# Lab 04A: Week 11 (Solutions)

## Contents



The **specific aims** of this lab are:

- perform simple and multiple regression
- identify which variables are significant and perform model selection
- check whether or not the assumptions for linear regression are met
- interpret the coefficients from a linear regression model
- use an estimated regression model to predict the outcomes for a new observation

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO6 Formulate, evaluate and interpret appropriate linear models to describe the relationships between multiple factors.

# 1 Questions

## 1.1 Wind

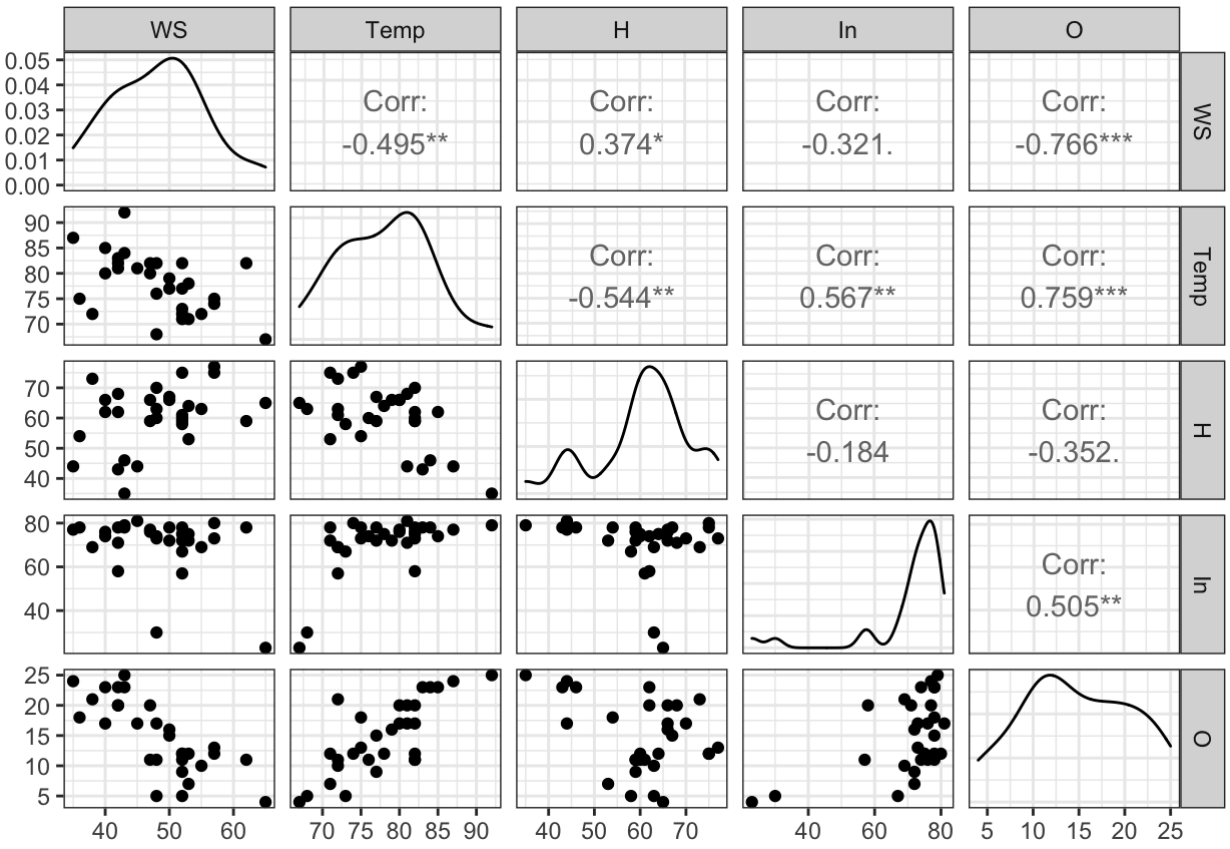
The data in `pollut.txt` are `WS` (wind speeds), `Temp` (temperature), `H` (humidity), `In` (insolation) and `O` (ozone) for 30 days.

```
pollut = read_csv("https://raw.githubusercontent.com/DATA2002/data/master/pollut.txt")
glimpse(pollut)
```

```
Rows: 30
Columns: 5
$ WS   <dbl> 50, 47, 57, 38, 52, 57, 53, 62, 52, 42, 47, 40, 42, 40,…
$ Temp <dbl> 77, 80, 75, 72, 71, 74, 78, 82, 82, 82, 82, 80, 81, 85,…
$ H    <dbl> 67, 66, 77, 73, 75, 75, 64, 59, 60, 62, 59, 66, 68, 62,…
$ In   <dbl> 78, 77, 73, 69, 78, 80, 75, 78, 75, 58, 76, 76, 71, 74,…
$ O    <dbl> 15, 20, 13, 21, 12, 12, 12, 11, 12, 20, 11, 17, 20, 23,…
```

1. Generate a pairs plot of the data using `pairs()` or the `ggpairs()` function from the **GGally** package (Schloerke et al. 2021).

```
# pairs(pollut)
library(GGally)
ggpairs(pollut) + theme_bw()
```




2. Perform a multiple regression of `ozone` on the other variables using `lm()`.

```
pollut_lm = lm(O ~ ., pollut)
# Or pollut_lm = lm(O ~ WS + Temp + H + In, pollut)
summary(pollut_lm)
```

```
Call:
lm(formula = O ~ ., data = pollut)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.5861 -1.0961  0.3512  1.7570  4.0712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.49370   13.50647  -1.147  0.26219
WS           -0.44291    0.08678  -5.104 2.85e-05 ***
Temp          0.56933    0.13977   4.073  0.00041 ***
H             0.09292    0.06535   1.422  0.16743
In            0.02275    0.05067   0.449  0.65728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.92 on 25 degrees of freedom
Multiple R-squared:  0.798, Adjusted R-squared:  0.7657
F-statistic: 24.69 on 4 and 25 DF,  p-value: 2.279e-08
```

3. Does it look like any variables can be dropped from the model? If you were doing backwards selection using the `drop1()` function which would you drop first? Write down a the workflow for a formal hypothesis test to see if the coefficient for insolation is significantly different to zero.

Yes, both humidity and insolation are **individually** insignificant at the 5% level of significance. We can't immediately drop both of them from the model as the p-values are only testing individual coefficients. If we were to drop one first, we would drop insolation as it has the largest p-value.

