Stat 206: Linear Models

Lecture 10

Nov. 2, 2015

Quantify Multicollinearity: Variance Inflation Factor

Under the standardized model:

$$\sigma^2(\hat{\boldsymbol{\beta}}^*) = \qquad\qquad\qquad\qquad .$$

- The kth diagonal element of the inverse correlation matrix $\mathbf{r}_{XX}^{-1}$ is called the **variance inflation factor (VIF)** for $\hat{\beta}_k^*$, denoted by VIF_k.
- The variance of the estimated regression coefficient $\hat{\beta}_k^*$:

$$\sigma^2(\hat{\beta}_k^*) = \qquad\qquad\qquad\qquad .$$

- The variance of the estimated regression coefficient $\hat{\beta}_k$ in the original model:

$$\sigma^2(\hat{\beta}_k) = \qquad\qquad\qquad\qquad .$$

Quantify Multicollinearity: Variance Inflation Factor

Under the standardized model:

$$\boldsymbol{\sigma}^{\mathbf{2}}(\hat{\boldsymbol{\beta}}^*) = \sigma^{*2}\mathbf{r}_{XX}^{-1}.$$

- The kth diagonal element of the inverse correlation matrix $\mathbf{r}_{XX}^{-1}$ is called the **variance inflation factor (VIF)** for $\hat{\beta}_k^*$, denoted by VIF_k.
- The variance of the estimated regression coefficient $\hat{\beta}_k^*$:

$$\sigma^2(\hat{\beta}_k^*) = VIF_k \sigma^{*2}.$$

- The variance of the estimated regression coefficient $\hat{\beta}_k$ in the original model:

$$\sigma^2(\hat{\beta}_k) = VIF_k \times \frac{\sigma^2}{\sum_{i=1}^n (X_{ik} - \bar{X}_k)^2}.$$

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad k = 1, \cdots, p-1,$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the rest of X variables $\{X_j : 1 \leq j \neq k \leq p-1\}$.

- If X_k is uncorrelated with the rest of the X variables, then $R_k^2 =$ and $VIF_k =$.
- If $R_k^2 > 0$, then VIF_k , indicating an variance for $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) due to the between X_k and the other X variables.
- If X_k has a perfect linear association with the rest of the X variables, i.e., X_k is their , then $R_k^2 =$, $VIF_k =$ and so the variance of $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) is .
- In practice, $\max_k VIF_k > 10$ is often taken as an indication that multicollinearity is high.

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad k = 1, \cdots, p-1,$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the rest of X variables $\{X_j : 1 \leq j \neq k \leq p-1\}$.

- If X_k is uncorrelated with the rest of the X variables, then $R_k^2 = 0$ and $VIF_k = 1$.
- If $R_k^2 > 0$, then $VIF_k > 1$, indicating an inflated variance for $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) due to the intercorrelation between X_k and the other X variables.
- If X_k has a perfect linear association with the rest of the X variables, i.e., X_k is their linear combination, then $R_k^2 = 1$, $VIF_k = \infty$ and so the variance of $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) is infinity (ill-defined).
- In practice, $\max_k VIF_k > 10$ is often taken as an indication that multicollinearity is high.

Body Fat

Correlation matrices.

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} 0.84 \\ 0.88 \\ 0.14 \end{bmatrix}.$$

X_1 and X_2 are highly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are not much correlated. Moreover,

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 708.84 & -631.92 & -270.99 \\ -631.92 & 564.34 & 241.49 \\ -270.99 & 241.49 & 104.61 \end{bmatrix}$$

So,

Each predictor is with the rest of the predictors.

Body Fat

Correlation matrices.

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} 0.84 \\ 0.88 \\ 0.14 \end{bmatrix}.$$

X_1 and X_2 are highly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are not much correlated. Moreover,

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 708.84 & -631.92 & -270.99 \\ -631.92 & 564.34 & 241.49 \\ -270.99 & 241.49 & 104.61 \end{bmatrix}$$

So,

$$R_1^2 = 0.9986, \quad R_2^2 = 0.9982, \quad R_3^2 = 0.9904.$$

Each predictor is highly intercorrelated with the rest of the predictors.

Variables in Model	$\hat{\beta}_1$	$\hat{\beta}_2$	$s\{\hat{\beta}_1\}$	$s\{\hat{\beta}_2\}$	MSE
Model 1: X_1	0.8572	-	0.1288	-	7.95
Model 2: X_2	-	0.8565	-	0.1100	6.3
Model 3: X_1, X_2	0.2224	0.6594	0.3034	0.2912	6.47
Model 4: X_1, X_2, X_3	4.334	-2.857	3.016	2.582	6.15

- The regression coefficient for X_1 (X_2) varies drastically depending on which other X variables are included in the model.
- The standard errors of the fitted regression coefficients are becoming inflated when more X variables are included into the model.
- MSE tends to decrease as additional X variables are added into the model.

R outputs

In Model 4, none of the three X variables is statistically significant by the T-tests. However, the F-test for regression relation is highly significant. Is there a paradox?

- From the general linear test perspective, each T-test is a test, testing whether the of an X variable is significant given X variables being included in the model.
- The three tests of the marginal effects of X_1, X_2, X_3 together are to testing whether there is a regression relation between Y and (X_1, X_2, X_3).
- The reduced model for each individual test contains X variables, whereas the reduced model for testing regression relation contains X variable.

