431 Lab 06 Sketch and Grading Rubric

Instructor: Dr. Thomas E. Love   Lab Author: Mr. Wyatt P. Bensken

Due: 2021-11-08 | Last Edit: 2021-11-08 08:47:45

Contents

1 Loading Packages

```
library(MatchLinReg);
library(patchwork);
library(car);
library(broom);
library(tidyverse)
```

2 Learning Objectives

1. Be able to work through a simple linear regression
2. Visualize and interpret the role of a potential confounder
3. Run a multivariable, model adjusting for this confounder, with an interpretation of the estimate and confidence interval
4. Expand this model and interpret the results and conclusion
5. Compare multiple linear models using various metrics

3 Packages and Functions

In Lab 06 we were hoping you would become more familiar with the following packages and functions:

Packages:

- `tidyverse`
- `MatchLinReg`
- `patchwork`
- `car`
- `GGally`
- `broom`

Functions:

- `%>%`
- `ggplot()`
- `boxCox()`
- `powerTransform()`
- `lm()`
- `summary()`
- `confint()`
- `plot()`
- `augment()`
- `group_by()`
- `summarize()`

4 The Data

In Lab 06 we'll be using the lindner dataset again that we saw in Lab 05. The data come from "an observational study of 996 patients receiving an initial Percutaneous Coronary Intervention (PCI) at Ohio Heart Health, Christ Hospital, Cincinnati in 1997 and followed for at least 6 months by the staff of the Lindner Center. The patients thought to be more severely diseased were assigned to treatment with abciximab (an expensive, high-molecular-weight IIb/IIIa cascade blocker); in fact, only 298 (29.9 percent) of patients received usual-care-alone with their initial PCI.". Information on the lindner dataset and its variables can be found at this site.[1,2]

[1] Rdocumentation. (n.d.). lindner: Lindner Center Data On 996 PCI Patients Analyzed By Kereiakes Et Al. (2000). Retrieved from https://www.rdocumentation.org/packages/MatchLinReg/versions/0.7.0/topics/lindner

[2] Kereiakes DJ, Obenchain RL, Barber BL, et al. Abciximab provides cost effective survival advantage in high volume interventional practice. Am Heart J 2000; 140: 603-610.

We'd like you to load the `lindner` data by either:

1. first installing the `MatchLinReg` package, then loading it, and finally running `data("lindner")`, or by
2. loading the `lab05_lind` data set provided for Lab 5.

5 Background

You're a statistician tasked with analyzing the `lindner` data. The principal investigator wants to examine the relationship between a predictor: the ejection fraction (`ejecfrac`) and an outcome: 6-month cardiac-related costs (`cardbill`), **among those patients who were alive at 6 months**. There are a number of data cleaning steps you'll need to do after reading in the data (which you should call `lindner`). This includes (a) select only those patients who were alive at 6 months (call this `lindner_alive`) , (b) you'll want to add an `id` to be able to properly identify patients since there are no unique identifiers, `row_number()` could be one approach, and (c) you'll want to partition your data to a 70% training (call this `lindner_alive_train`) and 30% test sample (call this `lindner_alive_test`), using `set.seed(431)`.

To start we'll load in our data, then create a new dataset of only those patients who were alive at 6 months. This brings us from 996 patients to 970 patients. We'll also need to add an ID variable, which we can just make the row number.

```
data("lindner")

lindner_alive <- lindner %>%
  filter(sixMonthSurvive == 1) %>%
  mutate(id = row_number())
```

We will now partition our data to get 70%, or 679, in the training data and the remaining 291 in the test data.

```
set.seed(431)
lindner_alive_train <- lindner_alive %>%
  slice_sample(., prop = 0.70)

lindner_alive_test <- anti_join(lindner_alive, lindner_alive_train, by = "id")
```

6 Question 1 (20 points)

Given the information above, work through an appropriate analysis of the data. Specifically do the following: (a) decide whether a square-root or log transformation of the outcome is more appropriate (only select between these two options), make said transformation, run a simple linear regression, and interpret and contextualize these results. The decisions regarding transformations as well as the build and interpretation of your model should be completed using just the training data set (`lindner_alive_train`) and should be called `model1`.

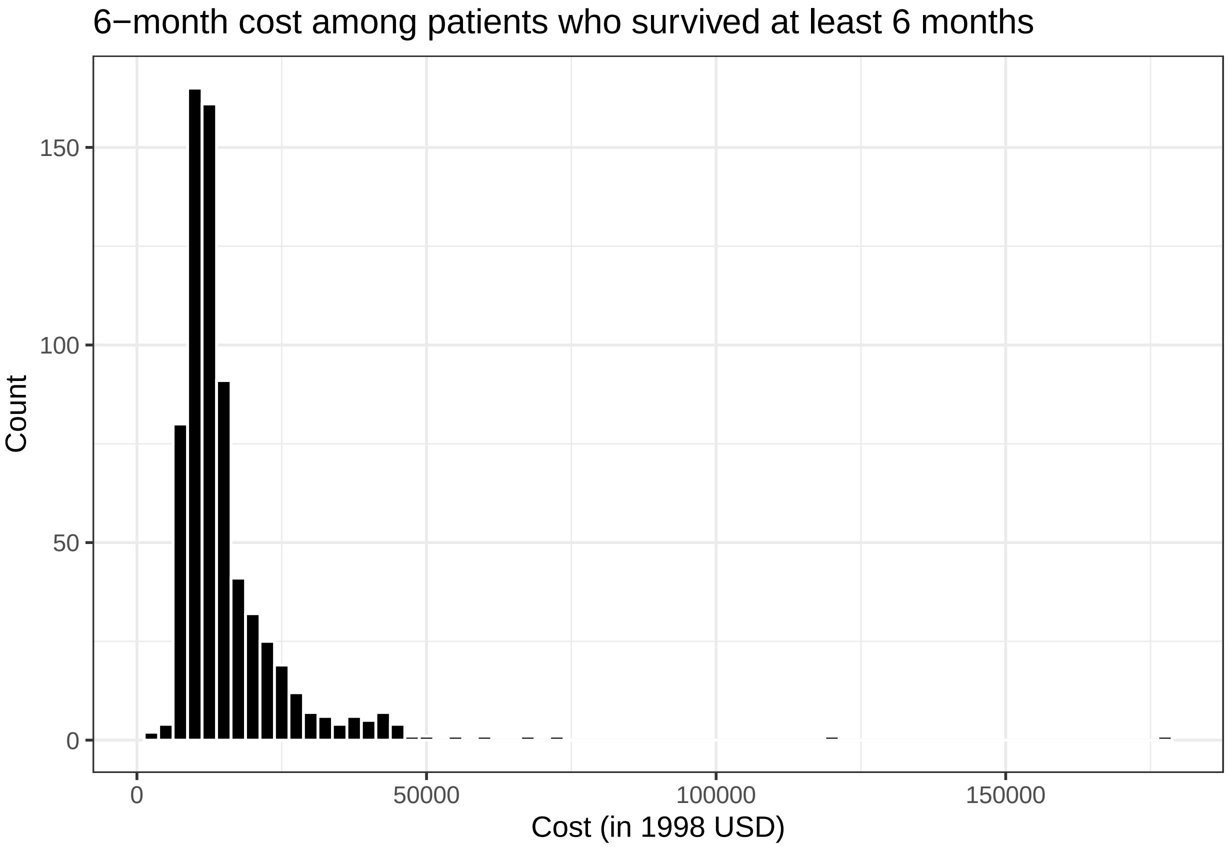
6.1 Assessing the Outcome

As stated in the background our outcome of interest here is the variable `cardbill` which are the cardiac related costs within 6 months of the initial PCI.

