Jiayuan Shi

MA684 HW4

1A.

From these formula,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | df | SS | MS | F |
| Model  Residual | k  n-k-1 | 7068.3  12580.0 | SSM/dfM  SSE/dfE | MSM/MSE |
| Total | n-1 | SSM+SSE | SST/dfT |  |

we can get the ANOVA for the regression:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | df | SS | MS | F |
| Model  Residual | 2  22 | 7068.3  12580.0 | 3534.15  571.8182 | 6.1805 |
| Total | 24 | 19648.3 | 818.6792 |  |

* R2 = SSModel / SSTotal = 7068.3/19648.3 = 0.359741

Interpretation: the proportion of variability in the selling price explained by the model is 35.9741%

* F-statistic = 6.1805, p-value = 0.0074

1B.

* If all other variables in the model are held constant, for each increase in the number of bedrooms, on average, the selling price expected to increase 20.32 thousands of dollars.

The interpretation of a slope in a simple regression does not need to have the condition of keeping all other variables in the model are held constant.

* 0.0734

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Parameter  Estimate | Standard  Error | t-statistic | p-value | 95% CI |
| Intercept  Number bedrooms  Age | 73.13  20.32  -1.24 | ---  7.76  0.66 | 2.62  2.88 | 0.016  0.0087 | 4.23, 36.41 |

So 95% confidence interval for the slope for age is

1C.

|  |  |  |
| --- | --- | --- |
| Variable | Mean | sd |
| Selling price  Number bedrooms  Age | 132.76  3.64  11.56 | 28.61  0.64  7.51 |

Based on both the standardized slopes, of the selling price, number of bedrooms is a stronger predictor than the age of home.

2.

> qolstudy <- read.csv("qolstudy.csv", header=TRUE)

> attach(qolstudy)

> reg <- lm(qol~age+sexfemale+educyears+income+married)

> summary(reg)

Call:

lm(formula = qol ~ age + sexfemale + educyears + income + married)

Residuals:

Min 1Q Median 3Q Max

-200.929 -46.459 3.031 53.155 157.333

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 44.64727 47.57475 0.938 0.3504

age 1.20965 0.57248 2.113 0.0373 \*

sexfemale 34.74903 15.28946 2.273 0.0253 \*

educyears 3.49619 2.85490 1.225 0.2238

income 0.09005 0.41617 0.216 0.8292

married 75.21971 17.42707 4.316 3.92e-05 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.87 on 94 degrees of freedom

Multiple R-squared: 0.2585, Adjusted R-squared: 0.219

F-statistic: 6.553 on 5 and 94 DF, p-value: 2.85e-05

2A. R2 = 0.2585. The proportion of variability in the quality of life explained by the model is 25.85%

F-statistic = 6.553, df = (5, 94), p-value = 2.85e-05 < 0.05

We can reject H0, which is and we can conclude that there is significant association between the quality of life and any of the independent variables.

2B. In this regression model, age, sex and marital status are significantly associated with quality of life.

2C. If all other variables in the model are held constant, for each year increase in the age, on average, the quality of life expected to increase 1.20965 units.

If all other variables in the model are held constant, compared to people who are unmarried, we would expect people who are married to be 75.21971 units, on average.

> confint(reg,"age")

2.5 % 97.5 %

age 0.07297716 2.346325

The confidence interval for the slope of age in this model is (0.07297716, 2.346325)

2D.

> lm.beta(reg)

age sexfemale educyears income

0.1958629 0.2054506 0.1166457 0.0199322

married

0.4006507

Based on all the standardized slopes, marital status is the strongest predictor of the quality of life.

2E.

> predict(reg,data.frame(age=30,sexfemale=0,married=0,educyears=16,income=50),interval='prediction')

fit lwr upr

1 141.3781 -13.65405 296.4102

A prediction interval for this quality of life is (-13.65405, 296.4102).

> predict(reg,data.frame(age=60,sexfemale=1,married=1,educyears=16,income=50),interval='confidence')

fit lwr upr

1 287.6363 249.0625 326.2102

A confidence interval for the mean quality of life for all such women id (249.0625, 326.2102)

3A.

