R Lab: Multiple Linear Regression (Solutions)

STAT 340: Applied Regression

Housing prices and log transformations.

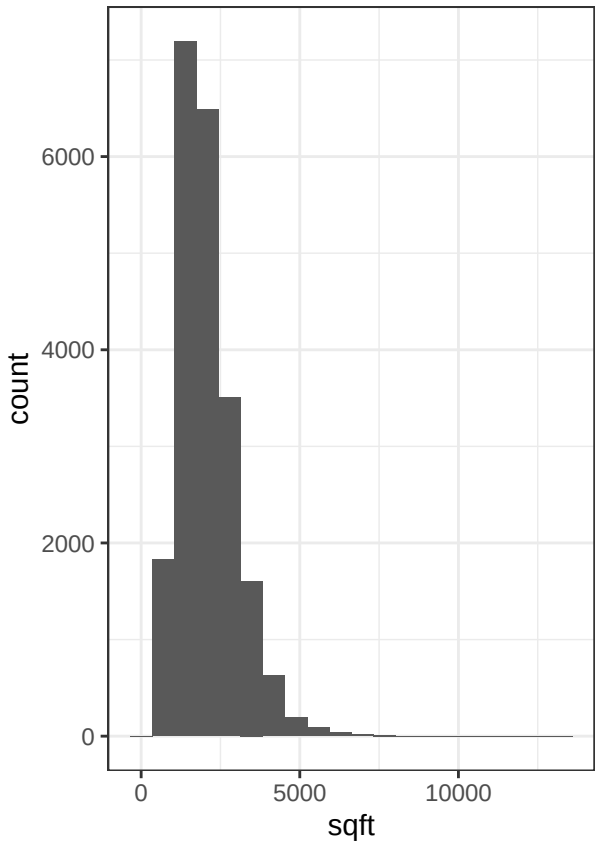
The dataset kingCountyHouses.csv contains data on over 20,000 houses sold in King County, washington (Kaggle 2018a). The dataset includes the following variables:

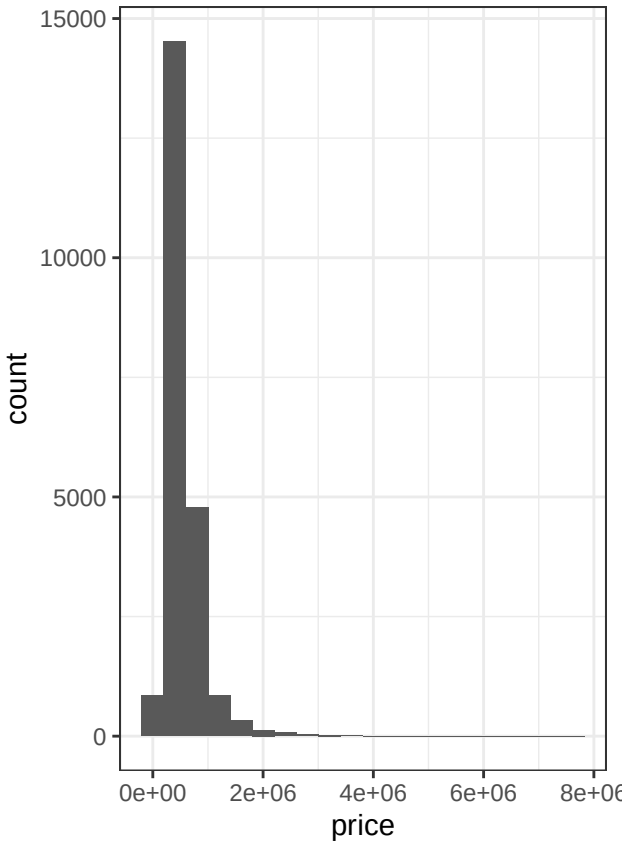
- `price` = selling price of the house
- `date` = date house was sold, measured in days since January, 2014
- `bedrooms` = number of bedrooms
- `bathrooms` = number of bathrooms
- `sqft` = interior square footage
- `floors` = number of floors
- `waterfront` = 1 if the house has a view of the waterfront, 0 otherwise
- `yr_built` = year the house was built
- `yr_renovated` = 0 if the house was never renovated, the year the house was renovated if else

We wish to create a linear model to predict the house's selling price.