Model 4

In Model 4, none of the three X variables is statistically significant by the T-tests. However, the F-test for regression relation is highly significant. Is there a paradox?

- From the general linear test perspective, each T-test is a marginal test, testing whether the marginal effect of an X variable is significant given **all other** X variables being included in the model.
- The three tests of the marginal effects of X_1, X_2, X_3 together are not equivalent to testing whether there is a regression relation between Y and (X_1, X_2, X_3).
- The reduced model for each individual test contains **all other** X variables, whereas the reduced model for testing regression relation contains no X variable.

◂ Model 4

Body Fat: Model 1

```
> summary(fit1)

Call:
lm(formula = Y ~ X1, data = fat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
X1            0.8572     0.1288   6.656 3.02e-06 ***
---

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared: 0.7111,     Adjusted R-squared: 0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06

> anova(fit1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
```

back

Body Fat: Model 2

```
> summary(fit2)

Call:
lm(formula = Y ~ X2, data = fat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
X2            0.8565     0.1100   7.786   3.6e-07 ***

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-squared: 0.771,      Adjusted R-squared: 0.7583
F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07

> anova(fit2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X2         1 381.97  381.97  60.617 3.6e-07 ***
Residuals 18 113.42    6.30
```

back

Body Fat: Model 3

```
> summary(fit3)

Call:
lm(formula = Y ~ X1 + X2, data = fat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
X1            0.2224     0.3034   0.733   0.4737
X2            0.6594     0.2912   2.265   0.0369 *
---

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared: 0.7781,     Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

> anova(fit3)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 54.4661 1.075e-06 ***
X2         1  33.17   33.17  5.1284    0.0369 *
Residuals 17 109.95    6.47
```

back

Body Fat: Model 4

```
> summary(fit4)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X2            -2.857      2.582  -1.106    0.285
X3            -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,     Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06

> anova(fit4)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 57.2768 1.131e-06 ***
X2         1  33.17   33.17  5.3931   0.03373 *
X3         1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

partial

multicollinearity

Body Fat

Prediction of new observations does not tend to get worse with additional (intercorrelated) X variables in the model.

