MATH 60604A Statistical Modelling

Chapter 2 Solutions

Juliana Schulz

Question 1

We are given the following output:

```
Call:
lm(formula = Price ~ Mileage, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-33.088 -10.372   2.868   9.233  20.868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.92843    3.16885  19.858  < 2e-16 ***
Mileage     -0.61269    0.07621  -8.039 5.26e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.89 on 58 degrees of freedom
Multiple R-squared:  0.527, Adjusted R-squared:  0.5189
F-statistic: 64.63 on 1 and 58 DF,  p-value: 5.263e-11
```

a) Interpret the coefficient for `Mileage`.

For every 1 unit increase in `Mileage`, that is, for every 1000 additional miles, the price of the car will decrease on average by 0.61269 (that is, by \$612.69.)

b) Report and interpret the value for the coefficient of determination.

$R^2 = 0.537$, we can interpret this as: the variable `Mileage` explains 52.7% of the variability in the price of the car.

Question 2

We are given the following output:

```
Call:
lm(formula = Price ~ Porsche, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-34.537 -12.057  -1.347  13.823  38.043

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.957      2.954  10.816 1.56e-15 ***
Porsche       18.580      4.178   4.447 4.00e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.18 on 58 degrees of freedom
Multiple R-squared:  0.2543,    Adjusted R-squared:  0.2414
F-statistic: 19.77 on 1 and 58 DF,  p-value: 3.997e-05
```

a) Write an expression for the underlying model being fit.

The model can be written in difference ways... Let $Y =$`Price`, $X =$`Porsche`, we can write

$$Y = \beta_0 + \beta_1 X + \epsilon \quad \text{or as} \quad E(Y|X) = \beta_0 + \beta_1 X.$$

In terms of the fitted model, we can write

$$\widehat{Y} = 31.96 + 18.58X \quad \text{or} \quad \widehat{E}(Y|X) = 31.96 + 18.58X$$

b) What is the predicted price for a Porsche?

For `Porsche`=1, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 = 50.54$ or \$50537

c) What is the estimated mean price for a Jaguar?

For `Porsche`=0, $\widehat{E}(Y|X = 0) = \hat{\beta}_0 = 31.96$ or \$31957

d) Based on the fitted model, is there a significant difference in the price of a Porsche compared to a Jaguar? Explain.

The parameter β_1 represents the difference in the mean price of a Porsche compared to a Jaguar:

$$\beta_1 = E(\texttt{Price}|\texttt{Porsche=1}) - E(\texttt{Price}|\texttt{Porsche=0}).$$

We're thus interested in testing

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

since if H_0 is true it means that there is no difference (on average) between the price of a Porsche and a Jaguar. The p-value for the above test is that corresponding to the parameter `Porsche`. Since the p-value is small ($4.00e-05$), for any reasonable α, $p < \alpha$ and thus we can reject H_0 and conclude that there is a significant difference between the price of a Porsche and that of a Jaguar.

Question 3

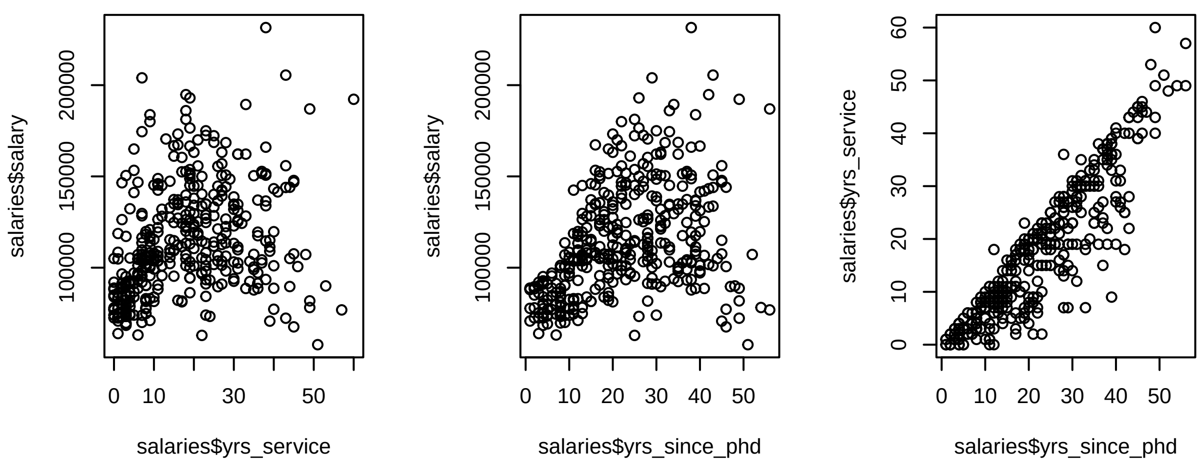
We first read the data:

```
salaries<-read.csv("Salaries.csv")
head(salaries)
```

```
       rank discipline yrs_since_phd yrs_service  sex salary
1      Prof          B            19          18 Male 139750
2      Prof          B            20          16 Male 173200
3  AsstProf          B             4           3 Male  79750
4      Prof          B            45          39 Male 115000
5      Prof          B            40          41 Male 141500
6 AssocProf          B             6           6 Male  97000
```

a) Produce scatterplots to visualize the pairwise relationships between the variables `salary`, `yrs_since_phd` and `yrs_service`. Comment.

```
par(mfrow=c(1,3))
plot(salaries$salary~salaries$yrs_service)
plot(salaries$salary~salaries$yrs_since_phd)
plot(salaries$yrs_service~salaries$yrs_since_phd)
```



Based on the scatterplots, we see a moderate positive relationship between `salary` and `yrs_service`, a moderate positive relationship between `salary` and `yrs_since_phd`, and a very strong positive relationship between `yrs_service` and `yrs_since_phd`. It seems that as the number of years of service increases, the salary tends to increase; and similarly for the number of years since PhD. Moreover, there is a very strong linear relation, naturally, between the number of years of service and the number of years since PhD indicating that as the number of years since PhD increases the number of years of service tends to increase.

b) Compute the correlations between the variables `salary`, `yrs_since_phd` and `yrs_service`. Are these correlations significant?

There are different ways to proceed in `R`:

```
# using cor and cor.test:
cor(salaries[,c(3,4,6)])
```

```
              yrs_since_phd yrs_service    salary
yrs_since_phd     1.0000000   0.9096491 0.4192311
yrs_service       0.9096491   1.0000000 0.3347447
salary            0.4192311   0.3347447 1.0000000
```

```
cor.test(salaries$salary,salaries$yrs_service)
```

```

    Pearson's product-moment correlation

data:  salaries$salary and salaries$yrs_service
t = 7.0602, df = 395, p-value = 7.529e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2443740 0.4193506
sample estimates:
      cor
0.3347447
```

```
cor.test(salaries$salary,salaries$yrs_since_phd)
```

```

    Pearson's product-moment correlation

data:  salaries$salary and salaries$yrs_since_phd
t = 9.1775, df = 395, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3346160 0.4971402
sample estimates:
      cor
0.4192311
```

```r
cor.test(salaries$yrs_since_phd,salaries$yrs_service)
```

```
	Pearson's product-moment correlation

data:  salaries$yrs_since_phd and salaries$yrs_service
t = 43.524, df = 395, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8909977 0.9252353
sample estimates:
      cor
0.9096491
```

```r
# alternative using rcorr from the Hmisc library
library(Hmisc)
```

```
Attaching package: 'Hmisc'


The following objects are masked from 'package:base':

    format.pval, units
```

```r
rcorr(as.matrix(salaries[,c(3,4,6)]))
```

```
              yrs_since_phd yrs_service salary
yrs_since_phd          1.00        0.91   0.42
yrs_service            0.91        1.00   0.33
salary                 0.42        0.33   1.00

n= 397


P
              yrs_since_phd yrs_service salary
yrs_since_phd                0           0
yrs_service    0                         0
salary         0             0
```

The sample correlations r confirm what we observed in the scatterplots. The p-values provided in the output correspond to the test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$, where ρ represents the true/population correlation between the corresponding variables. All of the p-values are small, so for any reasonable α, e.g. $\alpha = 0.01$, we can reject H_0 since $p < \alpha$. We can thus conclude that each pairwise correlation is significantly different from 0.

c) Fit a linear regression model for `salary` including the following explanatory variables (i) only `yrs_since_phd`, and (ii) only `yrs_service`. For both, provide the fitted model and comment on the significance of the variable effects (use $\alpha = 0.01$).

```r
# (i)
lm.i<-lm(salary~yrs_since_phd,data=salaries)
summary(lm.i)
```

```
Call:
lm(formula = salary ~ yrs_since_phd, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max 
-84171 -19432  -2858  16086 102383 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    91718.7     2765.8  33.162   <2e-16 ***
yrs_since_phd    985.3      107.4   9.177   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27530 on 395 degrees of freedom
Multiple R-squared:  0.1758,    Adjusted R-squared:  0.1737 
F-statistic: 84.23 on 1 and 395 DF,  p-value: < 2.2e-16
```

```r
# (ii)
lm.ii<-lm(salary~yrs_service,data=salaries)
summary(lm.ii)
```

```
Call:
lm(formula = salary ~ yrs_service, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max 
-81933 -20511  -3776  16417 101947 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  99974.7     2416.6   41.37  < 2e-16 ***
yrs_service    779.6      110.4    7.06 7.53e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28580 on 395 degrees of freedom
Multiple R-squared:  0.1121,    Adjusted R-squared:  0.1098 
F-statistic: 49.85 on 1 and 395 DF,  p-value: 7.529e-12
```

Based on the output, the fitted models are:

- $\widehat{\texttt{salary}} = 91719 + 985\ \texttt{yrs\_since\_phd}$
- $\widehat{\texttt{salary}} = 99975 + 780\ \texttt{yrs\_service}$

For each of the models, we can test whether the variable effects are significant by testing if the underlying parameter is significantly different from 0: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. For model (i) we see that the p-value for the parameter β_1 is less than $\alpha = 0.01$ ($p < 2e-16$) so that we can reject $H_0 : \beta_1 = 0$ and conclude that the variable `yrs_since_phd` has a significant linear effect on `salary`. Similarly, for model (ii) we see that the p-value for the parameter β_1 is less than α ($p = 7.53e-12$) so that we can reject $H_0 : \beta_1 = 0$ and conclude that the variable `yrs_service` has a significant linear effect on `salary`.

d) Carry out a residual analysis for model (i). Comment.

```
library(cowplot)

library("ggplot2")
res.dat<-data.frame(cbind(salaries$salary,salaries$yrs_since_phd,
                          lm.i$fitted,lm.i$residuals,rstandard(lm.i),rstudent(lm.i)))
names(res.dat)<-c("salary","yrs_since_phd","fitted","resid","rstand","rstud")
head(res.dat)
```