The first step is to examine the distribution of our outcome of interest to see if a transformation would be necessary. We can somewhat immediately see that there are some notable outliers, which give the distribution a fairly meaningful right-skew.

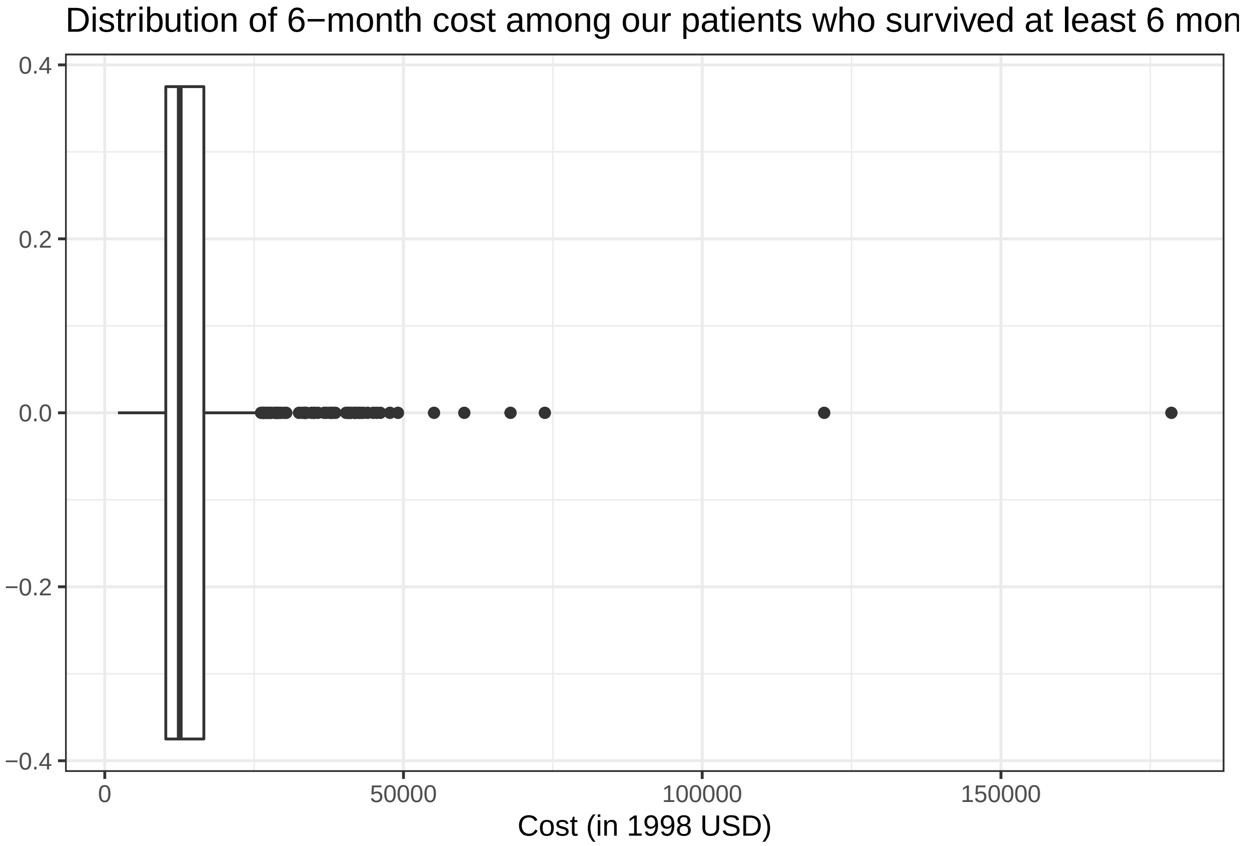
```
orig_hist <- ggplot(data = lindner_alive_train, aes(x = cardbill)) +
  geom_histogram(colour = "white", fill = "black", binwidth = 2500) +
  labs(x = "Cost (in 1998 USD)",
       y = "Count",
       title = "6-month cost among patients who survived at least 6 months") +
  theme_bw()

orig_hist
```



We can continue to evaluate this distribution using a scatterplot.

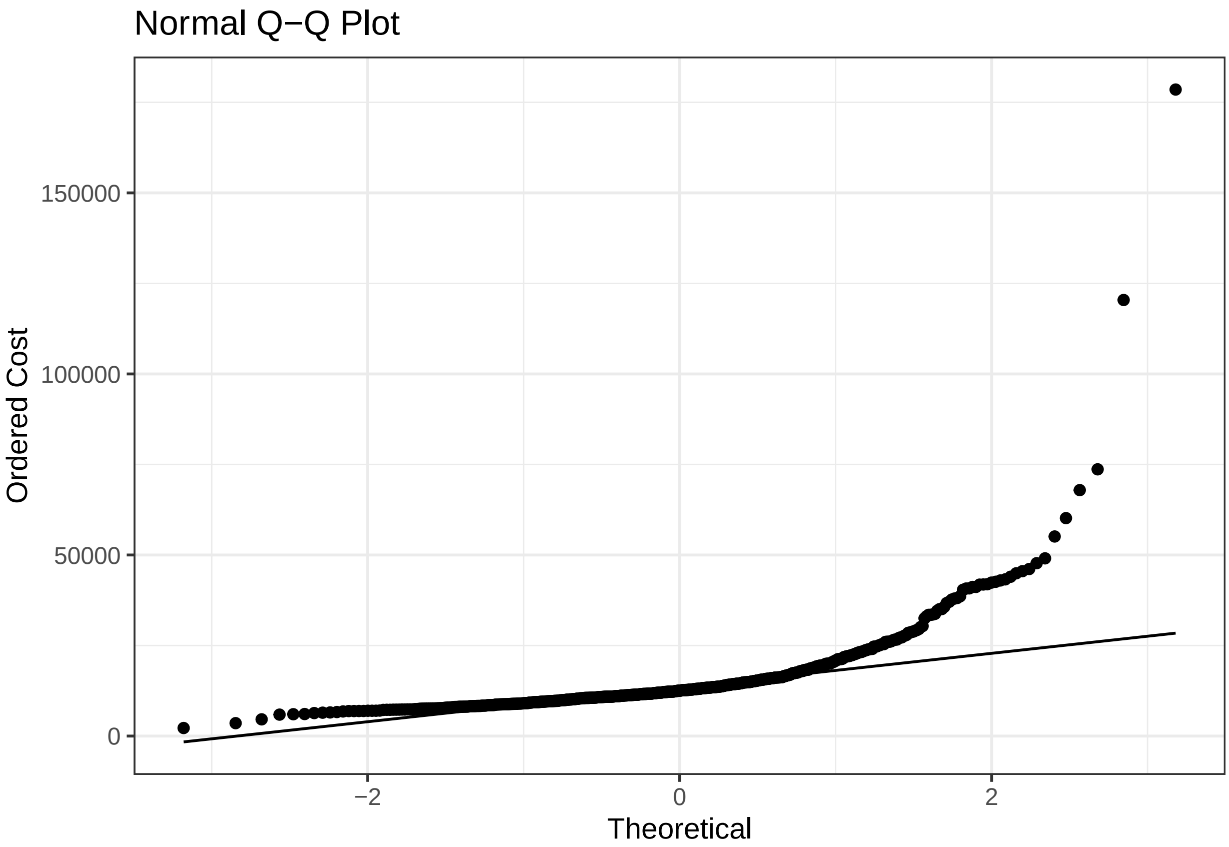
```
ggplot(data = lindner_alive_train, aes(x = cardbill)) +
  geom_boxplot() +
  labs(x = "Cost (in 1998 USD)",
       title = "Distribution of 6-month cost among our patients who survived at least 6 months") +
  theme_bw()
```