We can do a formal test to see if the coefficient of insolation is significant as follows.

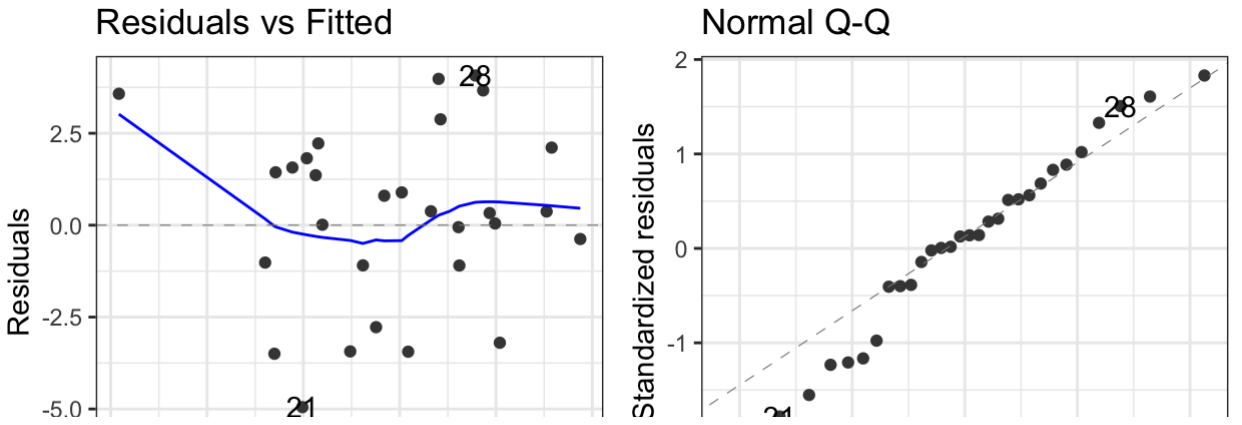
First we define the model with population parameters:

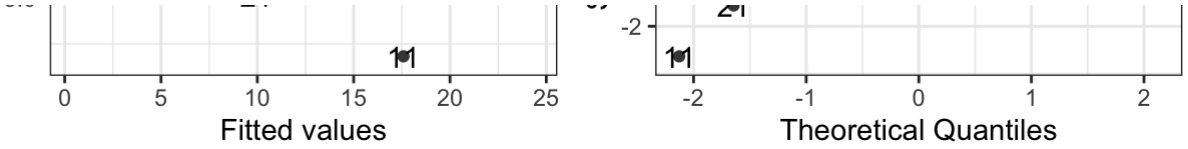
$$\mathrm{O} = \beta_0 + \beta_1 \mathrm{WS} + \beta_2 \mathrm{Temp} + \beta_3 \mathrm{H} + \beta_4 \mathrm{In} + \varepsilon$$

**Hypothesis:** $H_0\colon\ \beta_4 = 0$ vs $H_1\colon\ \beta_4 \neq 0$

**Assumptions:** The residuals $\varepsilon_i$ are iid $N(0, \sigma^2)$ and there is a linear relationship between $y$ and $x$.

```
library(ggfortify)
autoplot(pollut_lm, which = 1:2) + theme_bw()
```






- Linearity: there's no obvious pattern in the residual vs fitted values plot (e.g. no smiley face of frowny face) so it doesn't appear that we have misspecified the model
- Homoskedasticity: the residuals don't appear to be fanning out or changing their variability over the range of the fitted values so the constant error variance assumption is met.
- Normality: in the QQ plot, the points are reasonably close to the diagonal line. The bottom 7 or so points are not quite on the line, but it's not severe enough departure to cause too much concern. The normality assumption is at least approximately satisfied.

**Test statistic:** $T = \dfrac{\hat{\beta}_4}{\text{SE}(\hat{\beta}_4)} \sim t_{n-p}$ under $H_0$ where $p$ is the number of estimated coefficients (including the intercept) and $n$ is the sample size. This is also the degrees of freedom associated with the residual standard error in the R output (i.e. 25).

**Observed test statistic:** $t_0 = \frac{0.02275}{0.05067} = 0.449$ (from R)

**p-value:** $2P(t_{25} \geq |0.449|) = 0.65728$

**Conclusion:** Do not reject $H_0$ at the 5% level of significance as the p-value is greater than 0.05. Hence, there is no evidence to suggest that there is a significant linear relationship between ozone and insolation and it can be dropped from the model.

4. Rather than dropping variables using their individual p-values, we can instead consider using an information criterion. Use the `step()` function to perform selection using the AIC starting from the full model.

```
pollut_step = step(pollut_lm)
```

```
Start:  AIC=68.82
O ~ WS + Temp + H + In

       Df Sum of Sq    RSS    AIC
- In    1     1.719 214.81 67.056
<none>              213.09 68.815
- H     1    17.231 230.32 69.148
- Temp  1   141.424 354.51 82.086
- WS    1   222.041 435.13 88.233

Step:  AIC=67.06
```

```
O ~ WS + Temp + H

       Df Sum of Sq    RSS    AIC
<none>              214.81 67.056
- H     1     20.09 234.90 67.739
- Temp  1    216.28 431.09 85.954
- WS    1    226.96 441.77 86.688
```

```
pollut_step
```

```
Call:
lm(formula = O ~ WS + Temp + H, data = pollut)

Coefficients:
(Intercept)           WS         Temp            H
   -16.6070      -0.4462       0.6019       0.0985
```