```
> newX=data.frame(X1=25, X2=50, X3=29)
> predict.lm(fit1, newX, se.fit=TRUE)
$fit
       1
19.93356
$se.fit
[1] 0.6317416
...
> predict.lm(fit2, newX, se.fit=TRUE)
$fit
       1
19.19284
$se.fit
[1] 0.5758769
...
> predict.lm(fit3, newX, se.fit=TRUE)
$fit
       1
19.35566
$se.fit
[1] 0.6243083
...
> predict.lm(fit4, newX, se.fit=TRUE)
$fit
       1
19.19885
$se.fit
[1] 0.6194612
```

Interaction Regression Models

These are regression models with terms representing and various terms representing among respective predictors.

- Example. A model with three main effects and two interaction effects between X_1, X_2 and between X_1, X_3, respectively:

Interaction Regression Models

These are regression models with first-order terms representing *main effects* and various cross-product terms representing *interactions effects* among respective predictors.

- Example. A model with three main effects and two interaction effects between X_1, X_2 and between X_1, X_3, respectively:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \epsilon_i, \;\; i = 1, \cdots n.$$

Interpretation of Interaction Effects

Consider an interaction model with two quantitative predictors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, n.$$

- Response function:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- The change in the mean response with a unit increase in X_1 when X_2 is held constant is

- The change in the mean response with a unit increase in X_2 when X_1 is held constant is

- Therefore, when $\beta_3 \neq 0$, the effect of on the mean response depends on the level of

.

Interpretation of Interaction Effects

Consider an interaction model with two quantitative predictors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, n.$$

- Response function:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- The change in the mean response with a unit increase in X_1 when X_2 is held constant is

$$\beta_1 + \beta_3 X_2.$$

- The change in the mean response with a unit increase in X_2 when X_1 is held constant is

$$\beta_2 + \beta_3 X_1.$$

- Therefore, when $\beta_3 \neq 0$, the effect of X_1 (or X_2) on the mean response depends on the level of X_2 (or X_1).

Body Fat

Test whether second-order interaction terms between the three predictors should be included.

- 2nd-order interaction model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \epsilon_i.$$

- Test whether interaction terms may be dropped:

- The reduced model is the model.

$$SSR(X_1X_2, X_1X_3, X_2X_3 | X_1, X_2, X_3) = \qquad\qquad, d.f. = \qquad .$$

- SSE of the full model is 87.69 with $d.f. = 13$. F-statistic and pvalue are

- We conclude that the second-order interaction terms and a first-order model .

Body Fat

Test whether second-order interaction terms between the three predictors should be included.

- 2nd-order interaction model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \epsilon_i.$$

- Test whether interaction terms may be dropped:

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \;\; vs. \;\; \text{not all of the three equal zero.}$$

- The reduced model is the first-order model.

$$SSR(X_1X_2, X_1X_3, X_2X_3 | X_1, X_2, X_3) = 1.50 + 2.70 + 6.51 = 10.71, \; d.f. = 3.$$

- SSE of the full model is 87.69 with $d.f. = 13$. F-statistic and pvalue are

$$F^* = \frac{10.71/3}{87.69/13} = 0.53, \quad pvalue = P(F_{3,13} > 0.53) = 0.33.$$

- We conclude that the second-order interaction terms can be dropped and a first-order model suffices.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3, data = fat)

Residuals:
Min      1Q  Median      3Q     Max
-3.9267 -0.8728  0.1682  1.2929  3.3310

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 165.378761 136.401641   1.212    0.247
X1            5.325568   3.379692   1.576    0.139
X2           -4.816586   3.513565  -1.371    0.194
X3           -4.097260   3.122743  -1.312    0.212
X1:X2         0.008876   0.030850   0.288    0.778
X1:X3        -0.084791   0.073418  -1.155    0.269
X2:X3         0.090415   0.092001   0.983    0.344

Residual standard error: 2.597 on 13 degrees of freedom
Multiple R-squared:  0.823, Adjusted R-squared:  0.7413
F-statistic: 10.07 on 6 and 13 DF,  p-value: 0.0002959

> anova(fit)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 52.2238 6.682e-06 ***
X2         1  33.17   33.17  4.9173   0.04503 *
X3         1  11.55   11.55  1.7117   0.21343
X1:X2      1   1.50    1.50  0.2217   0.64552
X1:X3      1   2.70    2.70  0.4009   0.53760
X2:X3      1   6.51    6.51  0.9658   0.34366
Residuals 13  87.69    6.75
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Qualitative Predictors

Qualitative variables, a.k.a. **categorical variables**, represent certain characteristics of a subject.

- A qualitative variable has a fixed set of possible values/levels/classes.
- If a qualitative variable takes on exactly two values, it is called a **binary variable.**
- Examples.
  - Gender: male or female; binary variable.
  - Blood type: A, B, AB or O.
  - Smoke status: smoke or not smoke; binary variable.
  - Income level: high, medium or low.
  - Education level: high school, college, or advanced degree.

Indicators for Qualitative Variables

- To use a qualitative variable in a regression model as a predictor, we need to its classes.
- One popular approach is to use indicator variables (a.k.a. **dummy variables**).
  - An **indicator variable** is a variable only takes on the values .

Indicators for Qualitative Variables

- To use a qualitative variable in a regression model as a predictor, we need to quantitatively identify its classes.
- One popular approach is to use indicator variables (a.k.a. **dummy variables**).
    - An **indicator variable** is a variable only takes on the values 0 or 1.

To quantify a binary variable, we need indicator variable.

- Suppose the two classes are labelled as C_1, C_2. Then the indicator variable can be defined as

- For example, to code gender,

$$X = \begin{cases} 1 & \textit{if} \quad \textit{male} \\ 0 & \textit{if} \quad \textit{female} \end{cases}$$

- The above coding is , since we can arbitrarily choose the *reference class* – the class coded as 0.

To quantify a binary variable, we need **one** indicator variable.

- Suppose the two classes are labelled as C_1, C_2. Then the indicator variable can be defined as

$$X = \begin{cases} 1 & \textit{if} \quad C_1 \\ 0 & \textit{if} \quad C_2 \end{cases}$$

- For example, to code gender,

$$X = \begin{cases} 1 & \textit{if} \quad \textit{male} \\ 0 & \textit{if} \quad \textit{female} \end{cases}$$

- The above coding is **not unique**, since we can arbitrarily choose the *reference class* – the class coded as 0.

Insurance

An economist wanted to relate the speed with which a particular insurance innovation is adopted by an insurance firm (Y) to the size of the firm (X_1) and the type of the firm (X_2). He collected data on 20 insurance firms, 10 stock firms and 10 mutual firms.

- Y – – number of months elapsed before the firm adopted the innovation and X_1 – – the amount of total assets of the firm are quantitative variables.
- Type of the firm is a variable taking on two values: "stock" or "mutual". If we choose "mutual" as the reference class, then it can be quantified by an indicator variable:

Insurance

An economist wanted to relate the speed with which a particular insurance innovation is adopted by an insurance firm (Y) to the size of the firm (X_1) and the type of the firm (X_2). He collected data on 20 insurance firms, 10 stock firms and 10 mutual firms.

- Y – – number of months elapsed before the firm adopted the innovation and X_1 – – the amount of total assets of the firm are quantitative variables.
- Type of the firm is a binary variable taking on two values: "stock" or "mutual". If we choose "mutual" as the reference class, then it can be quantified by an indicator variable:

$$X_2 = \begin{cases} 1 & \text{if} \quad \text{stock} \\ 0 & \text{if} \quad \text{mutual} \end{cases}$$

A snapshot of the data.

Firm	Number_of_month_elapsed	Firm_size	Firm_Type	Indicator_Code
Y	X1	X2		
1	17	151	mutual	0
2	26	92	mutual	0
3	21	175	mutual	0
..	...	...	...	...
18	13	305	stock	1
19	30	124	stock	1
20	14	246	stock	1

Figure: Boxplots

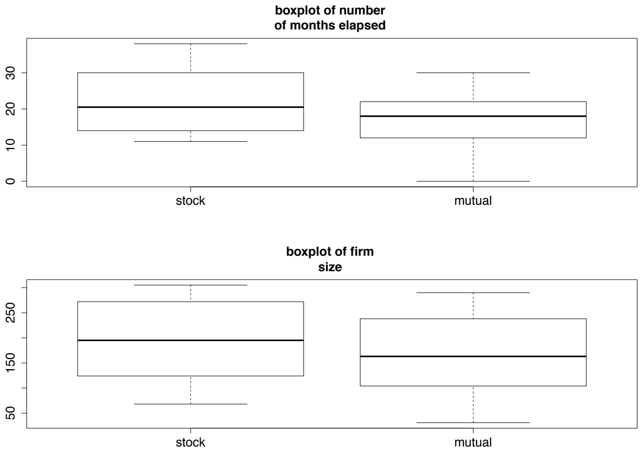
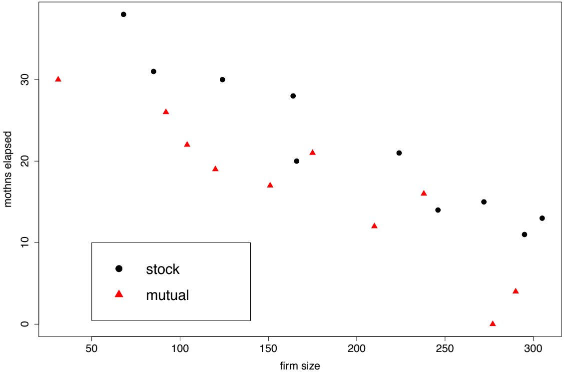
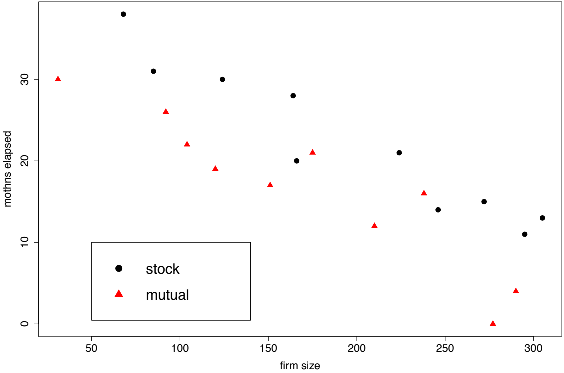


Figure: Scatter plot of months elapsed (Y) versus firm size (X_1).



The slope appears to be for the two types of firms, whereas the intercept appears to be . This means that a model would suffice.

Figure: Scatter plot of months elapsed (Y) versus firm size (X_1).



The slope appears to be similar for the two types of firms, whereas the intercept appears to be different. This means that a first-order model would suffice.

A first order model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- The response function (mean response):
  - For mutual firms, $X_2 = 0$ and the response function becomes
  - For stock firms, $X_2 = 1$ and the response function becomes
  - Both are with

A first order model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- The response function (mean response):

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

  - For mutual firms, $X_2 = 0$ and the response function becomes

$$E(Y) = \beta_0 + \beta_1 X_1 \quad \text{mutual firms}$$

  - For stock firms, $X_2 = 1$ and the response function becomes

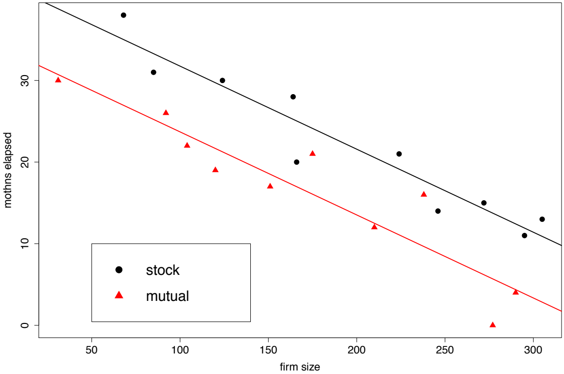
$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \quad \text{stock firms}$$

  - Both are straight lines with the same slope β_1 but with intercepts differing by β_2.

Insurance: First Order Model