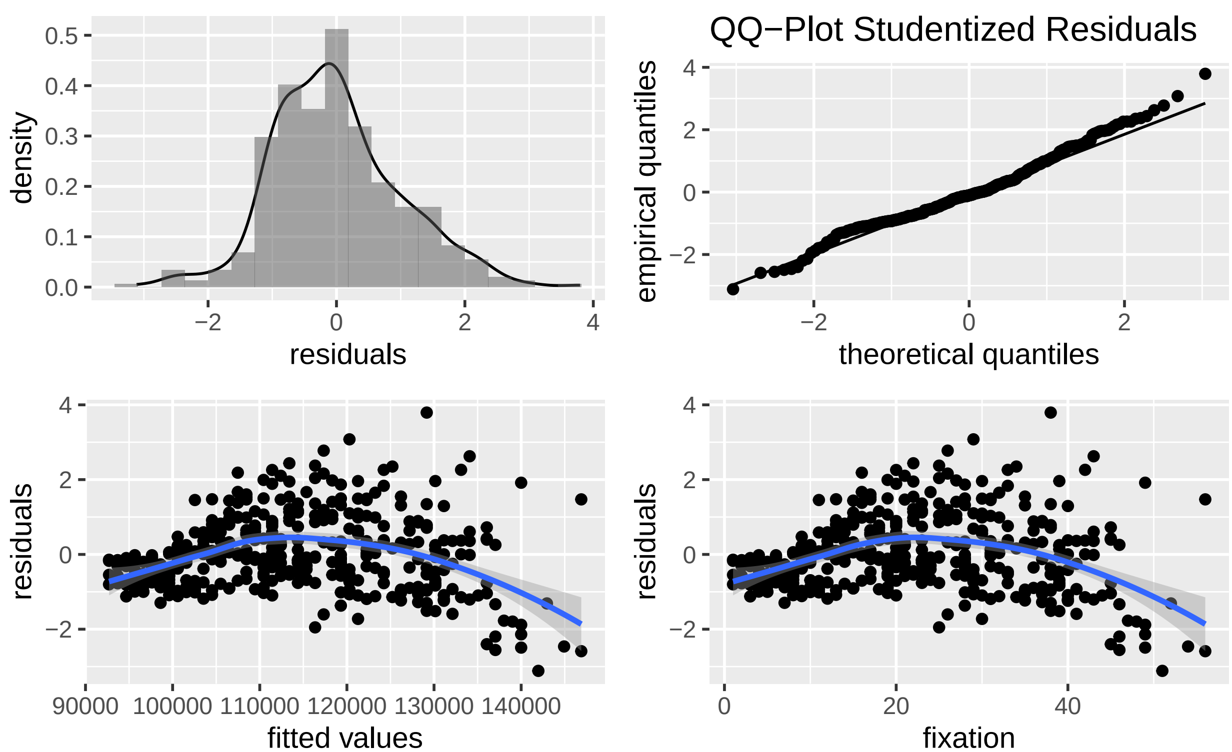
```
  salary yrs_since_phd    fitted       resid      rstand       rstud
1 139750            19 110440.19  29309.8142  1.06594386  1.06612820
2 173200            20 111425.53  61774.4721  2.24652715  2.25815415
3  79750             4  95660.05 -15910.0539 -0.58005575 -0.57956793
4 115000            45 136059.08 -21059.0810 -0.76883738 -0.76843874
5 141500            40 131132.37  10367.6296  0.37792197  0.37751154
6  97000             6  97630.74   -630.7382 -0.02298354 -0.02295444
```

```
# histogram rstud
plot1<-ggplot(data = res.dat, mapping = aes(x = rstud)) +
  geom_density() +
  geom_histogram(aes(y = ..density..), bins = 20, alpha = 0.5) +
  xlab("residuals")

# qqplot rstud
plot2<-ggplot(data = res.dat, mapping = aes(sample = rstud)) +
  stat_qq(distribution = qt, dparams = lm.i$df.residual) +
  stat_qq_line(distribution = qt, dparams = lm.i$df.residual) +
  labs(x = "theoretical quantiles",
       y = "empirical quantiles") +
  ggtitle("QQ-Plot Studentized Residuals")


# resid vs. fitted + smooth
plot3<-ggplot(data = res.dat,
       aes(x = fitted, y = rstud)) +
  geom_point() +
  geom_smooth() +
  theme(legend.position = "bottom") +
  ylab("residuals") +
  xlab("fitted values")

# resid vs. yrs_since_phd + smooth
plot4<-ggplot(data = res.dat,
       aes(x = yrs_since_phd, y = rstud)) +
```

```r
  geom_point() +
  geom_smooth() +
  theme(legend.position = "bottom") +
  ylab("residuals") +
  xlab("fixation")


plot_grid(plot1, plot2, plot3, plot4)
```

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
generated.
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The histogram of the residuals is roughly bell-shaped and is roughly symmetric about 0. The points on the QQ-plot fall roughly along the diagonal line. These plots suggest that the assumption of normally distributed error terms is reasonable. The plots of the residuals vs. the explanatory variable X (`yrs_since_phd`) show a funnel shape, that is, the variability in the residuals increases with with X. This suggests that there is heteroscedasticity. A similar pattern is seen in the plot of the residuals vs. the predicted values. There also seems to be a very slight curvature in the loess line, suggesting perhaps there is a quadratic relationship between the variables.

e) Using linear regression, test whether there is a significant difference between the salaries of female and male professors.

Let Y denote the `salary` and let X be an indicator variable where $X = 0$ if the subject is female and $X = 1$ if the subject is male. Consider the linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

In this model, $\beta_1 = E(Y|X = 1) - E(Y|X = 0)$ represents the difference in the mean salary for men vs. women. Thus, to test if there is a significant difference between the salaries of female and male professors, we can test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta \neq 0$. Fitting this model gives the following results:

```r
summary(lm(salary~sex,data=salaries))
```

```
Call:
lm(formula = salary ~ sex, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-57290 -23502  -6828  19710 116455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101002       4809  21.001  < 2e-16 ***
sexMale        14088       5065   2.782  0.00567 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared:  0.01921,   Adjusted R-squared:  0.01673
F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667
```

Based on the output, we see that the p-value for this test is 0.0057, thus for $\alpha = 0.05$ or even $\alpha = 0.01$, we can reject H_0 as $p < \alpha$. Thus we can conclude that there is a significant difference in the salaries of male and female professors, at the $\alpha = 0.01$ significance level.

Question 4

a) Fit a linear regression model for `salary` using both `yrs_since_phd` and `yrs_service` as predictors. Comment on the significance of the variable effects. Are the results from this model surprising based on what was observed in question Question 3? Explain.

```r
summary(lm(salary~yrs_since_phd+yrs_service,data=salaries))
```

```
Call:
lm(formula = salary ~ yrs_since_phd + yrs_service, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-79735 -19823  -2617  15149 106149

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    89912.2     2843.6  31.620  < 2e-16 ***
yrs_since_phd   1562.9      256.8   6.086 2.75e-09 ***
yrs_service     -629.1      254.5  -2.472   0.0138 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27360 on 394 degrees of freedom
Multiple R-squared:  0.1883,    Adjusted R-squared:  0.1842
F-statistic: 45.71 on 2 and 394 DF,  p-value: < 2.2e-16
```

The p-value for β_1 is less than $\alpha = 0.01$ so that we can reject $H_0 : \beta_1 = 0$ and conclude that `yrs_since_phd` has a significant linear effect on `salary`, even after adjusting for `yrs_service`. On the other hand, the p-value for β_2 is larger than α ($p = 0.0138$) so that we fail to reject $H_0 : \beta_2 = 0$ and thus conclude that the variable `yrs_service` does not have a significant linear effect on `salary`, once `yrs_since_phd` is included in the model. It seems that the variable `yrs_service` is redundant in the model once `yrs_since_phd` is included in the model. And indeed, in **Question 3** we saw that the correlation between `yrs_service` and `yrs_since_phd` was very high.

b) Fit a linear regression model for salary including sex and rank as predictors. (Use `Female` as the reference level for sex and `AsstProf` as the reference level for rank).

Let Y denote `salary`. The variable `sex` is binary (i.e. categorical with 2 levels) and thus we can include a single variable to represent it in the model: define X_{sex} as the indicator variable for `sex`, with $X_{sex} = 1$ for male and $X_{sex} = 0$ for female. Since `rank` is a categorical variable with 3 levels, we must introduce 2 dummy or indicator variables in the model. Using `AsstProf` as the reference level, we can define X_{assoc} to be the indicator for the level `AssocProf` and X_{prof} to be the indicator for the level `Prof`. That is,

Level	X_{assoc}	X_{prof}
`AsstProf`	0	0
`AssocProf`	1	0
`Prof`	0	1

We're thus interested in fitting the model

$$Y = \beta_0 + \beta_1 X_{sex} + \beta_2 X_{assoc} + \beta_3 X_{prof} + \epsilon$$

We can fit this model in R:

```r
# set reference level:
salaries$rank<-relevel(as.factor(salaries$rank),"AsstProf")
# model:
summary(lm(salary~sex+rank,data=salaries))
```

```
Call:
lm(formula = salary ~ sex + rank, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-69307 -15757  -1449  12359 104438

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      76645       4433  17.290  < 2e-16 ***
sexMale           4943       4026   1.228  0.22029
rankAssocProf    13061       4128   3.164  0.00168 **
rankProf         45519       3252  13.998  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 23620 on 393 degrees of freedom
Multiple R-squared:  0.3966,    Adjusted R-squared:  0.392
F-statistic: 86.09 on 3 and 393 DF,  p-value: < 2.2e-16
```

(i) Provide the fitted model and interpret all of the model parameters.

The fitted model is

$$\hat{y} = 76645 + 4943X_{sex} + 13061X_{assoc} + 45519X_{prof}$$

From this model, we have that

$$\begin{aligned}
\beta_0 &= E(Y|X_{sex} = 0, X_{assoc} = 0, X_{prof} = 0) \\
\beta_1 &= E(Y|X_{sex} = 1, X_{assoc}, X_{prof}) - E(Y|X_{sex} = 0, X_{assoc}, X_{prof}) \\
\beta_2 &= E(Y|X_{sex}, X_{assoc} = 1, X_{prof} = 0) - E(Y|X_{sex}, X_{assoc} = 0, X_{prof} = 0) \\
\beta_3 &= E(Y|X_{sex}, X_{assoc} = 0, X_{prof} = 1) - E(Y|X_{sex}, X_{assoc} = 0, X_{prof} = 0)
\end{aligned}$$

We can thus interpret the parameters as follows:

- β_0 represents the mean salary for female assistant professors, which is estimated as $\hat{\beta}_0 = 76645$.
- β_1 represents the difference in the mean salary for males vs. females of the same rank, that is, holding rank fixed, which is estimated as $\hat{\beta}_1 = 4943$. In other words, the difference in the salary of males vs. females of the same rank is on average \$4943.
- β_2 represents the difference in the mean salary for associate professors vs. assistant professors of the same sex, that is, holding sex fixed, which is estimated as $\hat{\beta}_2 = 13061$. In other words, the difference in the salary of associate vs. assistant professors of the same sex is on average \$13061.
- β_3 represents the difference in the mean salary for full professors vs. assistant professors of the same sex, that is, holding sex fixed, which is estimated as $\hat{\beta}_3 = 45519$. In other words, the difference in the salary of full vs. assistant professors of the same sex is on average \$45519.