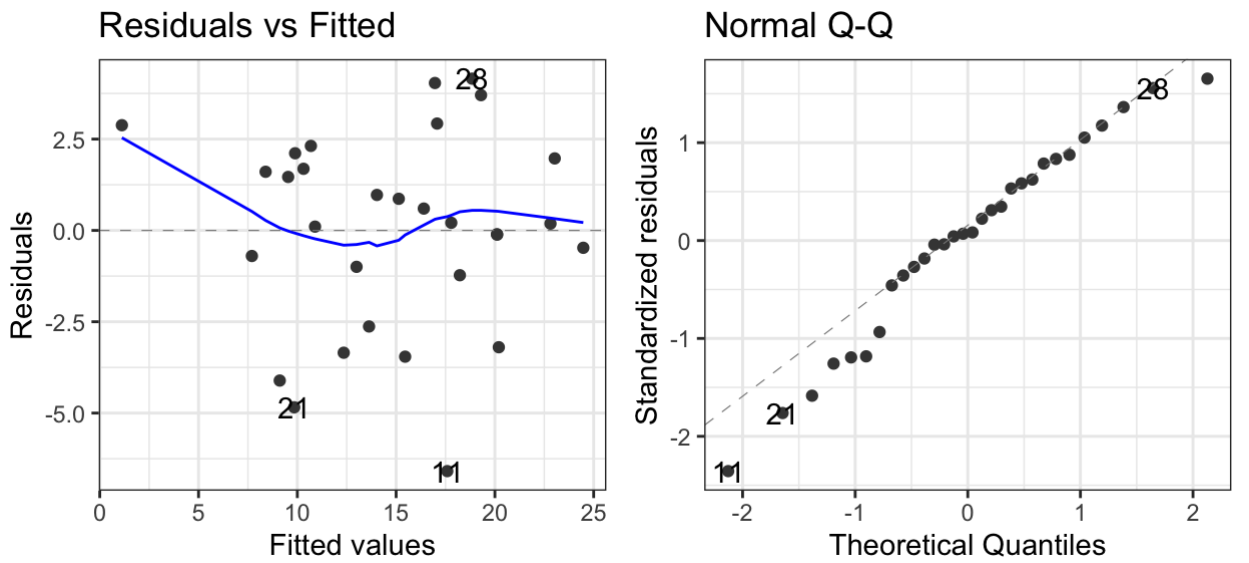
Backwards selection using the AIC has dropped insolation but decided to keep humidity in the model.

5. Write down the fitted model for the model selected by the step-wise procedure.

$$\widehat{\text{Ozone}} = -16.6070 - 0.4462 \times \text{WS} + 0.6019 \times \text{Temp} + 0.0985 \times \text{Humidity}$$

6. Check the linear regression assumptions for the stepwise model.

```
library(ggfortify)
autoplot(pollut_step, which = 1:2) + theme_bw()
```




- Linearity: there's no obvious pattern in the residual vs fitted values plot (e.g. no smiley face of frowny face) so it doesn't appear that we have misspecified the model

- Homoskedasticity: the residuals don't appear to be fanning out or changing their variability over the range of the fitted values so the constant error variance assumption is met.
- Normality: in the QQ plot, the points are reasonably close to the diagonal line. The bottom 7 or so points are not quite on the line, but it's not severe enough departure to cause too much concern. The normality assumption is at least approximately satisfied.

7. What proportion of the variability of ozone is explained by the explanatory variables in the step-wise selected model?

```
summary(pollut_step)
```

```
Call:
lm(formula = O ~ WS + Temp + H, data = pollut)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5887 -1.1686  0.1978  1.9004  4.1544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.60697   13.07154  -1.270    0.215
WS           -0.44620    0.08513  -5.241 1.78e-05 ***
Temp          0.60190    0.11764   5.117 2.47e-05 ***
H             0.09850    0.06316   1.559    0.131
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.874 on 26 degrees of freedom
Multiple R-squared:  0.7964,    Adjusted R-squared:  0.7729
F-statistic: 33.89 on 3 and 26 DF,  p-value: 3.904e-09
```

Looking at the $R^2$ value (multiple R-squared) from the summary output, 80% of the variability of ozone is explained by the regression on wind speed, temperature and humidity.

8. Use the model to estimate the average `ozone` for days when `WS=40`, `Temp=80` and `H=50`. Is a confidence interval or a prediction interval most appropriate here? Write down the interval you think is most appropriate.

```
newdata = data.frame(WS = 40, Temp = 80, H = 50)
predict(pollut_step, newdata, interval = "confidence")
```

```
      fit      lwr      upr
1 18.6218 16.70852 20.53509
```

```
predict(pollut_step, newdata, interval = "prediction")
```

```
      fit      lwr      upr
1 18.6218 12.41146 24.83215
```

Using the regression, the estimate average ozone for days when `WS=40`, `Temp=80` and `H=50` is 18.62.

A confidence interval is more appropriate here, because the question asked about estimating the **average ozone on days when.....**

If instead the question asked: **predict the ozone on a day when...** we'd use a prediction interval instead.

The 95% confidence interval for the estimated ozone level is (16.71, 20.54).

## 1.2 Diabetes

Efron et al. (2004) introduced the diabetes data set with 442 observations and 11 variables. It is often used as an exemplar data set to illustrate new model selection techniques. The following commands will help you get a feel for the data.

```
# install.packages('mplot')
data("diabetes", package = "mplot")
# help('diabetes', package='mplot')
```

```
glimpse(diabetes)  # glimpse the structure of the diabetes
pairs(diabetes)  # traditional pairs plot
GGally::ggpairs(diabetes)  # ggplotified pairs plot
boxplot(diabetes)  # always a good idea to check for gross outliers
boxplot(scale(diabetes))  # always a good idea to check for gross outliers
```

```
# OPTIONAL!!
# install.packages(c("pairsD3","heatmaply","skimr"))
pairsD3::shinypairs(diabetes) # interactive pairs plot of the data set
heatmaply::heatmaply(cor(diabetes))
skimr::skim(diabetes) # summary of the diabetes data
```

We can fit the null model (without any variables) and the full model as follows:

```
M0 = lm(y ~ 1, data = diabetes)  # Null model
M1 = lm(y ~ ., data = diabetes)  # Full model
```

We can compare the results side by side using the **stargazer** package (Hlavac 2018).

```
# stargazer::stargazer(M0, M1, type = 'latex', header = FALSE)
stargazer::stargazer(M0, M1, type = "html")
```

| | *Dependent variable:* | |
|---|---|---|
| | y | |
| | (1) | (2) |
| age | | -0.036 |
| | | (0.217) |
| sex | | -22.860*** |
| | | (5.836) |
| bmi | | 5.603*** |
| | | (0.717) |
| map | | 1.117*** |
| | | (0.225) |
| tc | | -1.090* |
| | | (0.573) |
| ldl | | 0.746 |
| | | (0.531) |
| hdl | | 0.372 |
| | | (0.782) |
| tch | | 6.534 |
| | | (5.959) |
| ltg | | 68.483*** |
| | | (15.670) |
| glu | | 0.280 |
| | | (0.273) |
| Constant | 152.133*** | -334.567*** |
| | (3.667) | (67.455) |
| Observations | 442 | 442 |
| $R^2$ | 0.000 | 0.518 |
| Adjusted $R^2$ | 0.000 | 0.507 |
| Residual Std. Error | 77.093 (df = 441) | 54.154 (df = 431) |
| F Statistic | | 46.272*** (df = 10; 431) |
| *Note:* | | *p<0.1;* ***p<0.05;*** p<0.01 |