It's evident we have some concerns regarding Normality. We can further see this in a Normal Q-Q plot.

```
orig_qq <- ggplot(data = lindner_alive_train, aes(sample = cardbill)) +
  geom_point(stat = "qq") +
  geom_qq_line() +
  labs(x = "Theoretical",
       y = "Ordered Cost",
       title = "Normal Q-Q Plot") +
  theme_bw()

orig_qq
```



Now, it's not immediately clear what type of transformation we will need but we'll certainly need one. As stated in the prompt, we are to decide between a square-root and a log transformation. From the below figure, it seems that the log-transformation may be the most appropriate.

```
orig_hist2 <- ggplot(data = lindner_alive_train, aes(x = cardbill)) +
  geom_histogram(colour = "white", fill = "black", bins = 35) +
  labs(x = "Cost",
       y = "Count",
       title = "`cardbill`") +
  theme_bw()

orig_qq2 <- ggplot(data = lindner_alive_train, aes(sample = cardbill)) +
  geom_point(stat = "qq") +
  geom_qq_line() +
  labs(x = "Theoretical",
       y = "Ordered Cost",
       title = "Normal Q-Q Plot",
       subtitle = "`cardbill`") +
  theme_bw()

log_hist <- ggplot(data = lindner_alive_train, aes(x = log(cardbill))) +
  geom_histogram(colour = "white", fill = "black", bins = 35) +
  labs(x = "log(cardbill)",
       y = "Count",
       title = "`log(Cost)`") +
  theme_bw()

log_qq <- ggplot(data = lindner_alive_train, aes(sample = log(cardbill))) +
```

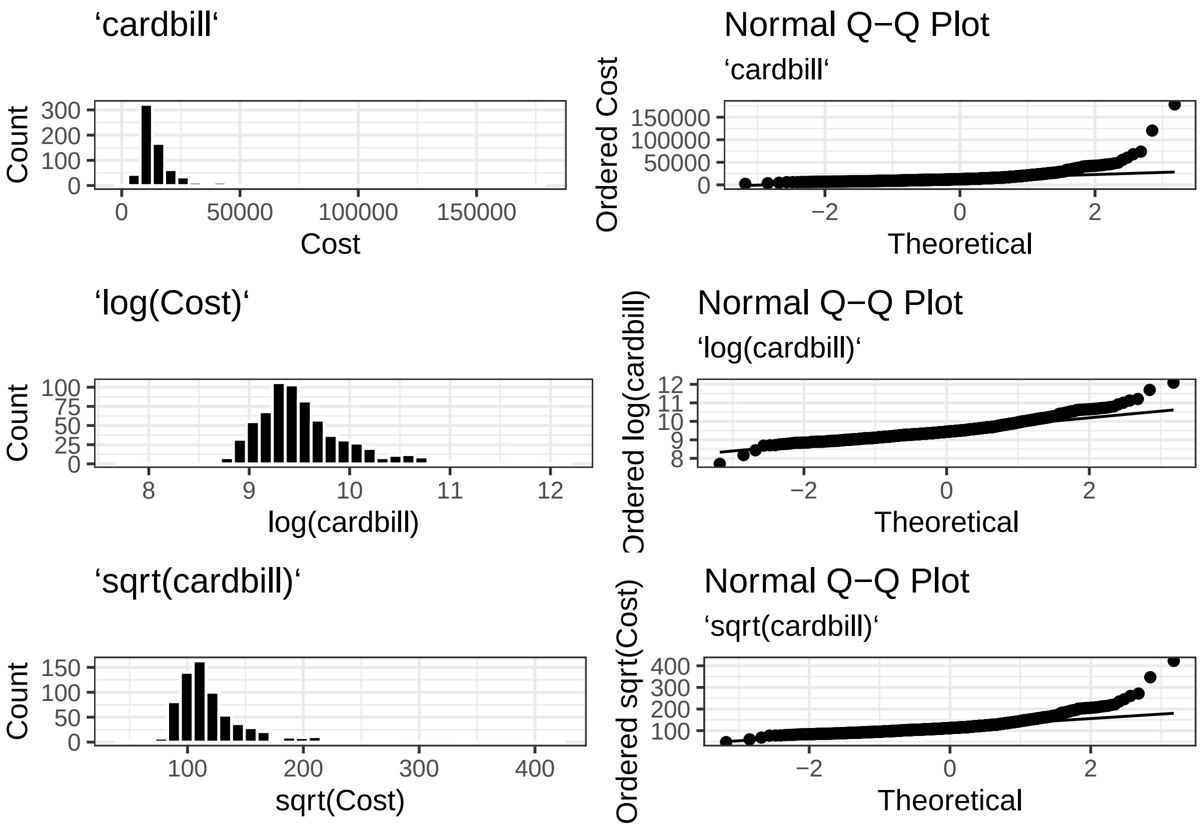
```r
  geom_point(stat = "qq") +
  geom_qq_line() +
  labs(x = "Theoretical",
       y = "Ordered log(cardbill)",
       title = "Normal Q-Q Plot",
       subtitle = "`log(cardbill)`") +
  theme_bw()

sqrt_hist <- ggplot(data = lindner_alive_train, aes(x = sqrt(cardbill))) +
  geom_histogram(colour = "white", fill = "black", bins = 35) +
  labs(x = "sqrt(Cost)",
       y = "Count",
       title = "`sqrt(cardbill)`") +
  theme_bw()

sqrt_qq <- ggplot(data = lindner_alive_train, aes(sample = sqrt(cardbill))) +
  geom_point(stat = "qq") +
  geom_qq_line() +
  labs(x = "Theoretical",
       y = "Ordered sqrt(Cost)",
       title = "Normal Q-Q Plot",
       subtitle = "`sqrt(cardbill)`") +
  theme_bw()


transformation_comb_fig <- (orig_hist2 | orig_qq2) /
                              (log_hist | log_qq) /
                              (sqrt_hist | sqrt_qq)

transformation_comb_fig
```



For ease in future analyses, I will setup a new variable which contains these transformed values.