(ii) In this model, is there a significant difference between the salaries of female and male professors? How do these results compare with that in **Question 3**? Explain.

Here we are interested in testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. The p-value corresponding to this test if $p = 0.2203$, so for any reasonable α (e.g. $\alpha = 0.05$), we fail to reject H_0 and conclude that there is not a significant difference in the salary of male and female professors after adjusting for rank. In the model in **Question 3**, which only included the explanatory variable `sex`, we found that there was a significant difference in the salary of male and female professors. However, in this model, once we adjust for rank, the difference in the salary of male and female professors is no longer significantly different from 0. Thus, it seems that the difference in salaries of male and female professors seen in **Question 3** can in fact be explained by a difference in their ranks as professors rather than their sex.

c) Fit a linear regression model for `salary` including `rank`, `discipline` and `yrs_service`. (Use `AsstProf` as the reference level for `rank` and `A` as the reference level for `discipline`).

Define Y and X_{assoc}, X_{prof} as in part (b). Let $X_B = 1$ if the discipline is B and $X_B = 0$ if the discipline is A. Let X_{yrs} denote the number of years of service (`yrs_service`). We're interested in the model:

$$Y = \beta_0 + \beta_1 X_{assoc} + \beta_2 X_{prof} + \beta_3 X_B + \beta_4 X_{yrs} + \epsilon.$$

Fitting this model in `R` gives the following results:

```r
lmod<-lm(salary~rank+discipline+yrs_service,data=salaries)
summary(lmod)
```

```
Call:
lm(formula = salary ~ rank + discipline + yrs_service, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-64198 -14040  -1299  10724  99253

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   72253.53    3169.48  22.797  < 2e-16 ***
rankAssocProf 14483.23    4100.53   3.532 0.000461 ***
rankProf      49377.50    3832.90  12.883  < 2e-16 ***
disciplineB   13561.43    2315.91   5.856 1.01e-08 ***
yrs_service     -76.33     111.25  -0.686 0.493039
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22670 on 392 degrees of freedom
Multiple R-squared:  0.4456,    Adjusted R-squared:   0.44
F-statistic: 78.78 on 4 and 392 DF,  p-value: < 2.2e-16
```

(i) Provide the fitted model. Interpret the coefficient associated with the variable `yrs_service`.

The fitted model is

$$\hat{y} = 72254 + 14483 X_{assoc} + 49378 X_{prof} + 13561 X_B - 76 X_{yrs}.$$

We can interpret the coefficient associated with the variable `yrs_service`, i.e. β_4, as follows: for every one year increase in years of service, the salary will increase on average by β_4, holding rank and discipline constant. In terms of the fitted model, for every additional year of service, the salary of the professor will decrease on average by $76, holding rank and discipline fixed.

(ii) Comment on the difference between the test results from the individual parameter effects in comparison to the global effect for the variable `rank`. What can you conclude?

```
library(car)
```

```
Loading required package: carData
```

```
Anova(lmod,type=3)
```

```
Anova Table (Type III tests)

Response: salary
                Sum Sq  Df  F value    Pr(>F)
(Intercept) 2.6700e+11   1 519.6871 < 2.2e-16 ***
rank        1.0454e+11   2 101.7391 < 2.2e-16 ***
discipline  1.7617e+10   1  34.2901 1.005e-08 ***
yrs_service 2.4187e+08   1   0.4708     0.493
Residuals   2.0140e+11 392
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable `rank` is categorical, with 3 levels: `AsstProf`, `AssocProf`, `Prof`. We saw that in order to incorporate this variable into a linear regression model, we had to include two dummy variables. When the rank `AsstProf` is used as the reference level, this meant including the indicator variables X_{assoc} and X_{prof}. When we look at tests for the individual parameters, we're testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$, i.e. we're doing a t-test for the individual regression coefficients. In this model, β_1 is the coefficient corresponding to X_{assoc} and β_2 is the coefficient corresponding to X_{prof}. When we test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, we're testing whether there's a significant difference in the mean salary of associate vs. assistant professors, after adjusting for discipline and years of service:

$$\begin{aligned}\beta_1 &= E(Y|X_{assoc} = 1, X_{prof} = 0, X_B, X_{yrs}) - E(Y|X_{assoc} = 0, X_{prof} = 0, X_B, X_{yrs}) \\ &= E(\texttt{salary }|\texttt{AssocProf },\texttt{rank, yrs\_service}) - E(\texttt{salary }|\texttt{AsstProf },\texttt{rank, yrs\_service})\end{aligned}$$

Similarly, when we test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, we're testing whether there's a significant difference in the mean salary of full vs. assistant professors, after adjusting for discipline and years of service:

$$\begin{aligned}\beta_2 &= E(Y|X_{assoc} = 0, X_{prof} = 1, X_B, X_{yrs}) - E(Y|X_{assoc} = 0, X_{prof} = 0, X_B, X_{yrs}) \\ &= E(\texttt{salary }|\texttt{Prof },\texttt{rank, yrs\_service}) - E(\texttt{salary }|\texttt{AsstProf },\texttt{rank, yrs\_service})\end{aligned}$$

The test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ has p-value $p = 0.0005$ while that for β_2 has p-value $p < 2e - 16$. In both cases, the p-values are very small, e.g. if we take $\alpha = 0.01$ both p-values are smaller than α. We can thus conclude that there is a significant difference in the mean salary of associate vs. assistant professors, after adjusting for discipline and years of service. Similarly, we can conclude that there is a significant difference in the mean salary of full vs. assistant professors, after adjusting for discipline and years of service.

The test for the global effect of the variable `rank`, on the other hand, tests $H_0 : \beta_1 = \beta_2 = 0$ vs. $H_1 :$ at least one of β_1 or β_2 is different from 0. This allows us to test whether the variable `rank` is useful in explaining the response variable, `salary`. The p-value for this test is very small, $p < 2.2e-16$, and for any reasonable α, $p < \alpha$. Thus, we can reject H_0 and conclude that the variable `rank` is indeed useful for explaining the salary of professors, even after adjusting for discipline and years of service.

(iii) Carry out a global test for the overall fit of the model.

Recall the summary output for the model:

```r
summary(lm(salary~rank+discipline+yrs_service,data=salaries))
```

```
Call:
lm(formula = salary ~ rank + discipline + yrs_service, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-64198 -14040  -1299  10724  99253

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   72253.53    3169.48  22.797  < 2e-16 ***
rankAssocProf 14483.23    4100.53   3.532 0.000461 ***
rankProf      49377.50    3832.90  12.883  < 2e-16 ***
disciplineB   13561.43    2315.91   5.856 1.01e-08 ***
yrs_service     -76.33     111.25  -0.686 0.493039
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22670 on 392 degrees of freedom
Multiple R-squared:  0.4456,    Adjusted R-squared:   0.44
F-statistic: 78.78 on 4 and 392 DF,  p-value: < 2.2e-16
```

The global test for the overall fit of the model tests

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs.} \quad H_1 : \text{At least one } \beta_j \text{ is different from } 0$$

Based on the output, the F-statistic for this test is 79.78, with a p-value of $p < 2.2e-16$, so for any reasonable α, $p < \alpha$ and we can reject H_0. Thus, the model is useful at explaining the salary of professors; at least one of the variables in the model is useful for explaining the variability in professors' salaries. From this test, we don't know which one, but we know at least one β_j is significantly different from 0.

(iv) Based on this model, estimate the difference between the mean salaries of associate and full professors. Provide a 95% C.I. for this estimated difference.

To do this, we can either refit the model using either `AssocProf` or `Prof` as the reference level for the variable `rank`, or we can use functions in the `emmeans` library in `R`:

```r
library(emmeans)
comp<-emmeans(lm(salary~rank+discipline+yrs_service,data=salaries),~rank)
confint(contrast(comp,method="pairwise",adjust="none"))
```

```
 contrast             estimate   SE  df lower.CL upper.CL
 AsstProf - AssocProf   -14483 4101 392   -22545    -6421
 AsstProf - Prof        -49378 3833 392   -56913   -41842
 AssocProf - Prof       -34894 3376 392   -41532   -28257

Results are averaged over the levels of: discipline
Confidence level used: 0.95
```

From this, we see that the difference in the mean salary of `AssocProf` and `Prof` is estimated as -34894 with corresponding 95% C.I. given by $(-41532, -28257)$. Note that from the fitted model, as given in the beginning of part (c), we can estimate the difference as $\hat{\beta}_1 - \hat{\beta}_2 = 14483.23 - 49338.50 \approx -34894$, although we cannot obtain a CI for this.

Question 5

We first read the data:

```r
credit<-read.csv("Credit.csv")
head(credit)
```

```
  ID  Income Limit Rating Cards Age Education Gender Student Married Ethnicity
1  1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian
2  2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian
3  3 104.593  7075    514     4  71        11   Male      No      No     Asian
4  4 148.924  9504    681     3  36        11 Female      No      No     Asian
5  5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian
6  6  80.180  8047    569     4  77        10   Male      No      No Caucasian
  Balance
1     333
2     903
3     580
4     964
5     331
6    1151
```

a) Fit a linear regression model to assess the simultaneous effects of the above mentioned variables on a client's credit rating. (Use `No` as the reference level for both `Married` and `Student`). Provide the fitted model.

```r
credit.mod<-lm(Rating~Income+Limit+Cards+Age+Student+Married,data=credit)
summary(credit.mod)
```

```
Call:
lm(formula = Rating ~ Income + Limit + Cards + Age + Student +
    Married, data = credit)

Residuals:
     Min       1Q   Median       3Q      Max
-22.9095  -7.1491  -0.5578   6.0976  26.2778

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.267e+01  2.395e+00   9.465   <2e-16 ***
Income      3.096e-02  2.413e-02   1.283   0.2001
Limit       6.640e-02  3.641e-04 182.365   <2e-16 ***
Cards       4.898e+00  3.737e-01  13.106   <2e-16 ***
Age         6.862e-03  3.032e-02   0.226   0.8211
StudentYes  2.809e+00  1.710e+00   1.643   0.1013
MarriedYes  2.073e+00  1.055e+00   1.964   0.0503 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.21 on 393 degrees of freedom
Multiple R-squared:  0.9957,    Adjusted R-squared:  0.9956
F-statistic: 1.52e+04 on 6 and 393 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\widehat{\texttt{Rating}} = 22.67 + 0.03\texttt{Income} + 0.07\texttt{Limit} + 4.90\texttt{Cards} + 0.01\texttt{Age} + 2.81\texttt{StudentYes} + 2.07\texttt{MarriedYes}$$

where `Rating` represents the credit rating, `Income` represents the income, `Limit` represents the credit limit, `Age` represents the person's age, `StudentYes` is an indicator variable which takes on the value 1 if the individual is a student and is 0 otherwise, and `MarriedYes` is an indicator variable which takes on the value 1 if the individual is married and is 0 otherwise.

b) Comment on the significance of the variable effects (use $\alpha = 0.05$).