1. Try doing backward selection using AIC first.

```r
step_back_aic = step(M1, direction = "backward", trace = FALSE)
summary(step_back_aic)
```

```
Call:
lm(formula = y ~ sex + bmi + map + tc + ldl + ltg, data = diabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-158.275  -39.476   -2.065   37.219  148.690

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
sex          -21.5910     5.7056  -3.784 0.000176 ***
bmi            5.7111     0.7073   8.075 6.69e-15 ***
map            1.1266     0.2158   5.219 2.79e-07 ***
tc            -1.0429     0.2208  -4.724 3.12e-06 ***
ldl            0.8433     0.2298   3.670 0.000272 ***
ltg           73.3065     7.3083  10.031  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.06 on 435 degrees of freedom
Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16
```

2. Explore the forwards selection technique, which works very similarly to backwards selection, just set `direction = "forward"` in the `step()` function. When using `direction = "forward"` you need to specify a scope parameter: `scope = list(lower = M0, upper = M1)`.

```r
step_fwd_aic = step(M0, scope = list(lower = M0, upper = M1), direction = "forward",
    trace = FALSE)
summary(step_fwd_aic)
```

```
Call:
lm(formula = y ~ bmi + ltg + map + tc + sex + ldl, data = diabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-158.275  -39.476   -2.065   37.219  148.690

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
bmi            5.7111     0.7073   8.075 6.69e-15 ***
```

```
ltg           73.3065     7.3083  10.031  < 2e-16 ***
map            1.1266     0.2158   5.219 2.79e-07 ***
tc            -1.0429     0.2208  -4.724 3.12e-06 ***
sex          -21.5910     5.7056  -3.784 0.000176 ***
ldl            0.8433     0.2298   3.670 0.000272 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.06 on 435 degrees of freedom
Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16
```

3. Try using the `add1()` and `drop1()` functions. The general form is `add1(fitted.model, test = "F", scope = M1)` or `drop1(fitted.model, test = "F")`

```
add1(step_fwd_aic, test = "F", scope = M1)
```

```
Single term additions

Model:
y ~ bmi + ltg + map + tc + sex + ldl
       Df Sum of Sq     RSS    AIC F value Pr(>F)
<none>              1271494 3534.3
age     1      10.9 1271483 3536.3  0.0037 0.9515
hdl     1     394.8 1271099 3536.1  0.1348 0.7137
tch     1    3686.2 1267808 3535.0  1.2619 0.2619
glu     1    3532.6 1267961 3535.0  1.2091 0.2721
```

```
drop1(step_fwd_aic, test = "F")
```

```
Single term deletions

Model:
y ~ bmi + ltg + map + tc + sex + ldl
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1271494 3534.3
bmi     1    190592 1462086 3594.0  65.205 6.687e-15 ***
ltg     1    294092 1565586 3624.2 100.614 < 2.2e-16 ***
map     1     79625 1351119 3559.1  27.241 2.787e-07 ***
tc      1     65236 1336730 3554.4  22.318 3.123e-06 ***
sex     1     41856 1313350 3546.6  14.320 0.0001758 ***
ldl     1     39377 1310871 3545.7  13.472 0.0002723 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. What if you try backwards selection using an individual p-value approach, i.e. using `drop1()` from the full model.

```
drop1(M1, test = "F")
```

```
Single term deletions

Model:
y ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1263986 3539.6
age     1        82 1264068 3537.7  0.0281 0.8670306
sex     1     44999 1308984 3553.1 15.3439 0.0001042 ***
bmi     1    179033 1443019 3596.2 61.0477 4.296e-14 ***
map     1     72100 1336086 3562.2 24.5852 1.024e-06 ***
tc      1     10600 1274586 3541.3  3.6144 0.0579476 .
ldl     1      5799 1269785 3539.7  1.9774 0.1603902
hdl     1       663 1264649 3537.9  0.2260 0.6347233
tch     1      3526 1267512 3538.9  1.2024 0.2734587
ltg     1     56016 1320001 3556.8 19.1005 1.556e-05 ***
glu     1      3080 1267066 3538.7  1.0504 0.3059895
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M2 = update(M1, . ~ . - age)
drop1(M2, test = "F")
```

```
Single term deletions

Model:
y ~ sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1264068 3537.7
sex     1     45830 1309898 3551.4 15.6624 8.850e-05 ***
bmi     1    179084 1443152 3594.2 61.2027 3.993e-14 ***
map     1     73847 1337915 3560.8 25.2376 7.432e-07 ***
tc      1     10569 1274637 3539.4  3.6120   0.05803 .
ldl     1      5751 1269820 3537.7  1.9656   0.16163
hdl     1       646 1264715 3535.9  0.2209   0.63856
tch     1      3543 1267611 3536.9  1.2107   0.27180
ltg     1     55964 1320032 3554.8 19.1258 1.535e-05 ***
glu     1      3001 1267069 3536.7  1.0257   0.31173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M3 = update(M2, . ~ . - hdl)
drop1(M3, test = "F")
```

```
Single term deletions

Model:
y ~ sex + bmi + map + tc + ldl + tch + ltg + glu
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1264715 3535.9
```

