Basic Regression

Matthew Seguin

2.

```
library(rio)
hprice <- import("hprice.xls")
```

a.

```
reg1 <- lm(price ~ sqrft + bdrms, data = hprice)
summary(reg1)
```

```
##
## Call:
## lm(formula = price ~ sqrft + bdrms, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.799  -39.219   -5.298   31.002  231.239
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.47429   28.32941  -0.582   0.5622
## sqrft         0.12809    0.01282   9.993   <2e-16 ***
## bdrms        14.41436    8.63525   1.669   0.0983 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.21 on 98 degrees of freedom
## Multiple R-squared:  0.6283, Adjusted R-squared:  0.6207
## F-statistic: 82.83 on 2 and 98 DF,  p-value: < 2.2e-16
```

We just ran the regression $price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + U$. The estimated intercept is $\hat{\beta}_0 = -16.47429$, estimated coefficient for $sqrft$ is $\hat{\beta}_1 = 0.12809$ and the estimated coefficient for $bdrms$ is $\hat{\beta}_2 = 14.41436$. The intercept has a very significant p-value, it would be unwise to leave the intercept out of a regression. As you can see the t-statistic for $\hat{\beta}_1$ is very large, and hence the p-value for the hypothesis $\beta_1 = 0$ is very small. We can reject the hypothesis $H_0 : \beta_1 = 0$, this implies that square feet has a significant effect on the selling price (an expected result). The p-value for $\hat{\beta}_2$ is non-negligible, so we can't reject the hypothesis $H_0 : \beta_2 = 0$ which implies that the number of bedrooms might not have a significant effect on the selling price (a rather counter intuitive result). The R^2 isn't high but it isn't low either, so the model has a decent fit overall. Lastly, the p-value for the F-statistic is extremely low which tells us that at least one of the slopes is nonzero (which we know is probably β_1, the coefficient for square feet).

b.

Under our regression model $price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + U$, if $sqrft$ is held constant and $bdrms$ increases by 1 (i.e. $bdrms_1 = bdrms_0 + 1$) we will get the following:

$$\Delta price = price_1 - price_0 = (\beta_0 + \beta_1 sqrft + \beta_2 bdrms_1) - (\beta_0 + \beta_1 sqrft + \beta_2 bdrms_0) = \beta_2(bdrms_1 - bdrms_0) = \beta_2 \approx \hat{\beta}_2$$

This was estimated via the results of our regression using OLS, the result ($\hat{\beta}_2$) is shown below.

```
as.numeric(reg1$coefficients[3])
```

```
## [1] 14.41436
```

So the estimated increase in price for a house with one more bedroom, holding square footage constant, is 14.41436 which is 14,414.36 dollars in the real world.

c.

Under our regression model $price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + U$, if $bdrms$ increases by 1 (i.e. $bdrms_1 = bdrms_0 + 1$) and the extra bedroom is 140 extra square feet (i.e. $sqrft_1 = sqrft_0 + 140$) we will get the following:

$$\Delta price = price_1 - price_0 = (\beta_0 + \beta_1 sqrft_1 + \beta_2 bdrms_1) - (\beta_0 + \beta_1 sqrft_0 + \beta_2 bdrms_0)$$

$$= \beta_1(sqrft_1 - sqrft_0) + \beta_2(bdrms_1 - bdrms_0) = 140\beta_1 + \beta_2 \approx 140\hat{\beta}_1 + \hat{\beta}_2$$

We can use our estimations from the results of our regression using OLS, the result is shown below.

```
as.numeric(140*reg1$coefficients[2] + reg1$coefficients[3])
```

```
## [1] 32.34678
```

So the estimated increase in price for a house with one more bedroom, which is 140 extra square feet, is 32.34678 which is 32,346.78 dollars in the real world. This is higher than our result from part b, which is expected because not only do we have an increase in the number of bedrooms but we also have an increase in square feet (which has a positive coefficient).

d.

From part a we saw that $R^2 = 0.6283$ for this regression (which uses square feet and the number of bedrooms), so we can say that square feet and the number of bedrooms account for approximately 62.83% of the variation in price.

e.