(a) Generate appropriate graphs and summary statistics detailing both `price` and `sqft` individually and then together. What do you notice?

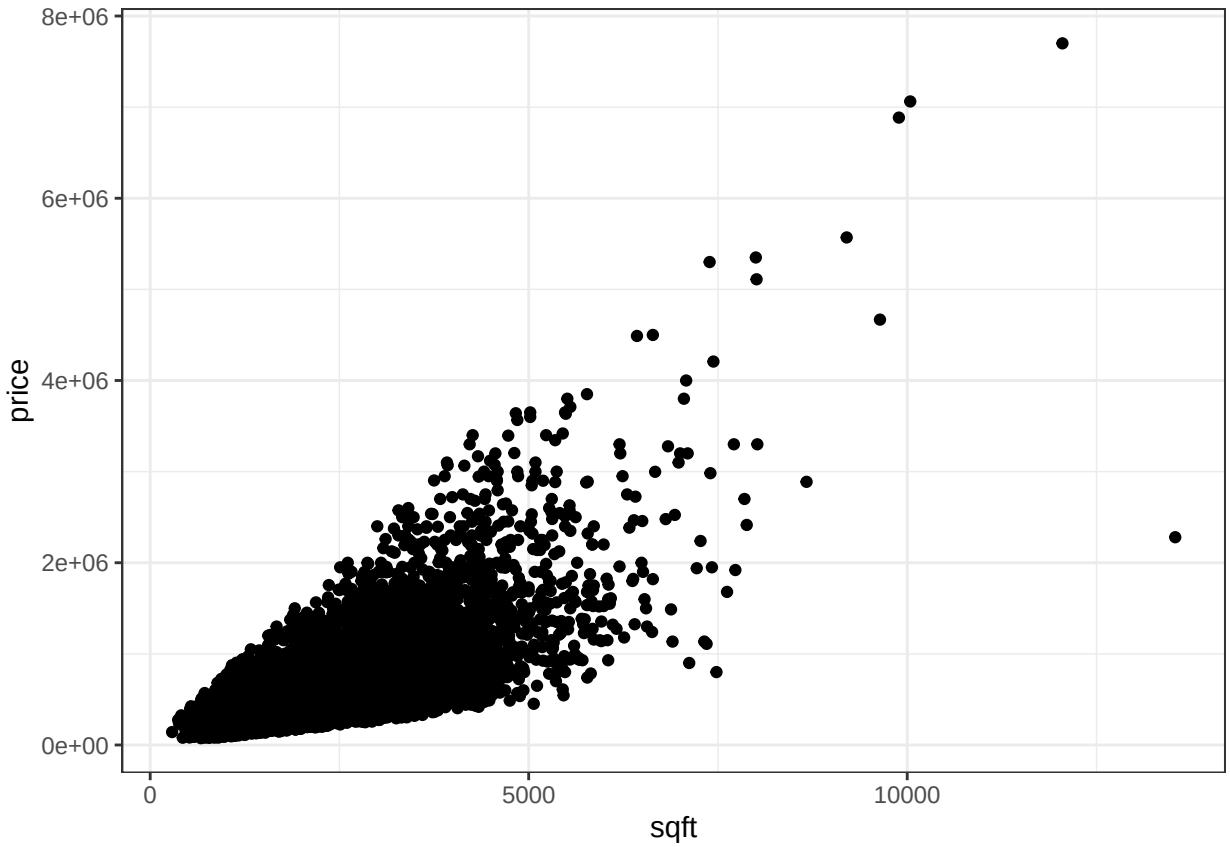
```
## Individual plots.
p1 <- ggplot(data=kch, aes(x=sqft)) +
        geom_histogram(bins=20) +
        theme_bw()

p2 <- ggplot(data=kch, aes(x=price)) +
        geom_histogram(bins=20) +
        theme_bw()

ggarrange(p1, p2, nrow=1)
```





When we examine the histograms for the predictor, `sqft` and the response, `price`, we see that the distributions for both of these variables are right skewed. We want to pay attention to the distribution `price` in particular, because if we adjust for our covariates, and the distribution of the residuals is not symmetric, this would mean we have a violation of normality. One thing we can do is consider a transformation, which is what is done in this lab. It does not matter if the distribution of `sqft` is skewed, unless this impacts the linearity assumption, in which case we may consider transforming it, but only after we have confirmed that the linearity assumption is still violated.

```
## Scatterplot.
ggplot(data=kch, aes(x=sqft, y=price)) +
  geom_point() +
  theme_bw()
```



Inspection of the scatterplot shows there is a positive relationship between `sqft` and `price`. There are some concerns about the equal variance assumption - the variance appears to increas as `sqft` increases, but we would want to confirm that by looking at the residual plot.