```
<none>              1264715 3555.9
sex     1     46381 1311096 3549.8 15.8794 7.920e-05 ***
bmi     1    178542 1443256 3592.3 61.1273 4.111e-14 ***
map     1     73533 1338248 3558.9 25.1756 7.655e-07 ***
tc      1     26839 1291554 3543.2  9.1890  0.002581 **
ldl     1      7505 1272219 3536.5  2.5694  0.109677
tch     1      3247 1267961 3535.0  1.1116  0.292320
ltg     1     97508 1362223 3566.7 33.3840 1.447e-08 ***
glu     1      3093 1267808 3535.0  1.0590  0.304011
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
M4 = update(M3, . ~ . - glu)
drop1(M4, test = "F")
```

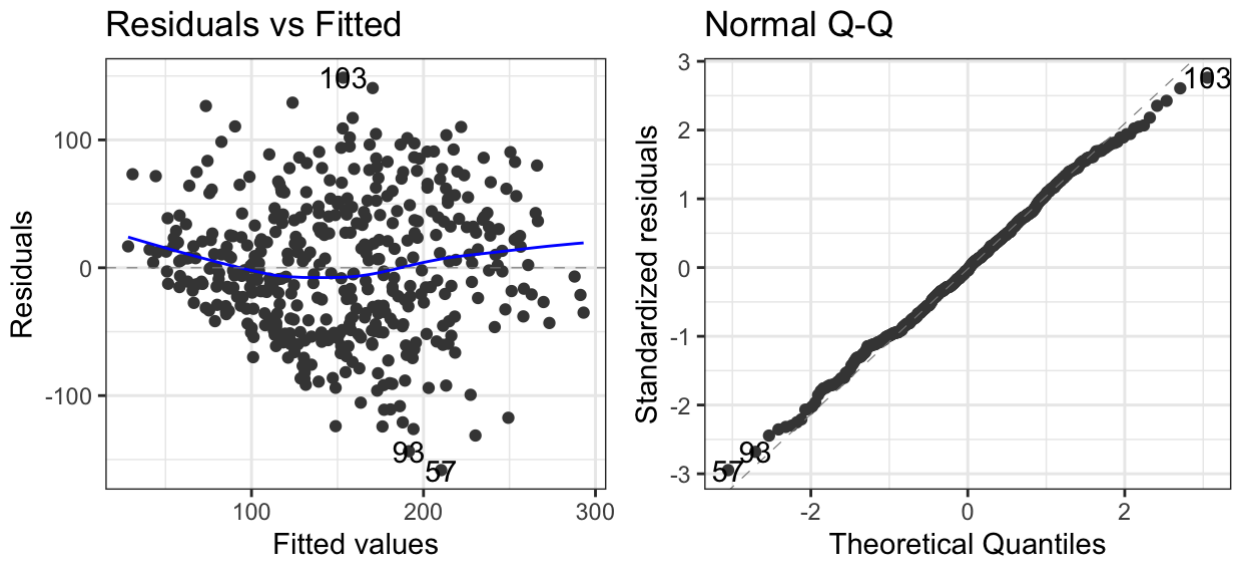
```
Single term deletions

Model:
y ~ sex + bmi + map + tc + ldl + tch + ltg
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1267808 3535.0
sex     1     44684 1312492 3548.3 15.2965 0.0001066 ***
bmi     1    189976 1457784 3594.7 65.0331 7.248e-15 ***
map     1     82152 1349960 3560.7 28.1225 1.818e-07 ***
tc      1     26378 1294186 3542.1  9.0298 0.0028101 **
ldl     1      7472 1275280 3535.6  2.5577 0.1104828
tch     1      3686 1271494 3534.3  1.2619 0.2619190
ltg     1    102520 1370328 3567.3 35.0950 6.399e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
M5 = update(M4, . ~ . - tch)
drop1(M5, test = "F")
```

```
Single term deletions

Model:
y ~ sex + bmi + map + tc + ldl + ltg
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1271494 3534.3
sex     1     41856 1313350 3546.6  14.320 0.0001758 ***
bmi     1    190592 1462086 3594.0  65.205 6.687e-15 ***
map     1     79625 1351119 3559.1  27.241 2.787e-07 ***
tc      1     65236 1336730 3554.4  22.318 3.123e-06 ***
ldl     1     39377 1310871 3545.7  13.472 0.0002723 ***
ltg     1    294092 1565586 3624.2 100.614 < 2.2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Are you satisfied with the model you have arrived at? Check the assumptions.

```
library(ggfortify)
autoplot(M5, which = 1:2) + theme_bw()
```




There does seem to be some fanning out of the residuals in the residual vs fitted value plot, indicating that there may be some heteroskedasticity in the our data.

In the normal QQ plot, the points are all reasonably close to the diagonal line, therefore we are confident that the normal assumption is at least approximately satisfied.

6. Write down your final fitted model and interpret the estimated coefficients.

```
M5
```

```
Call:
lm(formula = y ~ sex + bmi + map + tc + ldl + ltg, data = diabetes)

Coefficients:
(Intercept)          sex          bmi          map           tc
  -313.7666     -21.5910       5.7111       1.1266      -1.0429
        ldl          ltg
     0.8433      73.3065
```

$$\hat{y} = -313.8 - 21.6 \times \text{sex} + 5.7 \times \text{bmi} + 1.1 \times \text{map} - 1.0 \times \text{tc} + 0.8 \times \text{ldl} + 73.3 \times \text{ltg}$$

- On average, holding the other variables constant, a 1 $kg/m^2$ increase in BMI leads to a 5.7 unit increase in diabetes disease progression.
- On average, holding the other variables constant, a 1 mmHg increase in mean arterial blood pressure leads to a 1.1 unit increase in diabetes disease progression.
- On average, holding the other variables constant, a 1 mg/dL increase in total cholesterol leads to a

1.0 unit decrease in diabetes disease progression.

- On average, holding the other variables constant, a 1 mg/dL increase in low density lipoprotein leads to a 0.8 unit increase in diabetes disease progression.
- On average, holding the other variables constant, a 1 mg/dL increase in ltg leads to a 73.3 unit increase in diabetes disease progression.
- On average, the holding the other variables constant, the difference in diabetes disease progression between males and females in 21.6. If male = 1 and female = 2 then we can say that the disease progression is 26.1 units less for females than males.

*Note that it doesn't make sense to interpret the intercept in this model, as values of zero in many of the covariates are not possible.*

# 2 For practice after the computer lab

## 2.1 Predicting capital value

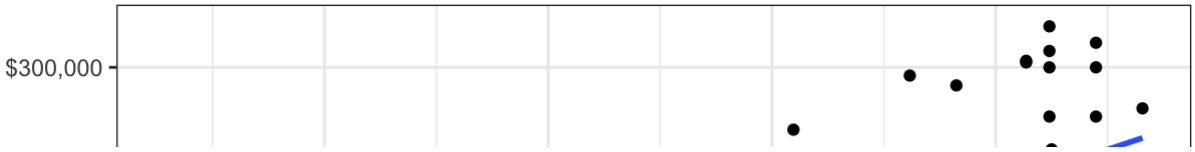
The data in `rentcap.txt` shows the capital value and annual rental value of 96 domestic properties in Auckland in 1991. The aim was to explore their relationship in the hope of being able to predict capital value from rental value.