For each of the variables in the model, we can test whether the corresponding coefficient is significantly different from 0, i.e. $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. Throughout, we'll use $\alpha = 0.05$. The variables `Income`, `Age`, `StudentYes` and `MarriedYes` all have p-values that are larger than 0.05, thus we fail to reject $H_0 : \beta_j = 0$ and conclude that each of these variables do not have significant (linear) effects on credit rating in the model that adjusts for income, limit, number of cards, age, student and marital status. On the other hand, the variables `Limit` and `Cards` have small p-values, $p < 0.05$, and thus we can reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j \neq 0$ and conclude that these two variables have significant (linear) effects on the credit rating, in the model that adjusts for income, limit, number of cards, age, student and marital status.

c) Comment on the value of R^2 for the model.

From the `R` output, we have that $R^2 = 0.9957$. We can interpret this value as 99.57% of the variability in the credit rating is explained by the model. This is a very high value of R^2, suggesting that the model is quite good (although perhaps we are actually overfitting....)

d) Predict the credit rating for a person with an income of 15, a limit of $5,000$, with 2 credit cards, who is a single, 27 years old student. Provide a 95% prediction interval for the predicted value.

```
new.dat<-as.data.frame(matrix(c(15,5000,2,27,"Yes","No"),ncol=6))
names(new.dat)<-c("Income","Limit","Cards","Age","Student","Married")
new.dat[,1:4] <- lapply(1:4, function(x)as.numeric(new.dat[[x]]))

predict(credit.mod,newdata=new.dat,interval="prediction")
```

```
       fit      lwr      upr
1 367.9339 347.4552 388.4126
```

We obtain a predicted credit rating value of 367.93 with corresponding 95% prediction interval $(347.46, 388.41)$.

e) Estimate the mean credit rating for individuals with an income of 100, a limit of $10,000$, 5 credit cards, is 55 years old, married and not a student. Provide a 95% C.I. for this estimate.

```
new.dat<-as.data.frame(matrix(c(100,10000,5,55,"No","Yes"),ncol=6))
names(new.dat)<-c("Income","Limit","Cards","Age","Student","Married")
new.dat[,1:4] <- lapply(1:4, function(x)as.numeric(new.dat[[x]]))

predict(credit.mod,newdata=new.dat,interval="confidence")
```

```
       fit      lwr      upr
1 716.7258 713.6904 719.7613
```

We obtain an estimated mean credit rating of 716.73 with corresponding 95% C.I. $(713.69, 719.76)$.

Question 6

Recall the given output:

```
Call:
lm(formula = Price ~ Age * CarType, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.8887  -6.8037  -0.9248   6.0900  23.0915 

Coefficients:
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          58.2268     4.7056  12.374  < 2e-16 ***
Age                  -4.8265     0.7522  -6.416 7.85e-09 ***
CarTypeJaguar        -1.2384     6.2742  -0.197  0.84401
CarTypePorsche        5.1483     5.3702   0.959  0.34048
Age:CarTypeJaguar    -0.2135     1.0668  -0.200  0.84186
Age:CarTypePorsche    2.7558     0.8119   3.394  0.00105 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.656 on 84 degrees of freedom
Multiple R-squared:  0.7172,    Adjusted R-squared:  0.7004
F-statistic: 42.61 on 5 and 84 DF,  p-value: < 2.2e-16
```

```
Anova Table (Type III tests)

Response: Price
             Sum Sq Df F value    Pr(>F)
(Intercept) 14277.6  1 153.115 < 2.2e-16 ***
Age          3839.0  1  41.170 7.852e-09 ***
CarType       198.1  2   1.062    0.3504
Age:CarType  2025.0  2  10.858 6.394e-05 ***
Residuals    7832.8 84
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let Y denote the price of the car, X_1 denote the age of the car, and define the following indicator variables as follows

	X_J	X_P
Jaguar	1	0
Porsche	0	1
BMW	0	0

The model here is

$$E(Y|X_1, X_J, X_P) = \beta_0 + \beta_1 X_1 + \beta_2 X_J + \beta_3 X_P + \beta_4 X_1 X_J + \beta_5 X_1 X_P$$

and thus

$$\begin{aligned} E(Y|X_1, \text{BMW}) &= E(Y|X_1, X_J = 0, X_P = 0) = \beta_0 + \beta_1 X_1 \\ E(Y|X_1, \text{Jaguar}) &= E(Y|X_1, X_J = 1, X_P = 0) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 \\ E(Y|X_1, \text{Porsche}) &= E(Y|X_1, X_J = 0, X_P = 1) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 \end{aligned}$$

a) Interpret the intercept in the model.

$\beta_0 = E(Y|X_1 = 0, \text{BMW})$ is the expected price of a new BMW, which is estimated as \$58227.

b) Interpret the coefficient for `Age` in the model.

$\beta_1 = E(Y|X_1 = x+1, X_J = 0, X_P = 0) - E(Y|X_1 = x, X_J = 0, X_P = 0)$ is the change in the mean price of a BMW associated with a one year increase in age, which is estimated as $\$-4826$. Thus, for every one year increase in the car's age, the price of a BMW decreases on average by $\$-4826$.

c) According to this model, what is the estimated price of a new Jaguar?

The estimated price of a new Jaguar is $\widehat{E}(Y|X_1, X_J = 1, X_P = 0) = \hat{\beta}_0 + \hat{\beta}_2 = \$56,988$

d) Write an expression for the mean price of a Porsche which is x years old.

The mean price of a Porsche of age x is

$$E(Y|X_1 = x, \text{Porsche}) = E(Y|X_1 = x, X_J = 0, X_P = 1) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x$$

e) Carry out a formal statistical test to assess whether the effect of age on the car's price depends on the type of car. Clearly write out the underlying hypotheses, the value of the test-statistic and the corresponding p-value. What can you conclude?

If the effect of age on the car's price depends on the type of car, then there is an *interaction*. So here, we're interested in testing whether the interaction is significant. The underlying hypotheses are then

$$H_0 : \beta_4 = \beta_5 = 0 \quad \text{vs.} \quad H_1 : \text{at least one of } \beta_4 \text{ or } \beta_5 \neq 0$$

The test statistic for this is $F = 10.86$ with a corresponding p-value $p = 6.394e-05$. Since the p-value is small (and will be smaller than any reasonable α), we can reject H_0 and conclude that at least one of β_4 or β_5 are significantly different from 0, that is, the interaction is significant in the model. Thus, the effect of age on the price of the car depends on the type of car.

Question 7

The data:

```
summary(NCbirths)
```

```
       X                ID             Plural           Sex       
 Min.   :   1.0   Min.   :   1.0   Min.   :1.000   Min.   :1.000  
 1st Qu.: 363.2   1st Qu.: 363.2   1st Qu.:1.000   1st Qu.:1.000  
 Median : 725.5   Median : 725.5   Median :1.000   Median :1.000  
 Mean   : 725.5   Mean   : 725.5   Mean   :1.037   Mean   :1.487  
 3rd Qu.:1087.8   3rd Qu.:1087.8   3rd Qu.:1.000   3rd Qu.:2.000  
 Max.   :1450.0   Max.   :1450.0   Max.   :3.000   Max.   :2.000  
                                                                  
     MomAge          Weeks          Marital         RaceMom     
 Min.   :13.00   Min.   :22.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:22.00   1st Qu.:38.00   1st Qu.:1.000   1st Qu.:1.000  
 Median :26.00   Median :39.00   Median :1.000   Median :1.000  
 Mean   :26.76   Mean   :38.62   Mean   :1.345   Mean   :1.831  
 3rd Qu.:31.00   3rd Qu.:40.00   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :43.00   Max.   :45.00   Max.   :2.000   Max.   :8.000  
                 NA's   :1                                      
   HispMom              Gained          Smoke         BirthWeightOz  
 Length:1450        Min.   : 0.0   Min.   :0.0000   Min.   : 12.0  
 Class :character   1st Qu.:20.0   1st Qu.:0.0000   1st Qu.:106.0  
 Mode  :character   Median :30.0   Median :0.0000   Median :118.0  
                    Mean   :30.6   Mean   :0.1446   Mean   :116.2  
                    3rd Qu.:40.0   3rd Qu.:0.0000   3rd Qu.:130.0  
                    Max.   :95.0   Max.   :1.0000   Max.   :181.0  
                    NA's   :40     NA's   :5                       
 BirthWeightGm         Low              Premie          MomRace         
 Min.   : 340.2   Min.   :0.00000   Min.   :0.0000   Length:1450       
 1st Qu.:3005.1   1st Qu.:0.00000   1st Qu.:0.0000   Class :character  
 Median :3345.3   Median :0.00000   Median :0.0000   Mode  :character  
 Mean   :3295.6   Mean   :0.08621   Mean   :0.1317                     
 3rd Qu.:3685.5   3rd Qu.:0.00000   3rd Qu.:0.0000                     
 Max.   :5131.4   Max.   :1.00000   Max.   :1.0000                     
```

We see that there are some missing values (`NA`) for certain variables. To remove these observations with missing values, we can use the `complete.cases` function:

```r
data<-NCbirths[complete.cases(NCbirths),]
summary(data)
```

```
       X                ID             Plural           Sex       
 Min.   :   1.0   Min.   :   1.0   Min.   :1.000   Min.   :1.000  
 1st Qu.: 363.0   1st Qu.: 363.0   1st Qu.:1.000   1st Qu.:1.000  
 Median : 726.0   Median : 726.0   Median :1.000   Median :1.000  
 Mean   : 725.8   Mean   : 725.8   Mean   :1.036   Mean   :1.489  
 3rd Qu.:1091.0   3rd Qu.:1091.0   3rd Qu.:1.000   3rd Qu.:2.000  
 Max.   :1450.0   Max.   :1450.0   Max.   :3.000   Max.   :2.000  
     MomAge          Weeks          Marital         RaceMom     
 Min.   :13.00   Min.   :22.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:22.00   1st Qu.:38.00   1st Qu.:1.000   1st Qu.:1.000  
 Median :26.00   Median :39.00   Median :1.000   Median :1.000  
 Mean   :26.79   Mean   :38.65   Mean   :1.345   Mean   :1.811  
 3rd Qu.:31.00   3rd Qu.:40.00   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :43.00   Max.   :45.00   Max.   :2.000   Max.   :8.000  
```

```
   HispMom              Gained          Smoke         BirthWeightOz  
 Length:1409        Min.   : 0.00   Min.   :0.0000   Min.   : 12.0  
 Class :character   1st Qu.:20.00   1st Qu.:0.0000   1st Qu.:106.0  
 Mode  :character   Median :30.00   Median :0.0000   Median :118.0  
                    Mean   :30.59   Mean   :0.1462   Mean   :116.4  
                    3rd Qu.:40.00   3rd Qu.:0.0000   3rd Qu.:130.0  
                    Max.   :95.00   Max.   :1.0000   Max.   :181.0  
 BirthWeightGm         Low              Premie          MomRace         
 Min.   : 340.2   Min.   :0.00000   Min.   :0.0000   Length:1409       
 1st Qu.:3005.1   1st Qu.:0.00000   1st Qu.:0.0000   Class :character  
 Median :3345.3   Median :0.00000   Median :0.0000   Mode  :character  
 Mean   :3301.1   Mean   :0.08446   Mean   :0.1285                     
 3rd Qu.:3685.5   3rd Qu.:0.00000   3rd Qu.:0.0000                     
 Max.   :5131.4   Max.   :1.00000   Max.   :1.0000                     
```

a) Consider a simple linear regression model, treating `BirthWeightGm` as the response variable, and `Sex` as the explanatory variable. Begin by fitting the model by directly including the sex variable exactly as is. Provide the estimated regression coefficients.