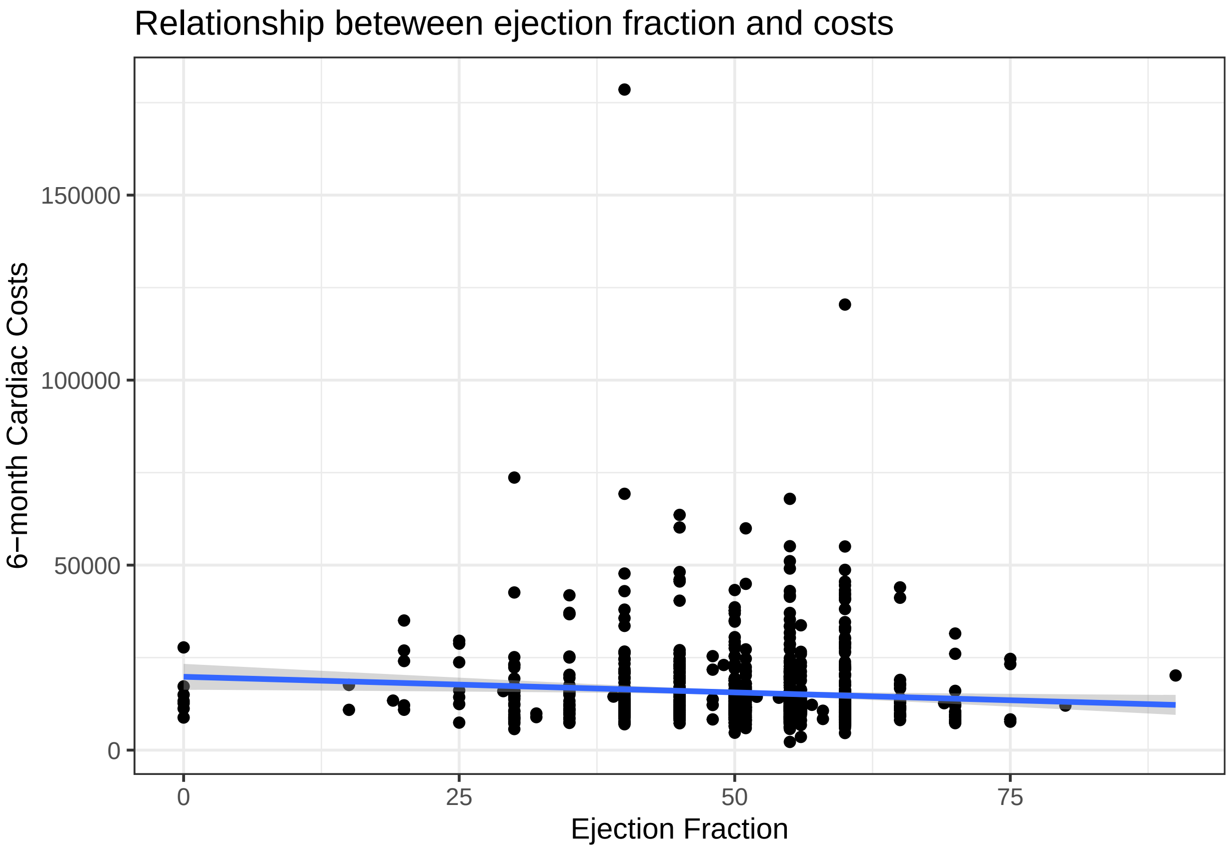
```
lindner_alive_train <- lindner_alive_train %>%
  mutate(cardbill_log = log(cardbill))
```

6.2 Linear Regression

As stated in the problem, we are interested in the effect of `ejecfrac` on our our costs (`cardbill`).

We can start by making a simple scatterplot, on our entire dataset, to get a look at this relationship. From this, we can not only see our outliers, but that there is perhaps not much of a relationship between ejection fraction and costs. However, we do see our linear regression line suggesting perhaps lower costs for patients with higher ejection fractions.

```
ggplot(data = lindner_alive, aes(x = ejecfrac, y = cardbill)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(x = "Ejection Fraction",
       y = "6-month Cardiac Costs",
       title = "Relationship beteween ejection fraction and costs") +
  theme_bw()
```



Now we can go ahead and build the model.

```
model1 <- lm(cardbill_log ~ ejecfrac, data = lindner_alive_train)
summary(model1)
```

```

Call:
lm(formula = cardbill_log ~ ejecfrac, data = lindner_alive_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7935 -0.2898 -0.0842  0.1848  2.5170 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.785055   0.086135  113.60  < 2e-16 ***
ejecfrac    -0.005239   0.001658   -3.16  0.00165 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4564 on 677 degrees of freedom
Multiple R-squared:  0.01454,	Adjusted R-squared:  0.01308 
F-statistic: 9.986 on 1 and 677 DF,  p-value: 0.001648
```

In addition to the point estimate, we can also obtain the 95% confidence interval for our estimate.

```
confint(model1)
```

```
                   2.5 %       97.5 %
(Intercept)  9.615931531  9.954177715
```

```
ejecfrac    -0.008494053 -0.001983783
```

Finally, we can use glance() to obtain some important model fit summaries.

```r
glance(model1) %>%
  select(r.squared, adj.r.squared, sigma, statistic, p.value, nobs,
         logLik:deviance) %>%
  knitr::kable(digits = 2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	nobs	logLik	AIC	BIC	deviance
0.01	0.01	0.46	9.99	0	679	-429.82	865.65	879.21	141

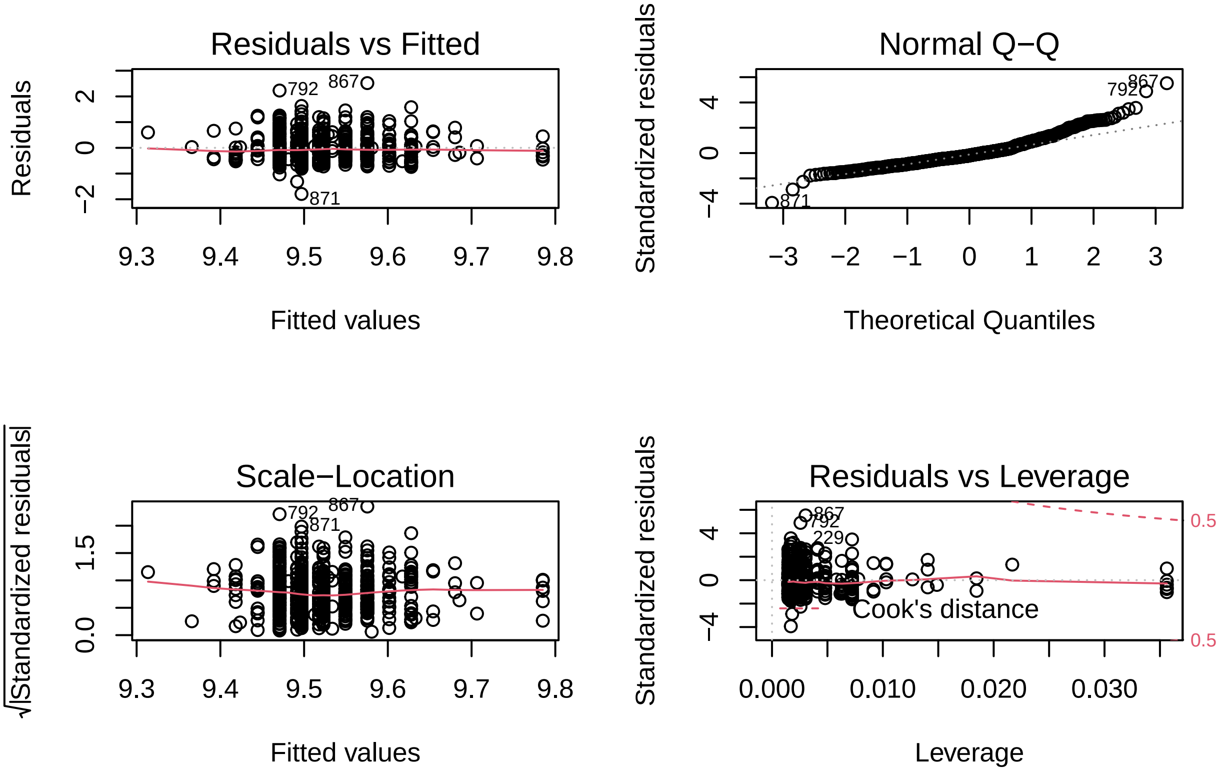
6.2.1 Residual Plots

In these residuals plots we are looking for the following:

1. Residual vs Fitted: we want to see a patternless figure, with no bend, and no thickening. In our figure there is perhaps some grouping, but no severe bend in our line so we are happy with this.
2. Normal Q-Q: here we can readily identify a few important points (792 and 867) and some concerns around the Normality of our data.
3. Scale-Location: We want to look at a non-flat loess smooth curve, but overall this seems okay.
4. Residuals vs. Leverage: While points 792 and 867, and perhaps 229, appeared to be important, but non violate our general guidelines for Cook's distance.

Overall, we have some concerns regarding normality, but otherwise are adequately happy with these plots.