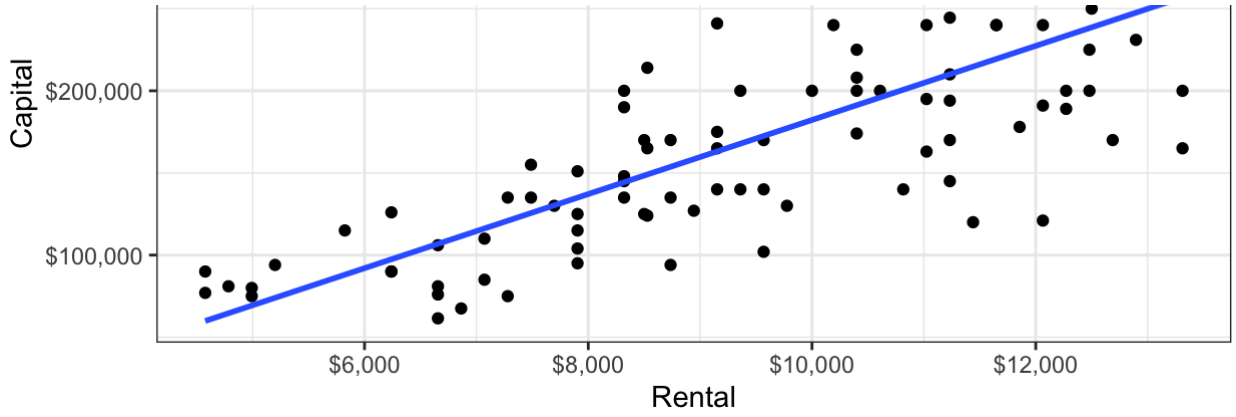
```
rent = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/rentcap.txt")
glimpse(rent)
```

```
Rows: 96
Columns: 3
$ Capital <dbl> 61500, 67500, 75000, 75000, 76000, 77000, 80000, 810…
$ ...2    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ Rental  <dbl> 6656, 6864, 4992, 7280, 6656, 4576, 4992, 6656, 4784…
```

1. Fit a simple linear regression to the data to assess whether the rental value has an influence on the capital value.

```
p = rent %>%
    ggplot() + aes(x = Rental, y = Capital) + geom_point() + geom_smooth(method = "lm",
    se = FALSE) + theme_bw() + scale_y_continuous(labels = scales::dollar) +
    scale_x_continuous(labels = scales::dollar)
p
```






```
rent.lm = lm(Capital ~ Rental, rent)
summary(rent.lm)
```

```
Call:
lm(formula = Capital ~ Rental, data = rent)

Residuals:
    Min      1Q  Median      3Q     Max
-107792  -31021    1272   23570   86825

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -43314.269  18021.939  -2.403   0.0182 *
Rental          22.555      1.825  12.359   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42500 on 94 degrees of freedom
Multiple R-squared:  0.6191,    Adjusted R-squared:  0.615
F-statistic: 152.8 on 1 and 94 DF,  p-value: < 2.2e-16
```

The slope coefficient for rental is significant ( $p < 0.0001$ ). We can interpret it as: on average, a $1 per annum increase in rental value results in a $22.56 increase in the capital value of the asset. Or it might be more natural to describe the relationship at a larger scale: on average, a $100 per annum increase in rental value results in a $2256 increase in the capital value of the asset.

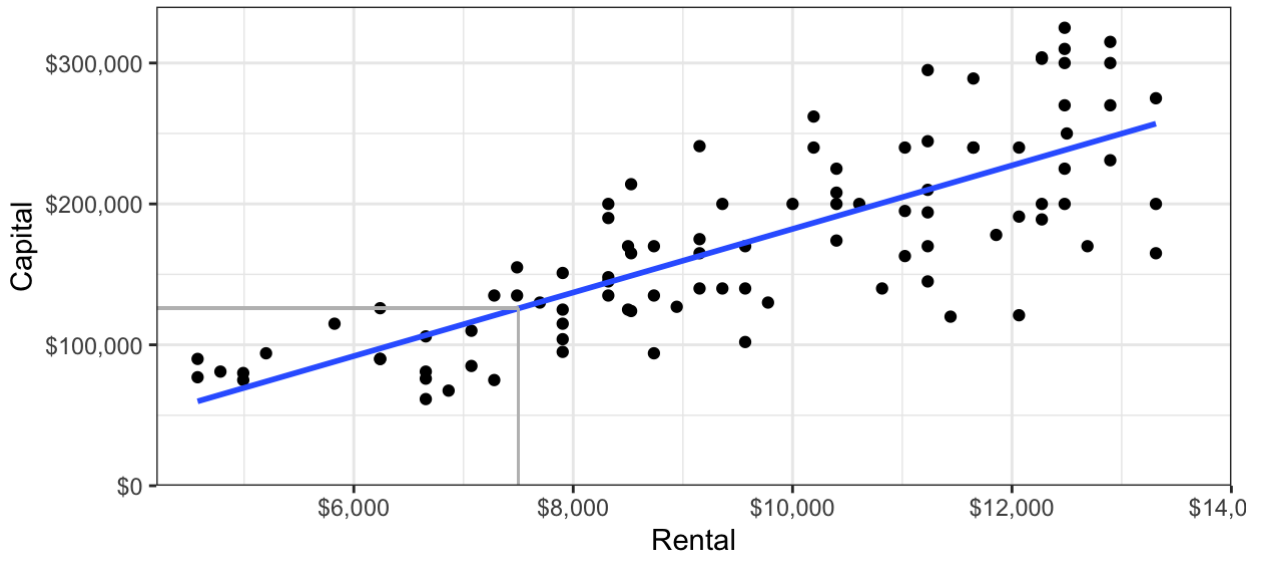
2. Obtain a predicted capital value from the rental value of 7500 and the corresponding 90% prediction interval for your predicted capital value.

```
predict(rent.lm, newdata = data.frame(Rental = 7500), interval = "prediction",
    level = 0.9)
```

```
       fit      lwr      upr
1 125849.8 54607.79 197091.8
```

We can visualise this prediction as on the plot as follows:

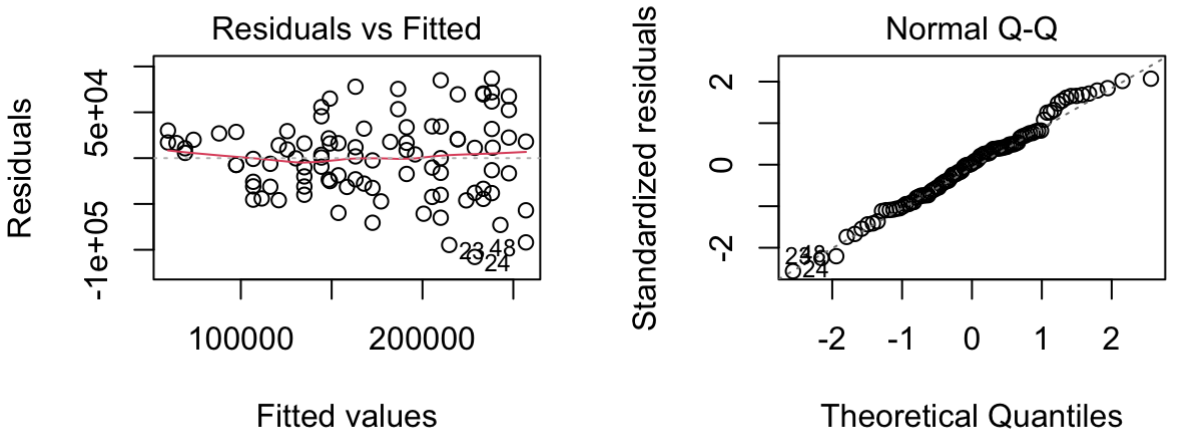
```
p + geom_segment(aes(y = 0, yend = 125849.8, x = 7500, xend = 7500), colour = "gray") +
    geom_segment(aes(y = 125849.8, yend = 125849.8, x = 4200, xend = 7500),
        colour = "gray") + scale_x_continuous(limits = c(4200, 14000),
    expand = c(0, 0), labels = scales::dollar) + scale_y_continuous(limits = c(0,
    340000), expand = c(0, 0), labels = scales::dollar)
```




3. Use various visualisations to comment on whether the assumptions for the prediction interval are satisfied. If not, find an appropriate transformation and re-fit the linear regression.

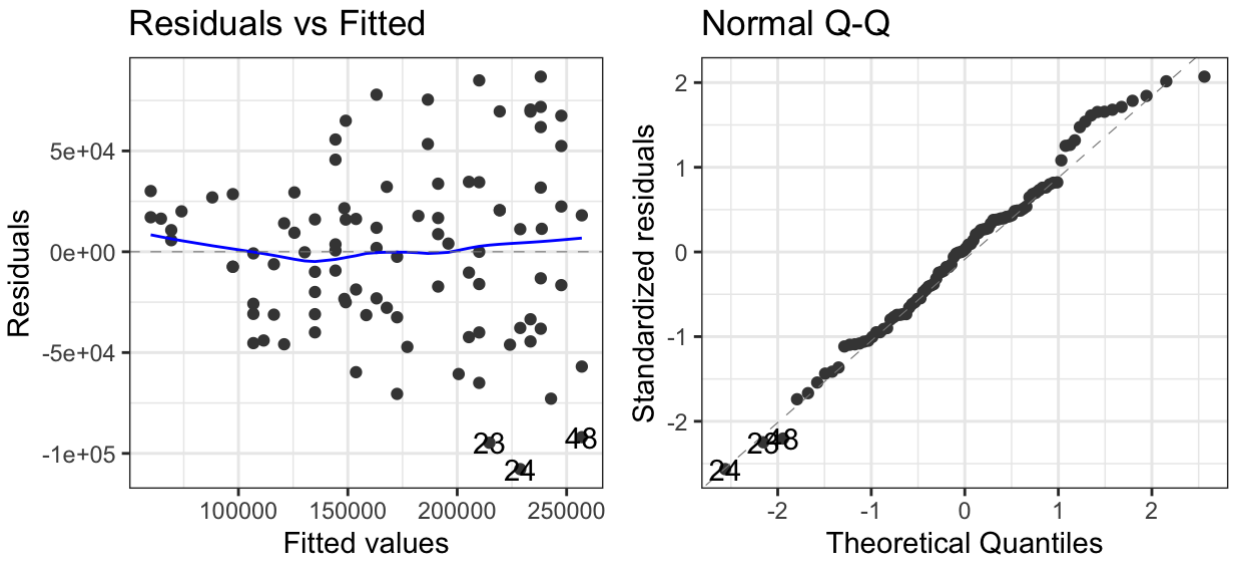
Using base graphics:

```
par(mfrow = c(1, 2))
plot(rent.lm, which = 1:2)
```




Using ggplot:

```
library(ggfortify)
autoplot(rent.lm, which = 1:2) + theme_bw()
```