```r
lm1<-lm(BirthWeightGm~Sex,data=data)
summary(lm1)
```

```
Call:
lm(formula = BirthWeightGm ~ Sex, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2929.93  -293.38    42.92   387.02  1800.62 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3391.34      52.46  64.642   <2e-16 ***
Sex           -60.61      33.40  -1.814   0.0698 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 626.7 on 1407 degrees of freedom
Multiple R-squared:  0.002334,  Adjusted R-squared:  0.001625 
F-statistic: 3.292 on 1 and 1407 DF,  p-value: 0.06982
```

We obtain $\hat{\beta}_0 = 3391.3373186$ and $\hat{\beta}_1 =$ -60.6060686.

b) Based on the model in (a), what is the estimated birth weight (in grams) for male babies? for female babies?

Let Y denote `BirthWeightGm`. According to the model

$$E(Y|male) = E(Y|sex = 1) = \beta_0 + \beta_1$$
$$E(Y|female) = E(Y|sex = 2) = \beta_0 + 2\beta_1$$

Thus, $\widehat{E}(Y|male) = 3330.73$ and $\widehat{E}(Y|female) = 3270.13$

c) Fit the same model again, this time, however, treating the sex variable as categorical (use Sex=1 as the reference level). Provide the fitted model.

```r
lm2<-lm(BirthWeightGm~as.factor(Sex),data=data)
summary(lm2)
```

```
Call:
lm(formula = BirthWeightGm ~ as.factor(Sex), data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2929.93  -293.38    42.92   387.02  1800.62 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3330.73      23.36 142.598   <2e-16 ***
as.factor(Sex)2   -60.61      33.40  -1.814   0.0698 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 626.7 on 1407 degrees of freedom
Multiple R-squared:  0.002334,	Adjusted R-squared:  0.001625 
F-statistic: 3.292 on 1 and 1407 DF,  p-value: 0.06982
```

Let Y denote the birth weight (in grams), and let $X = 1\{sex = 2\}$ be the indicator variable taking the value 1 if $sex = 2$ and 0 otherwise. The fitted model is then

$$\hat{y} = 3330.73 - 60.61x$$

d) Based on the model in (c), what is the estimated birth weight (in grams) for male babies? for female babies?

$$\widehat{E}(Y|male) = \widehat{E}(Y|X = 1) = \hat{\beta}_0 = 3330.73$$
$$\widehat{E}(Y|female) = \widehat{E}(Y|X = 2) = \hat{\beta}_0 + \hat{\beta}_1 = 3330.73 - 60.61 = 3270.13$$

Notice that we obtain the exact same estimated means as those from the model in a)!

e) Based on the model in (c), is there a significant difference in the birth weights of female babies in comparison to male babies? Formally carry out a statistical test, indicating the underlying hypotheses, the value of the test statistic, the p-value and your conclusion.

From the model in (c), the difference $E(Y|male) - E(Y|female) = \beta_1$. Thus, here we're interested in testing $H_0 : \beta_1 = 0$ vs. $H_1 \beta_1 \neq 0$. From the output from the model

```r
summary(lm2)
```

```
Call:
lm(formula = BirthWeightGm ~ as.factor(Sex), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2929.93  -293.38    42.92   387.02  1800.62

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3330.73      23.36 142.598   <2e-16 ***
as.factor(Sex)2   -60.61      33.40  -1.814   0.0698 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 626.7 on 1407 degrees of freedom
Multiple R-squared:  0.002334,  Adjusted R-squared:  0.001625
F-statistic: 3.292 on 1 and 1407 DF,  p-value: 0.06982
```

We obtain a test statistic of $T = -1.814$ with a corresponding p-value of 0.0698. Using $\alpha = 5\%$, $p > \alpha$ and thus we fail to reject H_0. That is, there is not a significant difference in the mean birth weights of female babies in comparison to male babies.

f) How do the estimated regression coefficients (i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$) compare in models (a) and (c)? Explain any differences and/or similarities in a few sentences.

Recall that model a) is of the form $E(Y|sex) = \beta_0 + \beta_1 sex$ where $sex = 1 = male$ and $sex = 2 = female$. This model leads to

$$E(Y|male) = \beta_0 + \beta_1$$
$$E(Y|female) = \beta_0 + 2\beta_1$$

Model c), on the other hand, is of the form $E(Y|X) = \beta_0^* + \beta_1^* X$ where $X = 1\{sex = 1\}$. This model then leads to

$$E(Y|male) = \beta_0^*$$
$$E(Y|female) = \beta_0^* + \beta_1^*$$

In comparing the two models, we see

$$E(Y|female) - E(Y|male) = \beta_1 = \beta_1^*$$

And indeed, we see that in both models a) and c), $\hat{\beta}_1 = \hat{\beta}_1^* = -60.61$.

On the other hand, the intercepts β_0 and β_0^* differ. In particular, $\beta_0 + \beta_1 = \beta_0^*$ and indeed we see that $\hat{\beta}_0 + \hat{\beta}_1 = 3330.73$ and $\hat{\beta}_0^* = 3330.73$.

Question 8

a) In one to two sentences, explain why we cannot simply include the variable `RaceMom` in the model directly as is.

Including the variable directly as is would imply it is treated as a continuous covariate, thus forcing a linear trend among the levels. Here, however, `RaceMom` is a nominal categorical variable, where there is no ordering among the levels. Thus, treating it as a continuous covariate would be nonsensical.

b) Fit a linear regression model, treating `BirthWeightGm` as the response variable and `RaceMom` as the explanatory variable (using level 1 as the reference level). Provide the fitted model.

```r
summary(lm(BirthWeightGm~as.factor(RaceMom),data=data))
```

```
Call:
lm(formula = BirthWeightGm ~ as.factor(RaceMom), data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2897.18  -301.73    51.22   391.42  1780.57 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         3350.778     20.899 160.332  < 2e-16 ***
as.factor(RaceMom)2 -214.047     40.325  -5.308 1.29e-07 ***
as.factor(RaceMom)3  -81.507    134.189  -0.607    0.544    
as.factor(RaceMom)4  221.322    440.121   0.503    0.615    
as.factor(RaceMom)5    4.221     54.587   0.077    0.938    
as.factor(RaceMom)7 -600.828    622.074  -0.966    0.334    
as.factor(RaceMom)8   39.624    134.189   0.295    0.768    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 621.7 on 1402 degrees of freedom
Multiple R-squared:  0.02176,	Adjusted R-squared:  0.01757 
F-statistic: 5.198 on 6 and 1402 DF,  p-value: 2.663e-05
```

As before, let Y denote the birth weight variable, `BirthWeightGm`, and let $X_j = 1\{RaceMom = j\}$, $j = 1, 2, \ldots, 8$. The fitted model is then

$$\hat{y} = 3350.778 - 214.047X_2 - 81.507X_3 + 221.322X_4 + 4.221X_5 - 600.828X_7 + 39.624X_8$$

(Note that here, level 1 is the reference level, and there are no observations with $RaceMom = 6$ in the dataset with complete observations only).

c) Based on the model from (b), comment on the significance of the regression parameters in the context of the problem.

From the model results provided in b), p-values for tests $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ are provided. Since the model includes only the categorical variable RaceMom, each $\beta_j = E(Y|RaceMom = j) - E(Y|RaceMom = 1)$, i.e., each β_j represents the difference betweeen the mean birth weight for a mother of race j in comparison to race 1 (white), $j = 2, 3, 4, 5, 7, 8$. The p-value corresponding to β_2 is 1.29×10^{-07} which is less than any reasonable choice of α. This implies that there is a significant difference between the mean birthweight of babies of mothers of race 2 (black) in comparison to race 1 (white). All other β_j have large p-values, which are larger than any reasonable α (say $\alpha = 5\%$). Thus, we fail to reject $H_0 : \beta_j = 0$ for $j = 3, 4, 5, 7, 8$. Thus, we conclude that there is not a significant difference in the mean birthweight for babies of white mothers in comparison to each of American Indian, Chinese, Japanese, Filipino, and Other Asian or Pacific Islander mothers, respectively.

d) Now consider a modified version of the `RaceMom` variable which takes the following levels:

1: if RaceMom=1
2: if RaceMom=2
3: otherwise (i.e., RaceMom=3,4,5,6,7, or 8)

Fit a linear regression model, again using `BirthWeightGm` as the response variable, and this time including this modified version of the race variable (use level 2 as the reference level). Provide the fitted model and interpret the regression parameters.

The new variable can be created in different ways, here is one approach:

```
attach(data)
race<-as.numeric(RaceMom==1)+2*as.numeric(RaceMom==2)+3*as.numeric(RaceMom>2)
```

The model can then be fit:

```
race<-relevel(as.factor(race),2)
lm4<-lm(BirthWeightGm~race)
summary(lm4)
```

```
Call:
lm(formula = BirthWeightGm ~ race)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-2897.18  -301.73    51.22   391.42  1780.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3136.73      34.46  91.030  < 2e-16 ***
race1         214.05      40.29   5.312 1.26e-07 ***
race3         211.85      55.92   3.789 0.000158 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 621.2 on 1406 degrees of freedom
Multiple R-squared:  0.0206,    Adjusted R-squared:  0.01921
F-statistic: 14.79 on 2 and 1406 DF,  p-value: 4.409e-07
```

Let Y denote `BirthWeightGm`, and let R_j denote the newly created race variable. The fitted model is

$$\hat{y} = 3136.73 + 214.05R_1 + 211.85R_2$$

From this model,

- $\beta_0 = E(Y|race = 2)$. According to the fitted model, the estimated intercept is 3136.73, and thus babies of black mothers weigh 3136.73 g on average.
- $\beta_1 = E(Y|race = 1) - E(Y|race = 2)$. According to the fitted model, the estimated difference in the birth weights of babies of white mothers is, on average, 214.05 grams larger than the birth weights of babies of black mothers.
- $\beta_2 = E(Y|race = 3) - E(Y|race = 2)$. According to the fitted model, the estimated difference in the birth weights of babies of mothers of other races is, on average, 211.85 grams larger than that for black mothers.

e) Based on the model in part (d), consider all possible pairwise tests comparing the mean birth weights of babies for mothers of the different levels of the modified race variable. What can you conclude?

