimated dependent variable. When heteroscedasticity is present, we can no longer trust our confidence intervals or hypothesis tests.