(b) Consider a simple linear regression model with `price` as the response variable and `sqft` as the explanatory variable (Model 1). Interpret the slope coefficient β_1. Are all conditions met for linear regression?

```
## Fit model and print summary
model1 <- lm(price ~ sqft, data=kch)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ sqft, data = kch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1476062  -147486   -24043   106182  4362067
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43580.743   4402.690  -9.899   <2e-16 ***
## sqft           280.624      1.936 144.920   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 261500 on 21611 degrees of freedom
## Multiple R-squared:  0.4929, Adjusted R-squared:  0.4928
## F-statistic: 2.1e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

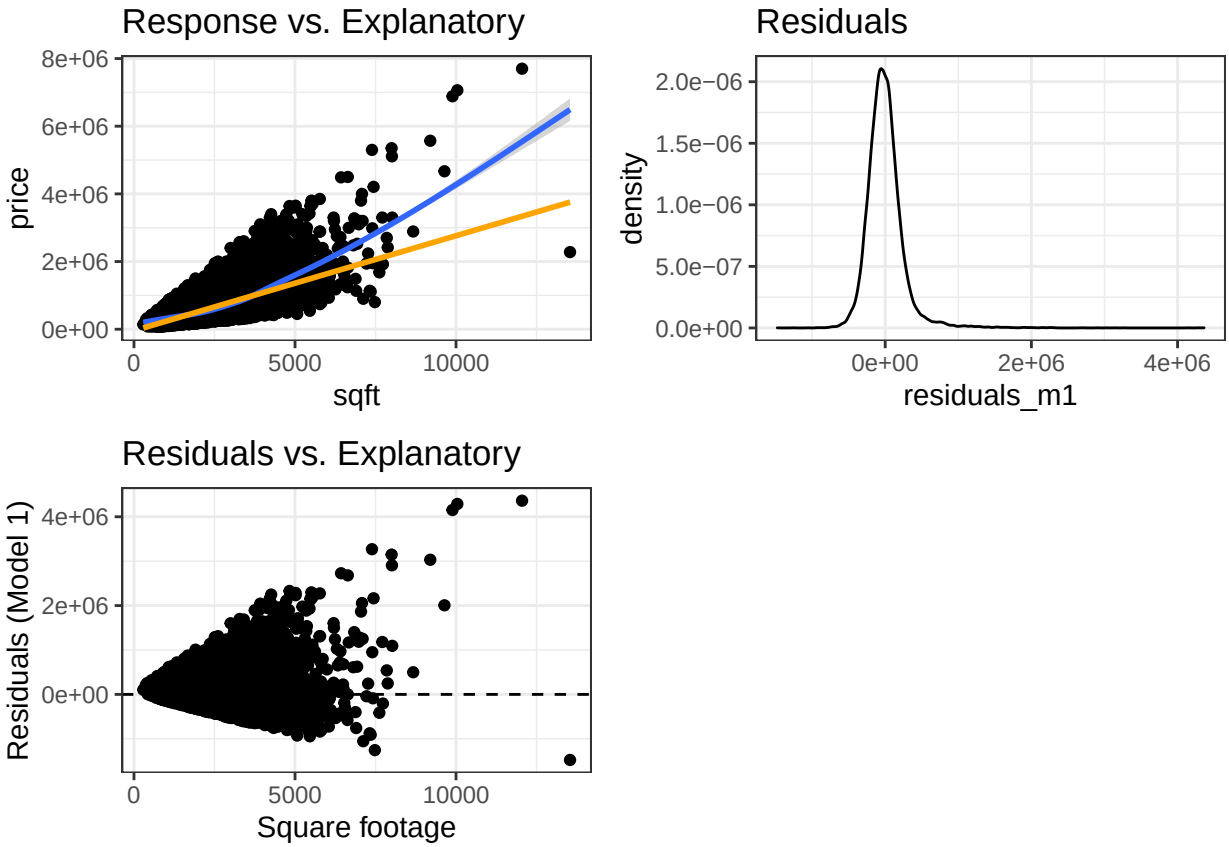
Interpretation: For a one square foot increase in house size, we expect, on average, a $280.62 increase in housing price for houses in King County. Note, we should be wary of this interpretation if we haven't evaluated whether the model assumptions are reasonable (see below).