```
> data=data.frame(read.table("insurance.txt"))
> names(data)=c("Y", "X1", "X2")
> fit=lm(Y~ X1+factor(X2), data=data)
> summary(fit)

Call:
lm(formula = Y ~ X1 + factor(X2), data = data)

Residuals:
Min      1Q  Median      3Q     Max
-5.6915 -1.7036 -0.4385  1.9210  6.3406

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     33.874069   1.813858   18.675 9.15e-13 ***
X1              -0.101742   0.008891 -11.443 2.07e-09 ***
factor(X2)stock  8.055469   1.459106   5.521 3.74e-05 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 3.221 on 17 degrees of freedom
Multiple R-squared: 0.8951,     Adjusted R-squared: 0.8827
F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

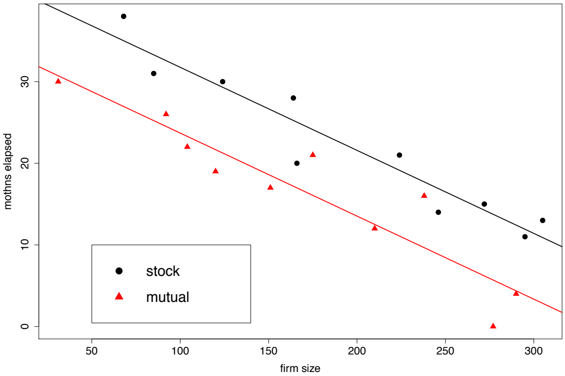
Figure: Response functions for the stock firms (black) and mutual firms (red).



Stock firms response line: .

Mutual firms response line: .

Figure: Response functions for the stock firms (black) and mutual firms (red).



Stock firms response line: $\widehat{E(Y)} = (33.874 + 8.055) - 0.1017X_1$.

Mutual firms response line: $\widehat{E(Y)} = 33.874 - 0.1017X_1$.