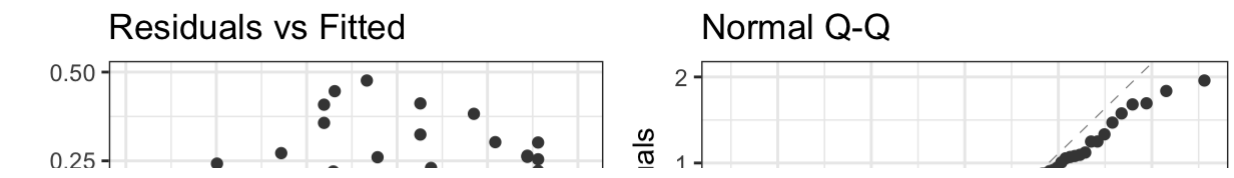


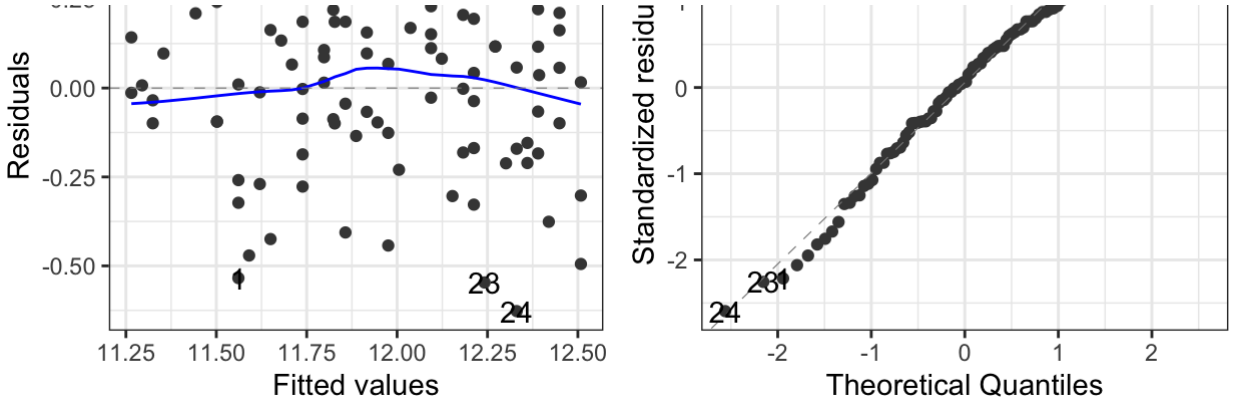

- The QQ plot shows a straight line which indicates that the normality assumption is reasonable. However, the residuals vs fitted plot shows a fan shaped plot which indicates that the assumption of homogeneous variance is violated. We can use a log transformed response and re-fit the linear regression. *Note: Most Box-Cox type transformations would work.*

```
tlm = lm(log(Capital) ~ Rental, rent)
summary(tlm)
```

```
Call:
lm(formula = log(Capital) ~ Rental, data = rent)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62718 -0.15788  0.01567  0.18625  0.47626 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.061e+01  1.037e-01  102.38   <2e-16 ***
Rental      1.423e-04  1.050e-05   13.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2445 on 94 degrees of freedom
Multiple R-squared:  0.6616,    Adjusted R-squared:  0.658 
F-statistic: 183.8 on 1 and 94 DF,  p-value: < 2.2e-16
```

```
autoplot(tlm, which = 1:2) + theme_bw()
```






The residuals against fitted values are now much better in that there is no longer any fanning out of the residuals across the range of fitted values. The normality assumption is still OK as the points are still close to the diagonal line.

```
sjPlot::tab_model(rent.lm, tlm, digits = 5, show.ci = FALSE)
```

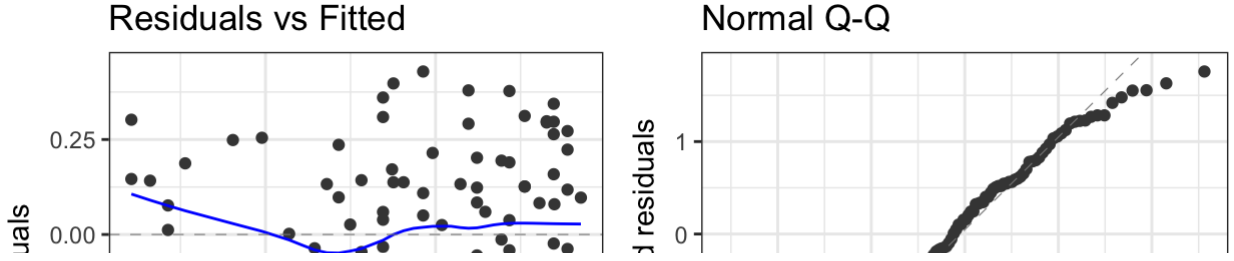
| | Capital | | log(Capital) | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *p* | *Estimates* | *p* |
| (Intercept) | -43314.26882 | **0.018** | 10.61380 | **<0.001** |
| Rental | 22.55521 | **<0.001** | 0.00014 | **<0.001** |
| Observations | 96 | | 96 | |
| $R^2$ / $R^2$ adjusted | 0.619 / 0.615 | | 0.662 / 0.658 | |

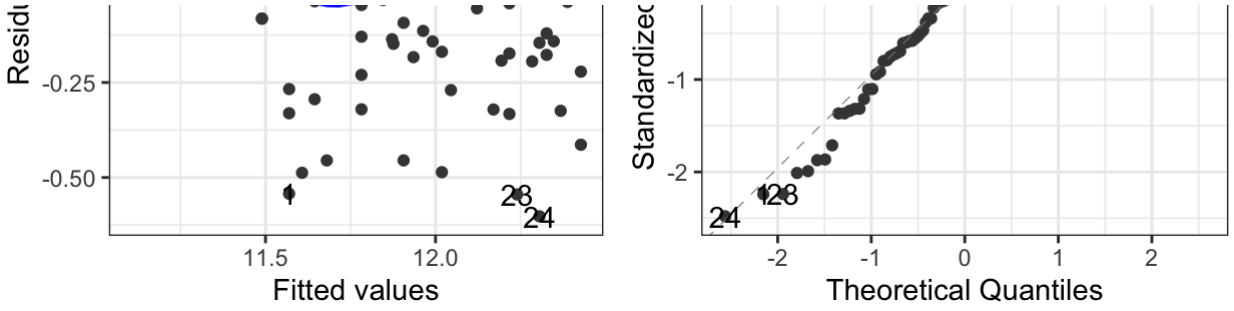
Note that we cannot directly compare the $R^2$ values from the model with Capital as the dependent variable and the model with log(Capital) as the dependent variable. This is because the $R^2$ measures the proportion of variation in the dependent variable explained by the model, but in these two models the dependent variable is different so the $R^2$ values can't be directly compared (even though in this case they look similar).

Our interpretation of the coefficient in the log(Capital) is: on average a $100 increase in the annual rental value leads to a 0.014% increase in capital value. This isn't a very natural way to interpret the relationship between rental value and capital value.

To improve the interpretability of the model, we could consider a log-log model.

```
ttlm = lm(log(Capital) ~ log(Rental), rent)
autoplot(ttlm, which = 1:2) + theme_bw()
```






```
sjPlot::tab_model(rent.lm, tlm, ttlm, digits = 5, show.ci = FALSE)
```

| | Capital | | log(Capital) | | log(Capital) | |
|---|---|---|---|---|---|---|
| *Predictors* | *Estimates* | *p* | *Estimates* | *p* | *Estimates* | *p* |
| (Intercept) | -43314.26882 | **0.018** | 10.61380 | **<0.001** | 0.67120 | 0.426 |
| Rental | 22.55521 | **<0.001** | 0.00014 | **<0.001** | | |
| Rental [log] | | | | | 1.23797 | **<0.001** |
| Observations | 96 | | 96 | | 96 | |
| $R^2$ / $R^2$ adjusted | 0.619 / 0.615 | | 0.662 / 0.658 | | 0.659 / 0.656 | |

The assumptions aren't as well met now, particularly with the residuals corresponding to all the low fitted values being above 0. However, if this was the model of choice we could interpret the coefficient of log(Rental) as a 1% increase in rental value leading to a 1.24% increase in capital value. This is related to the concept of elasticity in economics (i.e. capital value is elastic relative to rental value).

## References


Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." *The Annals of Statistics* 32 (2): 407–51. https://doi.org/10.1214/009053604000000067.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2021. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.