```
## Check conditions are met:
kch <- kch %>%
  mutate(
    residuals_m1=residuals(model1),
    fitted_m1=predict(model1)
  )

## Linearity
lin_m1 <- ggplot(data=kch, aes(x=sqft, y=price)) +
  geom_point() +
  geom_smooth() +
  geom_smooth(method="lm", color="orange", se=FALSE) +
  ggtitle("Response vs. Explanatory") +
  theme_bw()

## Independence
## No plot to assess -- check data description

## Nearly normal residuals
residp_m1 <- ggplot(data=kch, aes(x=residuals_m1)) +
  geom_density() +
  theme_bw() +
  ggtitle("Residuals")

## Equal variance
resid_m1 <- ggplot(data=kch, aes(x=sqft, y=residuals_m1)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype="dashed")  +
  ggtitle("Residuals vs. Explanatory") +
  ylab("Residuals (Model 1)") +
  xlab("Square footage") +
  theme_bw()

ggarrange(lin_m1, residp_m1, resid_m1, ncol=2)
```

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

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## adding dummy grobs
```



There are clearly problems with the assumptions for the linear model. First, although price and sqft are definitely positively correlated, linearity is potentially a problem here, as we can see from the plot in the upper left corner. The blue line looks more exponential than linear (a hint about a transformation!); the orange line is what we would get for the linear model. If linearity was reasonable, the orange and blue lines should be really similar. Second, the nearly normal residuals assumption is problematic, namely because the residual density plot (upper right) has a long right tail, so the skew persists, even after we account for square footage. Third, equal variance is a problematic assumption, as we suspected - looking at the bottom left plot, the residuals get larger as square footage increases. This "cornucopia" shape is a classic example of the equal variance assumption being violated.

We have to check how the data are collected to evaluate independence. From what I can tell, these data include homes sold in King County between May 2014 and May 2015. I haven't found anything to suggest this is a random sample, so independence could be violated. I might expect some spatial dependence, even, which we will talk about later in the course. For sake of the rest of the lab, though, we are going to proceed assuming independence is satisfied (but we should be careful in practice!)