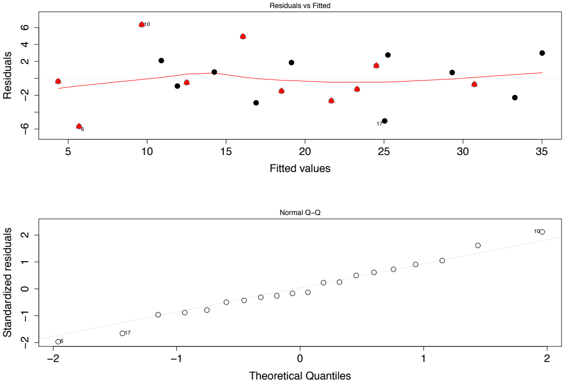


Figure: Residual plots.

No obvious violation of the linearity, constant error variance and normal error assumptions.

The economist was most interested in the effect of firm type on the speed to adopt an innovation.

- $\hat{\beta}_2 = 8.055$ means that for any given firm size, on average, it takes stock firms to adopt an innovation than mutual firms of the same size.
- A 95% confidence interval for β_2: $t(0.975; 17) = 2.11$

  With 95% confidence, we conclude that on average stock firms takes to adopt an innovation than mutual firms.
- The pvalue for testing whether $\beta_2 = 0$ is 3.74×10^{-5}. Therefore, β_2 is highly significant and firm type has a effect on the speed of adopting an innovation.

Why not simply fit two separate regression models for stock firms and mutual firms?

The economist was most interested in the effect of firm type on the speed to adopt an innovation.

- $\hat{\beta}_2 = 8.055$ means that for any given firm size, on average, it takes stock firms 8 more months to adopt an innovation than mutual firms of the same size.
- A 95% confidence interval for β_2: $t(0.975; 17) = 2.11$

$$8.055 \pm 2.11 \times 1.459 = [4.98, 11.13].$$

  With 95% confidence, we conclude that on average stock firms takes between 5 to 11 more months to adopt an innovation than mutual firms.
- The pvalue for testing whether $\beta_2 = 0$ is 3.74×10^{-5}. Therefore, β_2 is highly significant and firm type has a significant effect on the speed of adopting an innovation.

Why not simply fit two separate regression models for stock firms and mutual firms?

Summary. Interpretation of regression coefficients in a first order model with a quantitative variable (X_1) and an indicator variable (X_2):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, n.$$

- β_1 is the of the mean response line under both classes.
- β_0 is the under class 0 (i.e., the reference class).
- β_2 shows

  the mean response line is for class 1 for any given value of X_1.

Summary. Interpretation of regression coefficients in a first order model with a quantitative variable (X_1) and an indicator variable (X_2):

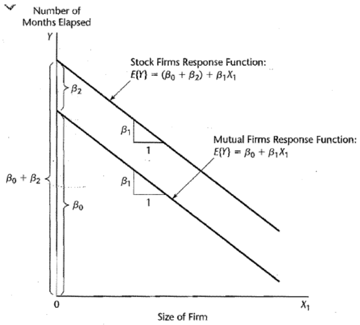
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, n.$$

- β_1 is the common slope of the mean response line under both classes.
- β_0 is the baseline intercept under class 0 (i.e., the reference class).
- β_2 shows how much higher (if positive) or lower (if negative) the mean response line is for class 1 for any given value of X_1.

In a first-order model without interaction, the effect of the qualitative variable is no matter the value of the quantitative variable.

Figure:

FIGURE 8.11 Illustration of Meaning of Regression Coefficients for Regression Model (8.33) with Indicator Variable X_2—Insurance Innovation Example.

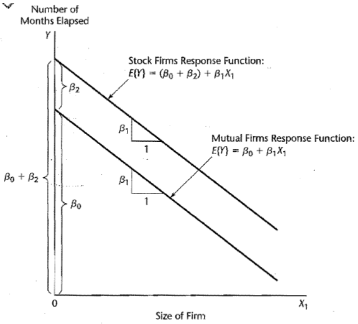


Response functions are with intercepts.

In a first-order model without interaction, the effect of the qualitative variable is the same no matter the value of the quantitative variable.

Figure:

FIGURE 8.11 Illustration of Meaning of Regression Coefficients for Regression Model (8.33) with Indicator Variable X_2—Insurance Innovation Example.



Response functions are parallel (same slope) with different intercepts.

Interactions between Quantitative and Qualitative Predictors

Interaction between qualitative and quantitative predictors can be introduced into the model through the usual manner, by

.

- Insurance company. A model with interaction between firm size and firm type:

  where X_1 is the amount of assets of a firm and X_2 is an indicator variable indicating the type of a firm.

Interactions between Quantitative and Qualitative Predictors

Interaction between qualitative and quantitative predictors can be introduced into the model through the usual manner, by adding cross-product terms.

- Insurance company. A model with interaction between firm size and firm type:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20,$$

  where X_1 is the amount of assets of a firm and X_2 is an indicator variable indicating the type of a firm.

- Response function (mean response):

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- For mutual firms, $X_2 = 0$ and thus $X_1 X_2 = 0$. The response function becomes

  which is a straight line with slope β_1 and intercept β_0.
- For stock firms, $X_2 = 1$ and thus $X_1 X_2 = X_1$. The response function becomes

  which is a straight line with slope $\beta_1 + \beta_3$ and intercept $\beta_0 + \beta_2$.