```r
par(mfrow=c(2,2))
plot(model1)
```



6.3 Interpreting and Contexualizing the Model

From this result of the model we have an estimate of -0.005239 (95% CI: -0.008494053, -0.001983783), which appears to be significant. Keep in mind this is on the log scale. To better understand what this means, we can create a scenario where we compare 2 patients: one of whom had an ejection fraction of 55 (Patient A) and one had an ejection fraction of 60 (Patient B).

Given our model, we expect that Patient A will have an inverse of the square root of the costs: $9.785055 - 0.005239(55) = 9.49691$ which we can then convert back by $9.49691 = log(cardbill)$ which yields $cardbill = 13,318.51$. On the other hand Patient B will have $9.785055 - 0.005239(60) = 9.470715$ which we can then convert back by $9.470715 = log(cardbill)$ which yields $cardbill = 12,974.16$. This finding is consistent with our scatterplot.

Overall, this suggests that those who have higher ejection fractions, accrue lower costs in the 6-months after the initial PCI. Strictly speaking we can interpret our coefficient as follows: for each ejection fraction point higher, we expect a change in the log of the 6-month cardiac costs of -0.005239 (95% CI: -0.008494053, -0.001983783).

6.4 Grading Rubric

Each student should start with the full 20 points. Listed below are key issues which we have evaluated this question on.

- If either of the following occur the student should lose 2 points for each:
  - No title, or a title which does not convey what the figure represents
  - Axes remain unlabeled or with the default labels
- If the figure provided does not allow one to interpret the need for a transformation the student should lose 5 points.
- If the student does not use visualizations to explore transformations, they should lose 5 points.
- If the student settles on the incorrect transformation, they should lose 5 points.
- If the student does not provide an interpretation of the coefficient and confidence interval, they should lose 3 points.
- If their interpretation is on the wrong scale (i.e. transformed outcome, but interpreting coefficient on original scale), they should similarly lose 3 points.
- If the TA identifies other issues, they are able to deduct points at their own discretion. These issues and deductions should be clearly documented in the grading sheet.

7 Question 2 (10 points)

> Now we want to examine the effect of a third variable, `abcix` or whether or not the patient had the abciximab augmentation, on the relationship between our main predictor and our outcome. Run, and discuss, a new linear regression which adjusts your original model for this variable. Call this `model2`. Again, this should be done on your training data set.

We can add this variable to our model fairly simply. Our estimates don't change all that much, but it does seem that, when controlling for ejection fractions, it seems that whether or not someone had the augmentation with abciximab (`abcix`) is associated with total costs. We can work out an example, very similarly to what we have done above. Further, even when controlling for whether or not someone had augmentation with abciximab, there is still an association with the ejection fraction

We will extend upon our above example, but now Patient A received abciximab and Patient C, who has the same ejection fraction as Patient A, did not receive abciximab. Given our model and the new coefficients:

- Patient A would have log of the costs: $9.605289 - 0.004210(55) + 0.180225(1) = 9.553964$ which we can then convert back by $9.553964 = log(cardbill)$ which yields $cardbill = 14100.48$.
- Patient B, on the other hand, would be calculated as $9.605289 - 0.004210(60) + 0.180225(0) = 9.352689$ which we can then convert back by $9.352689 = log(cardbill)$ which yields $cardbill = 11529.79$.

```
model2 <- lm(cardbill_log ~ ejecfrac + abcix, data = lindner_alive_train)
summary(model2)
```

```

Call:
lm(formula = cardbill_log ~ ejecfrac + abcix, data = lindner_alive_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6703 -0.2936 -0.1013  0.1994  2.6556 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.605289   0.092991 103.293  < 2e-16 ***
ejecfrac    -0.004210   0.001647  -2.556   0.0108 *  
abcix        0.180225   0.038226   4.715 2.94e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4494 on 676 degrees of freedom
Multiple R-squared:  0.04591,	Adjusted R-squared:  0.04309 
F-statistic: 16.26 on 2 and 676 DF,  p-value: 1.263e-07
```

```
confint(model2)
```

```
                  2.5 %        97.5 %
(Intercept)  9.422702952  9.7878753125
ejecfrac    -0.007443444 -0.0009758203
abcix        0.105170095  0.2552807281
```

Again, we can use glance() to obtain additional fit statistics:

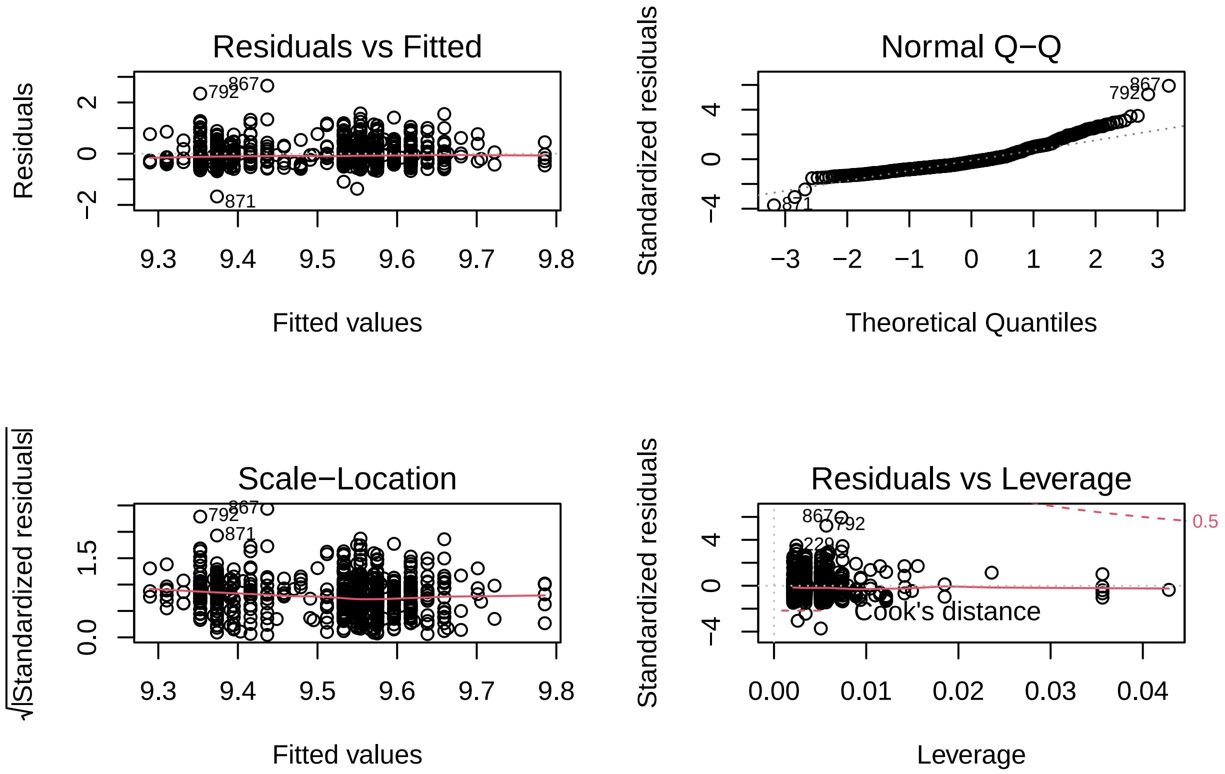
```
glance(model2) %>%
  select(r.squared, adj.r.squared, sigma, statistic, p.value, nobs,
         logLik:deviance) %>%
  knitr::kable(digits = 2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	nobs	logLik	AIC	BIC	deviance
0.05	0.04	0.45	16.26	0	679	-418.84	845.68	863.76	136.52