(c) Create a new variable, `logprice`, the natural log of `price`. Fit Model 2, where `logprice` is now the response variable and `sqft` is still the explanatory variable. Write out the regression equation in matrix form.

```
## Create new variable
kch <- kch %>%
  mutate(
    logprice=log(price)
  )
```

```r
model2 <- lm(logprice ~ sqft, data=kch)
summary(model2)
```

```
##
## Call:
## lm(formula = logprice ~ sqft, data = kch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97781 -0.28543  0.01472  0.26070  1.27628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.222e+01  6.374e-03  1916.9   <2e-16 ***
## sqft        3.987e-04  2.803e-06   142.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3785 on 21611 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4835
## F-statistic: 2.023e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Matrix form:

Let $\mathbf{y}$=log(price), a column vector of length 20,000; let $\mathbf{X} = \begin{bmatrix}\mathbf{1}_{20000} & \mathbf{x}_1\end{bmatrix}$, where $\mathbf{x}_1$=sqft, a column vector of length 20,000. Then, the linear model in matrix form is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\epsilon \sim Normal_{20000}(\mathbf{0}, \sigma^2\mathbf{I})$.

(d) How does `logprice` change when `sqft` increases by 1?

For a 1 square foot increase in house size, logprice increases by 3.987e-04. (Note, units don't make much sense on this scale - hence (e).)

(e) How does `price` change as `sqft` increases by 1? It may be helpful to recall that $log(a) - log(b) = log(\frac{a}{b})$.

For a 1 square foot increase in house size, price increases by a multiplicative factor of $e^{3.987e-04} \approx 1$.

The hint comes in here:

$\beta_1 = (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) = y|(x+1) - y|x = \log(price)|(x+1) - \log(price)|x = \log\left(\frac{price|(x+1)}{price|x}\right)$

$\Rightarrow e^{\beta_1} = \frac{price|(x+1)}{price|x} \Rightarrow (price|x) \times e^{\beta_1} = price|(x+1)$.

(f) Are all the conditions for linear regression met?

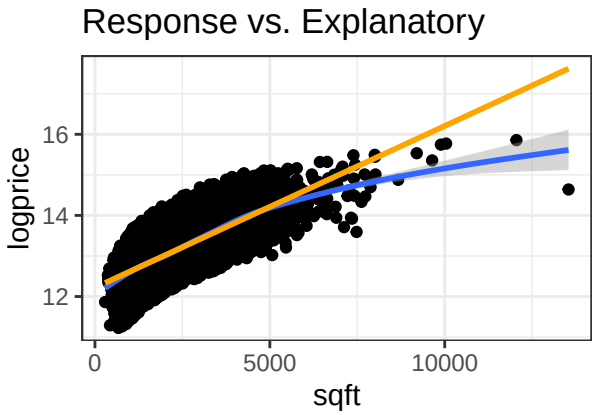
```r
## Check conditions are met:
kch <- kch %>%
  mutate(
    residuals_m2=residuals(model2),
    fitted_m2=predict(model2)
```

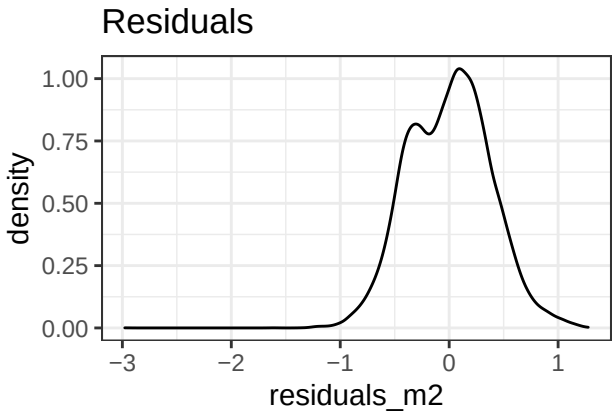
```
  )

## Linearity
lin_m2 <- ggplot(data=kch, aes(x=sqft, y=logprice)) +
  geom_point() +
  geom_smooth() +
  geom_smooth(method="lm", color="orange", se=FALSE) +
  ggtitle("Response vs. Explanatory") +
  theme_bw()

## Independence
## No plot to assess -- check data description

## Nearly normal residuals
residp_m2 <- ggplot(data=kch, aes(x=residuals_m2)) +
  geom_density() +
  theme_bw() +
  ggtitle("Residuals")

## Equal variance
resid_m2 <- ggplot(data=kch, aes(x=sqft, y=residuals_m2)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype="dashed")  +
  ggtitle("Residuals vs. Explanatory") +
  ylab("Residuals (Model 2)") +
  xlab("Square footage") +
  theme_bw()

ggarrange(lin_m2, residp_m2, resid_m2, ncol=2)
```

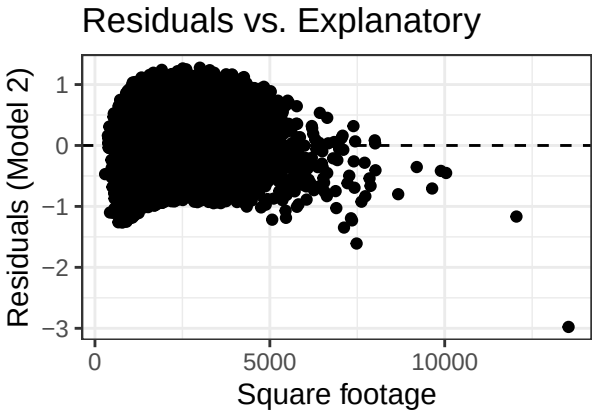
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## adding dummy grobs
```







No, the assumptions are not satisfied. Linearity, for one thing, is still problematic. Equal variance also appears to have an issue. You could keep trying to transform y, but we are going to go ahead and transform x, instead, and see if we can resolve the issues.