- Response function (mean response):

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- For mutual firms, $X_2 = 0$ and thus $X_1 X_2 = 0$. The response function becomes

$$E(Y) = \beta_0 + \beta_1 X_1 \quad \text{mutual firms,}$$

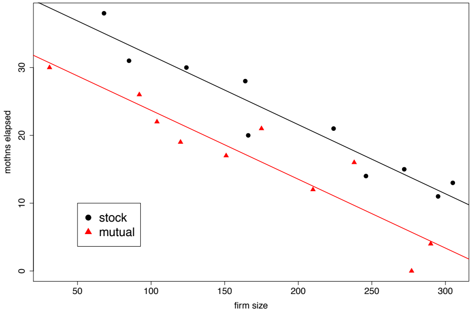
  which is a straight line with slope β_1 and intercept β_0.

- For stock firms, $X_2 = 1$ and thus $X_1 X_2 = X_1$. The response function becomes

$$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \quad \text{stock firms,}$$

  which is a straight line with slope $\beta_1 + \beta_3$ and intercept $\beta_0 + \beta_2$.

Figure: Response functions for the stock firms (black) and mutual firms (red) under the interaction model.



Stock firms: .

Mutual firms:

The two lines are because

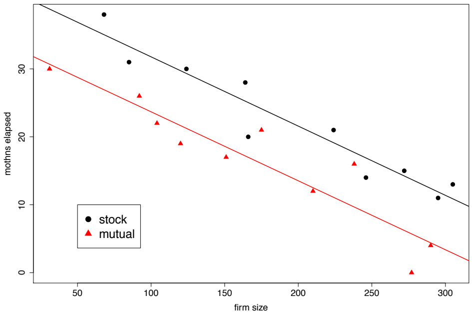
is very small compared to .

Insurance: Interaction Model

```
> fit2=lm(Y˜ X1+factor(X2)+X1:factor(X2), data=data)
> summary(fit2)
Call:
lm(formula = Y ˜ X1 + factor(X2) + X1:factor(X2), data = data)

Residuals:
Min      1Q  Median      3Q     Max
-5.7144 -1.7064 -0.4557  1.9311  6.3259

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        33.8383695  2.4406498  13.864 2.47e-10 ***
X1                 -0.1015306  0.0130525  -7.779 7.97e-07 ***
factor(X2)stock     8.1312501  3.6540517   2.225   0.0408 *
X1:factor(X2)stock -0.0004171  0.0183312  -0.023   0.9821
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 3.32 on 16 degrees of freedom
Multiple R-squared: 0.8951,     Adjusted R-squared: 0.8754
F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

β_3 and we conclude that there is
and the first-order model .

Figure: Response functions for the stock firms (black) and mutual firms (red) under the interaction model.



Stock firms: $\widehat{E(Y)} = (33.838 + 8.131) - (0.1015 + 0.00042)X_1$.
Mutual firms: $\widehat{E(Y)} = 33.838 - 0.1015X_1$. The two lines are nearly parallel because $\hat{\beta}_3 = -0.00042$ is very small compared to $\hat{\beta}_1 = -0.1015$.

Insurance: Interaction Model

```
> fit2=lm(Y~ X1+factor(X2)+X1:factor(X2), data=data)
> summary(fit2)
Call:
lm(formula = Y ~ X1 + factor(X2) + X1:factor(X2), data = data)

Residuals:
Min      1Q  Median      3Q     Max
-5.7144 -1.7064 -0.4557  1.9311  6.3259

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        33.8383695  2.4406498  13.864 2.47e-10 ***
X1                 -0.1015306  0.0130525  -7.779 7.97e-07 ***
factor(X2)stock     8.1312501  3.6540517   2.225   0.0408 *
X1:factor(X2)stock -0.0004171  0.0183312  -0.023   0.9821
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 3.32 on 16 degrees of freedom
Multiple R-squared: 0.8951,     Adjusted R-squared: 0.8754
F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

β_3 is not significant and we conclude that there is no interaction effect and the first-order model suffices.

Summary. Interpretation of regression coefficients in an interaction model with a quantitative variable (X_1) and an indicator variable (X_2).:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, n$$

- β_0 and β_1 are , respectively, of the response function for class 0 (i.e., the reference class).
- β_2 indicates

  is the intercept of the response function for class 1.
- β_3 indicates

  is the slope of the response function for class 1.

When interaction effects are present, the effect of the qualitative variable the value of the quantitative variable.

Summary. Interpretation of regression coefficients in an interaction model with a quantitative variable (X_1) and an indicator variable (X_2).:

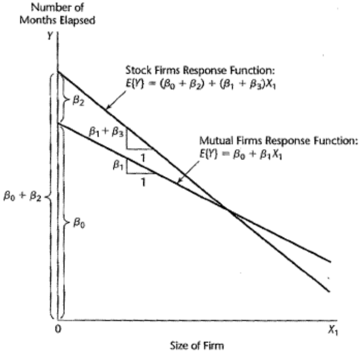
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, n$$

- β_0 and β_1 are baseline intercept and slope, respectively, of the response function for class 0 (i.e., the reference class).
- β_2 indicates how much greater (if positive) or smaller (if negative) is the intercept of the response function for class 1.
- β_3 indicates how much greater (if positive) or smaller (if negative) is the slope of the response function for class 1.