7.1 Residual Plots

As we have just two predictors, one of which is categorical, we have similar patterns in our residual plots. Again, from this we don't see anything too concerning beyond some issues of Normality.

```
par(mfrow=c(2,2))
plot(model2)
```



7.2 Grading Rubric

- If the student has successfully added `abcix` to the model, briefly discussed its impact, and has residual plots they should receive all 10 points.
- If the student does not include residual plots they should lose 2 points.
- If there is no discussion or interpretation of the model, the student should lose 4 points.
- Other issues which are noted can be deducted at the TAs discretion, and noted in the grading sheet.

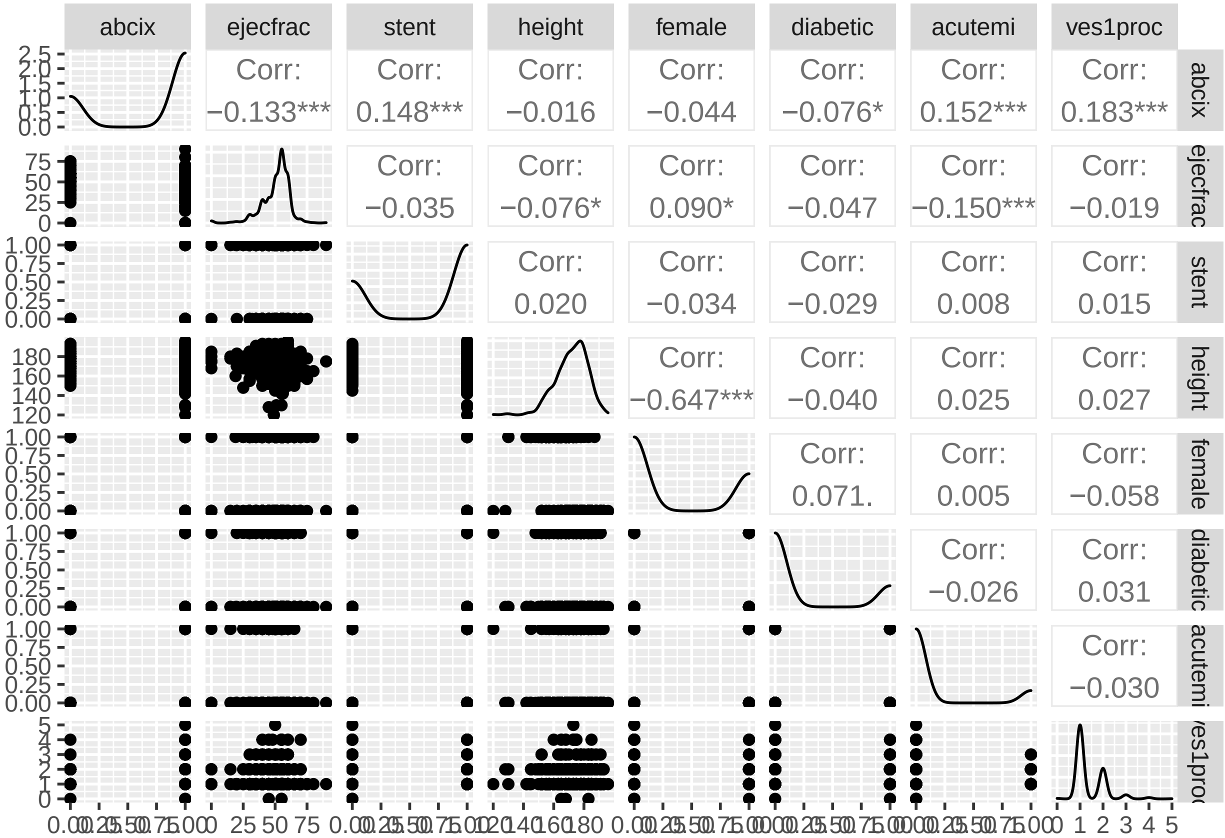
8 Question 3 (10 points)

> The investigator of the study has now asked you to add the following variables to your models: `stent`, `height`, `female`, `diabetic`, `acutemi`, and `ves1proc`. Assess the suitability of adding these variables (i.e. check for potential correlation issues), and then run and interpret this model. Call this `model3`.

We can start by building correlation matrix for all of our predictors. We can see here from the correlation coefficients that there are some strong correlations that may be worth considering before we build our full model. In this homework we just want you to note that these are of concern, but we'll go ahead and proceed. It seems, however, our variance inflation factor (VIF) is not so severe we'll need to worry about it.

```
GGally::ggpairs(dplyr::select(lindner_alive_train, abcix, ejecfrac, stent,
                              height, female, diabetic, acutemi, ves1proc))
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```



```
car::vif(lm(cardbill_log ~ abcix + ejecfrac + stent + height + female +
            diabetic + acutemi + ves1proc, data = lindner_alive_train))
```

```
   abcix ejecfrac    stent   height   female diabetic  acutemi ves1proc 
1.110448 1.049906 1.024075 1.728934 1.743425 1.017988 1.048900 1.043518 
```

8.1 Building The Model

Let's build our full model now, obtaining both the summary as well as the confidence intervals. In this model, we have now fully adjusted for the other covariates, and it seems our primary variable of interest (`ejecfrac`) is still associated with our cardiac costs. Our full regression equation is:

$$
\begin{aligned}
log(cardbill) = 9.4968274 - 0.0049088(ejecfrac) + 0.1512121(abcix) \\
+0.1021649(stent) - 0.0002582(height) + 0.0476881(female) - \\
0.0071010(diabetic) - 0.1182781(acutemi) + 0.1037275(ves1proc)
\end{aligned}
$$

```
model3 <- lm(cardbill_log ~ ejecfrac + abcix + stent + height + female +
               diabetic + acutemi + ves1proc, data = lindner_alive_train)
summary(model3)
```

```
Call:
lm(formula = cardbill_log ~ ejecfrac + abcix + stent + height +
    female + diabetic + acutemi + ves1proc, data = lindner_alive_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6846 -0.2839 -0.1215  0.1911  2.7330

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.4968274  0.3946259  24.065  < 2e-16 ***
ejecfrac    -0.0049088  0.0016430  -2.988 0.002914 **
abcix        0.1512121  0.0392184   3.856 0.000127 ***
stent        0.1021649  0.0362215   2.821 0.004936 **
height      -0.0002582  0.0021379  -0.121 0.903896
female       0.0476881  0.0474141   1.006 0.314884
diabetic    -0.0071010  0.0411011  -0.173 0.862884
acutemi     -0.1182781  0.0495801  -2.386 0.017328 *
ves1proc     0.1037275  0.0270869   3.829 0.000140 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4414 on 670 degrees of freedom
Multiple R-squared:  0.0876,    Adjusted R-squared:  0.0767
F-statistic: 8.041 on 8 and 670 DF,  p-value: 2.166e-10
```