(g) Create a new variable, `logsqft`, the natural log of `sqft`. Fit Model 3 where `price` and `logsqft` are the response and explanatory variables, respectively. Write out the regression line equation in matrix form.

```
## Create new variable
kch <- kch %>%
  mutate(
    logsqft=log(sqft)
  )

model3 <- lm(price ~ logsqft, data=kch)
summary(model3)
```

```
##
## Call:
## lm(formula = price ~ logsqft, data = kch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -606252 -170067  -33139  106342 6183772
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3451377      35169  -98.14   <2e-16 ***
## logsqft       528648       4651  113.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 290400 on 21611 degrees of freedom
## Multiple R-squared:  0.3742, Adjusted R-squared:  0.3742
## F-statistic: 1.292e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Let $\mathbf{y}$=log(price), a column vector of length 20,000; let $\mathbf{X} = \begin{bmatrix}\mathbf{1}_{20000} & \mathbf{x}_1\end{bmatrix}$, where $\mathbf{x}_1$=log(sqft), a column vector of length 20,000. Then, the linear model in matrix form is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\epsilon \sim Normal_{20000}(\mathbf{0}, \sigma^2\mathbf{I})$.

(h) How does predicted `price` change as `logsqft` increases by 1 in Model 3?

For a 1 unit increase in logsqft, the price should increase \$528,648, on average, in King County.

$e^{\beta_1} = \frac{price|(x+1)}{price|x} \Rightarrow (price|x) \times e^{\beta_1} = price|(x+1)$.

(i) How does predicted `price` change as `sqft` increases by 1%? As a hint, this is the same as multiplying `sqft` by 1.01.

It is useful to write this out.

$price|\log(1.01sqft) = \beta_0 + \beta_1(\log(1.01 \times sqft))$

$price|\log(sqft) = \beta_0 + \beta_1 \times \log(sqft)$

An increase of 1% can be expressed as:

$price|\log(1.01 \times sqft) - price|\log(1 \times sqft) = \beta_1 \times \log(1.01)$.

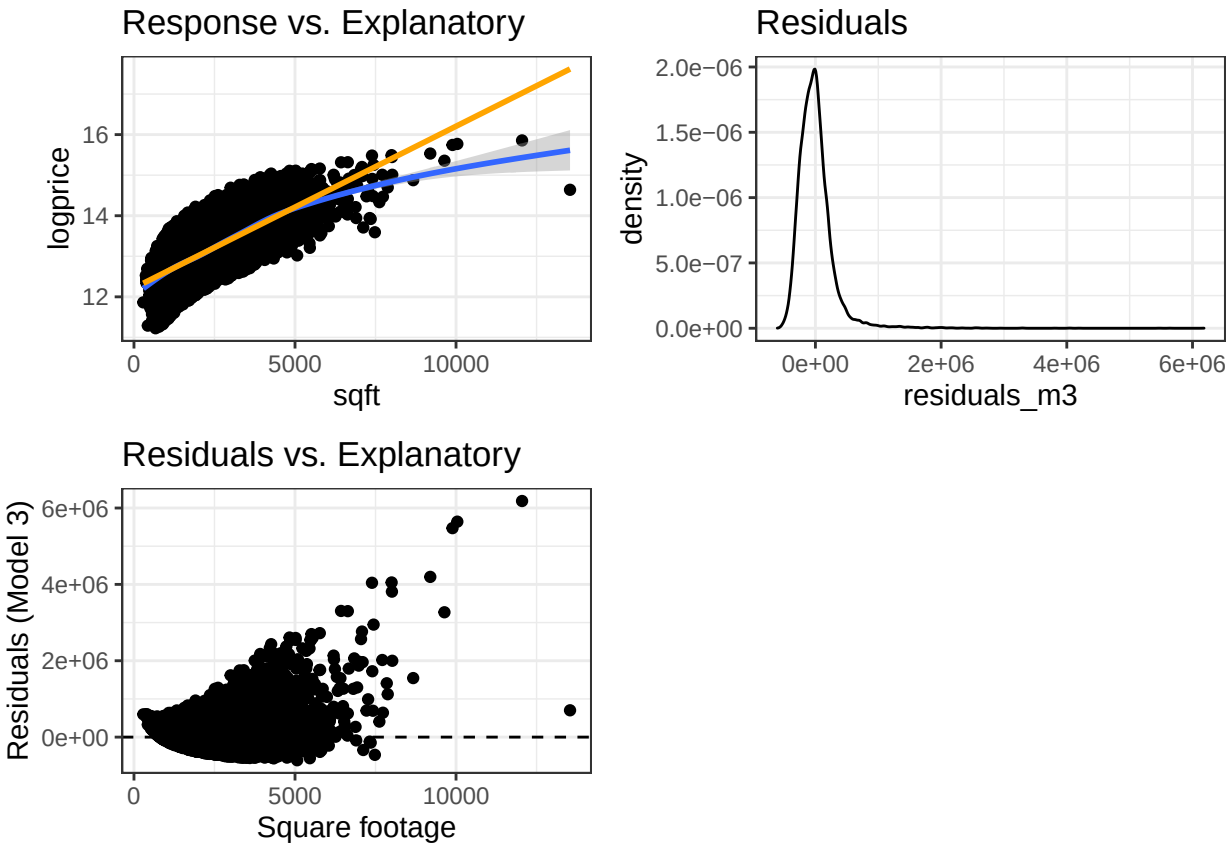
So, for a 1% increase in square footage, the predicted price increases by $\log(1.01) \times \$528648 = \$5260.22$ in King County.

(j) Are the conditions for linear regression met for Model 3? Why or why not?