When interaction effects are present, the effect of the qualitative variable depends on the value of the quantitative variable.

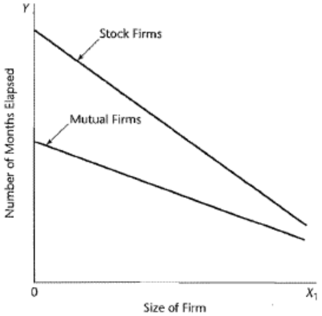
Possible Scenarios of Interaction

Figure: Disordinal interaction: response functions intersect within the scope of the model.



Mutual firms tend to adopt an innovation faster than stock firms for smaller firms, but for larger firms, stock firms tend to innovate more quickly.

Figure: Ordinal interaction: nonparallel response functions that do not intersect within the scope of the model.



Mutual firms tend to adopt the innovation faster than stock firms for all sizes of the firms in the scope of the model, but the difference is smaller for larger firms than for smaller firms.

Qualitative Variables with More than Two Classes

A qualitative variable with r classes, labeled as $C_l, \cdots, C_r$, need to be represented by indicator variables, each taking on the values :

For C_r (the reference class), .

The above quantification is not unique as we can choose a different *reference class*.

Qualitative Variables with More than Two Classes

A qualitative variable with r classes, labeled as $C_1, \cdots, C_r$, need to be represented by $\mathbf{r-1}$ indicator variables, each taking on the values 0 and 1:

$$
\begin{aligned}
X_1 &= \begin{cases} 1 & \text{if} \quad C_1 \\ 0 & \text{if} \quad \text{otherwise} \end{cases} \\
X_2 &= \begin{cases} 1 & \text{if} \quad C_2 \\ 0 & \text{if} \quad \text{otherwise} \end{cases} \\
&\vdots \\
X_{r-1} &= \begin{cases} 1 & \text{if} \quad C_{r-1} \\ 0 & \text{if} \quad \text{otherwise} \end{cases}
\end{aligned}
$$

For C_r (the reference class), $X_1 = \cdots = X_{r-1} = 0$.

The above qualification is not unique as we can choose a different *reference class*.

Real Estate

A city tax assessor was interested in predicting residential home sales prices (Y) as a function of various characteristics of the home including finished square footage (X_1) and the quality of construction (high, medium or low). Data was collected on 522 home sales made on 2002.

- Y – sales price and X_1 – square footage are quantitative variables.
- Quality of construction is a qualitative variable with three classes. Use "high quality" as the reference class, it can be quantified by indicator variables:

If a home is of high construction quality, then .

Real Estate

A city tax assessor was interested in predicting residential home sales prices (Y) as a function of various characteristics of the home including finished square footage (X_1) and the quality of construction (high, medium or low). Data was collected on 522 home sales made on 2002.

- Y – sales price and X_1 – square footage are quantitative variables.
- Quality of construction is a qualitative variable with three classes. Use "high quality" as the reference class, it can be quantified by two indicator variables:

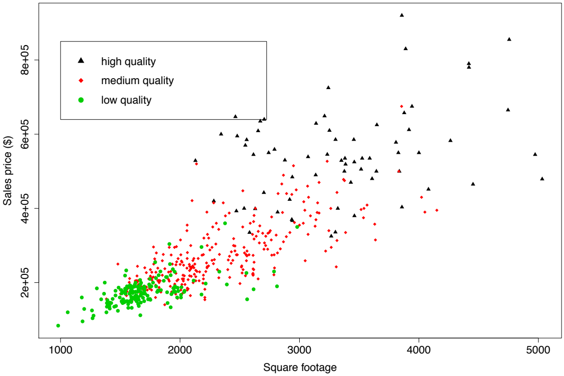
$$X_2 = \begin{cases} 1 & \textit{if} \quad \text{low quality} \\ 0 & \textit{if} \quad \textit{otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 & \textit{if} \quad \text{medium quality} \\ 0 & \textit{if} \quad \textit{otherwise} \end{cases}$$

If a home is of high construction quality, then $X_2 = X_3 = 0$.

Snapshot of data.

```
home sales_price square_footage constructrion_quality
Y          X1                                      X2          X3

1     360000        3032              medium              0           1
2     340000        2058              medium              0           1
3     250000        1780              medium              0           1
..     ...          ...                ...               ...         ...
9     195000        1622              low                 1           0
10    160000        1976              low                 1           0
11    190000        2812              low                 1           0
12    559000        2791              high                0           0
13    535000        3381              high                0           0
14    525000        3459              high                0           0
..     ...          ...                ...               ...         ...
521   124000        1480              low                 1           0
522   95500         1184              low                 1           0
```

Figure: Scatter plot of sales price versus square footage.



A first-order model :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 522.$$

- Response function (mean response):

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

- For high quality homes, $X_2 = X_3 = 0$, response function becomes:

  a straight line with slope β_1 and intercept β_0.

- For low quality homes, $X_2 = 1, X_3 = 0$, response function becomes:

  a straight line with slope β_1 and intercept $\beta_0 + \beta_2$.

- For medium quality homes, $X_2 = 0, X_3 = 1$, response function becomes:

  a straight line with slope β_1 and intercept $\beta_0 + \beta_3$.

A first-order model :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 522.$$

- Response function (mean response):

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

- For high quality homes, $X_2 = X_3 = 0$, response function becomes:

$$E(Y) = \beta_0 + \beta_1 X_1 \quad \text{high quality homes,}$$

  a straight line with slope β_1 and intercept β_0.