```r
library(emmeans)
comp<-emmeans(lm4,~race)
comp
```

```
 race emmean   SE   df lower.CL upper.CL
 2      3137 34.5 1406     3069     3204
 1      3351 20.9 1406     3310     3392
 3      3349 44.0 1406     3262     3435

Confidence level used: 0.95
```

```r
contrast(comp,method="pairwise",adjust="none")
```

```
 contrast      estimate   SE   df t.ratio p.value
 race2 - race1   -214.0 40.3 1406  -5.312  <.0001
 race2 - race3   -211.8 55.9 1406  -3.789  0.0002
 race1 - race3      2.2 48.7 1406   0.045  0.9640
```

The results above correspond to tests of the form $H_0 : \mu_j - \mu_k = 0$ vs. $H_1 : \mu_j - \mu_k \neq 0$ where $\mu_j = E(Y|race = j)$, for $j < k \in \{1, 2, 3\}$. The first two p-values correponding to comparisons of level 2 and 1 and levels 2 and 3 are both small (respectively given by < 0.0001 and 0.0002). Since the p-values are smaller than $\alpha = 5\%$ we can reject the underlying H_0 and conclude that there is a significant difference in the mean birth weight of babies for mothers of race black vs. race white, as well as between race black vs. race other. The last comparison, between levels 1 and 3, leads to a large p-value of 0.9640 and thus we fail to reject H_0. We can conclude that there is not a significant difference in the mean birth weight of babies for mothers of race white in comparison to other races.

Question 9

a) Provide the fitted model.

```r
lm5<-lm(BirthWeightGm~MomAge+Weeks+as.factor(Smoke)+as.factor(Marital))
summary(lm5)
```

```
Call:
lm(formula = BirthWeightGm ~ MomAge + Weeks + as.factor(Smoke) +
    as.factor(Marital))

Residuals:
     Min       1Q   Median       3Q      Max
-1836.26  -307.05    -5.25   324.18  1569.27

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2132.437    206.791 -10.312  < 2e-16 ***
MomAge                  9.204      2.460   3.742  0.00019 ***
Weeks                 135.690      5.000  27.136  < 2e-16 ***
as.factor(Smoke)1    -167.058     37.989  -4.398 1.18e-05 ***
as.factor(Marital)2   -95.630     31.766  -3.010  0.00265 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 496.8 on 1404 degrees of freedom
Multiple R-squared:  0.3744,    Adjusted R-squared:  0.3726
F-statistic:   210 on 4 and 1404 DF,  p-value: < 2.2e-16
```

Again, letting Y denote BirthWeightGm, the fitted model is

$$\hat{y} = -2132.437 + 9.204MomAge + 135.690Weeks - 167.058X_{Smoke} - 95.6301X_{Marital}$$

where $X_{Smoke} = 1\{Smoke = 1\}$ and $X_{Marital} = 1\{Marital = 2\}$

b) Comment on the value of R^2.

From the output in part a), $R^2 = 0.3744$ and thus we can say that 37.44% of the variability in Y is explained by the model.

c) Interpret the regression coefficients (i.e. the β_j) associated with the `MomAge` and `Smoke` variables.

- `MomAge`: $\hat{\beta}_1 = 9.204$. Thus, for every one year increase in the mother's age, the baby's birth weight will increase, on average, by 9.204 grams, holding all other variables fixed.
- `Smoke`: $\hat{\beta}_3 = -167.058$. Thus, on average, the birth weight of babies with smoker mothers is 167.058 grams lower in comparison to non-smoker mothers, holding all other variables constant.

d) Formally carry out a statistical test to verify whether `Marital` is significant in the model. Be sure to provide a conclusion in the context of the problem.

Here we're interested in testing $H_0 : \beta_4 = 0$ vs. $\beta_4 \neq 0$ where β_4 is the coefficient corresponding to the `Marital` variable. The test statistic is $T = -3.010$ with corresponding p-value 0.00265. Using $\alpha = 5\%$, $p < \alpha$ and thus we can reject H_0 and conclude that β_4 is significantly different from 0. Thus, there is a significant difference in the mean birth weights of babies of married mothers in comparison to unmarried mothers, even after adjusting for mother's age, weeks of gestation and smoker status.

e) Carry out a residual analysis. Comment on the results.

```r
res<-rstudent(lm5)
fitted<-lm5$fitted.values
data.res<-cbind(data,res,fitted)

library(cowplot)
library(ggplot2)

plot1<-ggplot(mapping = aes(x = res)) +
  geom_density() +
  geom_histogram(aes(y = ..density..), bins = 20, alpha = 0.5) +
  xlab("residuals")

plot2<-ggplot(mapping = aes(sample = res)) +
  stat_qq(distribution = qt, dparams = lm5$df.residua) +
  stat_qq_line(distribution = qt, dparams = lm5$df.residual) +
  labs(x = "theoretical quantiles",
       y = "empirical quantiles") +
  ggtitle("QQ-Plot Studentized Residuals")
```

```r
plot3<-ggplot(data=data.res,
       aes(x = fitted, y = res)) +
  geom_point() +
  geom_smooth() +
  theme(legend.position = "bottom") +
  ylab("residuals") +
  xlab("fitted values")

plot4<-ggplot(data=data.res,
       aes(x = MomAge, y = res)) +
  geom_point() +
  geom_smooth() +
  theme(legend.position = "bottom") +
  ylab("residuals") +
  xlab("Mom Age")

plot5<-ggplot(data=data.res,
       aes(x = Weeks, y = res)) +
  geom_point() +
  geom_smooth() +
  theme(legend.position = "bottom") +
  ylab("residuals") +
  xlab("Weeks")

plot6<-ggplot(data.res, aes(x=as.factor(Sex), y=res)) +
  geom_boxplot() +
  labs(title="Residuals",x="sex", y = "residuals")
tapply(data.res$res,as.factor(data.res$Sex),function(x) c(mean(x),var(x)) )
```

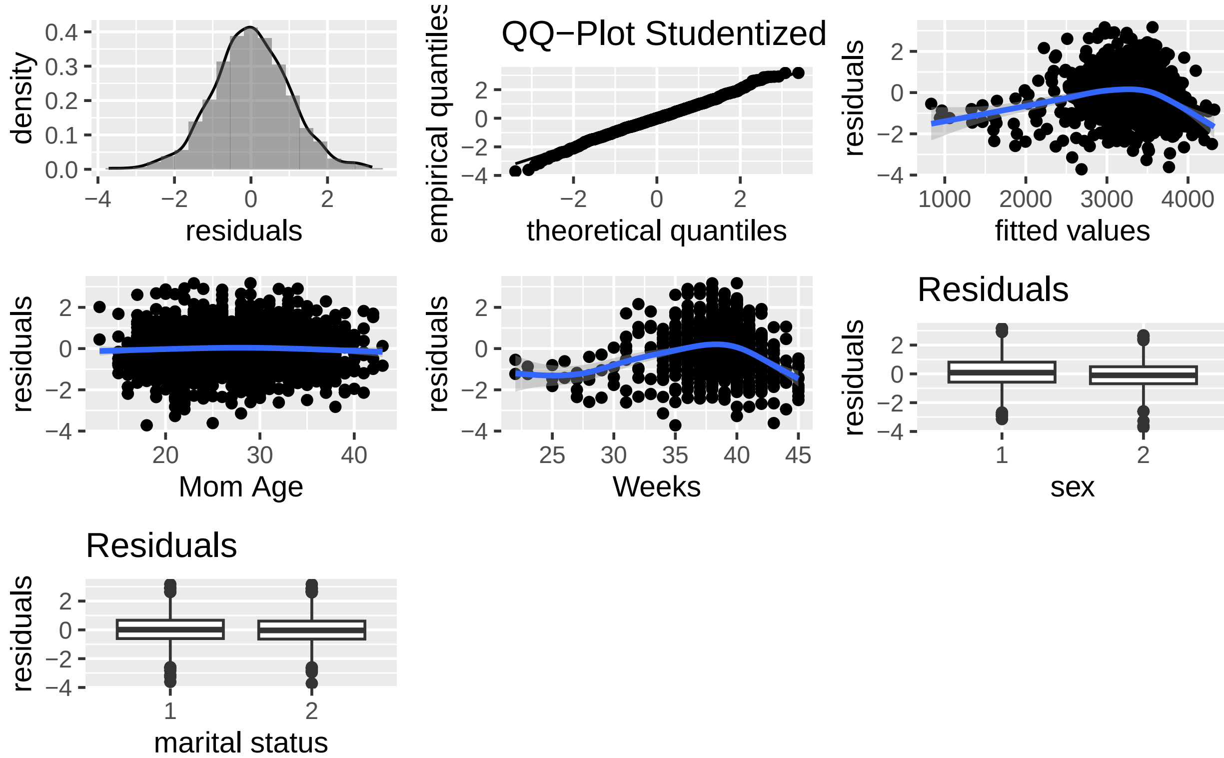
```
$`1`
[1] 0.09969538 1.11058004

$`2`
[1] -0.1047594  0.8705510
```

```r
plot7<-ggplot(data.res, aes(x=as.factor(Marital), y=res)) +
  geom_boxplot() +
  labs(title="Residuals",x="marital status", y = "residuals")
tapply(data.res$res,as.factor(data.res$Marital),function(x) c(mean(x),var(x)) )
```

```
$`1`
[1] -0.0002988021  0.9942968906

$`2`
[1] -0.000252549  1.021488509
```

```r
plot_grid(plot1, plot2, plot3, plot4, plot5, plot6, plot7)
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The histogram and QQ plot suggest that the normality assumption is reasonable: the histogram shows a bell shaped curve symmetric about 0 and the points in the QQ plot align relatively well along the diagonal line (suggesting a one-to-one relation between the theoretical quantiles and the empirical quantiles). The plot of the residuals vs. fitted values seems to show a funnel shape, suggesting there is heteroscedasticity. However, there are fewer observations below the value $\hat{y} = 2500$. Focusing on the bulk of the data (between $\hat{y} \in [2500, 4000]$, there is not really any trend. The plot of `Weeks` shows a similar funnel shape, as seen in the plot with the fitted values. The other plots and summaries do not show any issues or patterns, suggesting that the model is well specified. Overall, there could arguably be some evidence of heteroscedasticity here.

