8.2 Housing

Moshe Burnstein

`r Sys.Date()`

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

```
library(readxl)
housing_df <- read_excel('week-6-housing.xlsx')
head(housing_df)
str(housing_df)
unique(housing_df)
length(unique(housing_df))
summary(housing_df)
housing_df$`Sale Price`
summary(housing_df$'Sale Price')
unique(housing_df$'Sale Price')
```

The column "Sale Price" must be called with backticks.

```
sale_p <- housing_df$'Sale Price'
is.na(sale_p)
sum(is.na(sale_p))
ft_squared <- housing_df$sq_ft_lot
is.na(ft_squared)
sum(is.na(ft_squared))
```

I assigned the variable "sale_p" to the "Sale Price" column. There does not appear to be any NA values in "sale_p".

I assigned "ft_squared" for expedience. There are no NA values.

```
library(pastecs)
library(ggplot2)
stat.desc(sale_p)
```

```
stat.desc(ft_squared)
sales_histo <- ggplot(housing_df, aes(sale_p)) + geom_histogram(binwidth = 1289)
sales_histo
ft_squared_hist <- ggplot(housing_df, aes(ft_squared)) + geom_histogram(binwidth = 100)
ft_squared_hist
```

There are outliers in both "Sale Price" and "sq_ft_lot". The low price is $698 and the high is $4,400,000. Both are far from the mean $660,738. Square foot area has a minimum of 785 ft^2 and a maximum of 1,631,322 ft^2 with a mean of 2,229 ft^2.

```
sq_ft_lm <- lm(sale_p ~ ft_squared, housing_df)
summary(sq_ft_lm)
confint(sq_ft_lm)
price_lm <- lm(sale_p ~ ft_squared + year_built + bedrooms + square_feet_total_living + bath_full_count
summary(price_lm)
confint(price_lm)
better_lm <- lm(sale_p ~ ft_squared + year_built + square_feet_total_living + bath_full_count, data = h
summary(better_lm)
```

For "sq_ft_lm" the R^2=0.01435 and the adjusted R^2=0.01428. The 95% confidence interval is 0.729 to 0.973 ft^2.

For "price_lm" the R^2=0.2207 and the adjusted R^2=0.2204. ft_squared 1.559583e-01 3.878743e-01

```
year_built                2.152393e+03  3.006509e+03
bedrooms                 -1.925916e+04 -1.319389e+03
square_feet_total_living  1.586008e+02  1.768274e+02
bath_full_count           5.190619e+03  2.906486e+04
```

These represent the 95% confidence interval for these parameters.

Whereas in the simple regression "ft_squared" only accounts for 1.4% of the variation in sale price, in the multiple regression model the various predictors account for 22% of the variation in sale price. Removing bedrooms did not improve the model. I removed it because of its negative correlation.

```
sq_ft_lm$simple_lm <- rstandard(sq_ft_lm)
large_residual <- data.frame(sq_ft_lm$simple_lm>2 | sq_ft_lm$simple_lm<=-2)
sum(large_residual)
sq_ft_lm$dfbeta_simple <- dfbeta(sq_ft_lm)
price_lm$multiple_lm <- rstandard(price_lm)
large_residual_multiple <- data.frame(price_lm$multiple_lm>=2 | price_lm$multiple_lm<=-2)
sum(large_residual_multiple)
price_lm$dfbeta_multiple <- dfbeta(price_lm)
sq_ft_lm
price_lm
```

The standardized residuals for each model which are less than or equal to 2 or more than or equal to 2 are only 334.There is not great variation which would call the models into question

```
library(car)
dwt(price_lm)
```

The Durbin-Watson test strongly questions the assumption of independent errors. The value of 0.56 is not only not close to 2, but it is closing in on 0. The p-value of 0 implies that there is definitely correlation among the residuals.

```
vif(price_lm)
1/vif(price_lm)
mean(vif(price_lm))
```

Tests for multicollinearity

The vif stats show no reason for concern because the largest vif is much less than 10. The tolerance stats of 1/vif are good insomuch as none of the values are close to 0.1 and not even 0.2, which would have raised a red flag. The average vif is 1.6, and only values substantially greater than one would suggest that the regression may be biased.

```
plot(sq_ft_lm)
```

The q-q plot shows points very distant from the normal distribution line at the extremes, which indicates deviations from normality.

```
plot(price_lm)
```

This q-q plot also shows deviations from the normal distribution at the extremes.

```
hist(rstudent(sq_ft_lm))
```

Both histograms are not normal, bell-curved distributions.

```
rstudent_stat <- rstudent(sq_ft_lm)
rstudent_stat
dffit_simple <- dffits(sq_ft_lm)
dffit_simple
leverage_simple <- hatvalues(sq_ft_lm)
leverage_simple
cov_ratios_simple <- covratio(sq_ft_lm)
cov_ratios_simple
```

```
rstudent_multiple <- rstudent(price_lm)
rstudent_multiple
dffit_multiple <- dffits(price_lm)
dffit_multiple
leverage_multiple <- hatvalues(price_lm)
leverage_multiple
cov_ratios_multiple <- covratio(price_lm)
cov_ratios_multiple
```

The covariance ratios are all close to 1. They show no cause for concern.

There seems to be bias to the models, and therefore question as to generalizing to the population at large.