```r
## Check conditions are met:
kch <- kch %>%
  mutate(
    residuals_m3=residuals(model3),
    fitted_m3=predict(model3)
  )

## Linearity
lin_m3 <- ggplot(data=kch, aes(x=sqft, y=logprice)) +
  geom_point() +
  geom_smooth() +
  geom_smooth(method="lm", color="orange", se=FALSE) +
  ggtitle("Response vs. Explanatory") +
  theme_bw()

## Independence
```

```r
## No plot to assess -- check data description

## Nearly normal residuals
residp_m3 <- ggplot(data=kch, aes(x=residuals_m3)) +
  geom_density() +
  theme_bw() +
  ggtitle("Residuals")

## Equal variance
resid_m3 <- ggplot(data=kch, aes(x=sqft, y=residuals_m3)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype="dashed")  +
  ggtitle("Residuals vs. Explanatory") +
  ylab("Residuals (Model 3)") +
  xlab("Square footage") +
  theme_bw()

ggarrange(lin_m3, residp_m3, resid_m3, ncol=2)
```

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



There are still problems with linearity, equal variance, and nearly normal residuals. Let's try one more transformation.

(k) Fit Model 4, with `logsqft` and `logprice` as the response and explanatory variables, respectively. Write out the regression line equation in matrix form.

```
model4 <- lm(logprice ~ logsqft, data=kch)
summary(model4)
```

```
##
## Call:
## lm(formula = logprice ~ logsqft, data = kch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10511 -0.29300  0.01262  0.25701  1.33011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.729916   0.047062   143.0   <2e-16 ***
## logsqft     0.836771   0.006223   134.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3886 on 21611 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4555
## F-statistic: 1.808e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

(l) In Model 4, what is the effect on `price` corresponding to a 1% increase in `sqft`?

It is useful to write this out.

$\log(price)|\log(1.01sqft) = \beta_0 + \beta_1(\log(1.01 \times sqft))$

$\log(price)|\log(sqft) = \beta_0 + \beta_1 \times \log(sqft)$

An increase of 1% can be expressed as:

$\log(price)|\log(1.01 \times sqft) - \log(price)|\log(1 \times sqft) = \beta_1 \times \log(1.01)$.

To get this on the scale of price, we need to exponentiate:

$\Rightarrow \frac{price|\log(1.01 \times sqft)}{price|\log(1 \times sqft)} = 1.01e^{\beta_1}$.