Question 10

We first read the data:

```
data<-read.csv("GrinnellHouses_mod.csv")
head(data)
```

```
  X  Date          Address Bedrooms Baths SquareFeet   LotSize YearBuilt
1 1 16880 1020 Center St        3  1.00       1224 0.1721763      1900
2 2 16667     503 2nd Ave        3  1.00       1277 0.2066116      1900
3 3 16583   9090 Clay St        3  1.00       1079 0.1993572      1900
4 4 16700     320 Park St        3  2.00        912 0.2180000      1900
5 5 16702   1014 Pearl St        3  2.00       1488 0.1700000      1900
6 6 16877        501 High        4  1.75       2160 0.3126722      1880
  YearSold MonthSold DaySold CostPerSqFt OrigPrice ListPrice SalePrice SPLPPct
1     2006         3      20       22.06     35000     35000     27000   77.14
```

```
2      2005          8       19        24.08       35000       35000       30750    87.86
3      2005          5       27        38.92       45900       45900       42000    91.50
4      2005          9       21        54.82       59900       52500       50000    95.24
5      2005          9       23        33.60       50000       50000       50000   100.00
6      2006          3       17        25.00       71500       71500       54000    75.52
  winter old mid new
1      0   1   0   0
2      0   1   0   0
3      0   1   0   0
4      0   1   0   0
5      0   1   0   0
6      0   1   0   0
```

a) Model the sale price of a house in terms of the square footage, number of bathrooms, number of bedrooms, winter indicator, and age of the home (categorized as old/mid/new), including an interaction between the number of bedrooms and the winter indicator variable, as well as an interaction between the number of bedrooms and the age of the home (categorized as old/mid/new). Use level new as the reference level for the age of the home. Provide the model summary results directly obtained in R.

The linear regression model including SquareFeet, Baths, Bedrooms, winter, and the categorized age (old/mid/new), as well as the interaction between bedrooms and winter, and the interaction between bedrooms and age (old/mid/new):

```r
mod1<-lm(SalePrice~SquareFeet+Baths+Bedrooms*(winter)+Bedrooms*(old+mid),data=data)
summary(mod1)
```

```
Call:
lm(formula = SalePrice ~ SquareFeet + Baths + Bedrooms * (winter) +
    Bedrooms * (old + mid), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-238206  -23430    -795   20926  209653

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2016.923  16440.709   0.123 0.902396
SquareFeet          54.269      3.504  15.489  < 2e-16 ***
Baths            23246.083   3367.997   6.902 1.12e-11 ***
Bedrooms         16761.402   4789.601   3.500 0.000495 ***
winter          -41955.452  14172.592  -2.960 0.003174 **
old               4077.030  18062.280   0.226 0.821483
mid              -7486.937  19016.600  -0.394 0.693915
Bedrooms:winter  12506.082   4198.388   2.979 0.002991 **
Bedrooms:old    -26555.181   5036.311  -5.273 1.78e-07 ***
Bedrooms:mid    -10028.711   5448.894  -1.841 0.066105 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43650 on 722 degrees of freedom
```

```
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.697
F-statistic: 187.8 on 9 and 722 DF,  p-value: < 2.2e-16
```

b) Based on the model in part (a), write an expression for the fitted models for each category of home age, that is, for old homes, mid-aged homes, and new homes, respectively.

Let Y denote the sale price of the house, X_1 denote the square feet, X_2 the number of bathrooms, X_3 the number of bedrooms, X_4 the winter indicator, X_5 the indicator for an old home and X_6 the indicator for a mid-age home.

The model has the form

$$E(Y|X_1,\dots,X_6) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_4 X_3 + \beta_8 X_5 X_3 + \beta_9 X_6 X_3$$

- For old homes, the fitted model is

$$\begin{aligned}\widehat{E}(Y|X_1,X_2,X_3,X_4,X_5=1,X_6=0) &= (\hat\beta_0+\hat\beta_5) + \hat\beta_1 X_1 + \hat\beta_2 X_2 + \hat\beta_4 X_4 + \hat\beta_7 X_4 X_3 + (\hat\beta_3+\hat\beta_8) X_3 \\ &= 6093.953 + 54.269 X_1 + 23246.083 X_2 - 41955.452 X_4 + 12506.082 X_4 X_3 - 9793.779 X_3\end{aligned}$$

- For mid-aged homes, the fitted model is

$$\begin{aligned}\widehat{E}(Y|X_1,X_2,X_3,X_4,X_5=0,X_6=1) &= (\hat\beta_0+\hat\beta_6) + \hat\beta_1 X_1 + \hat\beta_2 X_2 + \hat\beta_4 X_4 + \hat\beta_7 X_4 X_3 + (\hat\beta_3+\hat\beta_9) X_3 \\ &= -5470.014 + 54.269 X_1 + 23246.083 X_2 - 41955.452 X_4 + 12506.082 X_4 X_3 + 6732.691 X_3\end{aligned}$$

- For new homes, the fitted model is

$$\begin{aligned}\widehat{E}(Y|X_1,X_2,X_3,X_4,X_5=0,X_6=0) &= \hat\beta_0 + \hat\beta_1 X_1 + \hat\beta_2 X_2 + \hat\beta_4 X_4 + \hat\beta_7 X_4 X_3 + \hat\beta_3 X_3 \\ &= 2016.923 + 54.269 X_1 + 23246.083 X_2 - 41955.452 X_4 + 12506.082 X_4 X_3 + 16761.402 X_3\end{aligned}$$

c) Based on the model in part (a), what is the estimated effect of the number of bedrooms on the sale price of an old home when the sale occurs in a winter month?

For an old home sold in a winter month, we have that

$$E(Y|X_1,X_2,X_3,X_4=1,X_5=1,X_6=0) = (\beta_0+\beta_4+\beta_5) + \beta_1 X_1 + \beta_2 X_2 + (\beta_3+\beta_7+\beta_8) X_3$$

And thus,

$$E(Y|X_1,X_2,X_3=x+1,X_4=1,X_5=1,X_6=0) - E(Y|X_1,X_2,X_3=x,X_4=1,X_5=1,X_6=0) = \beta_3+\beta_7+\beta_8$$

Which, according to the fitted model, is estimated as

$$16761.402 + 12506.082 - 26555.181 = 2712.303$$

d) Based on the model in part (a), interpret the regression coefficient associated with `SquareFeet` and the main effect of `Bedrooms`.

- `SquareFeet` (β_1): for every additional square foot, the sale price of the home will increase, on average, by \$54.27, when all other variables remain constant.
- main effect of `Bedrooms` (β_3): for a **new** home ($X_5 = X_6 = 0$) sold in a **non-winter** month ($X_4 = 0$), for every additional bedroom, the sale price of the home will increase on average by \$16761.40, when all other variables (square feet, number of bathrooms) remain fixed.

e) Based on the model in part (a), does the effect of the number of bedrooms on the sale price of the home depend on whether it was sold in a winter month? Justify your answer.

Here we are interested in testing whether the interaction between the winter month indicator and bedrooms variable is significant. That is, according to our model, we're interested in testing

$$H_0 : \beta_7 = 0 \quad \text{vs.} \quad H_1 : \beta_7 \neq 0$$

```r
summary(mod1)
```

```

Call:
lm(formula = SalePrice ~ SquareFeet + Baths + Bedrooms * (winter) +
    Bedrooms * (old + mid), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-238206  -23430    -795   20926  209653

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2016.923  16440.709   0.123 0.902396
SquareFeet          54.269      3.504  15.489  < 2e-16 ***
Baths            23246.083   3367.997   6.902 1.12e-11 ***
Bedrooms         16761.402   4789.601   3.500 0.000495 ***
winter          -41955.452  14172.592  -2.960 0.003174 **
old               4077.030  18062.280   0.226 0.821483
mid              -7486.937  19016.600  -0.394 0.693915
Bedrooms:winter  12506.082   4198.388   2.979 0.002991 **
Bedrooms:old    -26555.181   5036.311  -5.273 1.78e-07 ***
Bedrooms:mid    -10028.711   5448.894  -1.841 0.066105 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43650 on 722 degrees of freedom
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.697
F-statistic: 187.8 on 9 and 722 DF,  p-value: < 2.2e-16
```

From the model output, we see that the test statistic value is 2.979 with corresponding p-value 0.002991. Since the p-value is < 0.05, we can reject H_0 at the $\alpha = 5\%$ significance level and conclude that there is indeed a significant interaction between the number of bedrooms and the sale occurring in a winter month. That is, the effect of the number of bedrooms on the sale price of the house does significantly indeed depend on whether the sale occurred in a winter month.

f) Based on the model in part (a), is there a significant difference in the effect of the number of bedrooms on the sale price for mid-aged homes in comparison to new homes? Justify your answer.

According to the model, the difference in the effect of the number of bedrooms on the sale price for mid-aged homes in comparison to new homes is given by β_9 since for a new home

$$E(Y|X_1, X_2, X_3 = x+1, X_4, X_5 = 0, X_6 = 0) - E(Y|X_1, X_2, X_3 = x, X_4, X_5 = 0, X_6 = 0) = \beta_3$$

and for a mid-aged home

$$E(Y|X_1, X_2, X_3 = x+1, X_4, X_5 = 0, X_6 = 1) - E(Y|X_1, X_2, X_3 = x, X_4, X_5 = 0, X_6 = 1) = \beta_3 + \beta_9$$

Thus the difference is β_9. So we're interested in testing

$$H_0 : \beta_9 = 0 \quad \text{vs.} \quad H_1 : \beta_9 \neq 0$$

From the model output

```r
summary(mod1)
```

```
Call:
lm(formula = SalePrice ~ SquareFeet + Baths + Bedrooms * (winter) +
    Bedrooms * (old + mid), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-238206  -23430    -795   20926  209653

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2016.923  16440.709   0.123 0.902396
SquareFeet          54.269      3.504  15.489  < 2e-16 ***
Baths            23246.083   3367.997   6.902 1.12e-11 ***
Bedrooms         16761.402   4789.601   3.500 0.000495 ***
winter          -41955.452  14172.592  -2.960 0.003174 **
old               4077.030  18062.280   0.226 0.821483
mid              -7486.937  19016.600  -0.394 0.693915
Bedrooms:winter  12506.082   4198.388   2.979 0.002991 **
Bedrooms:old    -26555.181   5036.311  -5.273 1.78e-07 ***
Bedrooms:mid    -10028.711   5448.894  -1.841 0.066105 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43650 on 722 degrees of freedom
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.697
F-statistic: 187.8 on 9 and 722 DF,  p-value: < 2.2e-16
```

we see that the test statistic is given by -1.841 with p-value 0.066105. Using $\alpha = 5\%$, the p-value is larger than 0.05 and thus we fail to reject H_0. We can thus conclude that there is not a significant difference in the effect of the number of bedrooms on the sale price of the home for mid-aged homes in comparison to new homes.

g) Based on the model in part (a), does the effect of the number of bedrooms on the sale price of the home depend on the age of the home (categorized as old/mid/new)? Justify your answer.