Under our regression model $price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + U$, if $bdrms = 4$ and $sqrft = 2438$ we will get the following:

$$price = \beta_0 + 2438\beta_1 + 4\beta_2 \approx \hat{\beta}_0 + 2438\hat{\beta}_1 + 4\hat{\beta}_2$$

We can use our estimations from the results of our regression using OLS, the result is shown below.

```
as.numeric(reg1$coefficients[1] + 2438*reg1$coefficients[2] + 4*reg1$coefficients[3])
```

```
## [1] 353.4635
```

So the estimated price for a house with four bedrooms and 2438 square feet is 353.4635 which is 353,463.50 dollars in the real world.

f.

From the previous part the estimated price for a house with four bedrooms and 2438 square feet is $\hat{price} = 353.4635$, whereas the buyer actually payed $price = 300$. Therefore the residual here is $price - \hat{price} = 300 - 353.4635 = -53.4635$ which indicates that if this model is correct the buyer underpaid for the house. This regression is not the best, as indicated by the R^2, but it does have some significance so it is fairly indicative that the buyer did indeed underpay.

g.

The number of bedrooms almost certainly is a determinant of price which is indicated by the regression run before and by real world knowledge. Furthermore the number of bedrooms is clearly positively correlated with the square footage since if there are more bedrooms you need the space for them and hence there will be more square feet. This implies that if bedrooms is left out of the regression then the coefficient for square feet will have omitted variable bias. In the case of omitted variable bias we know $plim\ \hat{\beta}_1 = \beta_1 + \beta_2 \rho_{x,u} \frac{\sigma_u}{\sigma_x}$ (where ρ is the correlation between the included variable and the true residual term, and σ_u, σ_x are the standard deviations of the residual term and the included variable respectively). We know that $\rho > 0$ for square feet and the number of bedrooms and if bedrooms is left out it becomes part of the residual term we know that $\rho_{sqrft,u} > 0$ when the number of bedrooms is left out. Therefore we would have that the coefficient for square feet will be overestimated (that is it is positively biased) if the number of bedrooms is left out of the regression.
I will show the sample correlation between square feet and the number of bedrooms below as well as the new regression.

```
cor(hprice$bdrms, hprice$sqrft)
```

```
## [1] 0.5242245
```

```
reg2 <- lm(price ~ sqrft, data = hprice)
summary(reg2)
```

```
##
## Call:
## lm(formula = price ~ sqrft, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -115.237  -34.984   -5.694   26.693  237.071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.40322   22.63529   0.548    0.585
## sqrft        0.13931    0.01101  12.649   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.75 on 99 degrees of freedom
## Multiple R-squared:  0.6177, Adjusted R-squared:  0.6139
## F-statistic:   160 on 1 and 99 DF,  p-value: < 2.2e-16
```

As expected the sample correlation is fairly strong and positive between square feet and the number of bedrooms. In this new regression we got that $\hat{\beta}_1 = 0.13931$ which is greater than our estimate from the first regression ($\hat{\beta}_1 = 0.1280887$). This is consistent with what we predicted would happen in the case of omitted variable bias.

h.

If all the houses were one story and covered the entire lot then there would be perfect multicollinearity, but this is not the reality. Firstly, not every house covers the entire lot and the amount of the lot it covers is variable because some people like to have bigger backyards for example. Secondly, some houses are several floors which allows there to be more square feet in the house than even the lot has. So the lot size and the square feet of the house are definitely not perfectly multicollinear. The result of the regression is below.

```
reg3 <- lm(price ~ sqrft + bdrms + lotsize, data = hprice)
summary(reg3)
```

```
##
## Call:
## lm(formula = price ~ sqrft + bdrms + lotsize, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.600  -36.878   -6.243   30.597  209.349
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.287271  26.846344  -0.756 0.451672
## sqrft         0.121818   0.012268   9.930  < 2e-16 ***
## bdrms        13.734186   8.178776   1.679 0.096322 .
## lotsize       0.002129   0.000607   3.508 0.000686 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.02 on 97 degrees of freedom
## Multiple R-squared:  0.6702, Adjusted R-squared:   0.66
## F-statistic: 65.69 on 3 and 97 DF,  p-value: < 2.2e-16
```

The regression gives $\hat{\beta}_3 = 0.002129$ as the coefficient for lot size. This means that increasing the lot size by 1 square foot (while keeping all else constant) is expected to increase the price by 0.002129 which is a little over 2 dollars in the real world. Note however that in the real world people are much more likely to increase lot size by a more significant number than just one square foot. If the lot size increases but the square footage of the house does not this means that the outside space (for example the backyard) got larger. So the coefficient for lot size can be interpreted as the expected increase in price extra outside space gives, according to this regression outside space costs about 2.13 dollars per square foot.