- For low quality homes, $X_2 = 1, X_3 = 0$, response function becomes:

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \quad \text{low quality homes,}$$

  a straight line with slope β_1 and intercept $\beta_0 + \beta_2$.

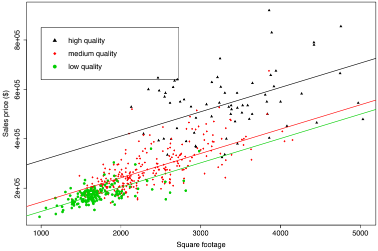
- For medium quality homes, $X_2 = 0, X_3 = 1$, response function becomes:

$$E(Y) = (\beta_0 + \beta_3) + \beta_1 X_1 \quad \text{medium quality homes,}$$

  a straight line with slope β_1 and intercept $\beta_0 + \beta_3$.

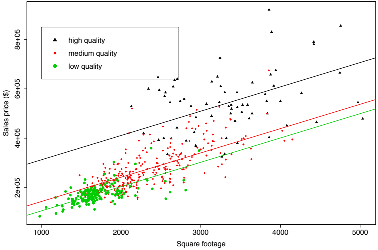
```
> table(data[,"quality"])

high    low medium
68    164    290
> fit=lm(sales˜Sq+factor(quality), data=data)
> summary(fit)

Call:
lm(formula = sales ˜ Sq + factor(quality), data = data)

Residuals:
Min      1Q  Median      3Q     Max
-230493  -31093   -7680   23993  326172

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.141e+05  2.014e+04   10.64   <2e-16 ***
Sq                     9.844e+01  5.557e+00   17.71   <2e-16 ***
factor(quality)low    -2.068e+05  1.295e+04  -15.97   <2e-16 ***
factor(quality)medium -1.689e+05  1.030e+04  -16.40   <2e-16 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 63640 on 518 degrees of freedom
Multiple R-squared: 0.7883,     Adjusted R-squared: 0.7871
F-statistic:   643 on 3 and 518 DF,  p-value: < 2.2e-16
```

Figure: Response functions for high, medium and low quality homes.



- High quality home:
  Medium quality home:
  Low quality home:
- The three response lines are with intercepts.

Figure: Response functions for high, medium and low quality homes.



- High quality home: $\widehat{E(Y)} = 2.141 \times 10^5 + 98.44X_1$; Medium quality home: $\widehat{E(Y)} = (2.141 - 1.689) \times 10^5 + 98.44X_1$; Low quality home: $\widehat{E(Y)} = (2.141 - 2.068) \times 10^5 + 98.44X_1$.
- The three response lines are parallel (same slope) with different intercepts.

- β_2 and β_3 measure   of low and medium construction quality, respectively, on the sales price (Y) for any given home size (X_1), compared with high construction quality (the reference class). Both are highly significant.
- *What if we want to measure the differential effect between medium quality construction and low quality construction on sales price?*
    - This is measured by   .
    - A point estimator is   with squared standard error:
    - The SE can be derived from the estimated variance-covariance matrix of $\hat{\beta}$: $MSE \times (\mathbf{X}'\mathbf{X})^{-1}$:

- β_2 and β_3 measure the differential effects of low and medium construction quality, respectively, on the sales price (Y) for any given home size (X_1), compared with high construction quality (the reference class). Both are highly significant.
- *What if we want to measure the differential effect between medium quality construction and low quality construction on sales price?*
  - This is measured by $\beta_3 - \beta_2$.
  - A point estimator is $\hat{\beta}_3 - \hat{\beta}_2$ with squared standard error:

$$s^2\{\hat{\beta}_3 - \hat{\beta}_2\} = s^2\{\hat{\beta}_3\} + s^2\{\hat{\beta}_2\} - 2s\{\beta_3, \beta_2\},$$

  - The SE can be derived from the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$: $MSE \times (\mathbf{X}'\mathbf{X})^{-1}$: $c = (0, 0, -1, 1)^T$, then $\hat{\beta}_3 - \hat{\beta}_2 = c^T\hat{\beta}$ and $s(\hat{\beta}_3 - \hat{\beta}_2) = \sqrt{c^T(MSE(\mathbf{X}'\mathbf{X})^{-1})c}$.

```
> coef=coef(fit)  ##estiamted regression coefficients
> b32=coef[4]-coef[3] ## beta3-beta2

> var.b=vcov(fit)  ## variance-covariance matrix of the estimated regression coefficients
> cons=matrix(c(0,0,-1,1)) ## contrast vector for beta3-beta2

>  sd.b32=sqrt(t(cons)%*%var.b%*%cons) ## standard error of hat{beta3}-hat{beta2}

> b32
factor(quality)medium
37921
> sd.b32
[,1]
[1,] 7102.962

> qt(0.975, 518) ##97.5% percentile of t with 518 (=522-4) d.f.
[1] 1.964554
>
> b32+qt(0.975,518)*sd.b32 ## upper limit of the 95% C.I. for beta3-beta2
[,1]
[1,] 51875.16
> b32-qt(0.975,518)*sd.b32 ## lower limit of the 95% C.I. for beta3-beta2
[,1]
[1,] 23966.85
```

We are 95% confident that, on average, medium quality homes have in between \$24000 to \$52000 higher sales price than low quality homes for any given home size.