Here we're interested in testing whether there's a significant interaction between the age of the home (categorized as old/mid/new) and the number of bedrooms. Since the age of the home is categorical with 3 levels, this test actually involves several parameters:

$$H_0 : \beta_8 = \beta_9 = 0 \quad \text{vs} \quad H_1 : \text{ at least one of } \beta_8 \text{ or } \beta_9 \neq 0$$

This test can be done using an F-test, where the complete model is the model considered in a), and the reduced model is that with the interaction between the age of the home (old/mid/new) and number of bedrooms is removed, i.e., where $\beta_8 = \beta_9 = 0$.

```
mod2<-lm(SalePrice~SquareFeet+Baths+Bedrooms*winter+old+mid,data=data)
summary(mod2)
```

```
Call:
lm(formula = SalePrice ~ SquareFeet + Baths + Bedrooms * winter +
    old + mid, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-256803  -23957   -1205   21944  199258

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      61378.880   9083.048   6.758 2.89e-11 ***
SquareFeet          52.754      3.572  14.771  < 2e-16 ***
Baths            25357.093   3401.886   7.454 2.59e-13 ***
Bedrooms          -530.207   2438.742  -0.217  0.82795
winter          -45384.082  14467.015  -3.137  0.00178 **
old             -86518.155   5454.375 -15.862  < 2e-16 ***
mid             -45577.680   5198.440  -8.768  < 2e-16 ***
Bedrooms:winter  13440.432   4286.267   3.136  0.00178 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44610 on 724 degrees of freedom
Multiple R-squared:  0.6865,    Adjusted R-squared:  0.6834
F-statistic: 226.4 on 7 and 724 DF,  p-value: < 2.2e-16
```

```
anova(mod1,mod2)
```

```
Analysis of Variance Table

Model 1: SalePrice ~ SquareFeet + Baths + Bedrooms * (winter) + Bedrooms *
    (old + mid)
Model 2: SalePrice ~ SquareFeet + Baths + Bedrooms * winter + old + mid
  Res.Df        RSS Df   Sum of Sq      F   Pr(>F)
1    722 1.3754e+12
2    724 1.4410e+12 -2 -6.5589e+10 17.215 4.97e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the `anova` function, we see that the test statistic is $F = 17.215$ with corresponding p-value 4.97×10^{-8}. Since the p-value is very small (smaller than any reasonable α), we can reject H_0 and conclude that there is indeed a significant interaction between the number of bedrooms and the age of the home (old/mid/new). That is, the effect of the number of bedrooms on the sale price of the home does indeed depend on the age of the home.

h) Test whether the sale occurring in a winter month (i.e., the variable `winter`) is globally significant in the model, using an F-test. Justify your answer. Use $\alpha = 1\%$.

Here we're interested in assessing whether the winter month is globally significant, thus we're interested in testing

$$H_0 : \beta_4 = \beta_7 = 0 \quad \text{vs.} \quad H_1 : \text{ at least one of } \beta_4 \text{ or } \beta_7 \neq 0$$

This test can be done using an F-test, where the complete model is the model considered in a), and the reduced model is that with the winter month variable removed from the model, i.e., where $\beta_4 = \beta_7 = 0$.

```r
mod3<-lm(SalePrice~SquareFeet+Baths+Bedrooms+Bedrooms*(old+mid),data=data)
summary(mod3)
```

```
Call:
lm(formula = SalePrice ~ SquareFeet + Baths + Bedrooms + Bedrooms *
    (old + mid), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-242548  -22468    -577   20757  211329

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9618.423  16011.815  -0.601   0.5482
SquareFeet       55.064      3.508  15.698  < 2e-16 ***
Baths         22519.444   3375.223   6.672 5.02e-11 ***
Bedrooms      20271.561   4666.278   4.344 1.60e-05 ***
old            6166.944  18123.829   0.340   0.7338
mid           -5296.521  19086.835  -0.277   0.7815
Bedrooms:old -27301.050   5053.934  -5.402 8.95e-08 ***
Bedrooms:mid -10718.339   5470.097  -1.959   0.0504 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 43860 on 724 degrees of freedom
Multiple R-squared:  0.697, Adjusted R-squared:  0.6941
F-statistic: 237.9 on 7 and 724 DF,  p-value: < 2.2e-16
```

```r
anova(mod1,mod3)
```

```
Analysis of Variance Table

Model 1: SalePrice ~ SquareFeet + Baths + Bedrooms * (winter) + Bedrooms *
    (old + mid)
Model 2: SalePrice ~ SquareFeet + Baths + Bedrooms + Bedrooms * (old +
    mid)
  Res.Df        RSS Df   Sum of Sq      F  Pr(>F)
1    722 1.3754e+12
2    724 1.3925e+12 -2 -1.7122e+10 4.4941 0.01149 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the `anova` function, we see that the test statistic is $F = 4.4941$ with corresponding p-value 0.01149. Since the p-value is larger than $\alpha = 1\%$, we fail to reject H_0. We can thus conclude that the winter month variable is not globally significant in the model.

Question 11

a) Fit a linear regression model to the data allowing to model the sale price in terms of a quadratic relationship with square footage, in interaction with the age of the home (categorized as old/mid/new). (That is, the model should include `SquareFeet` + `SquareFeet`^{2}, in interaction with the categorized home age, with no other variables in the model). Use the new category as the reference level. Provide the model summary results directly obtained in R.

Here we're interested in the model

$$E(Y|X_1, X_5, X_6) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta 3 X_5 + \beta_4 X_6 + \beta_5 X_1 X_5 + \beta_6 X_1 X_6 + \beta_7 X_1^2 X_5 + \beta_8 X_1^2 X_6$$

```r
mod4<-lm(SalePrice~(SquareFeet+I(SquareFeet^2))*(old+mid),data=data)
summary(mod4)
```

```
Call:
lm(formula = SalePrice ~ (SquareFeet + I(SquareFeet^2)) * (old +
    mid), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-180487  -25437      -4   21722  215802

Coefficients:
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.842e+04  2.507e+04  -0.735 0.462689
SquareFeet           1.534e+02  2.179e+01   7.039  4.5e-12 ***
I(SquareFeet^2)     -1.203e-02  4.124e-03  -2.918 0.003637 **
old                  1.295e+04  2.789e+04   0.464 0.642582
mid                  1.359e+04  3.216e+04   0.423 0.672759
SquareFeet:old      -8.606e+01  2.422e+01  -3.554 0.000404 ***
SquareFeet:mid      -3.844e+01  3.235e+01  -1.188 0.235090
I(SquareFeet^2):old  1.020e-02  4.574e-03   2.230 0.026038 *
I(SquareFeet^2):mid  7.437e-04  7.657e-03   0.097 0.922660
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45550 on 723 degrees of freedom
Multiple R-squared:  0.6736,    Adjusted R-squared:   0.67
F-statistic: 186.5 on 8 and 723 DF,  p-value: < 2.2e-16
```

b) From the model in part a), provide an expression for the fitted model for each category of home age, that is, for old homes, mid-aged homes and new homes, respectively.

Based on the model,

- for new homes:
  $$E(Y|X_1, X_5 = 0, X_6 = 0) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$
  which is estimated as
  $$-18420 + 153.4X_1 - 0.01203X_1^2$$
- for mid-aged homes:
  $$E(Y|X_1, X_5 = 0, X_6 = 1) = (\beta_0 + \beta_4) + (\beta_1 + \beta_6)X_1 + (\beta_2 + \beta_8)X_1^2$$
  which is estimated as
  $$-4830 + 114.96X_1 - 0.0112863X_1^2$$
- for old homes:
  $$E(Y|X_1, X_5 = 1, X_6 = 0) = (\beta_0 + \beta 3) + (\beta_1 + \beta_5)X_1 + (\beta_2 + \beta_7)X_1^2$$
  which is estimated as
  $$-5470 + 67.34X_1 - 0.00183X_1^2$$

c) What is the estimated difference in the mean sale price of a home with 2000 square feet in comparison to a home with 1000 square feet, for a house built after 1980? What about for a home built between 1950 and 1980? And a home built before 1950? Be sure to show your work!

Based on the answer in part b),

- for a new home:

$$\left\{-18420 + 153.4(2000) - 0.01203(2000^2)\right\} - \left\{-18420 + 153.4(1000) - 0.01203(1000^2)\right\} = 117310$$

- for a mid-aged home:

$$\left\{-4830 + 114.96(2000) - 0.0112863(2000^2)\right\} - \left\{-4830 + 114.96(1000) - 0.0112863(1000^2)\right\} = 81101.1$$

- for an old home:

$$\left\{-5470 + 67.34(2000) - 0.00183(2000^2)\right\} - \left\{-5470 + 67.34(1000) - 0.00183(1000^2)\right\} = 61850$$

d) Now fit a model using the log-transformed sale price (i.e. $\ln($`SalePrice`$)$) as the response variable, and as covariates the square footage (no quadratic term this time), categorized home age, and their interaction. Again, use new as the reference level for the home age. Provide an expression for the fitted model.

```r
mod5<-lm(log(SalePrice)~(SquareFeet)*(old+mid),data=data)
summary(mod5)
```

```
Call:
lm(formula = log(SalePrice) ~ (SquareFeet) * (old + mid), data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.89627 -0.17803  0.07252  0.25571  0.89157 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.146e+01  9.149e-02 125.227  < 2e-16 ***
SquareFeet      3.928e-04  4.236e-05   9.271  < 2e-16 ***
old            -9.243e-01  1.110e-01  -8.328 4.09e-16 ***
mid            -4.624e-01  1.172e-01  -3.946 8.73e-05 ***
SquareFeet:old  9.873e-05  5.485e-05   1.800   0.0722 .  
SquareFeet:mid  1.236e-04  6.547e-05   1.888   0.0594 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4079 on 726 degrees of freedom
Multiple R-squared:  0.529, Adjusted R-squared:  0.5257 
F-statistic: 163.1 on 5 and 726 DF,  p-value: < 2.2e-16
```

We can write the fitted model as

$$\widehat{E}\{\ln(Y)|X_1, X_5, X_6\} = 11.46 + 0.0003928X_1 - 0.9243X_5 - 0.4624X_6 + (9.873 \times 10^{-5})X_1X_5 + 0.0001236X_1X_6$$

e) Based on the model in part (d), interpret the coefficient corresponding to the main effect of `SquareFeet`.

For the linear regression model on the log-transformed response here,

$$E(Y|X_1, X_5, X_6) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_5 + \beta_3 X_6 + \beta_4 X_1 X_5 + \beta_5 X_1 X_6 + \sigma^2/2)$$

and thus the main effect for square feet, i.e., β_1 can be interpreted on the exponential scale as

$$\exp(\beta_1) = \frac{E(Y|X_1 = x+1, X_5 = 0, X_6 = 0)}{E(Y|X_1 = x, X_5 = 0, X_6 = 0)}$$

Here, $\exp(\hat{\beta}_1) = 1.000393$. Thus, for every additional square foot, the average sale price of a **new** home is **multiplied** by a factor of 1.000393.