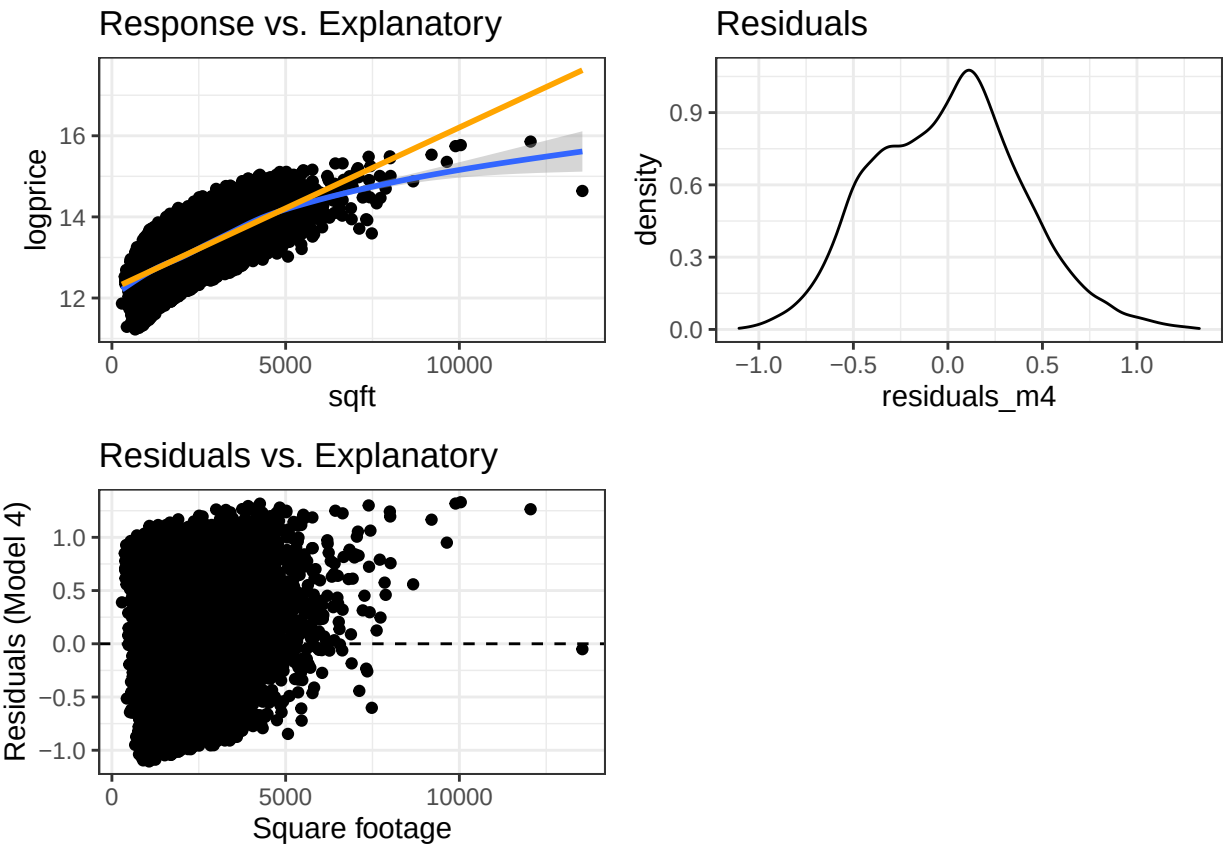
So, for a 1% increase in square footage, the predicted price increases by $1.01 \times e^{0.837} = 2.33$ times.

(m) Are the linear regression conditions satisfied in Model 4? Why or why not?

```
## Check conditions are met:
kch <- kch %>%
  mutate(
    residuals_m4=residuals(model4),
    fitted_m4=predict(model4)
  )

## Linearity
lin_m4 <- ggplot(data=kch, aes(x=sqft, y=logprice)) +
  geom_point() +
  geom_smooth() +
  geom_smooth(method="lm", color="orange", se=FALSE) +
  ggtitle("Response vs. Explanatory") +
```

```
  theme_bw()

## Independence
## No plot to assess -- check data description

## Nearly normal residuals
residp_m4 <- ggplot(data=kch, aes(x=residuals_m4)) +
  geom_density() +
  theme_bw() +
  ggtitle("Residuals")

## Equal variance
resid_m4 <- ggplot(data=kch, aes(x=sqft, y=residuals_m4)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype="dashed")  +
  ggtitle("Residuals vs. Explanatory") +
  ylab("Residuals (Model 4)") +
  xlab("Square footage") +
  theme_bw()

ggarrange(lin_m4, residp_m4, resid_m4, ncol=2)
```

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



Still not perfect, but definitely an improvement. We may need more advanced methods to analyze these data. Remember that independence is likely violated.

References

Kaggle. 2018a. "House Sales in King County, USA." https://www.kaggle.com/harlfoxem/housesalesprediction/home.

P. Roback and J. Legler. 2020. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R.*