```
confint(model3)
```

```
                   2.5 %       97.5 %
(Intercept)  8.721975000 10.271679734
ejecfrac    -0.008134947 -0.001682680
abcix        0.074206370  0.228217779
stent        0.031043588  0.173286227
height      -0.004456040  0.003939578
female      -0.045409876  0.140786171
diabetic    -0.087803523  0.073601427
acutemi     -0.215629145 -0.020927050
ves1proc     0.050542013  0.156913050
```

Using glance():

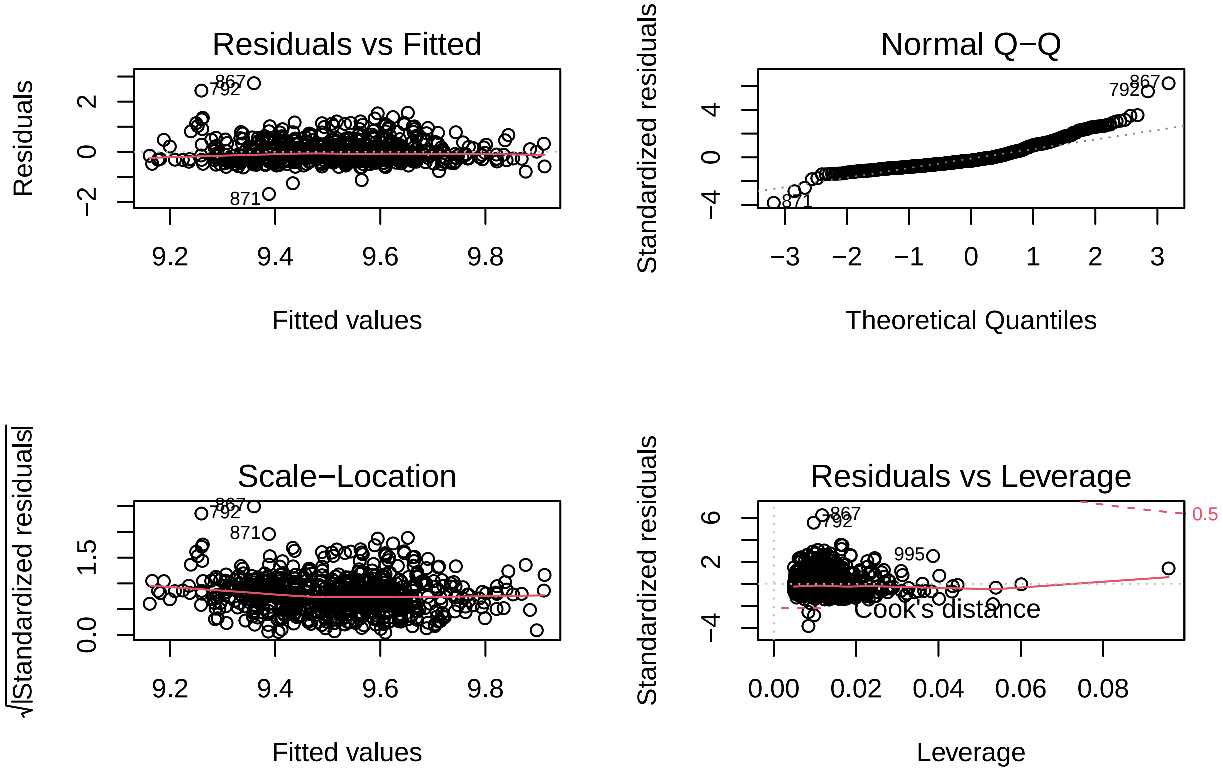
```
glance(model3) %>%
  select(r.squared, adj.r.squared, sigma, statistic, p.value, nobs,
         logLik:deviance) %>%
  knitr::kable(digits = 2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	nobs	logLik	AIC	BIC	deviance
0.09	0.08	0.44	8.04	0	679	-403.67	827.34	872.55	130.55

8.2 Residual Plots

Again, none of these suggest any grave concerns, but perhaps some concerns around Normality that would be important to consider.

```
par(mfrow=c(2,2))
plot(model3)
```



8.3 Grading Rubric

- If the student has evaluated collinearity, successfully added the additional variables to the model, briefly discussed their impact, and has residual plots they should receive all 10 points.
- If the student does either of the following they should lose 2 points each:
  - Does not include a scatterplot matrix of variance inflation factors for the addition variables
  - Does not include residual plots
- If there is no discussion or interpretation of the model, the student should lose 4 points.
- Other issues which are noted can be deducted at the TAs discretion, and noted in the grading sheet.

9 Question 4 (10 points)

> It's been suggested that the effect of `height` depends on `female`, which would suggest the desire to include an interaction term between these two variables in the model. Add this interaction term, run the model, and briefly discuss whether or not we see the interaction between these variables impacting `cardbill`. Call this `model4`.

Now, in this example we've decided to add an interaction term for `height` and `female`, while keeping all other variables in our model as we did in `model3`. Overall, it seems that the interaction term doesn't add much to our model. We don't need too much interpretation here, other than to recognize that we've added the interaction term but it doesn't seem to add much.

```
model4 <- lm(cardbill_log ~ ejecfrac + abcix + stent + height * female +
               diabetic + acutemi + ves1proc, data = lindner_alive_train)
summary(model4)
```

```

Call:
lm(formula = cardbill_log ~ ejecfrac + abcix + stent + height *
    female + diabetic + acutemi + ves1proc, data = lindner_alive_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6780 -0.2809 -0.1167  0.1892  2.7368

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.235154   0.471706  19.578  < 2e-16 ***
ejecfrac      -0.004975   0.001644  -3.026 0.002576 **
abcix          0.149896   0.039239   3.820 0.000146 ***
stent          0.100356   0.036265   2.767 0.005809 **
height         0.001245   0.002603   0.478 0.632620
female         0.822699   0.766855   1.073 0.283737
diabetic      -0.006369   0.041107  -0.155 0.876920
acutemi       -0.112929   0.049860  -2.265 0.023836 *
ves1proc       0.104242   0.027091   3.848 0.000131 ***
height:female -0.004640   0.004583  -1.013 0.311630
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4414 on 669 degrees of freedom
Multiple R-squared:  0.08899,	Adjusted R-squared:  0.07674
F-statistic: 7.261 on 9 and 669 DF,  p-value: 4.074e-10
```

```
confint(model4)
```

```
                     2.5 %       97.5 %
(Intercept)    8.308952113 10.161356702
ejecfrac      -0.008203786 -0.001746497
abcix          0.072848884  0.226942403
stent          0.029149123  0.171562405
height        -0.003865555  0.006355028
female        -0.683031899  2.328430736
diabetic      -0.087082503  0.074344814
acutemi       -0.210829489 -0.015028443
ves1proc       0.051047956  0.157435981
height:female -0.013638323  0.004357818
```

Using glance():