> t.test(x = SYSBP[CURSMOKE==c(0)], y = SYSBP[CURSMOKE==c(1)],alternative = c("two.sided"),

+ var.equal = TRUE,conf.level = 0.95)

Two Sample t-test

data: SYSBP[CURSMOKE == c(0)] and SYSBP[CURSMOKE == c(1)]

t = 4.3037, df = 198, p-value = 2.639e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

7.152078 19.249942

sample estimates:

mean of x mean of y

140.3556 127.1545

> sum(CURSMOKE==c(0))

[1] 90

> sum(CURSMOKE==c(1))

[1] 110

> sd(SYSBP[CURSMOKE==c(0)])

[1] 24.96562

> sd(SYSBP[CURSMOKE==c(1)])

[1] 18.36028

Systolic blood pressure in smokers and non-smokers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Non-smokers | Smokers | t-value | p-value |
| n  mean  standard deviation | 90  140.3556  24.96562 | 110  127.1545  18.36028 | 4.3037 | 2.639e-05 |

Because the p-value is less than 0.05, we can reject H0, which is mean blood pressure of smokers and non-smokers is the same. So we can conclude that there is a significant difference that between blood pressure of smokers and that of non-smokers.

3B.

> reg <- lm(SYSBP~CURSMOKE)

> summary(reg)

Call:

lm(formula = SYSBP ~ CURSMOKE)

Residuals:

Min 1Q Median 3Q Max

-44.356 -15.155 -2.905 10.845 72.845

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 140.356 2.275 61.699 < 2e-16 \*\*\*

CURSMOKE -13.201 3.067 -4.304 2.64e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.58 on 198 degrees of freedom

Multiple R-squared: 0.08554, Adjusted R-squared: 0.08092

F-statistic: 18.52 on 1 and 198 DF, p-value: 2.639e-05

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Slope | Std Error | t-value | p-value |
| Intercept  Current Smoking | 140.356  -13.201 | 2.275  3.067 | 61.699  -4.304 | < 2e-16  2.64e-05 |

This analysis shows a significant association between smoking and systolic blood pressure, because the p-value is less than 0.05, and we can reject the null hypothesis that there is no association between smoking and systolic blood pressure.

The t-statistic from this analysis is 4.304(two-sided), is the same as the t-statistic 4.3037 in from 3A above.

For a non-smoker, predicted systolic blood pressures = 140.356 + 0\*-13.201= 140.356

For a smoker, predicted systolic blood pressures = 140.356 + 1\*-13.201= 127.155

These predicted values are same as the mean blood pressures in 3A.

3C.

> reg <- lm(SYSBP~AGE+SEX+BMI+CURSMOKE)

> summary(reg)

Call:

lm(formula = SYSBP ~ AGE + SEX + BMI + CURSMOKE)

Residuals:

Min 1Q Median 3Q Max

-41.095 -13.273 -3.545 11.751 72.088

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 50.8388 12.2248 4.159 4.79e-05 \*\*\*

AGE 0.8534 0.1598 5.340 2.57e-07 \*\*\*

SEX 4.6354 2.8737 1.613 0.108

BMI 1.3282 0.3024 4.393 1.83e-05 \*\*\*

CURSMOKE -4.0980 3.0048 -1.364 0.174

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.1 on 195 degrees of freedom

Multiple R-squared: 0.2945, Adjusted R-squared: 0.28

F-statistic: 20.35 on 4 and 195 DF, p-value: 5.028e-14

Regression predicting systolic blood pressure from smoking status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Slope | Std Error | t-value | p-value |
| Intercept  Age  Sex  BMI  Current Smoking | 50.8388  0.8534  4.6354  1.3282  -4.0980 | 12.2248  0.1598  2.8737  0.3024  3.0048 | 4.159  5.340  1.613  4.393  -1.364 | 4.79e-05  2.57e-07  0.108  1.83e-05  0.174 |

Based on this analysis, there is not a significant association between current smoking and systolic blood pressure, because the p-value is bigger than 0.05, and we cannot reject the null hypothesis that there is no association between smoking and systolic blood pressure.

3D. The slope and p-value for Current Smoking are different in the multiple regression analysis in 3C and the simple regression analysis in 3B, because between smokers and non-smokers, there are still some other variables can influence the systolic blood pressure. For example, smokers are younger and have lower BMI in this data set, and so the differences in systolic blood pressure between smokers and non-smokers may in part be due to differences on these variables (there may be confounding).

I think multiple regression analysis is more appropriate, because in 3C, we use multiple regression as an adjusted analysis to control for confounding.