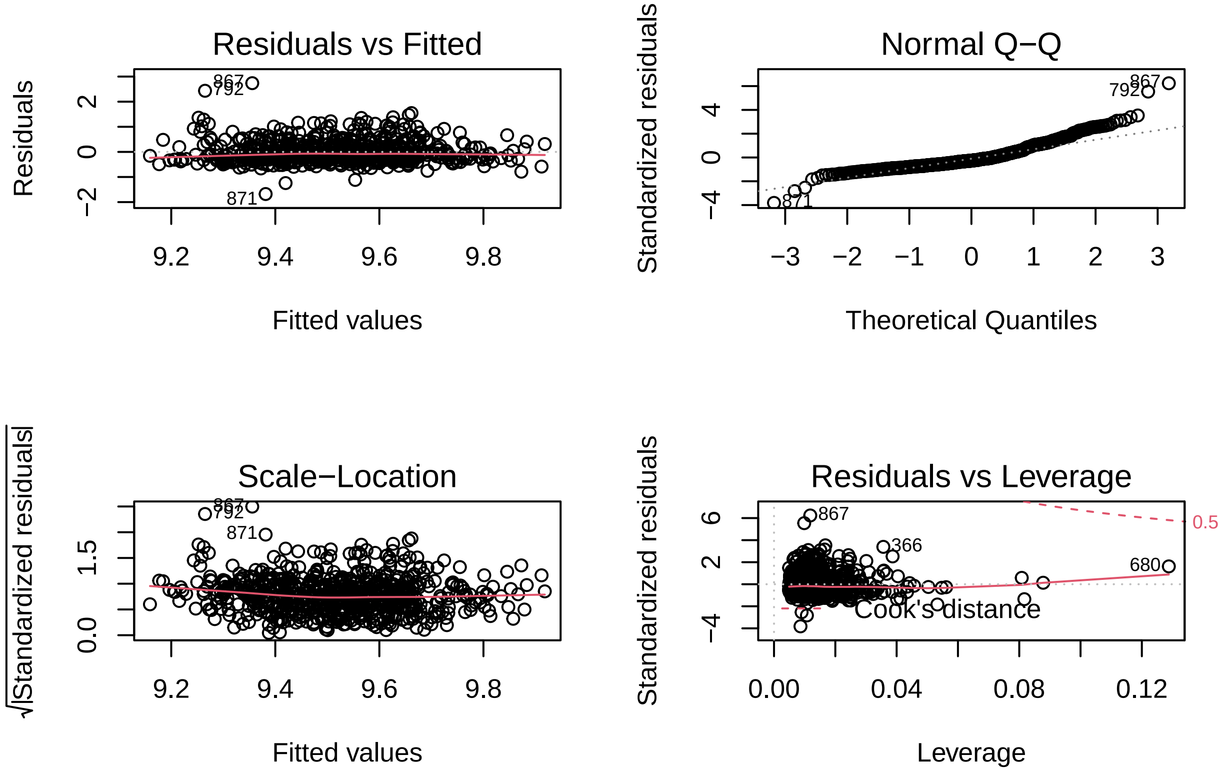
```
glance(model4) %>%
  select(r.squared, adj.r.squared, sigma, statistic, p.value, nobs,
         logLik:deviance) %>%
  knitr::kable(digits = 2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	nobs	logLik	AIC	BIC	deviance
0.09	0.08	0.44	7.26	0	679	-403.15	828.3	878.03	130.35

9.1 Residual Plots

Unsurprisingly, these look quite close to `model3` and overall suggest no major concerns other than Normality.

```
par(mfrow=c(2,2))
plot(model4)
```



9.2 Grading Rubric

- If the student has successfully added the interaction term to the model, briefly discussed its impact, and has residual plots they should receive all 10 points.
- If the student does not include residual plots they should lose 2 points.
- If there is no discussion or interpretation of the model, the student should lose 4 points.
- Other issues which are noted can be deducted at the TAs discretion, and noted in the grading sheet.

10 Question 5 (20 points)

> By now you should have created 4 models. Fit these models to the test data we had held-out earlier. Then, compare these models, using their adjusted R^2, AIC and BIC (from the training data), as well as their MAPE, RMSPE, and maximum prediction error (from the test data.) Which model performs best in which settings?

We can start by briefly summarizing our adjusted R^2 from within the training data.

Model	Adjusted R^2	AIC	BIC
Model 1	0.013	865.7	879.2
Model 2	0.043	845.7	863.8
Model 3	0.077	827.3	872.6
Model 4	0.077	828.3	878.0

From our adjusted R^2 values, we can see that the larger models perform better than the smaller, but that the addition of an interaction term doesn't help us too much. Overall, though, these models account for a fairly small amount of the variation in the logarithm of `cardbill`. When we look at the AIC and BIC we again see they are quite similar, with Model 3 having hte smallest AIC while model 2 has the smallest BIC.

Next, we'll evaluate the model's accuracy in our test sample. First, we'll need to fit each model to the new data.

```r
model1_test <- augment(model1, newdata = lindner_alive_test) %>%
  mutate(model = "Model 1")
model2_test <- augment(model2, newdata = lindner_alive_test) %>%
  mutate(model = "Model 2")
model3_test <- augment(model3, newdata = lindner_alive_test) %>%
  mutate(model = "Model 3")
model4_test <- augment(model4, newdata = lindner_alive_test) %>%
  mutate(model = "Model 4")
```

As we built these models to predict the log, we'll back transform to more easily compare our results.

```r
test_models <- bind_rows(model1_test, model2_test, model3_test, model4_test) %>%
  mutate(fit_cardbill = exp(.fitted),
         res_cardbill = cardbill - fit_cardbill) %>%
  select(model, id, cardbill, fit_cardbill, res_cardbill, everything()) %>%
  arrange(id, model)
```

Now we will summarize, and print our, our prediction errors.

```r
test_models %>%
  group_by(model) %>%
  summarise(MAPE = round(mean(abs(res_cardbill)),1),
            medAPE = round(median(abs(res_cardbill)),1),
            maxAPE = round(max(abs(res_cardbill)),1),
            RMSPE = round(sqrt(mean(res_cardbill^2)),1)) %>%
  knitr::kable()
```

model	MAPE	medAPE	maxAPE	RMSPE
Model 1	5794.4	3160.0	54864.6	9934.1
Model 2	5763.8	3433.5	56729.2	9863.0
Model 3	5747.3	3356.3	55854.1	9760.1
Model 4	5770.8	3377.4	55954.7	9786.4

We can see here that, overall, the models are incredibly close to one another. It is clear, though, Model 3 gives us the smallest MAPE, maxAPE, RMSPE although they are incredibly close to Model 4. Given the above adjusted R^2 results, and these prediction errors it seems that Model 3 may be the best fit for these data.

10.1 Grading Rubric

- If the student has successfully fit the data to the hold-out sample, discussed the adjusted R^2, AIC, BIC, and appropriately calculated the MAPE and RMSPE, and identified Model 3 as the 'best' they should receive all 20 points.
- If the student did not fit the data to the hold-out sample, they should lose 5 points.
- If the MAPE and RMSPE are not calculated, they should lose 5 points.
- If there is not a discussion of the adjusted R^2, AIC, and BIC, from the training data, the student should lose 3 points.
- If the student has settled on the wrong model, they should lose 3 points.
- Other issues which are noted can be deducted at the TAs discretion, and noted in the grading sheet.

11 Question 6 (20 points)

> Write a brief essay (150 words would be sufficient, but you can write more if you like) which relates what you've done in this assignment to what you learned in your reading of Spiegelhalter.

We don't write sketches for essay questions.

11.1 Grading Rubric

- Most students should receive between 15 and 20 points, assuming they have a well-written and sensible paragraph.